Validation de réponses dans un système de questions réponses

Arnaud Grappy

Université Paris Sud
École doctorale d'informatique

Mémoire de thèse

pour obtenir le grade de

DOCTEUR EN INFORMATIQUE DE L'UNIVERSITÉ DE PARIS SUD

Validation de réponses dans un système de questions réponses

Arnaud Grappy

soutenue le 8 novembre 2011 devant le jury composé de

Rapporteurs	Isabelle Tellier
	Patrice Bellot
Directrice	Brigitte Grau
Président du jury	François Yvon
Examinateurs	Olivier Ferret
	Thierry Poibeau

Résumé

Avec l'augmentation des connaissances disponibles sur Internet est apparue la difficulté d'obtenir une information. Les moteurs de recherche permettent de retourner des pages Web censées contenir l'information désirée à partir de mots clés. Toutefois il est encore nécessaire de trouver la bonne requête et d'examiner les documents retournés.
Les systèmes de questions réponses ont pour but de renvoyer directement une réponse concise à partir d'une question posée en langue naturelle. La réponse est généralement accompagnée d'un passage de texte censé la justifier. Par exemple, pour la question « Quel est le réalisateur d'Avatar ? » la réponse « James Cameron »peut être renvoyée accompagnée de « James Cameron a réalisé Avatar. ».

Cette thèse se focalise sur la validation de réponses qui permet de déterminer automatiquement si la réponse est valide. Une réponse est valide si elle est correcte (répond bien à la question) et justifiée par le passage textuel. Cette validation permet d'améliorer les systèmes de questions réponses en ne renvoyant à l'utilisateur que les réponses valides.

Les approches permettant de reconnaître les réponses valides peuvent se décomposer en deux grandes catégories :

- les approches utilisant un formalisme de représentation particulier de la question et du passage dans lequel les structures sont comparées ;
- les approches suivant une approche par apprentissage qui combinent différents critères d'ordres lexicaux ou syntaxiques.

Dans le but d'identifier les différents phénomènes sous-tendant la validation de réponses, nous avons participé à la création d'un corpus annoté manuellement. Ces phénomènes sont de différentes natures telle que la paraphrase ou la coréférence. On peut aussi remarquer que les différentes informations sont réparties sur plusieurs phrases, voire sont manquantes dans les passages contenant la réponse.

Une deuxième étude de corpus de questions a porté sur les différentes informations à vérifier afin de détecter qu'une réponse est valide. Cette étude a montré que les trois phénomènes les plus fréquents sont la vérification du type de la réponse, la date et le lieu contenus dans la question.

Ces différentes études ont permis de mettre au point notre système de validation de réponses qui s'appuie sur une combinaison de critères. Certains critères traitent de la présence dans le passage des mots de la question ce qui permet de pointer la présence des informations de la question. Un traitement particulier a été effectué pour les informations de date en détectant une réponse comme n'étant pas valide si le passage ne contient pas la date contenue dans la question. D'autres critères,

dont la proximité dans le passage des mots de la question et de la réponse, portent sur le lien entre les différents mots de la question dans le passage.

Le second grand type de vérification permet de mesurer la compatibilité entre la réponse et la question. Un certain nombre de questions attendent une réponse d'un type particulier. La question de l'exemple précédent attend ainsi un réalisateur en réponse. Si la réponse n'est pas de ce type alors elle est incorrecte. Comme cette information peut ne pas se trouver dans le passage justificatif, elle est recherchée dans des documents autres à l'aide de la structure des pages Wikipédia, en utilisant des patrons syntaxiques ou grâce à des fréquences d'apparitions du type et de la réponse dans des documents. La vérification du type est particulièrement efficace puisqu'elle effectue 80 % de bonnes détections. La vérification de la validité des réponses est également pertinente puisque lors de la participation à une campagne d'évaluation, AVE 2008, le système s'est placé parmi les meilleurs toutes langues confondues.

La dernière contribution a consisté à intégrer le module de validation dans un système de questions réponses, QAVAL. Dans ce cadre, de nombreuses réponses sont extraites par QAVAL et ordonnées grâce au module de validation de réponses. Le système n'est plus utilisé afin de détecter les réponses valides mais pour fournir un score de confiance à chaque réponse. Le système QAVAL peut ainsi aussi bien être utilisé en effectuant des recherches dans des articles de journaux que dans des articles issus du Web. Les résultats sont assez bons puisqu'ils dépassent ceux obtenus par un simple ordonnancement des réponses de près de 50 %.

Remerciements

Une thèse étant un travail de longue durée, je tiens tout d'abord à remercier tous ceux et celles qui m'ont permis de la mener jusqu'au bout que ce soit directement à travers des conversations scientifiques ou indirectement par leur soutien.

Je voudrais tout d'abord remercier ma directrice de thèse Brigitte Grau qui m'a soutenu tout au long de ces années. Je retiens particulièrement toutes les discussions intéressantes qui ont permis de mener à bien ce travail. Je voudrais également la remercier pour tout le travail que nous avons effectué et toujours dans la bonne humeur. Pour finir je voudrais la remercier pour les nombreuses relectures de ce manuscrit et des différents articles que nous avons écrits.

Je remercie également Anne-Laure Ligozat qui a encadré le stage antérieur à cette thèse qui m'a permis de découvrir le monde de la recherche et m'a donné envie de continuer à travers cette thèse. Je la remercie également pour tous les conseils et remarques que j'ai reçu.

Je remercie Isabelle Tellier et Patrice Bellot d'avoir accepté d'être les rapporteurs de ma thèse, et pour les rapports détaillés soulevant des questions très pertinentes.

Je remercie également l'ensemble des examinateurs : Olivier Ferret, Thierry Poibeau et François Yvon. Les remarques et questions posées à la suite de ma présentation ont été particulièrement intéressantes et enrichissantes.

Je tiens également à remercier toutes les personnes travaillant au LIMSI et plus particulièrement les membres du groupe ILES pour l'accueil convivial que j'ai reçu et pour m'avoir intégré parmi eux. Je commencerais par remercier mes collègues avec qui j'ai participé à l'amélioration des systèmes de questions réponses pour toute l'aide que j'ai reçue et les réunions intéressantes. Je voudrais également remercier mes collègues qui sont devenus de vrais amis (Tifanie, Guillaume, Mathieu et Kévin) et qui m'ont poussé à m'accrocher pour finir cette thèse malgrès des moments de doute. Je voudrais encore remercier mes différents collègues de bureau pour les nombreuses discussions informelles et éclats de rire que nous avons pu avoir pendant les diverses pauses.

Je voudrais également en profiter pour remercier tous mes amis pour les différentes soirées ou vacances que nous avons faites ce qui m'a permis de relativiser les quelques difficultés et à prendre du recul sur mon travail.

Je souhaiterais également remercier toute ma famille, et plus particulièrement mes parents et ma sœur qui m'ont soutenus pendant mes nombreuses années d'étude ce qui m'a permis d'arriver jusqu'à ce stade ainsi que pour les différents conseils qu'ils m'ont donnés.

Je voudrais enfin remercier ma copine, Gaëlle, pour tout l'amour que j'ai reçu et notamment quand j'en avais le plus besoin, lors des jours sans fin passés à améliorer le manuscrit et les moments de stress de la soutenance.

Pour finir j'aimerais remercier tous ceux qui m'ont signalés que mes lacets étaient défaits.

Table des matières

Table des figures

Introduction

De tout temps, la diffusion des connaissances a été une première nécessité. La préhistoire prend ainsi fin avec l'invention de l'écriture, le moyen-âge avec l'invention de l'imprimerie, les temps modernes arrivent avec l'école obligatoire de Jules Ferry. Aujourd'hui une révolution a été faite avec l'invention et la diffusion d'Internet. Il devient alors possible de rechercher n'importe quelle information à partir de chez soi en utilisant simplement un ordinateur voire certains téléphones.

Avec la croissance continue d'Internet, est également apparue la difficulté d'obtenir le plus simplement possible une information précise. Les moteurs de recherche tels que Google ou Yahoo ! permettent de retourner un ensemble de pages censées contenir l'information recherchée à partir de mots clés définissant la recherche. Par exemple, si l'information recherchée correspond à la question « *Quel pays a gagné la coupe du monde de football en 1998 ?* » la requête peut alors être « *pays gagné coupe du monde de football 1998* ». Le moteur de recherche retourne alors les pages concernant l'événement ainsi que, pour chacune d'elles un court passage de texte qui idéalement contiendrait l'information. La figure 1 présente la page Google consacrée à une telle recherche. Dans l'exemple, la réponse « la France » se situe dans le sixième passage de texte. Ce type de recherche est très largement utilisé. Toutefois il se heurte à deux problèmes :

- l'utilisateur doit trouver les mots clés pertinents ce qui n'est pas toujours facile ;
- il doit examiner les passages de texte et parfois même les pages censées contenir l'information recherchée ce qui peut être coûteux en temps.

Les systèmes de questions réponses ont pour but de pallier ces difficultés. En effet, ils permettent à l'utilisateur de poser une question en langue naturelle comme « *Qui a gagné la coupe du monde en 1998 ?* » et fournissent directement la réponse « La France ». Pour ce faire, le système cherche les documents censés contenir la réponse puis les analyse afin d'extraire la réponse la plus pertinente. Pour permettre à l'utilisateur de s'assurer que la réponse est bien correcte, un passage textuel lui est également fourni. Ce passage, appelé passage justificatif, est censé justifier la réponse. Par exemple, à la question précédente le passage peut être « *Le 12 juillet 1998, la France devient championne du monde de football* ». De nombreux systèmes de questions réponses existent, notamment sur le français, et obtiennent de bons résultats quand il s'agit de détecter des documents mais ont plus de difficultés à fournir une réponse correcte en première position. C'est pourquoi nous nous sommes intéressés au problème consistant à évaluer si une réponse est correcte ou non.

FIG. 1 – Résultats de la recherche Google

Problématique

La validation de réponses a pour but de déterminer automatiquement si la réponse est correcte en vérifiant notamment qu'elle est justifiée par le passage. On dira alors qu'une telle réponse est valide. Dans l'exemple précédent, « La France » est effectivement une réponse valide. En revanche, la réponse « le Brésil » issue du passage « *Le Brésil a gagné quatre fois la coupe du monde (1958 - 1962 - 1970 - 1994)* » ne l'est pas car la victoire du pays n'a pas eu lieu à la date souhaitée. Lors de son apparition, lors de la campagne d'évaluation AVE 2006 [Peñas, et al. 2006], cette tâche était présentée comme une manière d'évaluer automatiquement les systèmes de questions réponses. Ainsi, la réponse était dite valide si elle était correcte et justifiée par le passage. Dans le cadre de ce manuscrit, nous considérons cette validation comme un module d'un système de questions réponses visant à ne fournir à l'utilisateur que les réponses valides. Ainsi, si un système de questions réponses a le choix entre les deux réponses précédentes, la validation devra permettre de privilégier la première.

Il est à noter que bien souvent un certain nombre de phénomènes rendent la tâche difficile puisque le contenu du passage justificatif n'est pas écrit sous la même forme que la question. Ainsi, dans certains exemples, il est nécessaire de mettre en œuvre un mécanisme d'inférence pour arriver à établir une correspondance entre ces deux entités. Cette thèse se focalise donc sur l'étude de la validation de réponses et permet de répondre à trois questions :

- Qu'est ce qu'une réponse valide ?
- Comment modéliser la validation de réponses et la mettre en œuvre ?
- Comment l'intégrer dans un système de questions réponses ?

Dans ce cadre, les solutions proposées seront appliquées au français mais notre approche devra pouvoir s'appliquer à d'autres langues. Les méthodes devront donc être robustes aux différentes langues. Un autre aspect concerne le type de document dans lesquels sera effectuée la recherche.

Contributions

La réponse à ces questions a donné lieu à la mise en évidence des phénomènes liés à la tâche, la création d'un système de validation de réponses et son intégration dans un système de questions réponses.

La première contribution consiste en une caractérisation des phénomènes linguistiques et discursifs à traiter pour permettre la justification des réponses. Pour ce faire, nous avons annoté un corpus constitué d'un ensemble de documents contenant la réponse à une question et censés la justifier. Les annotateurs devaient décider du fait que la réponse était justifiée. Un document justifie une réponse s'il porte sur le sujet exprimé dans la question et précise l'information demandée. Aussi la deuxième tâche des annotateurs consistait à relever les phénomènes rendant possible la justification. Ces phénomènes correspondent à des composants classiques du traitement automatique des langues qui marquent :

- l'inférence d'une information absente du document ;
- son éloignement de la réponse avec entre autre des phénomènes d'anaphores ;
- la reformulation d'une partie de la question par exemple sous forme de paraphrase.

Le corpus ainsi créé se devait d'être le plus proche possible de ce que rencontrent les systèmes de questions réponses. Sa création a été effectuée de manière semi-automatique avec des recherches de documents automatiques à partir d'un ensemble de mots clés vérifiés manuellement. L'annotation a révélé que très souvent les phénomènes se combinaient rendant difficile une validation fondée sur un appariement sémantique global entre la question et le passage. De plus, un tel traitement peut nécessiter des ressources qui ne sont pas disponibles dans toutes les langues.

Ainsi, la seconde contribution porte sur la création d'un système de validation de réponses dont le principe consiste à décomposer les questions en ensembles de vérifications puis à mettre en œuvre différents niveaux de traitements selon les ressources disponibles ou la nature des textes. La décomposition des questions permet alors d'effectuer l'ensemble de ces vérifications de manière séparée. Ces vérifications portent entre autre sur l'action de la question, sa date ou une vérification du type de la réponse. Par exemple, la question « *Dans quelle grande capitale la tour Eiffel fut-elle érigée en 1889 ?* » se décompose en :

- la réponse est une capitale ;
- la réponse est une ville ;

– la tour Eiffel a été érigée ;
– l'action se déroule en 1889 ;
– la réponse correspond au lieu où a lieu l'action de la question.

Aucun système n'avait auparavant décomposé les questions de cette manière dans un cadre de validation de réponses. Notre approche permet d'isoler des phénomènes relevant de processus de résolution différents et d'apporter des solutions dédiées à chacun d'eux. De plus, on voit que certains éléments portent sur l'information donnée dans la question et d'autres sur la caractérisation de la réponse.

Une fois ces décompositions effectuées, différents traitements sont appliqués. Les premiers cherchent à vérifier que les informations contenues dans la question se trouvent également dans le passage justificatif. Pour cela, différents critères lexicaux sont considérés. Certains portent ainsi sur la présence des termes de la question dans le passage ce qui permet de vérifier que les informations mentionnées par la question se trouvent dans le passage. D'autres étudient leur proximité dans le passage afin de détecter que les mots sont liés. Une vérification toute particulière concerne la date contenue dans la question. Elle permet de voir que l'événement contenu dans le passage se situe au même moment que celui de la question, par exemple, à la question précédente, que l'action du passage a bien lieu en 1889. Ces critères ont été très souvent utilisés pour détecter la validité d'une réponse et les différentes méthodes existantes suivent un processus similaire.

Les secondes vérifications permettent de reconnaître que la réponse est compatible avec la question. Un certain nombre de questions attendent une réponse d'un certain type. La question précédente attend ainsi en réponse une ville qui est une capitale. Bien sûr, si la réponse n'est pas une capitale alors elle est incorrecte. Cette information étant souvent absente du passage justificatif elle doit se rechercher en dehors de celui-ci. Ainsi des documents externes tels que l'encyclopédie Wikipédia sont exploités. La plupart des études portant sur ce sujet utilisent des bases de données comme WordNet. Malheureusement ces données ne peuvent être exploitées que sur l'anglais. Notre étude peut être appliquée sur de nombreuses langues et dans un cadre plus large que la validation de réponses comme la recherche d'entités nommées. La combinaison des différentes vérifications est gérée par apprentissage.

La troisième contribution porte sur l'intégration de notre système au sein d'un système de questions réponses, QAVAL. Afin d'utiliser au mieux notre système, QAVAL extrait des réponses en grand nombre depuis de courts passages de texte et notre approche de validation de réponses a donné lieu à un module permettant d'ordonner les réponses afin que la bonne réponse se trouve en première position. Cela permet d'évaluer notre système de validation de réponses dans un cadre fonctionnel. Certains travaux effectuaient de l'ordonnancement de réponses grâce à des approches relativement similaires mais aucune ne s'était vraiment centrée sur l'utilisation d'une méthode de validation de réponses. De plus, certains critères que nous proposons sont spécifiques à ce travail.

Plan du document

Ce manuscrit commence par un état de l'art, dans le chapitre 1, dans lequel nous avons mis l'accent sur la définition de critères pertinents. Tout d'abord il présente plus en détail les systèmes de

questions réponses leurs différents modules, leur évaluation. Puis, en nous appuyant sur l'existant, nous formulons notre définition de la validation de réponses dans laquelle il est dit qu'une réponse est valide si elle est compatible avec la question et que les informations demandées par la question se trouvent dans le passage justificatif. Pour finir ce chapitre présente les différentes approches permettant de détecter les réponses valides. Certaines s'appuient sur un formalisme de représentation de la question et du passage, par exemple des arbres syntaxiques, et recherchent la similarité des deux structures. D'autres voient la validation comme un problème de classification et combinent différents critères souvent d'ordre lexical.

La création et l'annotation du corpus permettant de mettre en évidence les difficultés posées pour la validation de réponses sont décrites dans le chapitre 2. Ce chapitre présente aussi bien l'élaboration du corpus que le guide d'annotation et les résultats observés.

Le système de validation de réponses est présenté dans le chapitre 3. Tout d'abord une analyse des décompositions effectuées sur des questions montre que le type spécifique et les informations de date et de lieux sont des vérifications nécessaires pour de nombreuses questions.

L'analyse des passages justificatifs est effectuée grâce à un ensemble de critères visant pour la plupart à comparer le passage avec la question en tenant par exemple compte de la présence dans le passage des mots de la question et reprend des critères présents dans les systèmes déjà existants. Une première évaluation effectuée tient compte de la capacité du système à reconnaître les réponses valides grâce à un ensemble de données issu de campagnes d'évaluation.

La vérification du type spécifique est originale et est effectuée, tout d'abord, en définissant davantage le problème puis à l'aide de critères recherchant cette information dans des documents. Les évaluations ont montré que le système réalise une bonne performance puisqu'il effectue 80 % de bonnes détections.

Le système de questions réponses QAVAL, présenté dans le chapitre 4, a pour spécificité de pouvoir s'appliquer aussi bien aux documents issus d'articles de journaux qu'à ceux issus du Web malgré leurs caractéristiques spéciales. Sa présentation s'axe plus particulièrement sur le module d'ordonnancement des réponses qui traite d'un très grand nombre de réponses à valider et ordonner. Cet ordonnancement obtient de bons résultats puisque, quand la bonne réponse est contenue dans l'ensemble des candidats extraits, elle se trouve dans les cinq premières positions dans 75 % des cas.

Le document se termine naturellement par une conclusion globale et différentes perspectives à plus ou moins long terme.

Chapitre 1

La validation de réponses

Sommaire

Ce chapitre a pour but de présenter les différentes approches proposées pour valider des réponses ainsi que leur intégration dans les systèmes de questions réponses. Afin de bien comprendre en quoi consiste cette tâche, il est nécessaire de commencer par décrire les systèmes de questions réponses dans lesquels la validation sera placée. Les systèmes de questions réponses (SQR) recherchent la réponse à une question posée en langue naturelle dans un ensemble de documents. Nous verrons en section 1.1 les différents types de questions et de réponses ainsi que le fonctionnement de ces systèmes qui se modélisent souvent comme un mécanisme de filtres allant de nombreux documents à une réponse concise de quelques mots.

La validation de réponses a pour but, quant à elle, de vérifier que la réponse renvoyée est valide. Pour qu'une réponse soit valide, il faut qu'elle soit correcte et justifiée par le passage de texte duquel elle a été extraite. La section 1.2 présente plus formellement cette tâche et notamment les manières de l'évaluer qui s'intéressent davantage à la découverte de réponses valides qu'à celles de réponses non valides. La validation de réponses peut aussi être vue comme un cas particulier de l'implication textuelle. Celle-ci vise à reconnaître les cas où le sens d'un court texte, une affirmation en quelques mots, peut être déduit d'un texte plus long, quelques phrases. Cette problématique est applicable dans les systèmes de questions réponses mais se retrouve également dans d'autres tâches de traitement automatique des langues comme la paraphrase ou le résumé de texte.

Les approches pour la validation de réponses ont été préalablement conçues pour la détection de l'implication textuelle et certaines méthodes ont été appliquées aux deux tâches. Deux grands types d'approches existent : les approches analytiques s'appuyant sur un formalisme de représentation et des analyses profondes syntaxiques ou sémantiques, présentées section 1.4, et les approches fondées sur la combinaison par apprentissage de différents critères locaux d'ordre lexical et syntaxique, présentées en section 1.5.

Lors de la présentation de ces différentes approches, nous nous attacherons particulièrement à les mettre en relation avec notre problématique qui consiste à étudier la validation de réponses en français pour l'utiliser dans un système de questions réponses. Le fait de travailler sur le français réduit la possibilité d'utilisation de certaines méthodes. Par exemple, certaines approches nécessitent de disposer d'un analyseur syntaxique ou sémantique robuste. Ces ressources sont disponibles en anglais mais beaucoup moins libre d'accès en français sur lequel il n'existe notamment pas de base de connaissance sémantique analogue à WordNet [Fellbaum 1998].

Les approches analytiques recherchent une similarité globale entre la forme syntaxique de la question et celle du passage en considérant par exemple les transformations à effectuer pour passer d'une forme à une autre. Ce type de formalisme permet aussi de définir des vérifications telles que la recherche de liens syntaxiques communs au passage et à la question.

D'autres approches s'intéressent davantage à la forme sémantique de la question et des passages et, à partir de celle-ci, créent un formalisme logique à partir duquel un système de preuves peut être appliqué afin de détecter l'implication.

Les autres approches, ne s'appuyant pas sur un formalisme de représentation particulier, consistent à combiner différentes vérifications permettant de mesurer une similarité entre le passage et la question. Certaines portent par exemple sur la présence ou la répartition des termes de la question dans le passage en se fondant sur l'idée que si le passage justifie la réponse alors il doit contenir les mots de la question, ils doivent être suffisamment proches les uns des autres afin d'être effectivement reliés. D'autres vérifications portent sur la redondance de la réponse en se fiant à l'idée que si la réponse apparaît fréquemment dans des documents accompagnée des mots de la question alors elle est probablement correcte.

Après avoir présenté les systèmes de questions réponses ainsi que les systèmes de validation de réponses, nous étudierons en section 1.6 l'utilisation de cette validation dans les systèmes de questions réponses notamment pour ordonner des réponses candidates.

1.1 Les systèmes de questions réponses

Notre travail prend sa place dans le cadre des systèmes de questions réponses. Ces systèmes sont apparus lors de la campagne TREC en 1999 [Voorhees & Tice 1999]. Leur but est de répondre à des questions posées, en langue naturelle, par des utilisateurs en extrayant la réponse dans des documents. Un tel système prend en entrée une question sous forme textuelle comme « *Quelle est la date de sortie de Ratatouille ?* » et renvoie la réponse « 2007 » ainsi que le document duquel elle a été extraite et qui est censé la justifier « *Ratatouille, un film de Brad Bird sorti en salles en 2007* ».
Le but de cette section est de présenter plus précisément ces systèmes. Pour ce faire nous commençons par voir les entrées/sorties des systèmes. Leur fonctionnement est illustré en s'appuyant sur le système du LIMSI dont l'étude a servi de base à ce travail, le système FRASQUES [Ferret, et al. 2002].

1.1.1 Définition

Pour rechercher la réponse à une question posée en langue naturelle, l'information donnée sous forme de question permet de sélectionner un ensemble de documents susceptibles de contenir la réponse. Les documents sont ensuite analysés afin d'en extraire une réponse souvent courte.

Le dictionnaire Larousse nous apprend qu'une question est, entre autres, « *une demande faite pour obtenir une information, vérifier des connaissances* ». Cette définition nous fait ainsi apparaître qu'une question n'est pas forcément énoncée sous forme interrogative et que « *Citez un film de Steven Spielberg.* » est une question puisqu'elle attend une information comme « E.T. ».

En sortie, les systèmes renvoient une réponse (*Ce que quelqu'un dit, écrit ou fait pour répondre*). Celle-ci est généralement accompagnée d'un passage de texte permettant de la justifier. Par exemple à la question « *Qui a tué Henri IV ?* », un système parfait pourra renvoyer la réponse « Ravaillac » accompagnée du passage duquel la réponse a été extraite, « *Ravaillac a assassiné Henri IV.* ». Le fait d'associer un passage à la réponse a pour but de permettre à un utilisateur d'évaluer la pertinence de la réponse du système, sa validité. C'est ce processus de validation que nous étudierons. Nous considérerons une réponse comme valide si elle est appropriée (répond bien à la question), précise (donne toute l'information recherchée), concise (ne fournit pas d'informations supplémentaires) et justifiée par le passage de texte l'accompagnant.

Les recherches des différentes réponses sont généralement effectuées dans deux grands types de collections : les collections regroupant des articles de journaux et celles constituées de documents issus du Web. Les articles de journaux présentent l'avantage d'être généralement bien rédigés ce qui permet aux analyses syntaxiques et sémantiques de s'appliquer correctement.

Les documents Web correspondent à des pages Web préalablement collectées. Le fait de les collecter à l'avance permet de leur appliquer un certain nombre de prétraitements afin de rendre la recherche plus aisée. De par leur plus grand nombre, ces documents peuvent contenir davantage d'informations que les précédents. Toutefois ils peuvent se trouver sous un format difficile à utiliser sans prétraitement particulier puisque leur structure est conçue pour rendre un aspect visuel. De plus il arrive souvent qu'ils soient moins bien rédigés et contiennent plus de faute d'orthographes et de grammaire. C'est notamment le cas des blogs. Ces raisons font que les analyses en profondeur s'appliquent plus difficilement.

Certains systèmes, par exemple RITEL [Rosset 2008] ont été développés sur l'oral. Ainsi, les documents sur lesquels les recherches sont effectuées correspondent à des retranscriptions écrites de discussions orales. Ce type de documents peut également poser problème pour les traitements car là encore des structures spéciales sont considérées. Ce type de documents n'ont pas fait l'objet d'études spécifiques dans ce travail.

La réponse est extraite d'un passage de texte l'accompagnant et censé la justifier. Ce passage, généralement court, de l'ordre de quelques phrases, est utilisé afin de permettre à l'utilisateur d'évaluer la qualité de la réponse. La justification n'est pas toujours complète et peut faire appel à certaines connaissances. Ainsi à la question « *Quel pays l'Irak a t-il envahi en 1991 ?* », un passage de texte justificatif peut être « *En 1991, l'Irak envahit le Koweït.* ». Dans cet exemple, il n'est pas dit spécifiquement que le Koweït est un pays. Comme cette information fait partie des connaissances générales de chacun, nous pouvons supposer qu'elle est connue par l'utilisateur et qu'il n'est donc pas nécessaire qu'elle soit contenue dans le passage. Mais c'est une information à justifier par un système étudiant la validité de la réponse. D'autres mécanismes tels que l'inférence peuvent rendre une information difficile à vérifier. Par exemple, pour la question « *Qui a assassiné Henri IV ?* » et le passage « *Ravaillac a poignardé Henri IV* » nous pouvons inférer que le fait de poignarder quelqu'un entraîne sa mort même si ce n'est pas toujours le cas.

Plusieurs types de questions existent et sont souvent reliées à un type de réponses particulier ainsi qu'à des méthodes différentes pour y répondre. Falco [2009] reprend ainsi une typologie de questions issue des différentes campagnes d'évaluation des systèmes de questions réponses, typologie que nous allons suivre pour définir les types de questions suivantes :

- **les questions de définition** cherchent à obtenir la définition d'un objet « *Qu'est ce que l'Atlantis ?* » ou demandent une information sur une personne ou une organisation comme « *Qui est Nicolas Sarkozy ?* ». Pour ce type de questions, certaines réponses peuvent être courtes (« président de la république ») mais d'autres peuvent être beaucoup plus élaborées. La réponse à la dernière question pourrait ainsi parler de son parcours politique, de son enfance ... L'extraction des réponses à ce type de question est souvent fondée sur l'utilisation de patrons d'extraction comme « Nicolas Sarkozy est <Réponse> ». La validation de réponses peut elle aussi être simple en retenant la réponse la plus souvent extraite ou provenant d'un patron fiable ;
- **les questions factuelles** cherchent à obtenir une information précise concernant un événement ou une entité comme une date (« *En quelle année la France est-elle devenue championne du monde de football ?* »), un lieu (« *Dans quel ville se trouve la tour Eiffel ?* ») ou une personne (« *Qui a tué Henri IV ?* »). Certaines questions attendent une entité nommée d'un type particulier en retour et d'autres non (« *Citer un succès de Michael Jackson.* ») mais attendent généralement dans ce cas un groupe nominal. Le mécanisme visant à valider la réponse à une question de ce type peut être assez complexe et fait l'objet de cette thèse.
- **les questions de type listes** attendent un certain nombre de réponses, ce nombre étant spécifié (« *Qui sont les 7 nains ?* [1] ») ou non (« *Quels sont les pays de l'union européenne ?* »). Le mécanisme visant à valider une réponse peut être perçu comme une partie du traitement des questions factuelles de type liste pour lesquelles il est aussi nécessaire de s'assurer que toutes

[1] Dans l'histoire originelle ils n'étaient pas nommés. Simplet, Prof ou Grincheux sont des pures inventions de Walt Disney.

les réponses sont bien fournies ;

- **les questions booléennes** attendent généralement comme réponse soit OUI soit NON. C'est par exemple le cas de la question « *Existe t-il des petits hommes verts sur Mars ?* ». Le mécanisme visant à valider ces réponses peut se rapprocher de l'implication textuelle. En effet le passage justificatif correspond au texte et la forme déclarative de la question à l'hypothèse, ici « *Il existe des petits hommes verts sur Mars* » ;
- **les questions enchaînées** imitent un dialogue que pourrait avoir une personne avec une machine. Ainsi l'utilisateur commence par poser une première question (« *Où se trouve le musée du Louvre ?* »), le système trouve une réponse (« à Paris »). Et l'utilisateur pose une seconde question portant sur la réponse (« *Qui est le maire de cette ville ?* ») ou sur un élément de la question précédente (« *Quand fut-il construit ?* »). Une difficulté consiste alors à détecter l'élément auquel il est fait référence (la dernière réponse, un élément de la question précédente ...). La thèse de Séjourné [2009] traite plus particulièrement de ce problème ;
- **les questions complexes** attendent une réponse plus élaborée et dont la taille peut aller jusqu'à un paragraphe. Elles peuvent être séparées en plusieurs catégories. Certaines questions demandent ainsi l'opinion d'une personne concernant un événement (« *Que pensez vous du cumul des mandats ?* ») et d'autres une explication comme les questions commençant par les pronoms interrogatifs « comment » (« *Comment faire une omelette ?* ») et « pourquoi » (« *Pourquoi la terre tourne ?* »). Ces questions étant très différentes des autres et relevant d'un mécanisme de recherche particulier ne seront pas considérées par la suite.

Cette étude ainsi que les systèmes de questions réponses présentés se placent en domaine ouvert. Les questions peuvent faire référence à n'importe quel sujet et portent le plus souvent sur des connaissances encyclopédiques. A l'inverse, les systèmes s'appliquant en domaine fermé, par exemple le domaine médical, répondent à un certain type de questions. De plus, dans de tel domaine, il existe souvent des ressources particulières. Ainsi, en domaine médical l'ontologie MeSH [Lipscomb 2000] permet de répertorier toutes les maladies et traitements.

1.1.2 Fonctionnement

Les systèmes de questions réponses ont donc pour but d'extraire la réponse à une question dans un ensemble de documents. Pour ce faire, une première étape vise à extraire les informations principales de la question et notamment ses mots clés. Un moteur de recherche cherche ensuite les documents censés contenir la réponse à l'aide des mots clés obtenus. Puis les documents sont examinés afin de détecter des réponses candidates qui sont enfin ordonnées avant d'être fournies à l'utilisateur.

Afin de bien comprendre les systèmes de questions réponses, il est nécessaire d'expliquer leur fonctionnement plus en détail. De nombreux systèmes de questions réponses existent et effectuent tous des traitements différents. Toutefois, ils suivent une architecture globale similaire qui peut être décomposée en cinq étapes :

1. l'analyse de la question permettant d'en obtenir un certain nombre d'éléments utiles pour la suite tels que les mots clés et le type de réponse attendu ;
2. la recherche de documents qui fournit ces termes à un moteur de recherche afin de collecter un ensemble de documents susceptibles de contenir la réponse ;

3. la sélection de courts passages de textes, de une à quelques phrases, issus de ces documents et les plus à même de contenir la réponse c'est-à-dire les plus similaires à la question ;
4. l'extraction de la réponse depuis ces passages ;
5. l'ordonnancement des réponses afin que la meilleure se trouve en première position.

Le système peut donc se voir comme un mécanisme de filtre partant d'un grand ensemble de documents pour arriver à une réponse très concise de l'ordre de quelques mots. Pour comprendre le fonctionnement précis, nous suivrons dans la suite de cette section le fonctionnement du système FRASQUES [Ferret et al. 2002] dont l'architecture est rappelée dans la figure 1.1.

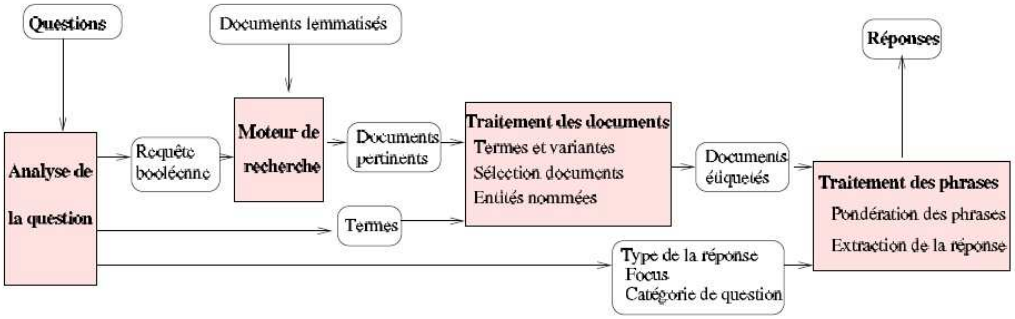

FIG. 1.1 – Architecture du système FRASQUES

1.1.2.1 Analyse des questions

La première étape d'un SQR consiste en l'analyse des questions afin d'extraire de nombreuses informations utilisées dans les étapes suivantes. Dans tous les systèmes de questions réponses, l'analyse des questions extrait un ensemble de mots importants de la question, dits mots clés. Lors de la recherche de documents, ces mots clés sont fournis au moteur de recherche afin d'obtenir les documents.

Chaque système extrait d'autres informations, différentes selon l'approche suivie. Le système FRASQUES extrait ainsi :

– **le type de réponse attendu**. Ce type correspond à l'entité nommée attendue en réponse et peut prendre les valeurs lieu, personne, date, organisation. Ces valeurs correspondent au type défini pour la classification des entités nommées MUC [Grishman & Sundheim 1995]. Ce critère est utilisé par tous les systèmes de questions réponses mais les types sont variables selon les systèmes et sont liés aux types d'entités nommées que le système sait reconnaître. Pour trouver ce type, l'analyse tient compte de la structure de la question et de listes de marqueurs. L'analyse syntaxique de la question permet notamment d'obtenir le type de réponse à des questions dont le pronom interrogatif est « Qui » puisque ces questions peuvent attendre une « personne » en réponse (« *Qui a eu cette idée folle un jour d'inventer l'école ?* ») mais peuvent aussi chercher à définir une personne (« *Qui est Jacques Chirac ?* »). Le type dépend du verbe suivant le pronom interrogatif ;

- **le type spécifique de réponse attendu** définit le type de la réponse attendue. Par exemple, la question « *Quel président a succédé à Jacques Chirac ?* » a comme type d'entité nommée attendu « personne » et comme type spécifique « président ». Ce type est directement présent dans certaines questions ;
- **le focus** est l'élément primordial et l'objet à propos duquel une information est demandée. De ce fait la réponse devrait lui être reliée dans les textes. Dans le système FRASQUES, il correspondait initialement à l'entité désignée par un nom sur laquelle porte la question. Il a été complété dans la thèse de El Ayari [2009] par l'événement sur lequel porte la question. Deux exemples sont ainsi présentés. À la question « *Où se situent les îles marquises ?* » le focus est « îles marquises » alors que pour « *Quand le pont de Normandie a-t-il été inauguré ?* » le focus est « inaugurer » ;
- **la catégorie de la question** détermine le type de relation liant la réponse et le focus ou le type spécifique. La méthode d'extraction de la réponse diffère d'une catégorie à l'autre. Ainsi, FRASQUES s'appuie sur l'utilisation de patrons d'extraction liés aux catégories de questions pour certaines questions et sur une reconnaissance d'entités nommées pour d'autres.

1.1.2.2 Recherche des documents pertinents

1.1.2.2.1 Prétraitement des documents Les systèmes de questions réponses cherchent le plus souvent la réponse à une question dans des documents déjà connus des systèmes. Il est ainsi possible de leur appliquer, préalablement à leur utilisation, un ensemble de traitements visant à rendre les processus suivants plus rapides ou à normaliser les documents.

Ainsi, FRASQUES rajoute pour chaque mot sa forme lemmatisée et sa catégorie morphosyntaxique (verbe, nom, adjectif ...) afin de rassembler différentes variations d'un même mot. Les entités nommées présentes dans les documents sont aussi indiquées afin de faciliter l'extraction des réponses. Les documents étant souvent longs, un autre prétraitement consiste également à les découper en plusieurs parties de tailles similaires afin d'améliorer la recherche d'information.

Lors de la recherche des documents, le mécanisme de recherche s'appuie sur un index indiquant pour chaque terme les documents dans lesquels il se trouve. Les index dépendent des systèmes de questions réponses les utilisant. Certains SQR comme FRASQUES utilisent un seul index dans lequel toutes les entités des documents sont indexées. Dans d'autres systèmes plusieurs index sont utilisés, QRISTAL [Laurent, et al. 2005] en considère 86. Par exemple, certains portent sur les entités nommées et d'autres sur les noms propres.

1.1.2.2.2 Découverte des documents pertinents Les mots importants de la question détectés lors de son analyse sont fournis à un moteur de recherche afin d'obtenir les documents pertinents. Ces documents peuvent être, comme nous l'avons vu précédemment, préalablement traités et indexés. Le moteur utilisé pour le système FRASQUES est Lucene [Hatcher & Gospodnetic 2004]. Afin de trouver un maximum de documents, la recherche tient compte des synonymes des mots de la requête et considère les mots sous leurs formes lemmatisées.

1.1.2.3 Détection et pondération des passages

Cette étape a pour but d'analyser les documents détectés afin d'en extraire les passages les plus à même de contenir la réponse. Les passages sont de petite taille puisqu'ils sont constitués d'une à trois phrases.

Pour sélectionner les meilleurs passages, une pondération est effectuée. Elle porte sur la présence des mots de la question dans les différents passages ce qui permet de reconnaître les passages traitant des mêmes informations que la question. Afin de refléter l'importance des termes, tous n'ont pas le même poids. Par exemple les entités nommées ont un poids plus élevé qu'un simple nom. De plus les variations des différents mots sont considérées et ont un poids plus faible que les termes présents sous la forme initiale. Les schémas de pondération varient d'un système à l'autre et sont liés aux informations provenant des questions.

1.1.2.4 Extraction de la réponse

Cette étape extrait un ensemble de réponses possibles depuis les passages obtenus.

L'extraction des réponses diffère suivant le système considéré. Dans le système FRASQUES, une réponse est extraite par passage, l'extraction variant en fonction de la question. Si la question attend en retour une entité nommée alors la réponse extraite sera l'entité du type attendu la plus proche des mots de la question. Par exemple, en considérant la question « *Qui a réalisé Robin des bois prince des voleurs ?* » et le passage « *Robin des bois prince des voleur de Kevin Reynolds avec Kevin Costner* », « Kevin Reynolds » et « Kevin Costner » sont marqués comme étant des personnes lors du prétraitement des documents. « Kevin Reynolds » est extrait car c'est la personne la plus proche des mots de la question.

Si la réponse n'attend pas en retour une entité nommée alors l'extraction s'effectue à l'aide de règles d'extraction appelées patrons d'extraction. L'exemple suivant permet de mieux appréhender cette notion. A la question « *Qu'est ce qu'un airbus ?* » le passage « *l'avion, un airbus* » indique qu'un airbus est un avion et le patron d'extraction « réponse, FOCUS » avec FOCUS correspondant à « un airbus » permet d'extraire la réponse du passage. Les patrons utilisés varient d'une catégorie de question à une autre et s'axent autour du focus ou du type spécifique. De telles règles ne peuvent pas s'appliquer dans tous les cas car elles modélisent des relations syntaxiques locales et ainsi, il n'est pas possible d'extraire de réponses lorsque leur formulation est plus complexe.

1.1.2.5 Ordonnancement de réponses

Le mécanisme d'extraction permet de retenir plusieurs réponses. Afin de proposer les meilleures réponses en tête il est nécessaire de les ordonner. Les critères permettant cet ordonnancement peuvent aller des plus simples, redondance de la réponse, aux plus complexes utilisant par exemple des modules de validation de réponses. La section 1.6 présente plus en détail les différentes approches permettant d'ordonner ou réordonner les réponses.

Le système FRASQUES utilise une pondération calculée à partir de différents scores fournis par les étapes antérieures : la sélection des passages fournit un premier score et l'extraction des réponses un second, qui tient compte soit de la distance des termes par rapport à la réponse soit de la confiance

dans le patron d'extraction utilisé. La réponse qui semble la meilleure est donc celle ayant obtenu les meilleurs scores totaux. À cela se rajoute le fait qu'une réponse soit extraite de plusieurs documents.

De manière générale, les points les plus problématiques sont la sélection des passages et l'extraction et l'ordonnancement de réponses. L'extraction de réponses pourrait être améliorée afin d'identifier davantage de réponses. Le module d'ordonnancement de réponses pourrait lui aussi être amélioré afin de considérer davantage le contenu des passages justificatifs.

1.1.3 Évaluation

De nombreuses campagnes d'évaluation permettent d'évaluer les systèmes de questions réponses parmi lesquelles on peut noter : CLEF [Forner, et al. 2010], TREC [Voorhees & Tice 1999] et EQueR [Grau 2005]. Lors de ces campagnes, les évaluateurs fournissent aux différents systèmes participant le même ensemble de questions et de documents desquels les réponses peuvent être extraites. Les participants fournissent une ou plusieurs réponses à chaque question. Ces campagnes permettent d'évaluer les différentes approches en comparant les résultats obtenus sur les mêmes données. Elles permettent également de constituer des corpus sur lesquels les systèmes peuvent s'entraîner.

Deux grandes catégories d'évaluations existent. La première part d'un ensemble de réponses attendues pour les différentes questions. L'évaluation consiste alors à savoir si la réponse obtenue est correcte et correspond à la réponse attendue. Dans ce cadre, la réponse est jugée pertinente si elle est extraite d'un document identifié comme pouvant la justifier.

Dans le second cas, les passages justificatifs sont seuls considérés afin de s'assurer que la réponse est justifiée et différentes notes peuvent être fournies : la réponse est correcte et justifiée, la réponse est correcte mais n'est pas justifiée et la réponse n'est pas correcte.

Afin d'effectuer les évaluations, différentes métriques sont utilisées. Les deux premières calculent la proportion de bonnes réponses données en première position ou parmi les N premières, souvent trois ou cinq. La proportion de bonnes réponses en première position correspond à la situation où le système ne renvoie qu'une réponse par question. La mesure est alors la proportion de bonnes réponses renvoyées (accuracy). L'utilisation des N meilleures réponses considère plusieurs réponses par question en tenant compte du rang, car il est plus intéressant pour un utilisateur qu'une bonne réponse soit en seconde position qu'en cinquième.

Dans le cas où il y a plusieurs réponses par question, la mesure le plus souvent utilisée est le MRR (Mean Reciprocal Rank). Elle favorise le fait que la bonne réponse soit renvoyée en première position et utilise le rang de la première réponse correcte. Sa formule est la suivante :

$$MRR = \frac{1}{Nbquestions} * \sum \frac{1}{rank_i}$$

Elle correspond à la moyenne de l'inverse du rang de la bonne réponse pour les différentes questions.

Avec cette mesure, si une bonne réponse est donnée en première position alors son poids sera de 1, si elle est donnée en seconde elle aura un poids de $\frac{1}{2}$ et ainsi de suite. Souvent seuls les premiers rangs (3 à 5) sont comptabilisés, car après de ces rangs une bonne réponse ne sera pas considérée par l'utilisateur.

Pour terminer sur l'évaluation, voyons maintenant trois exemples sur le français qui correspondent aux deux grands types d'évaluations : les campagnes EQueR et CLEF, portant sur les articles de journaux et la campagne créée dans le cadre du projet Quæro traitant des documents issus du Web.

La campagne EQueR [Grau 2005] était une campagne portant sur 500 questions et un ensemble de 1,5 Go de données concernant des articles du journal le Monde et le Monde diplomatique de 1992 à 2000 ainsi que les textes du Sénat de 1996 à 2000.

La campagne CLEF a eu lieu de 2004 à 2008 sur le français. En 2006 [Magnini, et al. 2006], la campagne était constituée de 200 questions et de 176 000 documents correspondant aux articles du journal Le Monde et aux dépêches SDA de 1994 et 1995.

L'évaluation du projet européen Quæro [Quintard, et al. 2010] a lieu depuis 2008 et porte sur un ensemble de 500 000 documents issus du Web. Ces documents sont extraits par le moteur de recherche Exalead[2]. Ils sont obtenus par un mécanisme visant à mémoriser les pages renvoyées par les utilisateurs du moteur de recherche.

Ces différentes campagnes permettent de comparer le comportement des systèmes sur des documents de natures différentes. Le système ayant obtenu les meilleurs résultats toutes langues confondues, LCC [Moldovan, et al. 2002], effectue 83 % de bonnes détections lors de la campagne d'évaluation sur l'anglais TREC 11 [Voorhees 2002]. Il se fonde sur un formalisme logique sur lequel il effectue un mécanisme par preuve. Un tel système a besoin de bases de connaissances sémantiques importantes permettant la transformation du texte.

Les résultats sur les autres langues européennes sont moins bons, ce qui est sans doute dû à l'absence de bases de connaissances similaires à WordNet. Le système ayant obtenu les meilleurs résultats sur le français, QRISTAL [Laurent, et al. 2010], effectue 68 % de bonnes détections sur les données de la campagne d'évaluation CLEF 2006 [Magnini et al. 2006]. Il tient compte d'une indexation multicritères, d'un grand nombre d'outils de traitement automatique de la langue tels que la résolution d'anaphore ainsi que d'un ensemble de règles produites par une analyse manuelle. Ces différentes évaluations sont effectuées sur des documents correctement écrits, le plus souvent des articles de journaux.

Le fait d'effectuer des recherches sur les documents issus de pages Web détériore les résultats puisque les textes ne sont pas aussi correctement rédigés ; cela pose particulièrement problème aux systèmes fondés sur des traitements linguistiques assez importants. Ainsi, lors des évaluations Quæro, seulement 27,5 % à 50 % des questions obtenaient une bonne réponse en première position sur le français, lors de l'évaluation 2008. Le système QRISTAL a obtenu les meilleurs résultats avec 50 % de bonnes détections.

Lors de sa participation à la campagne EQueR [Grau, et al. 2006], le système FRASQUES a découvert 26 % de bonnes réponses et 42 % de passages justificatifs corrects. L'analyse de ces résultats a montré qu'il était nécessaire d'améliorer l'extraction et l'ordonnancement des réponses dans ce système.

Les systèmes n'obtiennent pas les mêmes résultats sur toutes les questions. Les questions sur lesquelles les meilleurs résultats sont obtenus sont les questions de définition (« *Qui est Robert De Niro ?* »). Cela peut s'expliquer car bien souvent l'extraction de la réponse s'appuie sur des patrons

[2]Exalead : http ://www.exalead.com/search/

d'extraction qui s'appliquent bien. En revanche, les questions demandant la raison d'un événement ou la manière dont il a été réalisé sont celles pour lesquelles les systèmes sont les moins performants. Ces questions posent notamment problème pour la détection de la réponse candidate. Lors des différentes campagnes d'évaluation, ces questions sont cependant très peu présentes, les questions les plus fréquentes étant les questions factuelles et les questions de définition. Après la présentation du contexte général de ce travail, nous allons maintenant développer ce que signifie plus précisément valider des réponses et les approches existantes permettant d'y répondre.

1.2 Validation de réponses

En 2006, une nouvelle voie visant à améliorer les systèmes de questions réponses a été proposée lors de la campagne CLEF [Peñas et al. 2006]. Cette innovation consistait à reconnaître automatiquement que la réponse proposée par les systèmes est valide, dans le but d'automatiser l'évaluation des systèmes de questions réponses.

Lors de l'extraction de la réponse par les différents systèmes, un grand nombre de documents est examiné afin d'en extraire les réponses les meilleures possibles. Toutefois, comme de nombreuses réponses sont considérées, il n'est pas possible d'effectuer des vérifications coûteuses en temps à ce niveau. En revanche, en se limitant à quelques réponses il est possible d'effectuer de telles actions. Par exemple, un système de questions réponses pourrait renvoyer les données suivantes à la question « *Qui est le père de la reine Elisabeth 2 ?* ».

Question : Qui est le père de la reine Elisabeth II ?
Réponse : François Mitterrand
Passage : François Mitterrand et la reine Elisabeth II ont inauguré le tunnel sous la manche.

Dans l'exemple précédent la réponse donnée n'est clairement pas la bonne car le passage ne traite pas de la parenté de la reine Elisabeth II mais de sa rencontre avec le président français. Remarquons toutefois que cette mauvaise réponse peut être extraite par un système de questions réponses cherchant un nom de personne qui soit le plus proche des mots de la question. En examinant le passage et notamment son sens il serait possible de se rendre compte que la réponse n'est pas correcte. La validité de la réponse doit donc être détectée en examinant le passage de texte duquel la réponse a été extraite, qu'on appellera par la suite le passage justificatif.

Pour comprendre la validation de réponses nous introduisons la notion de triplet réponse.

Définition : On appelle **triplet réponse** un triplet constitué :
– d'une **question** ;
– d'une **réponse** proposée à cette question ;
– d'un fragment de texte, appelé **passage justificatif**, constitué de quelques phrases et duquel la réponse a été extraite.

Ainsi l'exemple suivant est un triplet réponse.

Question : Qui a assassiné Henri IV ?
Réponse : Ravaillac
Passage : Henri IV est passé de vie à trépas par les mains de Ravaillac.

Dans ce cas la réponse est effectivement correcte et justifiée par le passage. Ces deux notions nous donnent une première définition de la validation de réponses. Nous pouvons toutefois remarquer que le passage énonce bien l'assassinat d'Henri IV mais n'est pas formulé de la même manière.

Dans certains cas l'information demandée dans la question se trouve effectivement dans le passage mais il est besoin d'une réelle inférence pour la percevoir. Ce type d'inférence nécessite des connaissances externes et une réelle compréhension du passage.

Question : En quelle année Richard cœur de lion est-il mort ?
Réponse : 1199
Passage : Le 26 mars 1199, Richard cœur de lion assiège le château de Châlus Chabrol. Il est atteint par un carreau d'arbalète tiré par un chevalier de petite noblesse limousine, Pierre Basile. La flèche est retirée mais la gangrène gagne le corps du roi. Il succombe onze jours après sa blessure.

Dans ce cas, il y a une inférence temporelle pour reconnaître que la mort du roi Richard a lieu la même année que sa blessure. Il faut également détecter que le verbe succomber est synonyme de mourir.

Question : Quel président a succédé à Jacques Chirac ?
Réponse : Nicolas Sarkozy
Passage : Nicolas Sarkozy était le ministre de l'intérieur pendant le mandat de Jacques Chirac.

Dans cet exemple la réponse est effectivement correcte mais un système ou une personne ne connaissant pas cette information ne pourra le déduire du passage puisqu'il ne précise pas que Nicolas Sarkozy a succédé à Jacques Chirac. Nous considèrerons alors que la réponse n'est pas valide. Il ne suffit donc pas que la réponse soit pertinente pour qu'elle soit valide puisqu'il faut que l'information contenue dans la question se retrouve également dans le passage justificatif ou puisse en être déduit. Un passage qui aurait justifié la réponse peut être :

Passage : Nicolas Sarkozy pris la succession de Jacques Chirac.

Dans ce cas le passage justifie bien la réponse auprès d'un utilisateur puisqu'il énonce le passage de flambeau entre les différents présidents. Toutefois, le passage n'énonce pas que Nicolas Sarkozy est un président, cela fait partie des connaissances générales de chacun. Mais du point de vue d'un système, afin de reconnaître que la réponse est valide, celui-ci devra la rechercher en dehors du passage justificatif. De manière générale, un passage est dit justificatif si une personne lisant ce passage peut en déduire que la réponse est correcte. Comme cette personne possède de nombreuses connaissances faisant partie de sa culture générale, un certain nombre d'informations ne seront pas obligatoirement présentes dans le passage. En revanche un système de validation de réponses devra vérifier ces éléments. Une question pertinente peut être de savoir quels types d'informations sont connus de tous et donc ne devront pas obligatoirement se trouver dans le passage mais devront en revanche être vérifiés en dehors de ceux-ci.

L'implication textuelle est une tâche qui consiste à détecter si le sens d'un passage de texte peut être déduit de celui d'un autre passage. La validation de réponses et l'implication textuelle sont deux tâches fortement liées. La validation de réponses traite en effet de triplets réponse composés d'un

passage de texte, d'une question et d'une réponse. À l'aide de règles de réécriture la question et la réponse peuvent se transformer en une hypothèse.
Par exemple, pour la question « *Qui a assassiné Henri IV ?* » et la réponse « Ravaillac » l'hypothèse « *Ravaillac a assassiné Henri IV* » peut être obtenue en remplaçant simplement le pronom interrogatif « Qui » par la réponse. Une fois les données transformées le travail devient le même. En effet pour montrer que le passage justifie la réponse il faut prouver que l'hypothèse est impliquée par le passage.

De ce fait, nous allons nous appuyer sur les deux tâches (validation et implication) afin de proposer notre définition de la validation de réponses.

1.2.1 Validation et AVE

Les systèmes de validation de réponses furent évalués lors des campagnes d'évaluation AVE qui eurent lieu de 2006 à 2008 [Peñas et al. 2006 ; Peñas, et al. 2007 ; Rodrigo, et al. 2009]. Elles étaient organisées dans le cadre de la campane CLEF.

Lors de la première campagne d'évaluation les systèmes recevaient en entrée un ensemble de paires constituées d'un passage justificatif et d'un court passage de texte appelé hypothèse. Cette hypothèse correspond à la forme déclarative de la question à laquelle la réponse a été ajoutée. L'hypothèse a été construite afin de se rapprocher de la tâche de détection d'implication textuelle. Les systèmes devaient alors détecter si le sens de l'hypothèse pouvait être déduit de celui du passage justificatif en renvoyant OUI le cas échéant et NON sinon.

Les campagnes suivantes se sont appuyées sur un ensemble de triplets réponses (question, réponse, passage justificatif) que les systèmes devaient valider ou non en renvoyant soit OUI soit NON.

Dans cette campagne, la validation de réponses est vue comme une tâche visant à évaluer automatiquement les réponses extraites par des systèmes de questions réponses, donc les SQR eux mêmes. Ainsi, les organisateurs de cette campagne considèrent qu'une réponse est valide si elle est correcte et justifiée par le passage textuel. Toutefois, notons qu'une telle définition ne peut pas être retenue dans le cadre de l'intégration de cette tâche dans un SQR pour l'améliorer car on ne peut pas savoir si une réponse est correcte à l'avance.

Afin d'être au plus près des conditions réelles, les entrées correspondent aux résultats fournis par les participants à la campagne QA@CLEF. Cette dernière est une campagne d'évaluation des systèmes de questions réponses dans laquelle les participants doivent fournir la réponse à différentes questions. Le fait d'utiliser ces données permet de voir l'utilité de la validation de réponses pour les systèmes de questions réponses existants de manière plus appropriée que si le corpus avait été créé manuellement. De par cette spécificité, le corpus contient davantage de triplets réponses non valides que de triplets réponses valides (environ 23 % des réponses totales sur l'anglais). Cela s'explique par le fait que les systèmes de questions réponses sont encore à améliorer et que chaque système renvoie plusieurs réponses par question. Le lien entre ces deux tâches peut se retrouver dans la figure 1.2 reprise de [Peñas et al. 2007]. Elle montre ainsi que les sorties du système de questions réponses correspondent aux entrées d'un système de validation de réponses. Ces sorties sont également évaluées pour savoir si les réponses sont valides ou non ce qui permet de les comparer avec la valeur détectée par le système de validation.

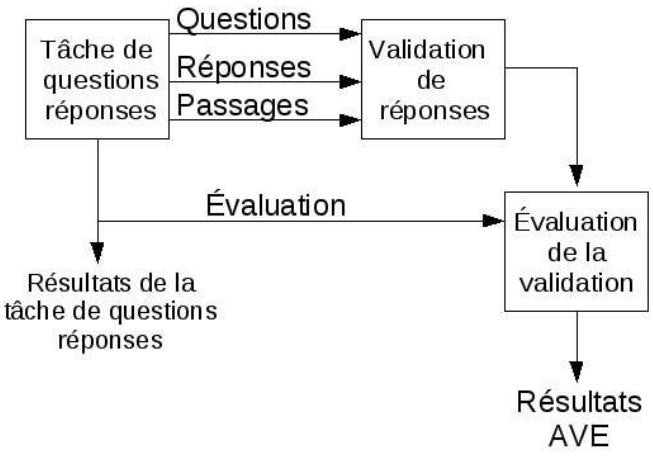

FIG. 1.2 – Les données de la campagne AVE repris de [Peñas et al. 2007]

La campagne AVE était présente sur différentes langues : l'anglais, l'allemand, le bulgare, l'espagnol, le français, l'italien et le portugais. Toutefois, comme les proportions de réponses OUI et NON varient d'une langue à l'autre, il est délicat de comparer deux systèmes traitant des langues différentes. Pour notre part, nous nous concentrerons exclusivement sur le français.

De par la méthode de création du corpus, le nombre de données à évaluer dépend de la participation à la campagne QA@CLEF. Ainsi en 2006, date de création de la campagne AVE, les participants à QA@CLEF étaient nombreux ce qui a permis d'obtenir, sur le français, un grand nombre d'exemples puisque 3267 triplets réponses furent proposés. Malheureusement, les années suivantes les systèmes de questions réponses participant à QA@CLEF étaient bien moins nombreux et moins de réponses leur ont été demandées. Ainsi très peu de nouvelles données ont été créées pour permettre d'évaluer les systèmes. En 2008 il y avait seulement 108 questions et 199 triplets réponses ce qui rend les résultats bien moins significatifs.

Lors de ces campagnes, la validation de réponses était principalement considérée comme une tâche visant à améliorer les systèmes de questions réponses en s'appliquant comme un filtre à la suite d'un système de questions réponses. Ce filtre a pour but de n'envoyer à l'utilisateur que les réponses valides. Une mauvaise réponse vue comme correcte à tort est donc plus pénalisant qu'une bonne réponse reconnue comme fausse.
C'est pour cette raison que les évaluations ne sont effectuées que sur les réponses OUI grâce à trois mesures : la précision, le rappel et la f-mesure.

- **la précision** est la proportion de réponses correctes parmi les réponses vues comme valides ;
- **le rappel** est le rapport entre le nombre de réponses reconnues correctement comme valides et le nombre total de réponses valides ;
- **la f-mesure** permet de combiner ces deux mesures.

$$précision = \frac{\#\ réponses\ OUI\ correctes\ renvoyées}{\#\ réponses\ OUI\ données} \qquad rappel = \frac{\#réponses\ OUI\ correctes\ renvoyées}{\#\ réponses\ OUI\ attendues}$$

$$f - mesure = \frac{2 * précision * rappel}{précision + rappel}$$

Pour mieux comprendre ces mesures, considérons l'exemple suivant : La base contient 10 triplets dont seulement 3 valides. Un système qui répondrait OUI à quatre réponses, 2 fois correctement et 2 fois non aura une précision (proportion de réponses correctes) de $\frac{1}{2}$ et un rappel (proportion de réponses trouvées) de $\frac{2}{3}$.

La campagne AVE a donc permis d'introduire la tâche de validation de réponses en détectant une réponse comme valide si elle est correcte et justifiée par le passage de texte l'accompagnant. Toutefois rien n'est dit sur ce que devait contenir le passage justificatif pour qu'il valide la réponse.

1.2.2 Validation de réponses et implication textuelle

La validation de réponses peut être perçue comme une continuité ou une tâche spécifique de l'implication textuelle. Celle-ci consiste à détecter si le sens d'un texte peut être inféré de celui d'un autre. Glickman [2006] définit l'implication textuelle comme suit :

définition : Un texte T implique une hypothèse H si, typiquement, un humain lisant T en déduit que l'hypothèse H est sûrement vraie.

Cette définition explique donc qu'il faut que le sens de l'hypothèse se retrouve dans celui du passage ou puisse en être inféré, sans pour autant qu'elle soit écrite de la même manière ni que la relation inverse soit vérifiée. Pour bien comprendre voyons un exemple.

Passage : Les médicaments qui ralentissent ou stoppent la maladie d'Alzheimer fonctionnent mieux si vous vous les administrez tôt
Hypothèse : La maladie d'Alzheimer est traitée grâce à des médicaments

Dans cet exemple le passage implique bien l'hypothèse. Toutefois la relation inverse n'est pas vérifiée puisqu'une partie des informations est manquante.

Passage : Ravaillac se jette sur Henri IV.
Hypothèse : Ravaillac a tué Henri IV.

Dans ce cas, le passage n'implique pas l'hypothèse puisqu'il ne fait pas directement mention de la mort d'Henri IV mais de l'acte la précédant.

1.2.2.1 La campagne RTE (Recognizing Textual Entailment)

Jusqu'en 2004, la détection de l'implication textuelle existait dans différents travaux de traitement automatique des langues, sans pour autant être perçue comme une tâche à part. Elle se trouve notamment présente dans les systèmes de questions réponses, l'évaluation de paraphrases et la comparaison de documents qui consiste à détecter si deux documents traitent du même sujet. La notion d'implication textuelle a été unifiée en 2004 lors de la campagne d'évaluation RTE ([Dagan, et al. 2005 ; Glickman 2006]) qui fait partie de la campagne PASCAL [3]. RTE s'intéresse au fait

[3] http ://pascallin.ecs.soton.ac.uk/

qu'une même information puisse être contenue dans des textes différents. Les systèmes participants reçoivent en entrée un ensemble de paires texte-hypothèse et doivent décider si le sens de l'hypothèse peut être déduit de celui du passage. Lors des premières campagnes, les systèmes devaient seulement détecter si le sens de l'hypothèse pouvait être déduit de celui du texte en renvoyant « OUI » le cas échéant et « NON » sinon. Puis, à partir de la troisième campagne [Giampiccolo, et al. 2007], trois valeurs sont à donner :

- **IMPLICATION** quand le texte implique l'hypothèse ;
- **CONTRADICTION** quand le texte contredit l'hypothèse ;
- **INCONNU** si ce n'est ni l'un ni l'autre.

Les données ont été créées à partir de chacune des tâches unifiées. Dans le cas des systèmes de questions réponses, les réponses à différentes questions ont été recherchées dans des documents. Le passage est obtenu par le module de recherche d'information du système. L'hypothèse correspond, quant à elle, à la forme déclarative de la question à laquelle la réponse est ajoutée. Nous pouvons remarquer que cette manière d'obtenir les données correspond à celle effectuée lors de la campagne AVE.

Deux types de données sont ainsi collectés, les premiers correspondent à la base d'apprentissage et les seconds à la base de tests. Deux ensembles d'environ 900 paires texte-hypothèse ont été collectés et contiennent autant d'exemples positifs (le passage implique l'hypothèse) que d'exemples négatifs (il n'y a pas d'implication).

La première mesure permettant d'évaluer les systèmes est le pourcentage de détections correctes que la valeur renvoyée soit OUI, NON ou INCONNU. Elle est relativement simple mais le fait d'avoir autant d'exemples de chaque catégorie la rend pertinente ce qui n'est pas le cas dans le cadre de la validation de réponses.

La seconde mesure part du principe que les valeurs attribuées par les systèmes sont accompagnées d'un score de confiance. Elle tient donc compte de la pertinence du score de confiance afin de privilégier une valeur bien évaluée, avec un bon score de confiance, à une mauvaise évaluation avec un faible score. Pour ce faire, les valeurs sont ordonnées par ordre décroissant et des mesures tenant compte de la proportion de réponses correctes à partir des différents rangs sont appliquées.

À partir de la cinquième campagne d'évaluation [Bentivogli, et al. 2009], une nouvelle tâche est apparue. Dans celle-ci, les données ne sont plus des paires texte-hypothèse et les systèmes doivent trouver les phrases impliquant une hypothèse parmi plusieurs proposées. Les hypothèses correspondent ici à des résumés que les évaluateurs ont effectués. Les systèmes doivent alors chercher les phrases impliquant l'hypothèse parmi celles ayant permis de créer ce résumé. Cette tâche est effectuée dans deux buts : être près des données réelles et analyser l'utilité des méthodes d'implication textuelle à travers une application, celle du résumé automatique. Cette tâche pilote de RTE5 est devenue la tâche principale dans la sixième campagne [Bentivogli, et al. 2010b].

Une autre tâche pilote est apparue lors de la campagne RTE6. Elle consistait à vérifier la sortie des systèmes de peuplement de base de connaissances. Les systèmes participant à cette tâche partent d'une entité, par exemple une personne, et recherchent certaines de ses caractéristiques, comme son lieu de naissance. Cette recherche est effectuée dans des documents et l'implication textuelle per-

met de vérifier que l'information extraite *« Chris Simox est canadien »* est correcte car extraite du document.

Dans le cadre de cette thèse, nous étudierons uniquement les méthodes proposées lors de la première tâche qui considère un couple texte-hypothèse et détecte si le sens de l'hypothèse peut être déduite de celui du passage. En effet, les données de cette tâche sont variées et contiennent davantage d'inférences et de phénomènes proches de ceux rencontrés lors de la validation de réponses.

1.2.3 Définition de la validation de réponses

Différentes tâches traitent donc de la validation de réponses. Tout d'abord, dans la campagne AVE, pour être valide une réponse doit être correcte et justifiée par le passage de texte l'accompagnant. Cette définition peut être donnée car, lors de cette campagne, la validation de réponses a pour but d'évaluer les systèmes de questions réponses. Dans un cadre de validation de réponses visant à améliorer les SQR, il n'est pas possible de savoir si la réponse est correcte ou non dans l'absolu ; la valeur de la réponse ne peut être connue à l'avance (sinon la recherche serait inutile). De ce fait elle sera considérée comme correcte si elle n'est pas manifestement fausse au vu de la question posée, et elle est justifiée par le passage. Dans bien des cas cette décision fait intervenir des connaissances non explicitées par le passage.

L'analyse de l'implication textuelle a montré que pour que le passage implique l'hypothèse il est nécessaire que les informations contenues dans l'hypothèse soient présentes dans le passage ou puissent en être déduites. Ces différentes informations nous permettent de proposer la définition suivante.

Définition : Un triplet réponse, constitué d'une question, d'une réponse et d'un passage justificatif, est valide si :

- la réponse extraite peut répondre à la question posée. Dans ce sens, il est nécessaire que les informations demandées ou données par la question et contraignant la réponse soient vérifiées. Par exemple, il faut que la réponse corresponde au type attendu par la question. Ainsi pour une question comme : « *Quel réalisateur a produit Super 8 ?* » la réponse proposée doit être un réalisateur. De plus, la question portant sur la précision d'une caractéristique de l'un de ces éléments, il s'agit de vérifier que la réponse lui est effectivement liée ;
- les informations contenues dans la question doivent être retrouvées à partir du passage justificatif. Cela permet entre autres de vérifier que le passage mentionne le même événement ou la même entité que la question, c'est-à-dire qu'ils traitent du même sujet.

Le premier type de vérifications concerne le type de la réponse et par conséquent ne se trouve pas forcément dans le passage. Ces vérifications correspondent souvent aux connaissances de culture générale. Afin de les vérifier, il peut être nécessaire d'effectuer des recherches en dehors du passage justificatif, par exemple dans d'autres documents ou dans des bases de connaissances.

Afin de détecter que le sens de la question se trouve dans le passage, différentes vérifications peuvent être mises en place. Les premières consistent à retrouver les entités de la question dans le passage en vérifiant par exemple que les termes de la question s'y trouvent. Les secondes vérifications peuvent montrer que les relations liant les mots de la question se trouvent également dans le passage justificatif ou peuvent en être déduits. Pour cela, une méthode possible peut s'appuyer sur la mise

en correspondance des relations syntaxiques reliant ces mots dans la question et dans le passage justificatif. Cette vérification pourrait s'appliquer quand les mêmes informations sont présentes mais rencontrerait davantage de difficultés dans un cadre d'inférence ou de paraphrase.

La validation de réponses ayant pour but d'améliorer les systèmes de questions réponses, il est également nécessaire d'étudier ses différentes intégrations possibles.

La première consiste à effectuer une validation en dernière étape d'un système de questions réponses en ne renvoyant que les réponses reconnues comme valides. Dans le cas contraire la réponse n'est pas renvoyée à l'utilisateur et une autre réponse, valide celle là, lui est préférée. Le système renvoie une valeur binaire. Les réponses renvoyées peuvent soit être issues d'un même système de questions réponses soit correspondre à la sortie d'un autre SQR, la validation servant à alors à choisir le système à appliquer [Téllez-Valero, et al. 2010] (cf. figure 1.3).

Dans ce cadre, on peut noter que les réponses sont évaluées à l'aide du rappel, de la précision et de la f-mesure des réponses valides, comme lors des campagnes AVE.

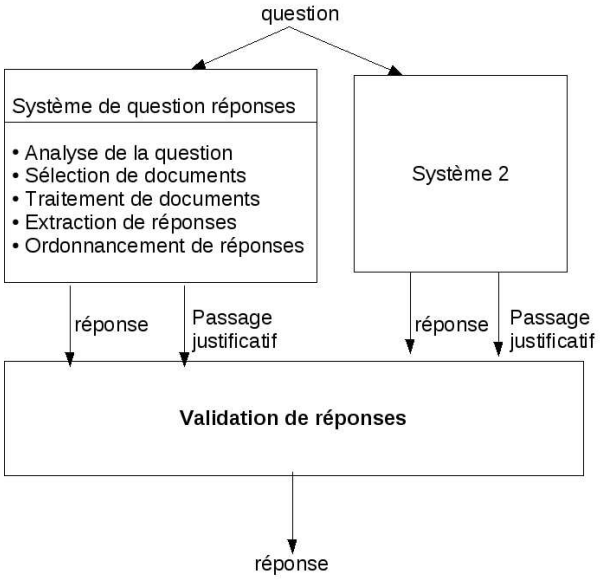

FIG. 1.3 – La validation de réponses comme filtre

La seconde utilisation de la validation consiste à valider ou ordonner les différentes réponses candidates (cf. figure 1.4). Le système renvoie alors un score indiquant sa confiance sur la validité des différentes réponses. L'ordonnancement de réponses a lieu une fois les différentes réponses extraites et vise à placer la meilleure en première position. La principale différence vient du fait qu'un score de confiance est à fournir pour chaque réponse afin de les ordonner. De plus, ces réponses n'ont pas été sélectionnées par un système de questions réponses et, de ce fait, elles sont plus souvent incorrectes car elles n'ont pas été filtrées par un module d'extraction de réponses.

Lors des évaluations AVE, une mesure d'évaluation tenait compte de l'ordonnancement de réponses. Pour cela les systèmes devaient reconnaître pour chaque question, la réponse ayant la plus de

chances d'être valide. Cette réponse est marquée comme sélectionnée. La mesure $qa\_accuracy$ est alors la proportion de réponses sélectionnées effectivement valides.

$$qa\_accuracy = \frac{\#réponses\ sélectionnées\ correctes}{\#questions}$$

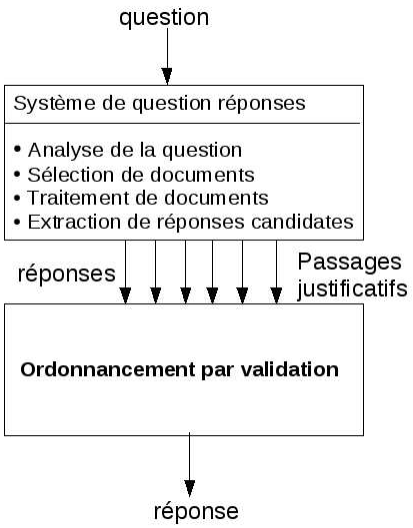

FIG. 1.4 – Utilisation de la validation de réponses pour l'ordonnancement de réponses

Nous avons donc présenté notre définition de la validation de réponses qui se base sur les informations données et attendues par la question ainsi que les différentes places possibles d'un système de validation dans un système de questions réponses. Afin de préciser ce que signifie la notion de « passage contenant la même information que la question ou permettant d'inférer cette information », nous repartirons des études faites à ce sujet.

1.3 Phénomènes sous tendant l'implication textuelle

De nombreuses études concernant l'implication textuelle ont été menées. La première d'entre elles, présentée dans [Glickman 2006], part du principe que, pour que le texte implique l'hypothèse, il faut qu'il contienne les mots de l'hypothèse à l'identique ou sous forme de variante. L'étude porte alors sur les implications lexicales et quatre phénomènes ont été révélés ainsi que leurs fréquences d'occurrences :

- les variations simples morpho-lexicales (30 %) ;
- les cas où un mot porte le sens d'un autre comme la synonymie ou l'hyponymie (33 %) ;
- les cas où une portion de phrase suggère le sens d'un mot (13 %) ;
- les cas où la référence a lieu à un niveau plus global (24 %).

Ces résultats montrent que l'appariement est délicat puisque plus d'un tiers des cas (13 % + 24 %) est très difficile à reconnaître automatiquement.

Une autre étude, présentée dans [Bar-Haim, et al. 2005], porte sur les différences entre les approches se plaçant au niveau lexical et celles utilisant la syntaxe des phrases. Dans l'approche lexicale, pour qu'il y ait implication, tous les termes de l'hypothèse doivent avoir une correspondance dans le texte, soit directe, soit avec des variations autorisées (par exemple d'ordre sémantique). Dans l'approche syntaxique, toutes les relations de l'hypothèse doivent se trouver dans le texte avec des variations autorisées, comme la paraphrase. L'étude consiste alors à simuler manuellement le fonctionnement idéal des deux types de systèmes et montre que les approches syntaxiques peuvent obtenir de meilleurs résultats que les approches lexicales (86 % contre 59 %).

Les études sur la syntaxe se sont poursuivies avec [Vanderwende & Dolan 2006]. Cette étude a pour but d'identifier les cas où la syntaxe suffit pour décider de l'implication textuelle. Ainsi, les annotateurs doivent détecter quatre cas :

- l'hypothèse peut être reconnue comme vraie grâce aux relations syntaxiques ;
- l'hypothèse est identifiable comme fausse grâce à la syntaxe car elle a un sens contraire à celui du texte ;
- la syntaxe ne suffit pas à décider si l'hypothèse est vérifiée ou non ;
- les analyses syntaxiques ne s'appliquent pas.

Les analyses ont montré que 37 % des détections pouvaient être effectuées en utilisant uniquement la syntaxe et 49 % en utilisant la syntaxe et un thésaurus. Cette étude permet donc de montrer que des vérifications supplémentaires sont à effectuer.

D'autres approches s'intéressent plus particulièrement aux formalismes permettant de détecter l'implication textuelle. Parmi elles, l'article [Clarke 2006] tient compte de deux fonctions ϕ et μ. La fonction $\phi(s)$ correspond à l'ensemble des contextes de la phrase s et μ comptabilise ces contextes. La notion de contexte varie d'une représentation à l'autre. À partir de ces notions, la formule suivante donne une valeur d'implication :

$$Ent(s1, s2) = \frac{\mu(\phi(s1) \bigcap \phi(s2))}{\mu(\phi(s1))}$$

Cette formule revient à la définition indiquant que pour que le texte (s2) implique entièrement l'hypothèse (s1) il faut que l'hypothèse soit vraie dans chaque contexte du texte, que chaque contexte de s1 soit dans s2.

Pour mieux comprendre cette notion, et la mettre en pratique, plusieurs méthodes sont également présentées. La première voit ϕ comme l'ensemble des mots d'une phrase. La méthode étudie donc la proportion de mots communs au passage et à l'hypothèse. Cette mesure peut aussi permettre de mesurer la proportion de relations sémantiques ou syntaxiques communes.

D'autres travaux visent à identifier les phénomènes qui rendent difficile la détection d'une implication textuelle. Les travaux présentés dans [Bentivogli, et al. 2010a] ont pour but de séparer les données en fonction des phénomènes linguistiques. Le mécanisme s'applique à chaque paire texte-hypothèse en trois temps :

- les phénomènes linguistiques sont identifiés dans la paire ;
- le texte est modifié pour être au plus proche de l'hypothèse à l'aide de règles correspondant au phénomène ;
- les paires sont groupées en type de phénomènes.

De nombreux phénomènes sont ainsi annotés et peuvent être regroupés en 5 grandes catégories :

– lexical *(hyponymie, synonymie)* (16 %) ;
– lexical-syntaxique *(verbalisation, paraphrase)* (9 %) ;
– syntaxe *(négation, apposition, voix passive)* (21 %) ;
– discours *(coréférence, anaphore)* (21 %) ;
– raisonnement *(méronymie, utilisation de connaissances externes plus complexes)* (32 %).

Afin d'évaluer la faisabilité de la méthode, un ensemble de 100 paires est analysé. Cela a permis de montrer que le phénomène le plus fréquent était le raisonnement, ce qui témoigne de la complexité de la tâche.

La première partie de ce chapitre a permis de définir la notion de validation de réponses et de la relier à la notion d'implication textuelle. La suite de ce chapitre présente les méthodes permettant de valider des réponses et reconnaître l'implication textuelle. Comme les tâches sont très proches, les méthodes le sont également. La différence est que, dans le cadre de la validation de réponses, l'information à vérifier porte aussi sur une partie de l'hypothèse, la réponse. Les systèmes peuvent ainsi effectuer des vérifications supplémentaires. Les méthodes proposées peuvent être regroupées en deux grandes catégories :

– les méthodes reposant sur un formalisme de représentation du texte et de l'hypothèse issu d'analyses profondes. Pour celles-ci, le texte et l'hypothèse peuvent ainsi se trouver sous forme d'arbre syntaxique, de graphe sémantique ou sous forme logique. La détection consiste alors à déterminer une correspondance globale entre ces deux formes ;
– les approches combinant un ensemble de caractéristiques locales permettant de détecter l'implication. Dans celles-ci, les critères sont le plus souvent d'ordre lexical et syntaxique et la combinaison est souvent effectuée par apprentissage.

1.4 Mise en correspondance de représentations structurées

Les approches présentées ici sont fondées sur l'utilisation d'une représentation du texte et de l'hypothèse reconnue par des analyses en profondeur, syntaxiques ou sémantiques. Pour cela, l'hypothèse doit se trouver sous forme de proposition. Ce type de données est disponible pour les méthodes détectant l'implication textuelle mais il est besoin d'un prétraitement pour la validation de réponses. Ce prétraitement consiste à créer l'hypothèse. Elle correspond à la forme déclarative de la question à laquelle la réponse est ajoutée. Cette hypothèse est souvent obtenue en appliquant des règles de transformation propres à chaque type de question.

1.4.1 Comparaison de représentations syntaxiques

Le premier traitement effectué par les différents systèmes consiste à analyser les textes et hypothèses grâce à des analyseurs syntaxiques. Différents types de représentations syntaxiques sont possibles et dépendent de l'analyseur choisi. Par exemple, certains tiennent compte d'arbres de dérivation et d'autres de relations de dépendances entre composants syntaxiques (groupe nominal, groupe propositionnel...) ou entre mots. De manière générale, les mécanismes de détection d'implication textuelle peuvent être comparés.

Un des analyseurs les plus communément utilisés sur l'anglais est Minipar [Lin 1998a]. Il crée des arbres syntaxiques dans lequel les nœuds correspondent aux mots et les liens aux relations syntaxiques

les liant tels que sujet, objet ou déterminant. Pour ce faire, il s'appuie sur le lexique du réseau lexical WordNet [Fellbaum & Miller 1998]. La figure 1.5 présente le traitement de la phrase « *John found a solution to the problem* ».

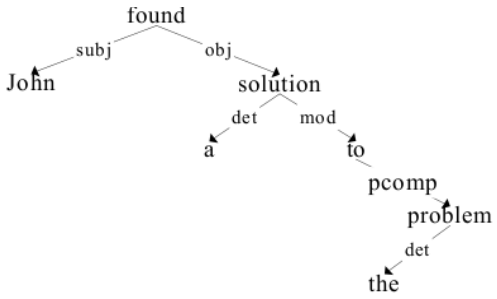

FIG. 1.5 – Analyse Minipar extraite de [Lin & Pantel 2001]

Le but des différentes approches est de déterminer une ressemblance entre l'arbre syntaxique de l'hypothèse et celui du texte. Deux grands types d'approches existent :

- les premières analysent les arbres du texte et de l'hypothèse afin de trouver leurs points communs ;
- les secondes transforment, ou évaluent les coûts de transformation, du texte ou de l'hypothèse afin de rendre les deux structures semblables. Pour ce faire, certaines approches portent sur les distances d'édition et d'autres sur les paraphrases.

1.4.1.1 Comparaison d'arbres syntaxiques

Les premières méthodes de comparaison des arbres syntaxiques n'utilisent pas pleinement la notion de graphe mais portent sur la détermination des relations syntaxiques communes au texte et à l'hypothèse.

De nombreux systèmes, [Inkpen, et al. 2006 ; Hickl, et al. 2006], évaluent la proportion de liens syntaxiques de l'hypothèse se trouvant également dans le passage sans faire de considération particulière sur le type de relation. Cette information vise à vérifier que les mots communs au passage et à l'hypothèse sont employés dans le même sens. Cette vérification vient le plus souvent en compléter d'autres portant par exemple sur les termes communs au passage et à l'hypothèse. Cette vérification revient à la notion de contexte en commun présentée dans [Clarke 2006].

Dans le cadre des systèmes de questions réponses FIDJI [Moriceau, et al. 2009] possède un mécanisme d'extraction de réponses fondé sur la reconnaissance des différentes relations portant sur la réponse potentielle.

Rus [2006] présente un système qui reconnaît une implication si plus de la moitié des liens syntaxiques de l'hypothèse se retrouvent dans le passage, ce qui permet d'effectuer 60 % de bonnes détections.

Volokh & Neumann [2010] étudient plus finement les relations et vérifient que les relations importantes de l'hypothèse se trouvent également dans le texte. Les relations autour du verbe sont es-

sentiellement considérées c'est-à-dire, les relations sujet-verbe, verbe-objet et verbe-complément. Si une de ces relations ou une de ses variantes n'est pas trouvée dans le texte alors l'hypothèse est reconnue comme n'étant pas impliquée par le passage. Pour l'exemple précédent, le système s'assure que les liens entre « John » et « found » et entre « found » et « solution » se trouvent dans le passage. Cette vérification peut sembler intéressante car bien souvent la relation d'intérêt a lieu au niveau du verbe. Les systèmes effectuent $\frac{2}{3}$ de bonnes détections. Mais encore faut-il disposer de ces relations de manière suffisamment fiable.

D'autres approches tiennent davantage compte de la notion d'arbre en effectuant une comparaison de nœuds et de relations. Par exemple, le système présenté dans [Marsi, et al. 2006], calcule des similarités entre les nœuds correspondant aux différents termes et le calcul de cette similarité tient notamment compte des relations de synonymie. Pour être sûr que les mots sont utilisés dans le même contexte, la comparaison des nœuds fils est aussi effectuée.

D'autres approches étudient les sous-arbres communs au texte et à l'hypothèse. Dans l'arbre de la figure 1.5, un sous-arbre correspond, par exemple, à la partie de phrase « *a solution to the problem* ». Les comparaisons, menées notamment dans [Schilder & Thomson McInnnes 2006] et [Wang & Neumann 2007], comptabilisent la proportion de sous-arbres de l'hypothèse se trouvant dans le passage. Cette comparaison permet de tenir compte des cas où les deux arbres ne peuvent pas être complètement alignés mais contiennent plusieurs parties similaires. Nous pouvons remarquer que la comparaison de sous-arbres implique une comparaison de termes communs puisque les termes sont les feuilles des différents arbres.

Dans les méthodes présentées jusqu'à présent, la comparaison entre les arbres syntaxiques était directe. L'idée proposée par Zanzotto & al. [Zanzotto, et al. 2006 ; Zanzotto & Moschitti 2006] est de comparer les paires texte-hypothèse à évaluer à des paires dont la valeur est déjà connue. Ainsi, en montrant que la paire (T1, H1) est similaire à la paire (T2, H2) et si T2 implique H2 alors T1 implique H1. Afin de comparer deux paires texte-hypothèse, il est nécessaire qu'elles suivent les mêmes règles d'écriture. Pour ce faire, la mesure de comparaison tient compte de deux éléments :

– T1 et H1 sont structurellement similaires à T2 et H2 ;
– les liens lexicaux entre T1 et H1 sont similaires à ceux entre T2 et H2.

Pour calculer le premier point, les mots similaires à T1 et H1 sont marqués, ce qui permet de comparer les structures des deux paires texte-hypothèse. Le second point est calculé en utilisant les sous-arbres communs. Avec ces deux mesures, il est possible de trouver une paire similaire à la paire à évaluer, ce qui permet de détecter la valeur d'implication de la paire qui est la même que celle de la paire détectée. Le système a participé à la campagne AVE en 2006 et a obtenu les seconds meilleurs résultats sur l'anglais avec une f-mesure de 0,41. La figure 1.6 issue de [Zanzotto & Moschitti 2006] présente la comparaison de deux paires texte-hypothèse. Nous pouvons voir qu'il y a un rassemblement des mots communs de T1 et H1 (end, year ...) et un rassemblement des structures similaires des textes T1, T3 et des hypothèses H1, H3. Par exemple il y a un rapprochement entre « All solid insurance companies » et « all wild mountains animals ».

1.4.1.2 Transformation d'arbres

Les derniers grands types d'approches étudient les transformations à appliquer au texte afin d'obtenir l'hypothèse ou au moins d'en être le plus proche possible et sont :

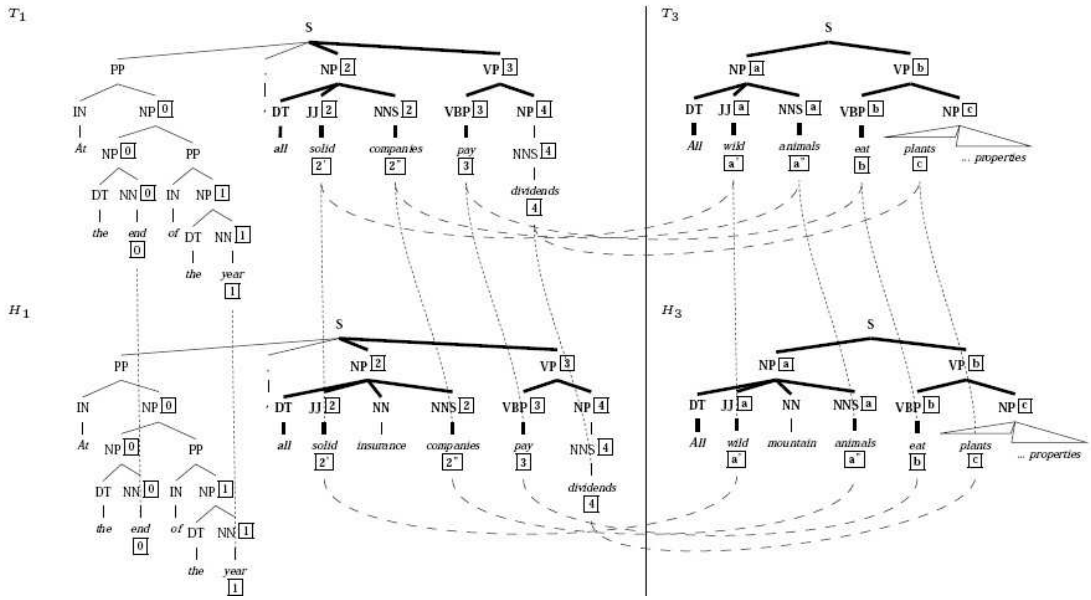

FIG. 1.6 – Comparaison de deux paires texte-hypothèse extrait de [Zanzotto & Moschitti 2006]

– le calcul de la distance d'édition qui porte sur un calcul de coût de transformation ;
– l'utilisation de paraphrases afin de modifier le texte.

Dans cette partie nous commencerons par voir le calcul de la distance d'édition. La distance d'édition est une mesure permettant de comparer deux objets en comptabilisant les transformations à effectuer au premier afin d'obtenir le second. Dans notre cas les deux objets à comparer sont des arbres syntaxiques et les transformations étudiées portent sur le texte afin d'obtenir l'hypothèse. Pour cela, trois transformations sont possibles :

– l'insertion qui ajoute un nœud de l'hypothèse dans le texte ;
– la suppression qui retire un nœud du texte ;
– la substitution qui remplace un nœud par un autre.

Notons que ces différentes opérations ne sont pas effectuées mais permettent de comparer les structures et que chacune de ces opérations a un coût. Pour détecter l'implication, la somme de ces coûts doit être inférieure à une certaine valeur. Cette valeur seuil tient souvent compte de la distance d'édition consistant à créer l'hypothèse à partir de rien et de celle de suppression totale du texte. Le score correspond alors à celui calculé sur deux structures totalement différentes.

Plusieurs systèmes, [Schilder & Thomson McInnnes 2006 ; Kouylekov & Negri 2010 ; Yashar, et al. 2009], utilisent cette méthode pour détecter l'implication textuelle ou la validation de réponses, [Kouylekov, et al. 2006]. Le système décrit dans [Yashar et al. 2009] a effectué ainsi 60 % de bonnes détections lors de la campagne RTE 5. La principale différence entre ces différents systèmes est la manière de tenir compte des coûts.

Le coût d'insertion semble assez délicat à mesurer puisqu'il dépend de l'importance du mot. En effet, plus un mot est important plus son ajout modifie le sens de la phrase. Ainsi, une des mesures utilisée pour évaluer l'importance d'un mot est la mesure IDF qui regarde l'apparition du mot dans un grand nombre de documents : plus le mot apparaît dans des documents différents moins il semble porteur de sens et plus il est possible de l'insérer sans changer le sens de la phrase. D'autres mesures tenant compte des spécificités de l'arbre, comme la profondeur du nœud ou son nombre de descendants, sont aussi utilisées.

Certains systèmes considèrent que le coût de suppression est nul puisque le texte est souvent bien plus grand que l'hypothèse, tandis que d'autres tiennent compte de l'importance du mot supprimé à l'aide des méthodes définissant le coût d'insertion.

Le coût de substitution dépend quant à lui de la ressemblance de ce mot avec celui déjà existant. L'idée étant que plus les mots sont similaires, plus la substitution est sans conséquence. Pour ce faire, une mesure se sert de WordNet afin d'évaluer le degré de synonymie entre les deux mots.

Afin de bien comprendre ces méthodes, voyons un exemple.

Texte : Henri IV a été assassiné par Ravaillac
Hypothèse : Henri IV est mort en 1610

Pour montrer qu'il n'y a pas implication dans l'exemple, le système peut tout d'abord substituer « a été assassiné » par « est mort » ce qui devrait être assez peu coûteux car ces termes sont reliés puis ajouter « 1610 » et retirer « Ravaillac » ce qui devrait avoir un coût élevé car ces termes sont porteurs de sens et absents respectivement du texte et du passage.

1.4.1.3 Utilisation de paraphrases

Une autre transformation possible tient compte des règles de paraphrases. Deux phrases sont dites paraphrases si elles ont le même sens. Jahna & Dragomir [2004] différencient la paraphrase et l'implication textuelle dans le sens où, pour que deux phrases P1 et P2 soient paraphrases, il faut qu'elles aient toutes les deux le même sens, c'est-à-dire que le sens de P1 soit contenu dans P2 et inversement. Dans le cas de l'implication textuelle, il suffit que la phrase P1 contienne les informations de P2 mais elle peut contenir des informations supplémentaires. Pour illustrer voyons un exemple :

Texte 1 : Ravaillac a assassiné Henri IV.
Texte 2 : Ravaillac a assassiné Henri IV le 14 mai 1610.
Hypothèse : Ravaillac a tué Henri IV.

Dans cet exemple, le texte 1 est paraphrase de l'hypothèse puisque assassiner est synonyme de tuer. Le texte 2 quant à lui implique l'hypothèse puisqu'il contient des informations supplémentaires portant sur la date de l'action qui ne peuvent pas être déduites de l'hypothèse. On parle alors de paraphrase sous phrastique car elle ne s'applique qu'à une partie de la phrase.

La plupart des approches traitant de la paraphrase recherchent les paraphrases d'une phrase ou d'une portion de phrase donnée. Parmi eux, Lin & Pantel [2001] recherchent des règles de paraphrases dans un ensemble de documents. Une règle de paraphrase est une transformation visant à transformer une phrase en une de ses paraphrases.
Dans l'exemple précédent, une règle possible consiste à remplacer « X a tué Y » par « X a assassiné

Y ». Pour ce faire, les chemins contenus dans les arbres syntaxiques sont analysés en se fondant sur l'hypothèse suivante : « Si deux chemins ont des contextes similaires (ici des extrémités en commun), le sens des chemins tend à être similaire ».
Pour tester la similarité des contextes, des recherches dans des textes sont effectuées sur le Web. Dans le cas de l'exemple précédent, il existe de nombreuses valeurs permettant d'instancier X (Achille, Oswald, Robert Ford...) et Y (Hector, Kennedy, Jessee James ...) et pour ces valeurs, « X a tué Y » apparaît quasiment aussi souvent que « X a assassiné Y », ce qui implique que les chemins sont paraphrases. De nombreuses règles de transformation ont ainsi été extraites.

Certaines méthodes détectant l'implication textuelle se servent de ces règles de paraphrase. Ainsi, Dagan & Glickman [2004] les utilisent dans un mécanisme transformant le texte pour le rendre le plus proche possible de l'hypothèse. Chacune de ces règles ayant un coût, la somme des poids indique la faisabilité de l'implication. Par exemple pour montrer qu'il y a implication entre « *John bought a novel* » et « *John purchased a book* » les transformations suivantes sont appliquées :

1. novel => book
2. bought a novel => bought a book
3. bought => purchased
4. bought a novel => purchased a book
5. John bought a novel => John purchased a book

Bar-Haim, et al. [2007] présentent un mécanisme similaire mais comportant en plus une comparaison entre les différents arbres obtenus. L'implication a lieu si un des arbres obtenu est similaire à l'hypothèse. 64 % de bonnes détections sont ainsi effectuées.

A l'inverse, Bosma & Callison-Bursh [2006] effectuent des transformations sur l'hypothèse afin de la rendre la plus semblable possible au texte. Chaque transformation correspond à une règle de paraphrase. Après avoir appliqué l'ensemble des transformations, un calcul de similarité est effectué entre la nouvelle hypothèse et le passage en tenant compte de la plus longue chaîne de mots communs au passage et à la nouvelle hypothèse. Ce système a participé à la campagne AVE en 2006 sur l'anglais. Les auteurs ont montré que tenir compte de cette transformation permet d'améliorer la f-mesure de 0,23 à 0,37 par rapport à utiliser simplement la plus longue chaîne commune.

Ce type d'approche peut sembler très intéressant car elle permet de bien comprendre les mécanismes à mettre en œuvre pour détecter l'implication textuelle. De plus elle concerne le sens des phrases ce qui permet effectivement de reconnaître les informations communes aux deux textes. Toutefois, elle nécessite la construction d'un ensemble de règles de paraphrases sur le français semblable à celles existant sur l'anglais.

Les approches présentées ci-dessus détectent l'implication textuelle ou la validation de réponses en se fondant sur la syntaxe, soit en effectuant des comparaisons d'arbres, soit en effectuant des transformations à l'aide d'un calcul de distance d'édition ou de règles de paraphrases. Cela permet d'identifier que les mots communs au texte et à l'hypothèse sont effectivement liés de la même manière. Toutefois, elles reposent sur l'utilisation d'analyseurs syntaxiques capables de bien relier les éléments de la phrase. Ces analyseurs s'appliquent seulement sur des documents rédigés correctement et, même dans ce cas, les relations de dépendances détectées sont souvent peu fiables, du moins en français ainsi que l'a montré l'étude de la thèse de Ligozat [2006].

1.4.2 Raisonnement sur des représentations logiques

Certains systèmes représentent le texte et l'hypothèse sous une forme logique du premier ordre. Par exemple, la phrase « le chat mange la souris » a « chat(x), manger(y), souris(z), rel(x, y, z) » comme représentation. Cette forme est obtenue à partir d'analyses syntaxiques et sémantiques profondes. Notons que de nombreux systèmes se servent également de bases de connaissances, telles que WordNet, et que les systèmes les plus performants utilisent une telle représentation. Toutefois, comme de telles bases de connaissances se trouvent uniquement sur l'anglais, elles semblent difficilement applicables sur le français sans recourir à un important travail de modélisation formelle des connaissances, ce qui n'est pas le choix fait dans cette thèse. Nous présenterons toutefois ces approches dans un but d'exhaustivité. Elles peuvent se séparer en deux grandes catégories :

- les systèmes par inférence qui appliquent un certain nombre de transformations pour rapprocher le texte et l'hypothèse, [De Salvo Braz, et al. 2005 ; Vanderwende, et al. 2006]. Ce mécanisme est très proche de celui présenté dans la section précédente puisque seul le formalisme change et ces systèmes ne seront donc pas détaillés.
- les systèmes par preuve logique, [Glöckner 2006 ; Tatu, et al. 2006]. Ils utilisent pleinement la représentation logique puisqu'ils prouvent, grâce aux approches classiques en Intelligence Artificielle, que le texte implique ou non l'hypothèse. Dans cette partie nous nous focaliserons donc sur ce type d'approche.

Afin de pouvoir effectuer la détection grâce à de l'inférence logique il est nécessaire de disposer d'un ensemble de règles et de prédicats. Bos & Markert [2006] tiennent par exemple compte de règles de transformation créées manuellement et d'une base de connaissance (KB) construite à partir des relations de synonymes WordNet. Puis la détection est effectuée à partir de règles d'implications classiques tel que $T \rightarrow H, T\ and\ KB \rightarrow H$. D'autres vérifications testent quant à elles la consistance des données avec des règles telles que $\neg(KB\ and\ T\ and\ \neg H)$ qui permettent de s'assurer que les connaissances contenues dans l'hypothèse ne sont pas inconsistantes avec celles du texte.

Le système COGEX [Tatu et al. 2006] traite de la validation de réponses en effectuant une implication sur le sens. Le système se fonde sur l'idée que le texte implique l'hypothèse s'il implique logiquement son sens. Dans ce formalisme, les prédicats correspondent aux noms, verbes et adjectifs. Les relations sont obtenues par une analyse syntaxique. Un ensemble d'axiomes venant de la ressource Extended WordNet [Mihalcea & Moldovan 2001] est également considéré.

La preuve est faite par réfutation. Pour ce faire, le passage (P) et l'hypothèse (H) sont mis sous forme logique et l'hypothèse est niée. Le mécanisme consiste alors à chercher une réfutation. L'idée classique de la logique est que $P \Rightarrow H$ est équivalente à $\neg(P)\ ou\ H$ qui est l'inverse de $P\ and\ \neg(H)$. Donc, si cette dernière formule est vraie alors il n'y a pas implication et, à l'inverse, si elle est fausse il y a implication.
Le mécanisme de preuve consiste à ajouter des axiomes de l'hypothèse jusqu'à obtenir une réfutation portant sur la dernière formule. S'il n'y a pas de réfutation détectée, une étape de relaxation est appliquée en retirant des prédicats. Chaque prédicat retiré ayant un poids il est possible de calculer un poids global qui sera comparé à une valeur seuil afin de détecter l'implication potentielle. C'est donc une méthode hybride qui combine un mécanisme par preuve et l'importance des différents termes de la question. Ce système a participé à la campagne AVE 2006 en anglais et en espagnol et a obtenu les meilleurs résultats sur ces deux langues avec une f-mesure respectivement de 0,44 et 0,61.

Le système de l'université d'Hagen ([Glöckner, et al. 2007], [Glöckner 2008]) contient lui aussi un mécanisme par preuve logique mais débute par une étape visant à normaliser le texte en modifiant les mots afin qu'ils soient au plus près de ceux de l'hypothèse. La preuve est effectuée par un mécanisme de relaxation récursive. Tout d'abord le système prend l'ensemble des prédicats du texte et de l'hypothèse et regarde s'il y a implication. Si ce n'est pas le cas il retire des prédicats de l'hypothèse. Ce mécanisme est effectué jusqu'à obtenir une implication. La décision finale est alors faite à partir des mots présents dans cette nouvelle hypothèse.

1.5 Méthodes fondées sur la combinaison de critères par apprentissage

Dans la partie précédente, différentes méthodes permettant de détecter l'implication textuelle ou la validation de réponses ont été présentées. Elles ont ceci en commun qu'elles s'appuient sur des analyses syntaxiques ou sémantiques profondes et s'appliquent le plus souvent sur l'anglais. Les approches présentées dans cette section sont applicables à différentes langues sans nécessiter un recours trop important à des ressources construites manuellement. Un certain nombre d'indications ou critères indiquant la validité de la réponse ou l'implication textuelle sont recherchées et combinées. La combinaison est souvent effectuée par apprentissage mais des combinaisons simples sont parfois considérées comme un ensemble de filtres [Pakray, et al. 2009] ou des pondérations de critères comparés à une valeur seuil, [Perini 2009].

Les méthodes présentées dans cette section posent la validation de réponses et l'implication textuelle comme un problème de classification en deux classes : la réponse est valide et la réponse n'est pas valide. Le problème peut ainsi se traiter en combinant différents critères par un modèle de classification. Deux modèles sont principalement utilisés : les arbres de décision ([Nicholson, et al. 2006 ; Newman, et al. 2006 ; Hickl et al. 2006]) et les SVM (séparateurs à vaste marge) ([Castillo 2009 ; Ferrández, et al. 2009 ; Kozareva, et al. 2006 ; Herrera, et al. 2006]). Les SVM cherchent une séparation linéaire des données entre les bons exemples et les mauvais en plaçant les points dans un espace contenant de nombreuses dimensions. Les arbres de décision cherchent récursivement le critère permettant de séparer au mieux les données. Le premier critère sépare les données en deux puis pour chaque sous partie le meilleur critère est recherché et les données sont ainsi récursivement séparées. Quand de nouveaux exemples de la base de tests se présentent, les valeurs pour les différents critères sont calculées puis la combinaison apprise est utilisée pour décider de la valeur de l'exemple considéré.

Les critères sont de différents ordres. Un certain nombre de critères sont ainsi d'ordre syntaxique et on retrouve des similarités d'arbres syntaxiques [Schilder & Thomson McInnnes 2006]. D'autres critères sont d'ordre lexical. Ceux-ci ont pour but de vérifier que les termes de l'hypothèse se trouvent dans le passage, de mesurer leur proximité dans le passage... Malheureusement il n'est pas possible de connaître clairement l'apport de chacun des critères car aucune étude n'a encore été menée dans ce sens. La suite de cette section présente ces critères que nous avons regroupés en deux grandes catégories :

- la similarité des énoncés (les termes communs ou similaires au passage et à l'hypothèse et la proximité des termes dans le passage) ;
- les vérifications propres à la validation de réponses.

1.5.1 Similarité des énoncés

1.5.1.1 Termes communs au passage et à l'hypothèse

Afin de vérifier que les informations contenues dans l'hypothèse se trouvent aussi dans le passage de texte, les mots de l'hypothèse sont recherchés dans le passage sous la même forme ou sous forme d'une variante.

L'évaluation de la présence des termes de l'hypothèse dans le passage peut s'appuyer sur :
- la comparaison de mots isolés ;
- la comparaison de termes complexes.

La vérification la plus simple, présentée notamment dans [J.Castillo 2008], comptabilise simplement le nombre de mots de l'hypothèse se trouvant à l'identique dans le texte. Des difficultés sont rencontrées dès qu'un terme se trouve sous forme d'une de ses variantes. Ainsi deux vérifications un peu plus complexes consistent à prendre les mots sous leurs formes lemmatisées. Ces traitements permettent donc de rassembler les différentes formes d'un même mot mais ne permettent pas d'aligner deux mots différents ayant un sens similaire comme « voiture » et « automobile ».

De nombreux systèmes s'appliquant sur l'anglais ont recours à WordNet [Fellbaum & Miller 1998] pour détecter le rapprochement entre deux mots. La méthode la plus simple, présentée notamment dans [Pakray et al. 2009] ou [Bahadorreza Ofoghi 2009], est fondée sur les relations de synonymie ou d'hyperonymie. Des approches plus complexes étudient les liens reliant les deux mots en partant du principe que plus les mots sont proches dans le réseau, plus leur sens peut être lié. L'approche présentée dans [Schilder & Thomson McInnnes 2006] consiste à calculer la longueur du plus court chemin reliant les deux mots à comparer. Une autre mesure utilisant ce réseau est la mesure Lin [Lin 1998b] qui est fondée sur la notion de plus proche ancêtre commun. Par exemple les mots « voiture » et « moto » ont tous les deux « véhicule » comme ancêtre. Le critère portant sur cette mesure suit l'hypothèse que plus l'ancêtre est proche des mots plus ils sont reliés.

Une autre ressource souvent utilisée est VerbOcean, [Chklovski & Pantel 2004]. Celle-ci porte exclusivement sur les verbes et leur construction et plus spécifiquement sur les relations qu'ils peuvent avoir entre eux.

Une fois la correspondance entre les termes détectée, il reste encore à mesurer le degré de recouvrement. Breck [2009] s'intéresse plus en détail à ce problème en détectant la non implication si au moins une information de l'hypothèse ne se trouve pas dans le texte. Plus spécifiquement si un nom commun, nom propre, verbe, adjectif, adverbe ou une expression numérique de l'hypothèse n'est pas présente dans le passage de texte. Cette détection obtient 61 % de bons résultats ce qui montre bien l'intérêt de l'idée mais aussi ses manques puisque toutes les informations ne sont pas forcément contenues dans un court passage. La mesure la plus classique comptabilise la proportion de mots de l'hypothèse dont un mot similaire est trouvé dans le passage ($\frac{|H \cap P|}{|H|}$). Cette mesure montre donc la proportion d'information de l'hypothèse contenue dans le passage. D'autres mesures similaires calculent la proportion de termes communs au passage et à l'hypothèse ($\frac{|H \cap P|}{|H \cup P|}$).

Le problème avec ces mesures est qu'elles ne tiennent pas compte de l'importance des mots. Pour cela, nous pouvons tout d'abord remarquer que la plupart des systèmes n'utilisent que les mots porteurs de sens tels que les noms, les verbes ou les adjectifs. En effet, un certain nombre de mots,

tels que les adverbes ou les prépositions, n'ont que peu d'importance pris seuls et leur présence commune au passage et à l'hypothèse ne signifie rien. Certains systèmes, comme MLENT [Kozareva et al. 2006], effectuent des vérifications propres à chaque catégorie morphosyntaxique en calculant la proportion de noms propres, de noms communs, de verbes et d'adjectifs de l'hypothèse présents dans le passage. Cela permet de bien séparer chaque catégorie et d'évaluer plus en détail l'importance des informations communes au passage et à l'hypothèse en se fiant à la remarque qu'un nom propre a plus d'importance qu'un adjectif.

La mesure TF*IDF est aussi souvent utilisée, entre autre par [Castillo 2009], pour tenir compte de l'importance des différents mots. Elle tient compte du nombre d'occurrences du mot dans le passage et de l'inverse de la fréquence du terme dans un grand ensemble de documents puisque le mot semble avoir moins d'importance s'il apparaît dans un grand ensemble de documents que s'il n'est présent que dans quelques uns.

Une autre mesure, le Token-Map, [Adams 2006], utilise une représentation matricielle. Une comparaison entre chaque terme de l'hypothèse et du passage est calculée en attribuant un score différent selon que les termes sont les mêmes, synonymes ou autre. Ces différentes valeurs sont placées dans une matrice ce qui permet ensuite de relier les différents termes de l'hypothèse avec un terme du passage. 63 % de bonnes détections sont effectuées.

La mesure « cosinus » est fréquemment utilisée par les systèmes, [Nicholson et al. 2006 ; Newman et al. 2006 ; J.Castillo 2008]. Elle considère le TF*IDF et deux vecteurs, l'un correspond au poids des mots dans le texte et l'autre aux poids dans l'hypothèse. La mesure permet alors de comparer ces deux vecteurs en utilisant le produit scalaire et la norme des vecteurs via un calcul de cosinus.

D'autres vérifications portent plus particulièrement sur les entités nommées. L'absence d'une entité nommée du passage alors qu'elle est présente dans l'hypothèse semble fortement indiquer qu'il n'y a pas d'implication. Cela est dû au fait que ces entités ont très peu de synonymes et se retrouvent généralement sous la même forme. Par exemple, un nom de personne peut difficilement se trouver sous une autre forme, à part peut être une anaphore. Par conséquent si la personne de l'hypothèse est absente du passage alors il n'y a probablement pas implication.

Rodrigo, et al. [2006] étudient ainsi l'utilité de ces entités pour un système de validation de réponses en effectuant plusieurs vérifications. La première considère une réponse comme non valide si au moins une des entités nommées de la question ne se trouve pas dans le passage. Un rappel de 0,83 est obtenu ce qui montre que très peu de réponses sont vues comme incorrectes à tort. La précision, plus basse (0,46), indique que ce critère n'est pas suffisant puisqu'il reconnaît trop de réponses valides à tort. Une autre expérience étudie l'apport de ce critère sur les entités nommées à un ensemble de critères lexicaux comportant entre autres la proportion de termes de la question se trouvant dans le passage. Pour cela, deux évaluations sont faites, l'une avec le critère et l'autre sans. Une amélioration de la f-mesure de 10 % témoigne de l'importance du critère.

De nombreux systèmes tiennent compte des entités nommées afin de détecter la validité des réponses ou l'implication textuelle. Ferrández et al. [2009] définissent ainsi deux critères : le premier teste si toutes les entités nommées de l'hypothèse sont contenues dans le texte, le second correspond à la proportion d'entités nommées de l'hypothèse présentes dans le texte. Dans une autre étude portant sur la validation de réponses [Ferrández, et al. 2008], une étape de filtre est appliquée avant tout autre

traitement, durant laquelle les réponses sont considérées non valides si le passage ne contient pas les entités nommées de la question.

D'autres types de termes, indiquant quant à eux la négation (pas, plus ...), permettent d'établir une correspondance entre l'hypothèse et le passage. La présence de l'un de ces termes dans le texte et non dans l'hypothèse peut indiquer que le passage contredit l'hypothèse, à condition qu'il porte sur la même action. Afin d'utiliser cette propriété, certains systèmes, [Malakasiotis 2009 ; Ferrández et al. 2009] utilisent une liste de mots spécifiques et d'autres, [Kozareva & Montoyo 2006], des bases de connaissances tels que les antonymes WordNet ou VerbOcean qui permettent d'identifier deux termes ayant un sens opposé. Les critères sont relativement simples puisqu'ils comptabilisent le nombre de négations présentes dans le passage ou dans l'hypothèse. Nairn, et al. [2006] s'intéressent particulièrement à la polarité en étudiant notamment les verbes modaux qui peuvent modifier la valeur de vérité du verbe les accompagnant. Par exemple, si « should » accompagne le verbe alors l'action portée par celui-ci n'a pas eu lieu.

1.5.1.2 Proximité des termes

La présence dans le passage des mots de l'hypothèse peut donc indiquer une implication. Cette partie porte sur la prise en compte des relations entre les termes de la question présents dans le passage en formant l'hypothèse suivante : si les termes de la question se trouvent dans le passage physiquement proches les uns des autres alors ils sont liés de la même manière que dans la question. Cette idée a été suivie par de nombreux systèmes qui en tiennent compte sous forme de critères pouvant être assez différents. Notons que les relations syntaxiques présentées en 1.4.1 permettent également de marquer ce lien. Cette nouvelle manière d'en tenir compte se passe d'une analyse syntaxique et peut ainsi s'appliquer sur davantage de cas comme par exemple sur des langues ou des types de documents sur lesquels les analyses ne s'appliquent pas ou mal.

Le premier type de vérifications porte sur les n-grammes. Un n-gramme est ici un ensemble de n mots consécutifs. Ainsi pour la phrase « Le chien court après le chat », un bigramme peut être « chien court » et un trigramme « Le chien court ». La plupart des vérifications traitant des n-grammes ([Herrera et al. 2006 ; Newman et al. 2006 ; M.A.Garcia-Cumbreras, et al. 2007]) commencent par compter la proportion de n-grammes de l'hypothèse se trouvant dans le passage avec différentes valeurs de n. Nous pouvons remarquer que quand n vaut 1 cela revient à étudier la proportion de termes de l'hypothèse présents dans le texte et quand n est grand la présence de l'hypothèse dans le texte est testée. Les mesures les plus simples ([J.Castillo 2008 ; Herrera et al. 2006]) étudient la proportion de bigrammes de l'hypothèse se trouvant dans le passage.

Certaines vérifications sont plus complexes comme la mesure BLEU ([Papineni, et al. 2002]). Cette mesure permet d'évaluer des systèmes de traduction automatique. Elle calcule la proportion de n-grammes du passage se trouvant dans l'hypothèse en tenant compte de quelques contraintes comme la différence de taille du passage et de l'hypothèse. Perez, et al. [2005] ont créé un système détectant l'implication textuelle à l'aide d'une valeur seuil ; si la valeur BLEU est supérieure au seuil alors il y a implication. Le système permet de détecter 53% de bons résultats ce qui correspond à ceux obtenus par la baseline consistant à reconnaître toutes les hypothèses comme impliquées par le texte les accompagnant.

Un autre critère souvent présent tient compte de la proximité de l'ensemble des mots de l'hypothèse dans le passage. Pour ce faire, les systèmes calculent la plus longue sous séquence de termes de l'hypothèse se trouvant dans le passage. La définition de cette séquence peut varier d'un système à l'autre. Ainsi, pour certains [Kozareva et al. 2006] il n'est pas besoin que les termes soient dans le même ordre dans le passage et dans l'hypothèse. Dans ce cas le système cherche alors une séquence constituée uniquement des mots de l'hypothèse. D'autres systèmes [Pakray et al. 2009] ne s'intéressent qu'aux séquences pour lesquelles les mots se trouvent dans le même ordre. Afin d'agrandir la chaîne commune ils considèrent que des mots bonus peuvent séparer les mots de l'hypothèse dans le passage.
Quelle que soit la méthode, le critère final tient compte de la taille de la séquence obtenue en calculant un poids correspondant au rapport entre la taille de la chaîne et celle de l'hypothèse. Voyons un exemple :

Passage : Ravaillac poignarda Henri IV en 1610.
Hypothèse : Henri IV fut tué par Ravaillac.
plus longue chaîne commune : Ravaillac BONUS Henri IV
score : $\frac{3}{6}$

Dans cet exemple l'hypothèse est constituée de 6 termes : « Henri », « IV », « être », « tuer », « par » et « Ravaillac ». Dans le passage, les mots « Henri » et « IV » sont adjacents et séparés d'un seul mot de « Ravaillac ». En utilisant un calcul de sous-chaîne tenant compte des mots quel que soit leur ordre et permettant un unique mot bonus, la séquence « Ravaillac BONUS Henri IV » est obtenue. Cette séquence contient donc 3 termes de l'hypothèse et la valeur tenant compte de ce critère est $\frac{3}{6}$.

Une comparaison entre cette mesure et le calcul de n-grammes communs peut être faite. Tout d'abord elle tient compte de l'ensemble des mots de l'hypothèse et non seulement de quelques termes ce qui permet de constater que l'ensemble des informations de l'hypothèse sont contenues ou non dans le passage. Elle permet aussi, dans certains cas, de ne pas contraindre l'ordre des mots ce qui permet de reconnaître une correspondance entre la voix active et la voix passive d'un même verbe ce qui n'est pas le cas avec le calcul de n-grammes communs.

Concernant la validation de réponses, nous avons vu que l'extraction des réponses recherchait souvent la réponse la plus proche des mots de la question. Des critères traitant de la proximité des termes de la question dans le passage semblent donc aussi pertinents quand l'un des termes peut être la réponse.

La distance d'édition, présentée sur les arbres syntaxiques en 1.4.1.2, consiste à transformer le passage afin d'obtenir l'hypothèse. Il est également possible de faire ces transformations sur les formes textuelles du passage et de l'hypothèse en appliquant la distance de Levenshtein. Elle correspond alors à une distance d'édition calculée en appliquant les opérations d'édition, de suppression et de substitution sur les mots.

1.5.2 Vérifications propres à la validation de réponses

1.5.2.1 Vérification du type attendu

Nous avons vu jusqu'à présent des critères permettant de traiter aussi bien de la validation de réponses que de l'implication textuelle puisque ces deux tâches ont de nombreux points communs. Toutefois, la validation de réponses permet de s'appuyer sur davantage d'informations que l'implication.

Téllez-Valero & al. ([Téllez-Valero, et al. 2007 ; Téllez-Valero, et al. 2008]) s'intéressent particulièrement à la validation de réponses à l'aide de critères propres aux systèmes de questions réponses. Ainsi le système commence par analyser les questions comme le fait un système de questions réponses afin de détecter les termes les plus pertinents de la question qui devraient normalement se trouver proches de la réponse. Un premier ensemble de critères étudie la présence de ces termes (focus, type spécifique) dans le passage en supposant par exemple que si le focus est absent du passage alors ce dernier a de bonnes chances de ne pas justifier la réponse puisque par définition le focus doit se trouver dans le passage justificatif.

Un autre type de critères traite de la redondance de la réponse en se fiant à l'idée que plus une réponse a été extraite plus elle semble pertinente. Lors de l'ordonnancement des réponses par les systèmes de questions réponses, la redondance est l'un des critères forts. Il semble donc naturel qu'il le soit également pour la validation de réponses.

Un autre trait tient compte de la compatibilité de la réponse avec le type d'information demandé dans la question. L'analyse de la question a permis de déterminer les questions qui attendent un certain type d'entité nommée en réponse. Le système vérifie donc que le type d'entité nommée de la réponse à valider correspond à celui attendu par la question. En effet, si la question attend une personne en réponse et si la réponse proposée est une date alors il est plus que probable que la réponse soit incorrecte.

Le système INAOE [Téllez-Valero et al. 2008] présenté lors de la campagne AVE 2008 s'applique sur l'espagnol. Il combine par apprentissage, grâce à une combinaison d'arbres de décision, ces critères avec d'autres portant sur le recouvrement lexical du passage et de la question tels que la proportion de termes communs et obtient une f-mesure de 0,39.

Higashinaka & Isozaki [2008] vérifient que la réponse est compatible avec la question en utilisant entre autres une correspondance d'entités nommées. Pour ce faire, les entités nommées attendues par la question sont identifiées à partir d'une analyse des questions qui reconnaît 17 types. Le type d'entité nommée de la réponse est comparé à celui attendu par la question. D'autres vérifications portent sur les questions « quel » comme « *Dans quelle ville se situent les pyramides ?* » . Le système examine les définitions de WordNet afin de vérifier, entre autre, qu'un mot clé de la question apparaît dans la définition de la réponse : c'est-à-dire si « ville » se trouve dans la définition de « Le Caire ».

Huang, et al. [2009] effectuent aussi des vérifications sur la réponse afin de savoir si le passage est susceptible de contenir la réponse. Pour ce faire, une entité nommée du type attendu par la question est recherché dans le passage. Ce type de vérification est aussi réalisé grâce à un ensemble de dictionnaires construit dans le cadre du projet Ephyra [Schlaefer, et al. 2007]. Ces dictionnaires répertorient les instances d'un type donné ; par exemple celui correspondant au type acteur contiendra

des noms comme « Tom Hanks » ou « Sean Connery ». Ainsi, on peut vérifier qu'une instance du type spécifique se trouve bien dans le passage même si elle n'a pas été annotée.

1.5.2.2 Redondance

Un certain nombre de systèmes exploitent la redondance du Web pour valider les réponses. L'idée est que si la réponse est correcte alors elle se trouve dans plusieurs documents accompagnée des mots clés de la question. Par exemple, à la question « *Dans quel pays se situent les pyramides ?* » il y aura plus de documents contenant « pyramides » et « Égypte » que de documents contenant « pyramides » et « Canada ». Ce type de vérification se rapproche bien sûr de la redondance de la réponse trouvée dans la collection interrogée.

Un des systèmes mettant en œuvre cette idée est présentée dans [Magnini, et al. 2002b]. Il commence par créer une requête correspondant à la question en reliant les mots clés de la question, leurs synonymes et le lemme des verbes par le mot clé OR. Puis, le nombre de documents contenant la requête, contenant la réponse et contenant l'ensemble requête et réponse sont comptabilisés. La proportion de documents correspondant à chacune des requêtes est calculée. Puis trois combinaisons sont effectuées à partir de ces trois valeurs pour indiquer le degré de liaison entre la réponse et les mots de la question grâce à des probabilités d'apparition. La mesure PMI calcule le rapport entre la probabilité de trouver les deux termes ensemble et le produit des probabilités séparées. La mesure CCP tient compte quant à elle des probabilités conditionnelles. La dernière, plus complexe, MLHR, est particulièrement pertinente dans les cas de données éparses.

$$PMI = \frac{P(Question\ et\ réponse)}{P(Question)*P(réponse)} \qquad CCP = \frac{P(réponse|question)}{P\left(réponse\right)^{\frac{2}{3}}}$$

$$MLHR = -2log\lambda$$

$$\lambda = \frac{L(p,k1,n1)L(p,k2,n2)}{L(p1,k1,n1)L(p2,k2,n2)} \qquad L(p,k,n) = p^k(1-p)^n$$

$$p1 = \frac{P(Questionetrponse)}{P(rponse)} \qquad p2 = \frac{P(Question-rponse)}{P(-rponse)} \qquad p = \frac{P(Question)}{nbdocuments}$$

Ce travail a été poursuivi dans [Magnini, et al. 2002a] qui compare cette méthode avec une approche fondée sur le contenu des pages Web trouvées. Cette dernière calcule les pages contenant les mots clés et la réponse ainsi qu'une valeur tenant compte de la distance entre la réponse et les différents mots clés dans ces documents. Le mécanisme d'évaluation consiste à reconnaître la bonne réponse parmi plusieurs, toutes présentes dans le même passage. Les résultats ont montré que les deux approches étaient équivalentes et détectaient correctement 82 % de réponses et possèdent un taux de recouvrement de plus de 90 %.

Le même type d'approche a été employé par Tonoike, et al. [2004]. Ce système se place dans le cadre du célèbre jeu télévisé « Qui veut gagner des millions ? » dont le but est de reconnaître pour une question donnée la bonne réponse parmi plusieurs proposées. Le système utilise pour ce faire la redondance du Web. Une requête est créée à partir des mots clés de la question. Le nombre de documents Web contenant la réponse, la question et les deux ensembles sont calculés et combinés

grâce à deux mesures. Les mesures correspondent aux proportions de documents contenant la réponse et les mots clés de la question par rapport aux documents contenant juste la réponse ou juste les mots clés de la question. Plus de 73% de bons résultats sont ainsi obtenus.

Ce type de vérifications permet également d'évaluer le lien entre deux mots dans un cadre d'implication textuelle. Glickman [2006] utilise ainsi une approche visant à aligner les termes du texte et de l'hypothèse à l'aide d'une probabilité tenant compte de l'apparition des termes dans les mêmes documents.

1.5.3 Conclusion

Un ensemble de travaux considèrent la validation de réponses ou l'implication textuelle comme un problème de classification et ont défini des critères variés. De manière à en faire une synthèse, nous les avons récapitulés dans le tableau 1.1. Le critère le plus communément utilisé est la proportion de termes de la question présents dans le passage, suivi par le calcul de la plus longue chaîne commune au passage et à la question. En revanche, aucun système n'évalue les critères pris indépendamment. Une telle étude permettrait de déduire les critères à mettre en place afin d'avoir le meilleur système possible.

Lors de la campagne AVE 2006, le système ayant obtenu les meilleurs résultats sur le français, MLENT [Kozareva et al. 2006], effectue une combinaison de critères par apprentissage. Il tient compte de :

- la proportion de n-grammes communs au passage et à l'hypothèse, avec n allant de 1 à la taille de l'hypothèse ce qui permet de calculer également la proportion de termes de la question présents dans le passage ;
- la taille de la plus longue séquence commune au passage et à l'hypothèse. Les mots de cette séquence ne sont pas forcément consécutifs mais doivent être dans le même ordre ;
- la proportion de termes de l'hypothèse ne se trouvant pas dans le passage ;
- la correspondance d'entités numériques.

Au final ce système obtient une f-mesure de 0,47.

Le système UNED [Herrera et al. 2006] est celui ayant obtenu les meilleurs résultats, toutes langues confondues, en suivant ce type d'approche. Il s'applique sur différentes langues. Ses meilleurs résultats sont obtenus sur l'espagnol avec une f-mesure de 0,56. Il tient tout d'abord compte de la présence des entités nommées de la question dans le passage en supposant que si aucune correspondance n'est détectée alors la réponse n'est pas valide. Les triplets réponses restants sont ordonnés à l'aide d'un SVM et de critères tenant compte de la proportion de n-grammes de la question dans le passage et de la proportion de termes communs.

1.6 Vérification et ordonnancement de réponses

Nous avons vu jusqu'à présent les méthodes détectant la validation de réponses une fois la réponse extraite qui se posent comme une décision binaire. La validation a alors été conçue comme un moyen d'évaluer automatiquement la réponse des sorties des systèmes de questions réponses. Toutefois elle

Critère	Système
Termes communs	
Mots communs	[Kozareva et al. 2006 ; Herrera et al. 2006] [Rodrigo et al. 2006 ; M.A.Garcia-Cumbreras et al. 2007] [Téllez-Valero et al. 2007 ; Téllez-Valero et al. 2008] [Ferrández et al. 2008 ; J.Castillo 2008] [Ferrandez, et al. 2007]
Entités nommées	[Rodrigo et al. 2006 ; Herrera et al. 2006] [Rodriguo, et al. 2007 ; Ferrández et al. 2008] [Kozareva et al. 2006 ; Téllez-Valero et al. 2008]
Similarité de termes grâce à WordNet	[M.A.Garcia-Cumbreras et al. 2007] [J.Castillo 2008 ; Ferrández et al. 2008]
Distinctions morphosyntaxiques	[Téllez-Valero et al. 2007 ; Ferrández et al. 2008] [Téllez-Valero et al. 2008]
Mesure cosinus	[Ferrández et al. 2008 ; J.Castillo 2008]
TF*IDF	[Ferrández et al. 2008 ; J.Castillo 2008]
Programmation dynamique	[Ferrández et al. 2008]
Distance de Jaro	[Ferrández et al. 2008]
Proximité des termes	
Plus longue chaîne commune	[Kozareva et al. 2006 ; Bosma & Callison-Bursh 2006] [M.A.Garcia-Cumbreras et al. 2007] [Téllez-Valero et al. 2007 ; Ferrandez et al. 2007]
N-grammes	[Kozareva et al. 2006 ; Herrera et al. 2006] [M.A.Garcia-Cumbreras et al. 2007 ; Ferrandez et al. 2007] [J.Castillo 2008 ; Rodrigo et al. 2006]
Distance de Levenshtein	[Ferrandez et al. 2007 ; Ferrández et al. 2008] [J.Castillo 2008]
Mesure rouge	[Ferrandez et al. 2007]
Spécifiques validation de réponses	
Vérification du type de la réponse	[Téllez-Valero et al. 2007 ; Téllez-Valero et al. 2008]
Contraintes sur la réponse	[Téllez-Valero et al. 2007 ; Ferrández et al. 2008] [Téllez-Valero et al. 2008]
Redondance de la réponse	[Téllez-Valero et al. 2008]

TAB. 1.1 – Critères de validation de réponses

peut également permettre d'ordonner ou réordonner les réponses globalement avant de les présenter à l'utilisateur. Dans ce cadre, deux traitements sont possibles :

– l'ordonnancement des passages justificatifs. Cette étape a lieu avant que les réponses soient extraites et a pour but de placer le passage le plus à même de contenir la réponse désirée en première position sans tenir compte de l'extraction de la réponse. Les systèmes vérifient alors que le passage répond à la question ;

– l'ordonnancement des réponses qui s'applique une fois un ensemble de réponses candidates extraites.

La principale différence entre la validation de réponses placée en sortie du système de questions réponses et l'ordonnancement de réponses vient de la tâche elle même [Ravichandran, et al. 2003]. En effet, les systèmes de validation de réponses détectent les réponses valides en donnant la valeur OUI le cas échéant et la valeur NON sinon, tandis que les systèmes d'ordonnancement de réponses ordonnent les différentes réponses mais globalement les critères de décision sont les mêmes et seule la manière de les utiliser change.

1.6.1 Ordonnancement en fonction d'une représentation syntaxique

Le premier type d'approches étudie la syntaxe par des mécanismes proches de ceux présentés en section 1.4.1. On retrouve ainsi des approches calculant la distance d'édition afin de trouver des ressemblances entre les arbres syntaxiques de la question et des passages. D'autres étudient quant à elles la correspondance entre relations syntaxiques présentes dans la question et les différents passages. Ces approches recherchent des similarités entre la question et des parties du passage justificatif et n'effectuent donc pas un appariement global entre la question et le passage entier. De manière générale le passage est comparé à la question ce qui fournit un score de similarité qui permet ensuite d'ordonner les réponses.

Le système présenté par Bernard [2011] traite ainsi de l'ordonnancement de réponses en utilisant une représentation des composants minimaux des phrases sous forme d'arbre représentant des syntagmes nominaux et verbaux. Le passage associé à une réponse ayant la meilleure correspondance avec la question est recherché en utilisant une mesure de distance d'édition sur ces composants. Pour ce faire, le système commence par marquer les points communs au passage et à la question puis modifie le passage en effectuant des transformations propres au calcul de distance d'édition auquel est ajouté le rattachement qui consiste à déplacer une partie de l'arbre en le rattachant à un nouvel endroit. Ces différentes opérations permettent de montrer le degré de similarité entre la question et le passage. Les réponses sont ensuite ordonnées grâce à ce score afin que les passages les plus similaires à la réponse se trouvent en première position.

Le système présenté par Cui, et al. [2005] part du principe que les relations syntaxiques se trouvant dans la question doivent aussi se retrouver dans les passages pertinents. Pour ce faire, un alignement entre chaque relation de la question et un chemin présent dans le passage est recherché. Quand deux relations sont alignées, un score de similarité correspondant à la probabilité de transformer un chemin en un autre est calculé tenant compte de l'ensemble des relations contenues dans un chemin. Cette méthode montre son intérêt puisque le MRR passe de 0,50 à 0,78.

Shen & Klakow [2006] effectuent un travail assez similaire qui consiste en deux ordonnancements. Le premier vérifie que les relations syntaxiques de la question portant sur la réponse dans la question se retrouvent bien dans le passage et le second vérifie que les relations liant les termes de la question sont effectivement présentes dans le passage. Pour effectuer cette vérification les relations sont comparées grâce à un score statistique tenant compte de leur occurrence en corpus. Ainsi, les différentes réponses sont ordonnées grâce à un score tenant compte de la présence des différentes relations et du type de relation de la question présentes dans le passage.

Le système présenté par Moschitti, et al. [2007] combine une représentation sous forme d'arbre syntaxique à des informations sémantiques à travers une nouvelle représentation d'arbre. L'arbre du texte et celui de l'hypothèse sont comparés à l'aide d'un calcul de noyau d'arbre. Cette notion revient à comparer les différents nœuds tant au niveau de leur contenu que de leur structure.

1.6.2 Ordonnancement grâce à une combinaison de critères

D'autres systèmes combinent différents critères par apprentissage ([Suzuki, et al. 2002 ; Ravichandran et al. 2003 ; Martin, et al. 2001]). Comme pour la validation de réponses, un certain nombre de critères concernent des correspondances entre la question et les différents passages. Suzuki et al. [2002] voient la sélection de réponse comme un problème de classification visant à différencier les réponses correctes de celles incorrectes. Les différents systèmes contiennent ainsi des critères déjà présentés dans la validation de réponses tels que le calcul de termes communs au passage et à l'hypothèse (avec ou sans utilisation de WordNet), la proximité des termes de la question dans le passage ou l'utilisation des entités nommées notamment pour vérifier que les entités nommées de la question se trouvent dans les passages. Afin d'effectuer l'ordonnancement, un score de confiance est fourni par le système de classification à chaque réponse. Une réponse plus confiante sera ainsi mieux placée qu'une autre dont le score est plus faible.

Le système d'IBM [Martin et al. 2001] combine un grand nombre de critères proches de ceux considérés par les systèmes participant à la tâche AVE. Parmi ceux-ci certains étudient les passages afin d'évaluer la correspondance de termes, d'autres portent sur la correspondance entre l'entité nommée attendue par la question et celle de la réponse, d'autres encore vérifient les relations dans lesquelles la réponse apparaît en regardant par exemple si le verbe de la question est lié à la réponse. Le dernier type de critères s'applique aux questions de définition et vérifie que la réponse est reliée dans WordNet au focus, le mot pour lequel une définition est demandée.

Le système présenté dans [Suzuki et al. 2002] effectue une combinaison de critères grâce à un SVM. Les critères traitent de l'apparition des mots de la question dans le passage tels que la proportion de termes de la question présents dans le passage, la proportion d'entités nommées communes ou la présence du focus dans le passage. L'évaluation est effectuée sur 1 358 questions pour lesquelles au moins une réponse correcte a été extraite. Le MRR passe alors de 0,47 pour la baseline à 0,7 ce qui constitue une bonne amélioration.

Une mesure particulièrement utile pour l'ordonnancement de réponses traite de la redondance de la réponse. L'idée étant que si la réponse a été extraite de nombreuses fois depuis des documents pouvant être différents alors elle a plus de chances d'être correcte qu'une réponse extraite une seule fois.

Clarke, et al. [2001] s'intéressent plus particulièrement à cette notion et traitent notamment de la redondance dans les documents. A cet effet, les termes de la question sont pondérés avec un poids tenant compte du nombre de documents contenant le terme ainsi que de la fréquence du terme. Cela permet d'ordonner les passages en tenant compte des poids des termes de la question contenus dans le document et ainsi de sélectionner les meilleurs passages. Un certain nombre de réponses en sont extraites grâce à des patrons d'extraction. Comme seules les questions attendant une personne en réponse sont considérées, les règles d'extraction sont assez simples et consistent à retenir les groupes

de 2 mots commençant par une majuscule. Les réponses extraites sont pondérées puis ordonnées en tenant compte du nombre de documents trouvés et de la distance entre la réponse et les mots de la question à l'aide d'une mesure logarithmique. 70 % de bonnes réponses sont ainsi détectées.

Harabagiu & Hickl [2006] s'intéressent particulièrement à l'intégration d'un système d'implication textuelle dans un système de questions réponses. L'implication textuelle est détectée grâce à un alignement lexical avec entre autres la similarité cosinus et la distance de Levenshtein. Une vérification de paraphrases s'appuyant sur des données du Web est également considérée. Les critères sont ensuite combinés par apprentissage ce qui permet de fournir un score de confiance. Trois utilisations différentes de ce module sont présentées :

- au niveau des réponses : le calcul d'implication textuelle permet de rejeter les réponses ne correspondant pas à l'implication et d'ordonner les réponses restantes grâce à un score de confiance ;
- au niveau des passages : les passages sont ordonnés par le score d'implication textuelle et seuls les meilleurs passages sont sélectionnés ;
- l'implication est utilisée afin de détecter des questions similaires et plus facilement traitables.

Les tests sont effectués sur un ensemble de 500 questions dont 67 % ont un type d'entité nommée attendu. La baseline qui consiste à ne pas utiliser le module d'implication textuelle obtient un MRR autour de 0,30. La meilleure utilisation a lieu au niveau des passages puisque le MRR passe alors à 0,55 pour les questions avec un type d'entité nommée attendu et 0,42 pour les questions sans. Une vérification hybride obtient de meilleurs résultats avec un MRR de 0,56. Nous pouvons toutefois noter que la reformulation des questions nécessite d'avoir un ensemble de questions simplifiées en grand nombre afin d'être applicable à chaque nouvelle question.

Pour finir, Echihabi, et al. [2004] comparent trois méthodes permettant de sélectionner une réponse en calculant la proportion de réponses correctes parmi les cinq premières fournies :

- la première consiste à utiliser des connaissances telles que la vérification portant sur le type attendu en réponse, les relations sémantiques communes ou l'utilisation de paraphrases afin de vérifier que le passage est similaire à la question. 57 % des réponses sont placées dans les cinq premières positions ;
- la seconde traite plus particulièrement des patrons d'extraction de réponses. Les patrons sont appris automatiquement à l'aide de recherches sur le Web portant sur les occurrences de la question et de la réponse. Un des critères permettant de reconnaître que la réponse est correcte vérifie qu'un tel patron s'applique sur le document duquel la réponse a été extraite. La fréquence de la réponse, la présence des termes de la question dans le passage et la correspondance au niveau du type spécifique sont également utilisés. Dans ce cas 36 % des questions ont une réponse correcte parmi les cinq premières extraites ;
- la troisième étudie la probabilité de transformer le passage en la question à l'aide d'un ensemble de permutations portant sur l'arbre syntaxique du passage. Ici, 31 % des questions ont une réponse correcte parmi les cinq premières.

Une combinaison des différents critères grâce un SVM a ensuite été appliqué. Les tests ont été faits sur un ensemble de 413 questions pour lesquelles il existe au moins une réponse correcte. Un MRR de 0,58 sur les cinq premières réponses a ainsi été obtenu.

Le système développé par [Téllez-Valero et al. 2010] tient compte de la validation de réponses

afin d'ordonner des réponses extraites par différents systèmes de questions réponses. Tout d'abord, les SQR recherchent en parallèle la réponse à la question posée. Puis, le module de validation de réponses pondère les réponses suivant un score de confiance. Les réponses sont alors ordonnées grâce à ce score. Le système de validation suit une approche par apprentissage avec des critères tenant compte de la correspondance du type d'entité nommée de la réponse, une similarité des différentes réponses et la proportion de termes et d'entités nommées communs au passage et à la question en tenant compte notamment de leur catégorie morphosyntaxique. L'apprentissage est effectué grâce à une combinaison d'arbres de décision fournie par la méthode bagging qui renvoie un score de confiance sur la validité de la réponse. L'évaluation est effectuée sur un ensemble de 190 questions dont la réponse est extraite par 17 systèmes de questions réponses. Le meilleur système obtient une accuracy (proportion de bonnes réponses correctes en première position) de 53 % et la combinaison permet de passer ce score à 58 %. Cette approche montre donc qu'il est judicieux de tenir compte de la validation de réponses pour combiner les résultats de différents systèmes. Toutefois il semble plus difficile d'avoir un aussi grand ensemble de systèmes de questions réponses et nous ne pouvons donc pas suivre une telle approche.

1.7 Conclusion

Ce chapitre a permis de présenter ce qu'étaient les systèmes de questions réponses ainsi que leur fonctionnement qui peut se décomposer en plusieurs phases : tout d'abord la question est analysée afin d'obtenir les informations pertinentes puis des documents censés contenir la réponse sont cherchés. Ils sont ensuite traités afin d'en sélectionner un fragment de texte. Les réponses sont enfin extraites et ordonnées à partir de cet ensemble de passages textuels. Le problème à résoudre au travers de ces processus est de trouver un passage de texte portant sur le sujet énoncé dans la question et apportant la réponse à cette question. Ces processus mettent en œuvre une succession d'heuristiques et d'approximations qui ne permettent pas de garantir que la réponse proposée est correcte. Aussi la notion de validation de réponses a été étudiée plus spécifiquement et a conduit à définir une réponse valide comme l'extrait de texte qui répond à la question posée et qui est justifié par un passage permettant de vérifier que les informations contenues dans la question puissent en être inférées. La validation de réponses est connexe à l'implication textuelle qui détecte les cas où le sens d'un texte peut être déduit de celui d'un autre. Ainsi les approches conçues pour répondre à ces deux tâches sont souvent similaires.

Parmi celles-ci, certaines sont fondées sur la comparaison de structures syntaxiques ou sémantiques de la question et du passage. Ces structures sont comparées en étudiant leurs propriétés communes ou les transformations permettant de passer d'une forme à l'autre. Les meilleures approches sont celles qui mettent en œuvre un système de preuve sur des représentations logiques. Ces différents systèmes obtiennent de bons résultats mais reposent sur des analyses en profondeur utilisant des bases de connaissances importantes développées manuellement. De telles analyses ne sont pas robustes à différents types de documents ou aisément applicables à une nouvelle langue. Ainsi les systèmes rencontrent certaines difficultés à s'appliquer aux documents provenant du Web. Les systèmes de validation ou d'ordonnancement de réponses présentés traitent tous des collections de journaux. Les systèmes de questions réponses qui ont été appliqués au Web utilisaient le filtre d'un moteur de recherche et aucun ne s'est heurté au problème de traitement des documents Web.

Le type d'approche qui présente plus de robustesse et ne demande pas la définition manuelle de connaissances effectue un certain nombre de vérifications vues comme des critères fournis à un système d'apprentissage. Parmi ces critères, certains traitent de la présence des termes de la question dans le passage. Les termes importants de la question sont souvent marqués et traités spécifiquement. D'autres critères étudient la proximité des termes de la question dans le passage. D'autres encore sont spécifiques aux systèmes de validation de réponses et vérifient, par exemple, que le type attendu par la question en réponse correspond au type de la réponse proposée. Les performances restent encore à améliorer et des critères supplémentaires à trouver. En effet, au delà des termes communs, les aspects les plus étudiés concernent la prise en compte des relations et des structures syntaxiques. Les modèles proposés se montrent prometteurs quand ils sont appliqués à des documents bien rédigées. Mais ils ne peuvent constituer une réponse pour analyser des articles provenant du Web. Une recherche de nouveaux critères pourrait consister à étudier plus précisément les informations rendant une réponse valide.

La validation de réponses est une tâche complexe. Afin d'y apporter une solution, il est nécessaire de bien comprendre ce qui en fait sa complexité. Dans ce but, nous avons réalisé une étude pour déterminer les phénomènes linguistiques à appréhender pour valider des réponses, à l'image de ce qui a été fait pour l'implication textuelle.

Chapitre 2

Corpus de justification de réponses

Sommaire

Afin d'étudier la validation de réponses et plus particulièrement la justification de réponses, il est nécessaire de comprendre les phénomènes en jeu puisque bien souvent le passage ne correspond pas directement à une forme déclarative de la question à laquelle la réponse est ajoutée.

L'étude présentée dans ce chapitre cherche à répondre aux deux questions suivantes : « Quels phénomènes rendent la justification de réponses une tâche difficile ? » et « Dans quelles proportions apparaissent-ils ? ». Afin d'y répondre un corpus de justifications a été créé [Grappy, et al. 2010].

Le corpus se doit d'être au plus près de ce que rencontre un système de questions réponses. Il est donc constitué de questions avec leurs réponses ainsi que de documents contenant la réponse et pouvant la justifier. Dans cette étude nous considérons le document en entier et pas seulement un passage de texte car généralement il contient toutes les informations permetant de justifier la réponse.

La création du corpus part d'un ensemble de questions et de leurs réponses associées et a pour but de sélectionner les documents à étudier. Une recherche est effectuée grâce à un moteur de recherche

utilisant les mots clés de la question comme le ferait un système de questions réponses classique. Afin que la réponse soit contenue dans les documents, elle est ajoutée à l'ensemble des mots permettant de créer la requête. Pour ne pas trop biaiser la justification en la rendant trop triviale, la requête ne contient qu'une partie des mots clés de la question.

Les documents sont ensuite annotés par différents experts. Deux annotations sont à effectuer : la première concerne l'état de la justification afin d'indiquer si la réponse est totalement justifiée, partiellement justifiée ou non justifiée au sein de la phrase contenant la réponse. Le but est d'évaluer en quelle mesure une phrase suffit pour justifier une réponse. La seconde s'applique quand la justification n'est que partielle et consiste à marquer pourquoi elle n'est pas complète. Différents phénomènes sont ainsi identifiés tels que l'absence d'une information dans la phrase justificative, la reformulation d'un élément ou le fait qu'une information soit distante de la réponse. Ainsi, avec ces deux annotations, il est possible de voir combien de passages nécessitent un traitement particulier ainsi que les traitements les plus importants.

Ce travail a pris sa place dans le cadre du projet CONIQUE [Moriceau, et al. 2008a]. Ce projet avait pour but d'étudier l'inférence dans un système de questions réponses. Il a donné lieu à la réalisation de différents systèmes : le premier, le système de questions réponses FIDJI, part du principe que les informations nécessaires pour justifier une réponse peuvent être réparties sur plusieurs passages [Moriceau et al. 2009]. Le second vérifie que les informations temporelles de la question se trouvent dans le passage et s'intéresse notamment au rapprochement de différentes expressions calendaires [Battistelli, et al. 2008].

Différentes personnes ont pris part à la création de ce corpus. Tout d'abord, comme ce travail entre dans le cadre du projet CONIQUE [Moriceau et al. 2008a], les membres de ce groupe ont travaillé sur l'élaboration et l'annotation du corpus. Cette étude fait aussi suite aux travaux menés par Barbier [2009] qui avait commencé les traitements et en particulier avait proposé la méthode de création du corpus. Avant de présenter la constitution du corpus, nous allons faire une revue des corpus existants et montrer pourquoi nous avons été amenés à en proposer un nouveau.

2.1 Corpus existants

Les données les plus fréquentes dans les systèmes de questions réponses viennent des campagnes d'évaluation qui permettent de collecter un ensemble de questions avec leurs réponses attendues. Dans certains cas l'évaluation porte également sur les passages de textes censés justifier les réponses. Ces données sont très utiles pour les systèmes car elles constituent un ensemble sur lequel ils peuvent s'évaluer et s'améliorer. Toutefois elles ne permettent pas, en l'état, de comprendre la difficulté d'obtenir une réponse correcte ni les différents phénomènes à traiter afin de la justifier.

La campagne AVE, (cf. section 1.2), proposait aux systèmes participants de détecter les cas où la réponse à une question est correcte et justifiée par un passage textuel. Les données de cette campagne permettent de différencier les passages justificatifs de ceux ne l'étant pas et constituent ainsi un premier pas dans la création d'un corpus de justification de réponses. Ces données pourraient servir à détecter les phénomènes rendant cette justification délicate. Toutefois, deux remarques peuvent être faites :

– les passages fournis aux participants sont relativement courts (n'excédant pas 250 caractères) ce qui ne permet pas de couvrir tous les phénomènes pouvant être rencontrés comme par exemple la relation anaphorique avec un élément éloigné ;
– comme les données ont été extraites par les systèmes de questions réponses, certains passages ont été filtrés ce qui ne permet pas d'obtenir tous les cas possibles.

Nous avons également vu précédemment que la campagne AVE est assez proche de la campagne RTE. Dans cette dernière il faut détecter les cas où le sens d'un passage textuel peut être déduit du sens d'un autre passage. Quelques études ont été menées afin de mieux comprendre cette tâche.

Vanderwende & Dolan [2006] ont effectué une étude sur les données de la campagne afin de détecter les cas où les seules informations syntaxiques et lexicales sont à utiliser afin d'observer l'implication. En se fiant uniquement à une correspondance syntaxique et un thèsaurus permettant d'avoir les variations de certains termes, 49 % des cas peuvent être reconnus. Il est donc nécessaire de tenir compte de variations syntaxiques telles que les constructions appositives présentes notamment dans l'exemple suivant :

Passage : la centrale Alameda, à l'ouest du Colorado, fut créée en 1592
Hypothèse : la centrale Alameda est à l'ouest du Colorado

Bentivogli et al. [2010a] présentent une approche permettant de séparer les phénomènes lexico-syntaxiques présents dans les paires texte-hypothèse afin d'étudier l'impact de chaque phénomène séparément. Le but final est de créer des ensembles de paires texte-hypothèse monothématiques, contenant un seul phénomène. Cinq grands types de phénomènes sont considérés :
– lexical (des acronymes, des synonymes ou des hyperonymes) ;
– lexical-syntaxique (paraphrases, verbalisation ...) ;
– syntaxique (voix active/voix passive, négation ...) ;
– discours (coréférence, apposition ...) ;
– raisonnement (utilisation de connaissances externes, métonymie ...).

La détection est effectuée par une approche en deux temps : tout d'abord, les phénomènes présents dans chaque paire sont identifiés puis le texte est transformé afin d'obtenir une forme similaire plus proche de celles de l'hypothèse. Pour ce faire, des règles de réécriture sont utilisées. Pour finir, les paires sont ordonnées en fonction du phénomène.

Passage : Doris Lessing received the 2007 Nobel Prize in Literature
Hypothèse : Doris Lessing won the Nobel Prize in Literature in 2007.

Par exemple, pour la paire de l'exemple précédent, un phénomène de synonymie a lieu entre « won » et « received ». Ce phénomène est marqué et le passage devient « *Doris Lessing won the 2007 Nobel Prize in Literature* ».

Le raisonnement est le phénomène le plus fréquent (33 %). La création de ce corpus pose le problème du lourd travail manuel à effectuer pour annoter les données ce qui limite le nombre de données pouvant être considérées. Ainsi, seules 90 paires sont examinées. Le corpus créé dans ce chapitre est relativement semblable à ce dernier à l'exception du fait qu'il concerne les systèmes de validation de réponses ce qui peut différer car les données ne sont pas les mêmes. Il peut notamment y avoir davantage de phénomènes distants tels que l'anaphore.

Bernhard, et al. [2011] s'intéressent aux dérivations morphologiques dans les systèmes de questions réponses et montrent que les phénomènes correspondent souvent à des adjectifs dénominaux (chilien, du Chili) et des nominalisations de verbe (inaugurer, inauguration).

Bédaride & Gardent [2010] s'intéressent à la détection d'une sous partie du corpus RTE centrée sur l'implication syntaxique afin d'évaluer les systèmes uniquement sur cet axe. La méthode de création s'appuie sur un ensemble de règles de transformation telles que le passage à la voix passive. Ces règles comportent des informations concernant le type de prédicat et d'attribut de chaque verbe. A partir de ces données, des transformations sont effectuées sur un ensemble de phrases afin d'en obtenir de nouvelles syntaxiquement différentes. Une intervention humaine permet enfin de détecter si le sens de la nouvelle phrase peut être impliqué de celui de la précédente.

D'autres corpus dédiés à des tâches particulières de questions réponses existent. C'est par exemple le cas du travail présenté dans [Rosset & Petel 2006] qui a pour but la création d'un ensemble de questions orales utilisables ensuite dans un système de questions réponses. Garcia-Fernandez [2010] s'intéresse à la formulation des questions et des réponses en essayant d'identifier la réponse la plus appropriée pour la question. Dans ce cas, la réponse n'est pas concise et reprend certaines informations de la question comme le fait un être humain. Le corpus créé porte donc sur les réponses données par différentes personnes.

D'autres corpus s'intéressent quant à eux à l'apport de certains phénomènes linguistiques pour les systèmes de questions réponses. C'est par exemple le cas de [Boldrini, et al. 2009] qui traite des anaphores dans le but d'améliorer les interactions entre l'utilisateur et le système.

Bien que de nombreuses études existent, il n'en existe pas sur le français portant sur la justification de réponses et indiquant les différents phénomènes qui nous intéressent. Les phénomènes que nous voulons mesurer sont ceux qui rendent difficiles un appariement question-passage du point de vue des éléments qui doivent s'aligner. Ainsi ils vont davantage concerner l'absence d'une information que les différents types de variations syntaxiques présentes.

2.2 Création du corpus

Le corpus souhaité doit permettre de quantifier les phénomènes de traitement automatique des langues à traiter pour justifier des réponses. Il se doit donc d'être représentatif de ceux-ci. Cette section commence par présenter les contraintes dont il faut tenir compte pour la création du corpus puis détaille la création du corpus en elle-même.

2.2.1 Les contraintes

Les systèmes de questions réponses recherchent les réponses dans des documents et plus particulièrement dans des passages relativement courts. La première remarque est que souvent, comme ces passages sont de petite taille, ils ne contiennent pas toutes les informations nécessaires pour justifier la réponse. Les évaluations des systèmes de questions réponses à TREC [Voorhees & Tice 1999] sont effectuées en vérifiant que la réponse est justifiée dans le document en entier. Dans notre étude, nous avons choisi de considérer les documents en entier car cela peut permettre d'observer davantage de

phénomènes tels que l'anaphore. La taille des documents peut donc varier d'un document à l'autre et passer de quelques lignes à quelques pages.

Les questions utilisées doivent être au plus près de celles rencontrées par les systèmes de questions réponses. Comme ceux-ci traitent davantage des questions conçues pour participer aux campagnes d'évaluation que des questions posées directement par des utilisateurs des systèmes, elles sont relativement bien formulées. Par exemple une question posée directement par une personne pourrait être « *C'est quand déjà que Nelson Mandela est sorti de prison ?* » alors que la même question utilisée lors d'une campagne est « *En quelle année Nelson Mandela est-il sorti de prison ?* ». Ainsi, nous avons directement utilisé des questions provenant de campagnes d'évaluation. Ces questions sont issues, pour la plupart, de la campagne EQueR [Ayache, et al. 2006].

Afin de pouvoir rencontrer des phénomènes différents sur un maximum de questions, certaines questions de cette campagne ne sont pas conservées. C'est le cas de toutes les questions de définition, des questions booléennes et des questions trop "simples" qui consistent à dater ou à localiser un élément sans avoir d'autres informations à vérifier comme « *Où est Paris ?* ». Pour ce type de questions quelques phénomènes peuvent rendre difficile la justification. Toutefois, ce sont les mêmes pour toutes les questions du même type (question de définition, lieu simple ...) ce qui rend ces questions peu intéressantes pour la création du corpus si elles figurent en trop grand nombre.

Lors de la création des campagnes de questions réponses, les questions sont parfois fortement inspirées des informations présentes dans les documents aussi bien au niveau de la forme que du fond. Par exemple, une question portant sur l'âge de Nelson Mandela lors de sa libération peut se trouver être « *A quel âge Nelson Mandela sortit-il de prison ?* » si dans le passage le verbe sortir est employé. Une étude portant sur les passages fournis en réponse par les systèmes participant à la campagne de questions réponses EQueR [Grau 2005] montre que 40% des passages correspondant à un triplet réponse valide contiennent tous les mots de la question sans variations. Dans 35% des cas un seul mot est absent ou sous forme de variante ; il correspond dans 30 % des cas au type spécifique et dans 52 % au verbe principal.

Afin que les passages ne soient pas trop proches des questions, les réponses seront cherchées dans un autre ensemble de documents que celui ayant servi, dans la campagne d'évaluation, à créer les questions.

Afin que l'étude soit significative, il est nécessaire de considérer un nombre suffisant de questions que nous avons estimé à 300 ce qui est supérieur au nombre de questions utilisées lors de campagnes d'évaluation des systèmes de questions réponses. Comme certaines questions de la campagne EQueR n'ont pas été retenues, d'autres questions issues de différentes campagnes CLEF ont été également considérées. Au final, le corpus contient 290 questions. Dans ces campagnes, les documents concernés correspondent aux articles des journaux Le Monde et Le Monde diplomatique et les dépêches SDRT. Comme le corpus de documents doit être différent, les documents pertinents sont recherchés dans l'encyclopédie Wikipédia française. Cela a l'avantage de pouvoir permettre la diffusion libre du corpus.

Afin que les données à annoter soient au plus près des données traitées par un système de questions réponses, les documents sont recherchés automatiquement à partir des mots clés de la question. Bien souvent, les documents ne contiennent pas la réponse. Les différentes réponses aux questions

concernées étant connues, elles sont utilisées afin de s'assurer que chaque document contient effectivement la réponse.

2.2.2 Sélection des documents

Après avoir présenté les contraintes portant sur le corpus, voyons maintenant sa création. Elle est effectuée de manière semi-automatique grâce à une succession de quatre étapes permettant d'extraire, pour chaque question, les documents à annoter grâce à un moteur de recherche appliqué sur des documents contenant des mots lemmatisés ce qui permet de reconnaître certaines variations d'un terme. La difficulté est alors de détecter les mots clés pertinents pour constituer la requête. Ce travail a été en grande partie effectué par Barbier [2009]. Afin d'illustrer le mécanisme mis en place, nous reprenons son exemple avec la question « *Quel volcan a détruit l'ancienne cité de Pompéi* ? ».

Tout d'abord une analyse automatique de la question a lieu afin d'en extraire les mots significatifs choisis selon leur catégorie morphosyntaxique (verbe, nom, adjectif). A la question de l'exemple, les mots "volcan", "détruire", "ancien", "cité" et "Pompéi" sont retenus. Ces mots sont ensuite ordonnés suivant leur importance. Dans un premier temps cet ordre suit les catégories morphosyntaxiques. Un nom propre est plus important qu'un nom commun qui est lui-même plus important qu'un adjectif. L'ordre obtenu pour l'exemple est le suivant {Pompéi, détruire, cité, volcan, ancien}. Puis une étape de réordonnancement est effectuée manuellement sur ces données. Cela permet de différencier deux mots de même catégorie morphosyntaxique (par exemple des noms communs) ainsi que de corriger les erreurs ayant pu avoir eu lieu. Dans notre exemple l'ordre final est « Pompéi> détruire > volcan> cité> ancien » . L'ordonnancement manuel a permis de marquer « volcan » comme plus important que « cité » qui est plus souvent utilisé.

Des requêtes sont ensuite créées à partir de ces mots ordonnés. Dans un système de questions réponses classique, le fait que tous les mots soient présents dans le passage indique que ce dernier est pertinent puisqu'il est susceptible de contenir la réponse. Afin de couvrir un maximum de variations dans les documents, de nombreuses requêtes sont créées. Chaque requête est effectivement construite à partir des mots clés de la question sans tous les contenir. Si un mot est absent d'une requête alors il n'est pas forcément contenu tel quel dans les documents. Ainsi, si la réponse est justifiée, alors l'information correspondant à ce mot doit se trouver sous une forme différente, par exemple un synonyme. Afin d'obtenir un corpus le plus représentatif possible, les documents ne doivent pas contenir trop de mots de la question. Dans ce but, ceux-ci sont pondérés avec un poids tenant compte de l'importance de leur catégorie morphosyntaxique et la somme des poids des mots contenus dans chaque requête ne doit pas dépasser une certaine valeur seuil. Les requêtes sont alors créées en choisissant les mots selon leur ordre. Afin que chaque mot soit utilisé, les mots n'étant pas considérés dans une requête seront considérés dans une autre.

Une nuance a toutefois lieu pour les noms propres. Ils correspondent aux informations les plus importantes de la question et pour lesquels il existe peu de variations. Ces mots sont donc volontairement présents dans chacune des requêtes afin de limiter le nombre de documents non pertinents. Dans ce but, la réponse est également ajoutée à l'ensemble des mots de chaque requête. Ainsi même si les documents ne justifient pas la réponse, ils la contiendront tous. Globalement, de nombreux documents susceptibles de justifier la réponse sont renvoyés.

Pour l'exemple précédent, les requêtes suivantes sont calculées :

1. Pompéi, cité, volcan, ancien (mot retiré détruire), Vésuve
2. Pompéi, détruire, cité (mot retiré volcan), Vésuve
3. Pompéi, détruire, volcan (mot retiré ancien), Vésuve

Nous voyons que chaque mot est absent d'au moins une requête, à part le nom propre Pompéi. L'adjectif ancien présent dans la première requête (celle sans verbe) n'est pas contenu dans les autres car le poids d'un verbe est supérieur à celui d'un nom et il est donc absent par filtre sur le poids des mots. Pour les 290 questions considérées, une moyenne de 1,6 requêtes par question a été obtenue.

2978 documents sont trouvés par le moteur de recherche Lucene auquel a été fourni chacune des requêtes. Le nombre de documents varie d'une question à une autre et peut aller de quelques uns à une centaine.

2.3 Guide d'annotation

Après avoir créé les données, l'étape suivante consiste à les annoter. Pour ce faire, une interface graphique a été reprise et modifiée et un guide d'annotation fourni à sept experts du domaine a été créé par les membres du projet CONIQUE. L'annotation et son guide ont été effectués en différentes étapes. Tout d'abord les annotateurs se sont mis d'accord sur un guide d'annotation en utilisant quelques exemples puis la répartition des données a été faite et chaque annotateur a annoté ses données. Une réunion s'en est suivie durant laquelle les difficultés rencontrées par les annotateurs ont été présentées et le guide a été modifié afin de tenir compte des remarques et suggestions faites. Dans un dernier temps, les données ont été réannotées en suivant ces modifications. Voyons le guide d'annotation.

2.3.1 Réponse justifiée ou Non

Le premier choix à faire est de savoir si une réponse est justifiée ou non. Les différentes annotations ont été effectuées en partant de la phrase contenant la réponse. Cette notion correspond au passage minimal utilisé par les systèmes de questions réponses lors de l'extraction des réponses. Quatre valeurs permettent d'annoter les réponses :

N/A : correspond aux cas où aucune valeur n'a été donnée car l'annotateur n'a pas encore traité cet extrait. Cette valeur sera modifiée dès qu'une annotation sera effectuée.

OUI : La réponse est complètement justifiée. Toutes les informations que la question donnait sont présentes dans la même phrase que la réponse et de manière analogue à celle de la question. Il n'y a pas de variation sémantique, morphologique ou anaphorique, seules des variations syntaxiques simples sont admises. Ces variations ne sont pas considérées car elles ne posent pas de problème particulier.

Question : En quelle année la catastrophe de Tchernobyl a-t-elle eu lieu ?
Réponse : 1986
Phrase : 20 ans après la catastrophe de Tchernobyl survenue le 26 avril 1986 ...

Partiellement : La réponse est partiellement justifiée car certaines informations ne sont pas présentes dans la phrase contenant la réponse ou apparaissent sous forme de variantes. Elles sont donc soit absentes du document, soit éloignées de la réponse, soit sous une autre forme (cf. section 2.3.2) ;

NON : La réponse n'est pas valide c'est-à-dire qu'elle est incorrecte ou que le document ne la justifie pas. Dans notre cas, seul ce cas est possible car la bonne réponse est forcément contenue dans le document. En pratique cela se rencontre souvent quand la question attend une date en réponse mais que dans le passage la date ne porte pas sur l'entité à dater.

Question : En quelle année Jacques Chirac est-il devenu président ?
Réponse : 1995
Passage : Le 16 juillet 1995, dans une allocution, à l'occasion du 53e anniversaire de la rafle du Vélodrome d'Hiver, Jacques Chirac reconnaît « la faute collective » de la France.

2.3.2 Réponses partiellement justifiées

Quand une réponse est partiellement justifiée, il faut indiquer les informations manquantes ou reformulées. Bien souvent, il ne s'agit pas uniquement d'un seul phénomène et tous sont à signaler. Tout d'abord, comme plusieurs occurrences de la réponse peuvent se trouver dans le document, il est nécessaire de marquer la réponse concernée en indiquant son numéro de ligne ainsi que la nature des phénomènes (paraphrase, anaphore ...). Afin de pouvoir étudier plus précisément les phénomènes repérés, les annotateurs doivent également indiquer les mots sur lesquels portent les transformations, le mot présent dans la question et le mot présent dans le document. Ces informations sont bien sûr utiles pour l'étude de la justification de réponses mais peuvent aussi être utilisées pour la création d'un corpus propre à chaque type de phénomène. Étudions maintenant, plus en détail, les phénomènes à annoter.

Le type de réponse : Certaines questions attendent une réponse d'un type particulier, par exemple la question « *Quel premier ministre français s'est suicidé ?* » attend en réponse un premier ministre. Or, il arrive que l'information de type ne soit pas aux alentours de la réponse. Elle peut soit ne pas être présente dans le document car l'information semble triviale et connue de tous, soit être présente mais être relativement éloignée, dans une phrase différente. Dans ces deux cas, il faut indiquer que le type de réponse est absent de la phrase de la réponse ainsi que le type de la réponse (premier ministre). Notons toutefois que quand le type spécifique correspond au type d'entité nommée attendue en réponse (lieux, date, organisation), l'information de type n'est pas à signaler. En effet, bien que ces informations soient très peu souvent indiquées explicitement dans les textes, elles sont généralement connues par un système de reconnaissance des entités nommées qui les a marquées lors de l'analyse des passages ou du prétraitement des documents.

Question : Quel premier ministre français s'est suicidé ?
Réponse : Pierre Bérégovoy
Passage : Pierre Bérégovoy s'est suicidé le 1 mai 1993.

Information manquante : Il est souvent nécessaire de vérifier plusieurs informations pour justifier une réponse. Par exemple, afin de s'assurer que la réponse « Pierre Bérégovoy » est valide à la question « *Quel premier ministre s'est suicidé en 1993 à Nevers ?* » il est nécessaire de montrer que :

- la réponse est un premier ministre ;
- il s'est suicidé ;
- l'action a eu lieu à Nevers ;
- l'action se déroule en 1993.

Certains documents peuvent ne contenir qu'une partie des informations à justifier. Ainsi la date ou le lieu peuvent être absents du document sans que cela ne modifie l'événement si l'on fait référence à un événement unique. Dans un tel cas, la réponse est partiellement justifiée car ce ne sont pas des informations fondamentales. En revanche, si l'action, se suicider, avait été absente la réponse n'aurait pas été justifiée. En considérant le passage précédant à la question « *Quel premier ministre s'est suicidé en 1993 à Nevers ?* » et en supposant que l'information de lieu n'est pas présente dans le document alors il faudra indiquer cette absence ainsi que l'information manquante.

A ce propos, on peut noter que le système de questions réponses FIDJI [Moriceau et al. 2009] combine les informations obtenues dans différents documents. Pour cet exemple, un document peut par exemple indiquer que Pierre Bérégovoy est premier ministre et un autre qu'il s'est suicidé.

Sémantique/Paraphrase : Certains documents peuvent contenir une information donnée sous une forme différente de celle de la question et dans ce cas il faut indiquer le terme de la question ainsi que sa reformulation.

Passage : Pierre Bérégovoy se tua d'une balle dans la tête.

Coréférence : Dans certains cas, un groupe nominal de la question utile pour justifier la réponse se trouve assez distant de la réponse, dans une phrase différente, mais un groupe de mots présent dans la phrase contenant la réponse lui fait référence. Il y a donc un phénomène de coréférence. Un autre cas possible est celui où la réponse elle-même est reprise par une expression référentielle par exemple dans la phrase « *Il s'est suicidé en 1993* », le pronom « il » fait référence à Bérégovoy. Pour marquer ces phénomènes, l'antécédent et le coréférent doivent être indiqués. Une information intéressante peut être de marquer l'emplacement, le numéro de ligne, de l'antécédent. Cela permettra de collecter ces données pour travailler plus spécifiquement sur la coréférence. Si une succession d'anaphores permet de relier les deux termes alors seules les extrémités sont à mentionner. Par exemple, dans l'exemple suivant, un phénomène d'anaphore a lieu pour lequel il faut marquer le coréférent (il) et l'antécédent (Luke Skywalker).

Question : Qui est la sœur de Luke Skywalker ?
Réponse : la princesse Leïa.
Passage : Luke Skywalker est un jeune jedi. Il a pour sœur la princesse Leïa.

Contexte : Dans certains cas une information à vérifier peut se trouver dans une phrase différente de celle de la réponse sans avoir de coréférent dans la phrase de la réponse. Il se peut en effet que même distante, la présence de cette information porte également sur la phrase contenant la réponse.

Par exemple, la page Wikipédia consacrée à 1993 a comme titre « 1993 » comme sous titre « *Décès en 1993* » et contient la phrase « *1° mai : Pierre Bérégovoy, homme politique, ancien premier ministre français.* ». Dans ce cas, ces informations distantes indiquent que la mort de Bérégovoy a eu lieu en 1993. Pour signaler un tel phénomène il faut indiquer le phénomène ainsi que le mot en contexte.

Curiosité : Dans certains cas rares la réponse peut se trouver partiellement justifiée mais le fait qui rend cette justification partielle est une combinaison complexe des points précédents et relève de connaissances externes ou de raisonnement. On appellera cela une curiosité puisque la réponse est justifiée mais semble difficile à retrouver automatiquement. Dans l'exemple suivant, un mécanisme d'inférence a lieu puisque la date de la création de la zone euro correspond à la date du passage à la monnaie unique.

Question : Quelle est la date du passage à la monnaie unique ?
Réponse : 1999
Passage : La zone euro est une zone monétaire qui regroupe les pays de l'Union Européenne qui ont adopté l'euro comme monnaie unique. La zone euro a été créée en 1999 par onze pays...

2.4 Outil d'annotation

2.4.1 Interface graphique

Afin d'effectuer les annotations, une interface graphique a été réalisée, adaptée de [Barbier 2009]. Elle présente la question, la réponse et le document à un annotateur (cf. figure 2.1). L'annotateur devant parcourir un ensemble de questions, deux composants sont mis en œuvre : un système de flèches afin de passer d'une question à la suivante (Q») ou à la précédente («Q) et un système permettant directement d'arriver à la question souhaitée en indiquant son numéro. Ce dernier mécanisme est particulièrement utile car chaque annotateur a des questions différentes à traiter.

Un système similaire permet de passer d'un document à l'autre pour la même question. Ici deux mécanismes sont également utilisés : le premier par simple flèche permet d'arriver au triplet réponse voulu (R+) et (R-). Le second s'applique quand le document a été traité et correspond au bouton « Suivant » qui permet de sauvegarder les informations annotées et de passer au suivant.

Comme les documents peuvent être longs, un mécanisme de clic permet d'arriver directement aux différentes instances de la réponse sans avoir à parcourir tout le document.

Intéressons-nous maintenant à la saisie d'une annotation. Afin d'indiquer les réponses partiellement justifiées il faut marquer la réponse avec son numéro de ligne, 3 dans l'exemple présenté dans la figure. Dans ce but chaque phrase correspond à une ligne différente et le numéro de la ligne est également précisé. De plus il faut indiquer les phénomènes rendant la justification partielle. Pour chaque phénomène, il faut ainsi marquer son type ainsi que les informations reliées qui correspondent le plus souvent au terme présent dans la question et à celui présent dans le passage. Au niveau de l'interface graphique, un ensemble de boîtes est utilisé. Une boîte correspond à chaque réponse et dans chacune d'elles un autre ensemble tient compte de chaque phénomène. Les sous-boîtes contiennent alors des

boutons permettant de marquer le type de phénomène et des zones de textes afin de marquer les informations. Dans l'exemple, une boîte correspond à la réponse de la ligne 3 qui contient deux sous-boîtes une pour indiquer le terme du passage, l'autre pour le terme de la question.

Afin de faciliter le traitement, nous avons aussi modifié l'affichage du texte afin de surligner la réponse en rouge et les termes de la question en bleu.

2.4.2 Exemples

Pour bien comprendre les annotations ainsi que l'interface graphique voyons maintenant deux exemples. Le premier correspond au triplet réponse suivant :

Question : Où Marcos fut-il dictateur ?
Réponse : Philippines
Passage : ... La présidence de Marcos aux Philippines.

Dans cet exemple, la justification aurait été complète si au lieu de « présidence » on avait eu « dictature » puisqu'on ne s'intéresse pas aux dérivations morphologiques. Comme, sous certaines considérations, « être dictateur » a le même sens que « la présidence de », la justification est notée comme partielle. L'information à marquer est donc la paraphrase et il faut indiquer les deux groupes de mots « présidence de » et « dictateur » ce qui se traduit graphiquement par les informations données dans la figure 2.1. Cet exemple permet de bien comprendre l'interface graphique avec les annotations, la réponse partiellement justifiée et l'indication d'une relation de paraphrase entre les deux termes. De plus nous pouvons voir que la requête contient seulement le terme « Marcos » car la question ne contient que deux mots non vides et que la requête ne doit pas contenir tous les termes de la question tout en contenant ses noms propres.

Un exemple plus complexe correspond au triplet réponse suivant (cf. figure 2.2) :

Question : Dans quelle grande capitale la Tour Eiffel fut-elle érigée en 1889 ?
Réponse : Paris
Phrase 1 : La *tour Eiffel* est une tour de fer puddlé *construite* par Gustave Eiffel et ses collaborateurs pour l'Exposition universelle de *1889*.
Phrase 2 : Situé à l'extrémité du Champ-de-Mars, en bordure de la Seine, ce *monument parisien*, symbole de la France et de sa *capitale* est l'un des sites les plus visités du pays.

La réponse est donnée par le terme « parisien » qui est une de ses variations dérivationnelles. De plus, les informations sont réparties entre les deux phrases. Ainsi, dans la phrase 2, la « tour Eiffel » est remplacée par un coréférent, monument. Le terme 1889 ne se trouve pas dans la phrase 2 mais l'information est connue à l'aide du contexte du passage. La construction du monument est présentée d'une manière différente à celle présente dans la question (érigée vs construite) et uniquement dans la phrase 1.

Les résultats des différentes annotations sont sauvegardés dans un fichier XML. A chaque requête est associé un fichier dans lequel est indiqué le contenu des documents ainsi que l'ensemble des annotations le concernant (état de la réponse, phénomènes rencontrés et les différentes remarques).

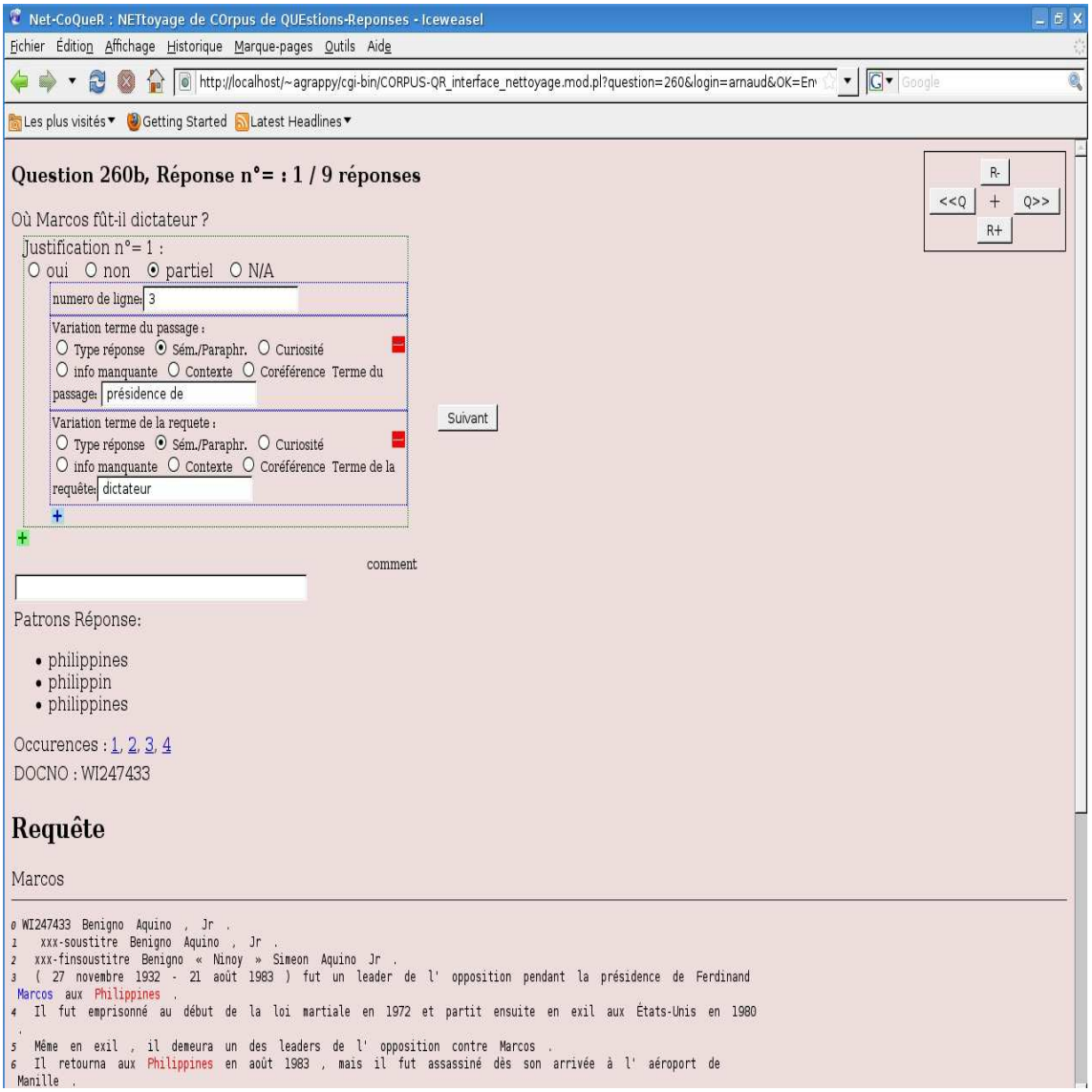

FIG. 2.1 – Annotation d'une justification partielle

2.5 Analyse du corpus

2.5.1 L'annotation

L'annotation a été effectuée par sept experts du domaine qui ont eu pour mission de détecter les réponses justifiées dans les passages et les phénomènes rendant la justification partielle. Le corpus contient 290 questions et 2978 passages.

La première étape a consisté en une répartition des questions afin que chaque annotateur annote à peu près autant de questions. Tout d'abord, les données sont séparées en deux groupes. Dans le

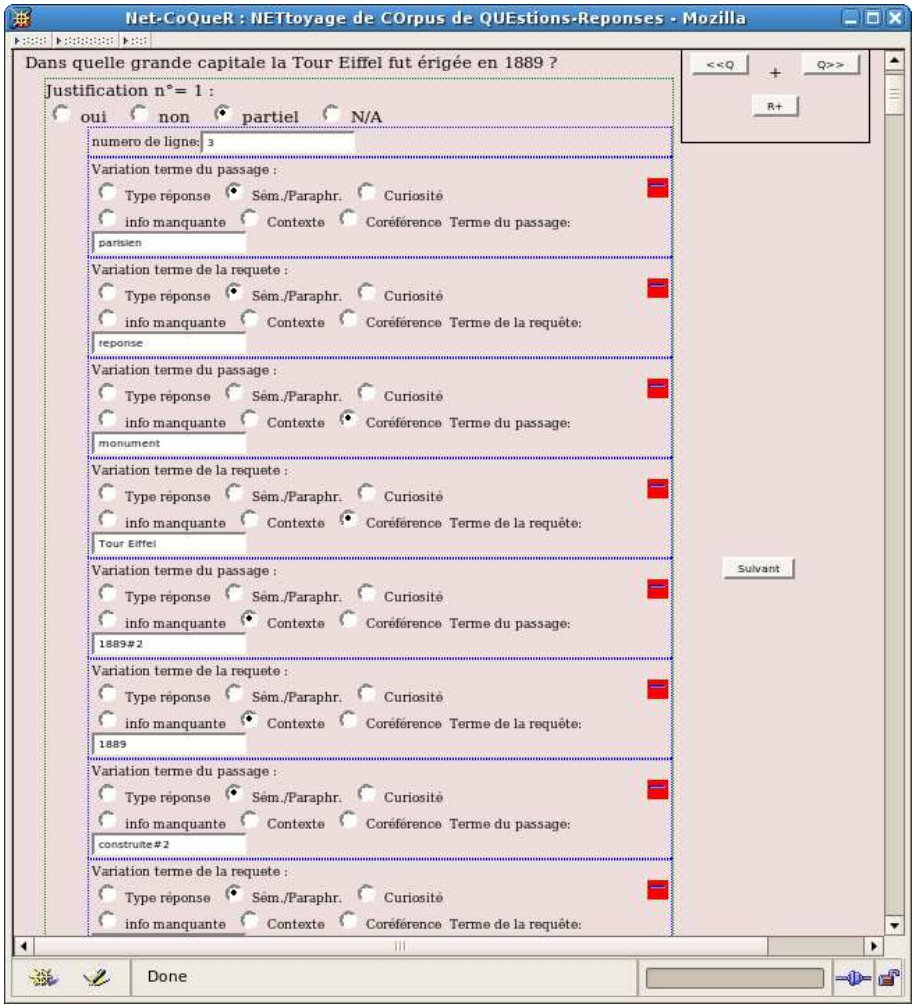

FIG. 2.2 – Annotation d'une justification complexe

premier se trouvent les questions qui seront annotées par un seul annotateur et dans le second celles qui seront annotées par plusieurs annotateurs. Ce second ensemble permet de savoir si les annotateurs s'entendent sur les annotations à effectuer. Il doit être assez grand pour être significatif de l'ensemble des données. Toutefois il ne faut pas que les annotateurs aient trop de données à annoter. Cet ensemble contient 80 questions ce qui correspond à un peu plus d'un quart des données.

Les questions annotées par un seul annotateur ont été réparties équitablement entre les participants qui ont donc dû traiter 30 questions chacun. La base d'annotation commune a aussi été répartie de manière à ce que chaque couple de participants ait trois questions en commun. Tous les couples d'annotateurs possibles ont été formés afin d'avoir une mesure d'accord entre annotateurs la plus significative possible.

2.5.2 Résultats globaux

Le premier résultat présente la répartition des justifications (cf. tableau 2.1). Pour la calculer, nous avons comptabilisé le nombre de documents contenant au moins une réponse justifiée (OUI), le nombre de documents contenant au moins une réponse partiellement justifiée sans en contenir de complètement valide (PARTIEL) et les autres documents (NON).

	#	%
OUI	201	6,8%
PARTIELLE	679	22,8%
NON	2098	70,4%

TAB. 2.1 – Répartition des justifications

Nous pouvons noter que la plupart des documents ne contiennent pas de réponses justifiées même partiellement puisque seuls 30 % des documents justifient la réponse. De plus, le corpus contient trois fois plus de réponses partiellement justifiées que de réponses totalement justifiées. Cette information montre qu'un système de questions réponses, pour être performant, doit tenir compte des différents phénomènes relevés.

Comme les systèmes de questions réponses se placent au niveau des questions et non à celui des documents, il est nécessaire d'observer cette répartition pour chaque question. Ainsi, il y a 80 questions (27,5%) pour lesquelles il existe au moins une réponse totalement justifiée, 125 (43 %) avec une justification partielle associée et 85 (29,5 %) questions n'ont pas de réponse justifiée. Ces informations montrent qu'en tenant compte des différents phénomènes, bien plus de questions peuvent avoir une réponse correcte associée (256 %) par rapport à ne pas en tenir compte.

Il y a 679 passages contenant une justification partielle. L'étape suivante consiste à étudier les phénomènes rendant la justification partielle. Le tableau 2.2 présente le nombre d'occurrences de chaque type de phénomène ainsi que le pourcentage de passages contenant chacun d'eux.

Types de phénomènes	**#**	**% / 679**
Paraphrase	400	59%
Information manquante	173	25%
Contexte	121	18%
Coréférence	86	12,5%
Type de réponse manquant	78	11,5%
Curiosité	142	21%
Total	1000	

TAB. 2.2 – Répartition des différents phénomènes pour les réponses partiellement justifiées

Nous pouvons tout d'abord voir qu'un même passage possède différents phénomènes reliés, en moyenne 1,5 phénomènes par passage. La variation sémantique et de paraphrase est le phénomène

le plus rencontré puisqu'il est présent dans près de 60% des passages. Cette variation correspond souvent à un synonyme d'un verbe. La valeur élevée d'information manquante (25% des passages) et de type manquant (11%) indiquent qu'il faudrait chercher certaines parties d'une question dans d'autres documents. La proportion de documents rencontrant un problème de type semble faible ; cependant il est à lier avec le fait qu'il ne s'applique pas à toutes les questions. Sur les annotations faites sur les questions mentionnant un type spécifique, ce phénomène se rencontre dans 16 % des cas.

Nous avons aussi considéré la répartition des réponses partiellement justifiées et totalement justifiées en fonction du type de questions (cf. figure 2.3). La catégorie la plus présente correspond aux questions factuelles de types « Quel » comme « *Quel est le président des États Unis ?* » car ce sont les questions demandant le plus d'informations à vérifier et donc les plus utiles pour la création de notre corpus. Nous pouvons aussi remarquer que le fait de détecter les réponses partielles permettrait d'améliorer les résultats obtenus, et ce, quelle que soit la catégorie de la question.

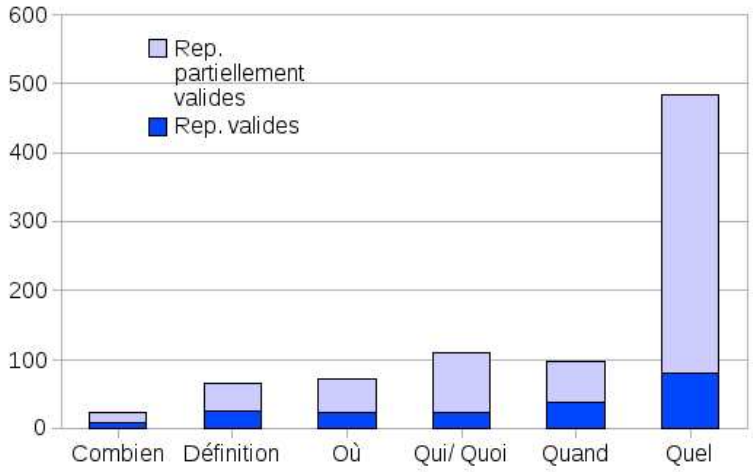

FIG. 2.3 – Répartition des justifications en fonction de la catégorie de la question

Une étude de corpus menée par Barbier [2009] sur des articles de journaux en anglais étudie également la justification de réponses et plus particulièrement deux phénomènes : la variation sémantique et la répartition spatiale des éléments. Pour le premier cas, il a été montré que les termes se retrouvent dans 62 % des cas sous la même forme dans le passage et dans l'hypothèse. Ainsi dans 38 % des cas une variation peut être rencontrée ce qui se rapproche du phénomène de variations sémantiques (voir sa thèse pour une analyse fine des types de variations). Le deuxième axe montre que seules 72 % des justifications se limitent à une phrase et qu'une justification de 3 phrases pourrait couvrir 89 % des cas.

2.5.3 Accord entre annotateurs

Il est nécessaire de comparer les différents résultats fournis par différents annotateurs afin d'évaluer la cohérence globale des résultats. C'est pour cette raison qu'une partie des données est annotée par plusieurs annotateurs.

La première cohérence recherchée porte sur l'état des passages. Contiennent-ils une réponse pleinement justifiée ? partiellement justifiée ? ou non justifiée ? Pour calculer cet accord, la mesure kappa est utilisée [Cohen 1960]. Elle a pour but de mesurer l'accord entre deux juges en se basant sur les probabilités d'apparition. Deux mesures sont utilisées : la probabilité d'accord entre deux annotateurs (Po) et la probabilité aléatoire d'accord (Pe). Ces probabilités sont calculées grâce à une matrice représentant les données observées. Le tableau 2.3 montre un exemple de matrice utilisée pour calculer la cohérence sur l'état des passages. Il correspond à l'accord entre deux annotateurs.

	OUI	NON	PARTIELLE	TOTAL
OUI	4	0	**2**	6
NON	0	37	**1**	38
PARTIELLE	0	**1**	2	3
TOTAL	4	38	5	47

TAB. 2.3 – Exemple de matrice de confusion pour calculer la mesure kappa

$$\begin{aligned} Po &= \frac{1}{\#exemples} * \#accords \\ &= \frac{1}{47} * (4 + 37 + 2) \\ &= 0,91 \end{aligned} \qquad \begin{aligned} Pe &= \frac{1}{\#exemples^2} * \sum_{i=0}^{n} n_{i.} n_{.i} \\ &= \frac{1}{47^2} * (6 * 4 + 38 * 38 + 3 * 5) \\ &= 0,67 \end{aligned}$$

$$\begin{aligned} kappa &= \frac{Po - Pe}{1 - Pe} \\ &= \frac{0,91 - 0,67}{1 - 0,67} \\ &= 0,73 \end{aligned}$$

Afin de mesurer l'accord entre tous les annotateurs nous avons calculé la moyenne des différents kappas correspondant à deux annotateurs et obtenu la valeur 0,63 ce qui correspond à un kappa élevé. Le principal désaccord a lieu quand un évaluateur donne la valeur PARTIELLE et l'autre NON :

certains annotateurs considèrent que lorsque trop de connaissances sont requises la réponse n'est pas justifiée tandis que d'autres la marquent comme une curiosité.

Un calcul similaire a été effectué afin de mesurer l'accord sur les phénomènes rencontrés (paraphrase, anaphore, ...) quand deux annotateurs voient la réponse comme PARTIELLE. La mesure obtenue dans ce cas est de 0,59 ce qui correspond à un kappa modéré. Les principales différences viennent du couple « élément en contexte » , « élément manquant ». Ces deux informations sont données quand un élément est absent d'une phrase et que seuls certains annotateurs ont vu l'information dans une phrase antérieure.

2.6 Conclusion

Ce chapitre a présenté la création d'un corpus permettant d'évaluer la justification de réponses en pointant plus spécifiquement les phénomènes rencontrés lors de la validation de réponses. L'élaboration de ce corpus est originale puisqu'aucun travail n'avait encore été fait sur des documents complets. Les différents phénomènes rencontrés sont les suivants : paraphrase, anaphore, mot absent du passage ou en contexte et type de la réponse absent. Cette étude a montré qu'il fallait tenir compte de ces phénomènes linguistiques afin de justifier près de 43 % des questions. Parmi ceux-ci, la variation sémantique est celui le plus présent, il apparaît dans 59 % des passages. Deux autres phénomènes sont aussi importants : l'absence d'un mot du document et l'absence du type de la réponse. Ces informations nous fournissent une première indication sur le système de validation de réponses à mettre en place en nous permettant d'évaluer l'importance de chaque phénomène.

Les annotations effectuées permettent de définir notre système de validation de réponses, dans le chapitre 3. Celui-ci s'appuiera sur deux grands types de vérifications des informations contenues dans la question : celles devant se trouver dans le passage et celles à rechercher en dehors du document. Ces informations correspondent le plus souvent au type spécifique car c'est une information relevant de la culture générale. Mais d'autres informations peuvent manquer, aussi notre modèle devra-t-il permettre de décider si l'information manquante est primordiale ou non. Comme nous l'avons vu, une information se situe souvent dans le document sans être contenue dans la phrase correspondant à la réponse. C'est pourquoi nous effectuerons la vérification des informations données dans la question sur des passages, dont il faudra déterminer la taille.

Chapitre 3

Le système de validation de réponses

Sommaire

3.1 Définition de la méthode choisie

Nous avons vu ce qu'est la validation de réponses, les méthodes pour la traiter ainsi que les phénomènes caractérisant cette tâche. Il est maintenant temps de déterminer notre approche permettant de reconnaître la validité d'une réponse.

Tout d'abord, nous considérons une réponse valide si toutes les informations permettant de la justifier peuvent être inférées du passage justificatif ou sont connues par un utilisateur. Par exemple, à la question « Quel président succéda à Jacques Chirac ? » le passage « Nicolas Sarkozy succéda à Jacques Chirac à la tête de la France » justifie bien la réponse « Nicolas Sarkozy » sans qu'il mentionne que « Nicolas Sarkozy » soit un président. Néanmoins, bien que ce type d'information soit connu de tous, il est nécessaire pour un système de les vérifier. Un bon passage devra avoir une taille suffisante afin de couvrir les phénomènes distants mais il devra également être suffisamment restreint pour que la réponse soit reliée aux termes de question.

Nous avons aussi vu que trois types d'approches permettent de détecter la validation de réponses : les premières s'appuient sur les formes syntaxiques du passage et de la question, les secondes sur les formes logiques et les dernières sur la combinaison de différentes vérifications par apprentissage. Les deux premières nécessitant des analyses en profondeur s'appliquent sur l'anglais mais plus difficilement sur le français ce qui est notamment dû à l'absence de bases de connaissances ou de réseau sémantique disponibles pour tous comme WordNet. Aussi, à moins de développer ce type de ressource, et ce en fonction de la langue et des documents traités, ce qui a été fait pour le système Qristal [Laurent et al. 2005], ce type d'approche ne peut être mis en œuvre. Nous avons donc choisi de suivre une approche plus facilement transposable et robuste, une approche par apprentissage. Cela permet de combiner différentes approches portant sur la vérification de différents types d'informations. De plus, le modèle prendra ses décisions en fonction de l'importance des différents critères, importance qui aura été déterminée automatiquement au vu d'exemples. En effet, il est difficile d'évaluer a priori l'importance de certains termes de la question, en dehors du contexte du document. La question restante est donc « Quels sont les critères à utiliser ? ». Comme les critères peuvent correspondre à différentes vérifications, la question peut donc se transformer en : quelles sont les informations devant être vérifiées ? Et comment les vérifier ?

Nous considérons qu'une question est composée d'éléments atomiques à justifier. Un élément atomique correspondant à un élément minimal contenant une seule information demandée ou contenue dans la question comme son action, le lieu et la date de cette action ou le type de la réponse. Une fois cette décomposition effectuée, il reste à vérifier chacune des informations, ce qui peut être fait séparément les unes des autres et dans des ressources différentes. Par exemple la question « Dans quelle grande capitale la Tour Eiffel fut-elle érigée en 1889 ? » est décomposée en :

- la réponse est une grande capitale et une ville (type de la réponse) ;
- la tour Eiffel a été érigée (événement) ;
- l'action a eu lieu en 1889 (date de l'événement) ;
- la réponse est le lieu de l'événement.

Nous pouvons remarquer que deux grands types de vérifications sont à effectuer :

- certaines portent sur l'événement mentionné par la question parmi lesquelles on trouve : la

vérification de l'événement en lui même, les vérifications de lieu et de date et les autres compléments circonstanciels possibles ;
– d'autres sur la réponse comme la vérification du type et le lien avec les mots de la question.

Deux types de vérifications seront effectuées : certaines informations seront recherchées dans le passage justificatif et d'autres en dehors.

Les premières vérifications reposent sur l'analyse du passage justificatif et permettent d'évaluer s'il contient bien les informations permettant de valider la réponse. En effet, pour que ce passage justifie la réponse il faut qu'il porte sur l'événement mentionné. Dans ce but, deux types de vérification pourraient être mis en place : le premier portera sur la présence, dans le passage, des termes de la question et le second sur les liens entre ces termes.

Pour effectuer la première vérification, nous tiendrons compte des variations linguistiques que l'on pourra trouver dans les passages. De plus, nous considèrerons non seulement les mots isolés, mais aussi des termes complexes.

Une vérification particulière porte sur les informations de date qui devront être mentionnées dans le passage. Si ce n'était pas le cas alors le passage aurait de bonnes chances de porter sur un événement analogue mais se déroulant à un autre moment et le triplet réponse serait probablement non valide. Cette vérification peut constituer un filtre car peu de variations sont possibles pour une même date. Ce même traitement pourrait également être employé pour les vérifications du lieu de l'événement de la question. Malheureusement, comme un lieu peut se trouver sous différentes formes comme l'un de ses méronymes, ce même type de contrainte sur sa présence semble peu pertinente et aucune vérification particulière n'a été effectuée. Pour la rendre pertinente il faudrait pouvoir établir un lien entre le lieu de la question et un lieu qui se trouverait dans le passage.

La vérification sur les relations entre termes porte sur les relations entre termes de la question et la relation de la question avec l'un des termes du passage. Généralement, cette vérification est effectuée en s'appuyant sur les relations syntaxiques. Comme les analyseurs ne sont pas très fiables pour ces notions, et notamment lorsque les documents ne respectent pas la grammaire de la langue, nous utiliserons des critères de surface. Afin d'évaluer un degré de liaison global des différents mots dans le passage, nous nous fonderons sur la proximité des mots dans la forme de surface du passage en supposant que plus les mots sont proches plus ils sont liés. Ce type de proximité pourra s'appliquer sur les termes de la question, mais aussi par rapport à la réponse potentielle.

Une autre stratégie consiste à s'appuyer sur des critères syntaxiques locaux, qui peuvent s'appliquer même si la totalité de la phrase est agrammaticale. Ainsi, lors de l'extraction des réponses, une méthode classique consiste à utiliser des patrons d'extraction. Ces patrons permettent de reconnaître un lien entre la réponse et le terme le plus pertinent de la question, le focus. Il semble donc pertinent de tenir compte de ces constructions particulières, lors de la validation de réponses, pour vérifier que la réponse est liée aux termes de la question. De la même manière, en recherchant les termes complexes, nous vérifions que certains mots sont reliés de manière analogue dans la question et le passage.

La vérification externe que nous avons traitée porte sur le type spécifique. En effet de nombreuses questions attendent une réponse correspondant à un type particulier que le passage ne mentionne pas forcément bien qu'il soit considéré comme justificatif de la réponse. Si la réponse n'est pas une instance du bon type alors le triplet réponse est non valide. Ces vérifications consisteront à exploiter

de grands corpus et les capacités des systèmes de reconnaissances d'entités nommées à typer des termes.

La figure 3.1 présente le traitement effectué afin de reconnaître que la réponse « Paris » est valide pour la question « Dans quelle grande capitale fut érigée la tour Eiffel en 1889 ? » et le passage « Erigée en 1889, la tour Eiffel est le monument phare de Paris ».

Afin d'évaluer la méthode proposée, deux types d'évaluation ont été effectués : la validation de réponses avec des données issues de la campagne d'évaluation AVE et la vérification du type de la réponse en elle-même.

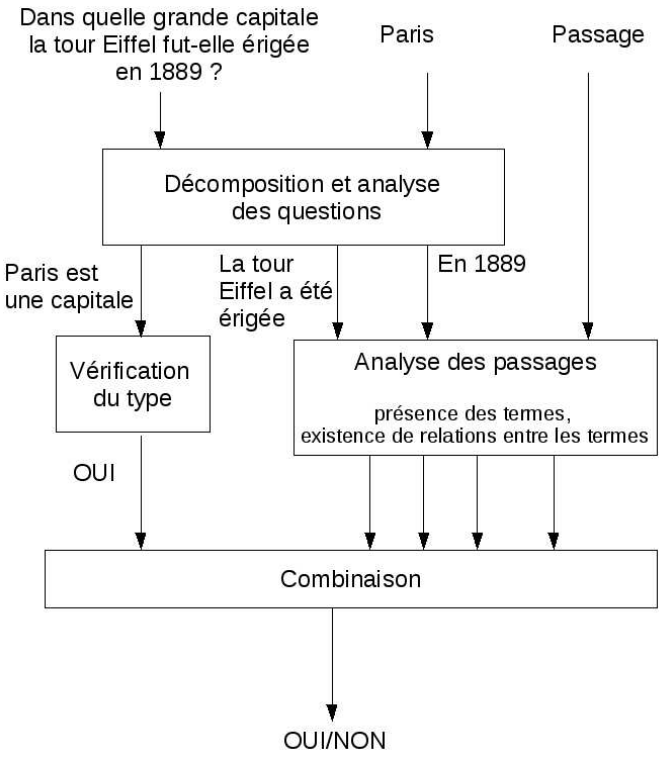

FIG. 3.1 – Traitements appliqués pour détecter la validité d'une réponse

3.2 Décomposition de questions

Afin de trouver les différentes vérifications à effectuer, les questions sont décomposées en éléments atomiques qu'il est possible de vérifier séparément. Cela ne signifie pas que les éléments soient indépendants mais que le processus de vérification de l'un peut être différent du processus de l'autre. La méthode de décomposition doit pouvoir couvrir toutes les vérifications possibles. Pour cela, un ensemble de règles syntaxiques s'appuyant entre autre sur la présence de mots clés a été créé afin de simplifier chaque question en une question minimale et un ensemble d'autres vérifications.

La première évaluation de cette décomposition permet de mettre en évidence l'importance des différents phénomènes ainsi que leurs occurrences dans l'ensemble ayant servi à créer les règles de décomposition. Une autre évaluation porte sur les règles et permet d'évaluer leur robustesse en les appliquant à un nouveau corpus, un nouvel ensemble de questions.

3.2.1 État de l'art

Notre travail consiste donc à identifier les différentes justifications à effectuer afin de valider la réponse. Le système de questions réponses FIDJI [Moriceau et al. 2009] effectue un travail relativement similaire puisqu'il identifie les relations présentes dans la question grâce à un analyseur syntaxique. Ces relations peuvent être au niveau de la réponse ou entre des mots de la question. La méthode cherche ensuite la réponse dans plusieurs documents en partant du principe que toutes les relations n'ont pas besoin de se trouver dans un même passage. Une vérification du type précis de la réponse est aussi effectuée en recherchant dans les documents des structures syntaxiques indiquant cette relation. Bien que cette méthode tienne compte d'une décomposition des questions afin de calculer leur validité, elle ne porte pas sur le type des informations contenues dans chacune des vérifications. De plus cette approche repose sur une écriture manuelle de règles pour normaliser des relations syntaxiques.

D'autres systèmes s'intéressent, plus particulièrement, aux questions complexes. Les questions complexes sont des questions contenant plusieurs vérifications principales comme « *Qui était le président américain quand la guerre du Vietnam s'est achevée ?* » pour laquelle il faut trouver la date de la fin de la guerre pour que la réponse corresponde bien au président des Etats Unis de cette époque. Pour traiter ce type de questions, les systèmes les décomposent souvent. Ainsi Saquete, et al. [2004] décomposent l'exemple ci-dessus en deux : « *Qui était le président américain ?* » et « *Quand la guerre du Vietnam s'est-elle achevée ?* » puis cherche les différentes réponses à ces questions et réduit l'ensemble de réponses concernant la première question en tenant compte de la seconde réponse. Hartrumpf [2008] commence, quant à lui, à chercher une réponse à la seconde question puis simplifie la question originelle en « *Qui était le président américain en 1967 ?* » avant d'en chercher la réponse.

Bentivogli et al. [2010a] effectuent une décomposition en phénomènes linguistiques (coréférence, paraphrase ...) présents lors d'une implication textuelle afin de rassembler les exemples en fonction du phénomène. Une analyse manuelle a permis de reconnaître l'ensemble des phénomènes sur des exemples. De la même manière, Cabrio & Magnini [2011] définissent l'implication textuelle fondée sur les composants comme la tâche visant à décomposer l'implication en types de problèmes à résoudre puis à reconnaître séparément chacun d'eux.

3.2.2 Réduction de questions

La décomposition d'une question a pour but de la rendre la plus courte possible c'est-à-dire d'arriver à une forme d'hypothèse minimale. L'hypothèse minimale est une reformulation de l'ensemble question + réponse réduite souvent à la forme sujet+verbe+objet ne contenant qu'un seul type d'élément à vérifier. Celle-ci se limite souvent à l'action contenue dans la question d'origine avec ses arguments. Cette action peut faire intervenir la réponse (« *Qui fut assassiné en 1968 ?* » se réduit en « *REPONSE fut assassiné.* »). Dans ce cas il faut montrer que la réponse a subi l'action. Mais il arrive

aussi que la réponse ne soit pas concernée par cette action. Par exemple, la question « *Quand le pont de Normandie a-t-il été inauguré ?* » se réduit en « *Le pont de Normandie a été inauguré.* » et la réponse sert alors à dater cet événement. Dans les cas où la question ne contient pas d'action (« *Où se trouve Paris ?* »), l'élément minimal est juste un groupe nominal (« Paris »). De manière générale cette notion d'hypothèse minimale se rapproche de celle de focus définie dans [El Ayari 2009].

Cette décomposition est à mettre en relation avec la répartition des informations au sein d'un document. Le plus souvent l'information minimale devra se trouver dans la même phrase. Les autres éléments, comme la date ou le lieu de l'action, peuvent être plus distants. Le type peut, quant à lui, être absent du document.

Le mécanisme consiste à simplifier la question en retirant itérativement des vérifications à effectuer jusqu'à obtenir l'hypothèse minimale. La méthode s'appuie sur une analyse syntaxique utilisant les entités nommées de la question, son analyse en syntagmes ainsi qu'un ensemble de mots clés tels que le pronom interrogatif de la question. La détection de ces phénomènes est effectuée à partir de règles établies à l'aide d'une analyse de corpus portant sur 839 questions provenant des campagnes d'évaluations CLEF 2006, CLEF 2007, CLEF 2008 et EQueR.

Voyons maintenant les différentes informations à valider. Elles peuvent se regrouper en trois catégories : les compléments circonstanciels portant sur l'événement de la question notamment de temps et de lieu et les informations caractérisant la réponse en déterminant son type et la relation qu'elle doit entretenir avec la forme minimale de la question. La dernière catégorie traite des questions complexes comportant plusieurs actions dominantes (« *Quel est le nom du physicien qui a été expulsé d'Allemagne et qui a fait de la recherche, jusqu'à sa mort en 1955, à l'institut des Études Avancées de Princeton ?* ») .

3.2.2.1 Compléments circonstanciels

- **les compléments de date** permettant de situer temporellement l'action de la question. Cette information peut correspondre à la réponse (« *Quand Martin Luther King a-t-il été assassiné ?* ») ou être contenue dans la question (« *Où ont eu lieu les Jeux olympiques d'hiver en 1994 ?* »). Elle est souvent distante de la réponse dans les documents. Deux méthodes de détection s'appliquent en fonction du cas :
  - quand la question attend une date en réponse, elle est tout d'abord reformulée afin de tenir compte du type spécifique à l'aide de règles de transformation et d'une liste de types correspondant à une date. « *En quelle année le général De Gaulle est-il décédé ?* » se transforme ainsi en « *Quand le général De Gaulle est-il décédé ?* ». Puis le pronom interrogatif est retiré et devient une information externe. Notons que certaines questions ainsi réduites ne correspondent alors plus qu'à un seul groupe nominal. Ainsi « *Quand a eu lieu la chute du mur de Berlin ?* » se réduit en « la chute du mur de Berlin ». Une liste de verbes a été créée afin de détecter ceux qui marquent une relation et ne désignent pas un événement porté par un verbe et peuvent donc être supprimés.
  - le second cas se présente lorsqu'une information de date est précisée par la question et permet de situer temporellement son action (« *Quel pays a remporté la coupe du monde de football en 1998 ?* »). La détection de telles relations consiste à sélectionner les entités nommées temporelles de la question précédées d'un mot clé tel que « en » ou « depuis ». Une

fois l'expression détectée elle est retirée de la question. D'autres mots clés peuvent servir à indiquer une information de date ne correspondant pas à une entité nommée mais à un événement. Ainsi dans la question « *Combien de personnes ont perdu la vie lors de l'attentat du World Trade Center ?* » le mot clé « lors de » permet de marquer l'information temporelle de « l'attentat du Word Trade Center ».

– **les compléments de lieu** : Ce type d'information consiste à situer spatialement un événement. Ce phénomène ainsi que la manière de le définir sont très proches des vérifications temporelles. En effet, cette information peut être la réponse (« *Où Jean-Pierre Lafon est-il ambassadeur de France ?* ») ou se trouver dans la question (« *Qui a épousé Bill Gates à Hawaï ?* »). Quand le lieu correspond à la réponse, la détection est faite en utilisant le pronom interrogatif de la question et en marquant son type spécifique lors d'une réécriture de la question. Par exemple la question « *Dans quel musée se trouvent les Noces de Cana ?* » est transformée en « *Où se trouvent les noces de Cana ?* ». Pour ce faire le type spécifique a été comparé à un ensemble de types collectés correspondant à un lieu. Comme pour les questions de date, il se peut que la question reformulée ne contienne pas d'action à proprement parler comme dans la question précédente.
La détection des vérifications spatiales contenues dans la question est effectuée en reconnaissant les entités nommées de lieu précédées par certains mots clés. Par exemple à la question « *Qui a épousé Bill Gates à Hawaï ?* », « Hawaï » est extrait de la question puisque c'est une entité nommée de type lieu précédé du mot clé « à »

– **les compléments de but** : cette vérification sert à donner la raison d'une action. Ce qui correspond aux questions ayant comme pronom interrogatif « Pourquoi » ;

– **les compléments de manière** : cette catégorie d'information explique la manière dont une action a été effectuée. Elle peut donc se trouver dans la réponse (« *Comment Dominique Voynet est-elle allée à L' Elysée ?* ») dans les questions ayant « Comment » comme pronom interrogatif. La question précédente se transforme alors en « *Dominique Voynet est allée à l'Elysée.* ». L'information peut aussi être précisée par la question (« *Combien de journalistes Diego Maradona blessa-t-il avec une carabine ?* »). Dans ce cas la question est décomposée en utilisant des mots clés comme « avec » ;

– **combien de fois** : Parmi les autres vérifications de type compléments circonstanciels, celle observée le plus souvent lors de l'analyse des différentes questions porte sur le nombre de fois qu'une action a été effectuée. Ce qui se trouve dans des expressions telles que « Combien de fois ». Par exemple, la question « *Combien de fois un missile atteint-il le mur du son ?* » peut se réduire en vérifiant que le missile atteint le mur du son et que cela a lieu le nombre de fois indiqué dans la réponse. Elle ne s'applique cependant pas pour toutes les questions commençant par le mot combien comme dans « *Combien de personnes ont perdu la vie lors de l'attentat du 11 septembre ?* » car ces questions ne traitent pas du même phénomène. Dans ce type de questions, la réponse correspond souvent, combiné à l'élément devant être quantifié, au sujet ou au complément d'objet du verbe de la question.

3.2.2.2 Informations sur la réponse

- **le type spécifique** que la question attend en retour et dont le terme est présent dans la question. Par exemple la question « *Quel président a succédé à Jacques Chirac ?* » a « président » comme type spécifique. Pour certaines questions, un type plus précis peut également être reconnu. Il correspond au groupe nominal le plus grand présent dans la question et explicitant le type. Ainsi à la question « *Dans quel film de Kevin Reynolds Kevin Costner a-t-il joué ?* », le type spécifique est « film » alors que le type plus précis est « film de Kevin Reynolds ». L'extraction de ce type s'appuie sur un ensemble de règles reposant sur l'analyse syntaxique de la question en ajoutant au type spécifique les groupes prépositionnels le précédant ou le suivant ;
- **les autres vérifications** : D'autres éléments peuvent venir spécifier la réponse et portent le plus souvent sur des caractéristiques d'une personne telles que son âge « *Qui fut assassiné le 4 novembre 1995 à l'âge de 73 ans ?* » ou sa nationalité. Elles peuvent se trouver dans une question mais aussi correspondre à la réponse « *A quel âge Roland Ratzenberger est-il mort ?* ». La détection de ce phénomène s'appuie sur la syntaxe de la question pour reconnaître le type spécifique de réponses attendu et une comparaison de ce type avec une liste indiquant les types propres à ces vérifications. Des mots clés sont aussi utilisés quand la vérification se trouve dans la question.

3.2.2.3 Questions complexes

Dans certains cas, la complexité de la question impose de la décomposer en plusieurs questions principales. C'est par exemple le cas de « *Quel est le nom du physicien qui a été expulsé de l'Allemagne et qui a fait de la recherche, jusqu'à sa mort en 1955, à l'Institut des Etudes Avancées de Princeton ?* » qui se décompose en « *Quel est le nom du physicien qui a été expulsé de l'Allemagne ?* » et « *Quel est le nom du physicien qui a fait de la recherche, jusqu'à sa mort en 1955, à l'institut des Études Avancées de Princeton ?* ». Afin de valider les réponses à ces questions il faudra vérifier les deux faits. Un système de questions réponses pourrait par exemple chercher les réponses à ces deux questions séparément puis calculer l'intersection des deux ensembles qui constituera la réponse à la question globale. La décomposition de ces questions s'appuie sur l'analyse de surface de la question. Par exemple, la question précédente, est de la forme « (.*)qui(.*) et qui (.*) » avec les deux parties contenant un verbe. De manière générale, il faut qu'il y ait un moyen de séparer une question en différentes propositions.

Dans d'autres cas, la complexité intervient au niveau des relations temporelles comme « *Combien de kilos Rafaël Dinelli a-t-il perdu après être tombé dans l'océan Indien ?* » pour lesquelles il faut chercher une réponse pour les deux évènements (Rafaël Dinelli a perdu N kilos, il est tombé dans l'océan indien) puis relier les deux actions de manière similaire à ce qui est fait pour les compléments circonstanciels de date.

3.2.2.4 Récapitulatif

La figure 3.2 présente la décomposition des questions avec la question minimale et l'ensemble des autres vérifications ainsi que les deux grands types de vérifications : les compléments circonstanciels et les informations supplémentaires.

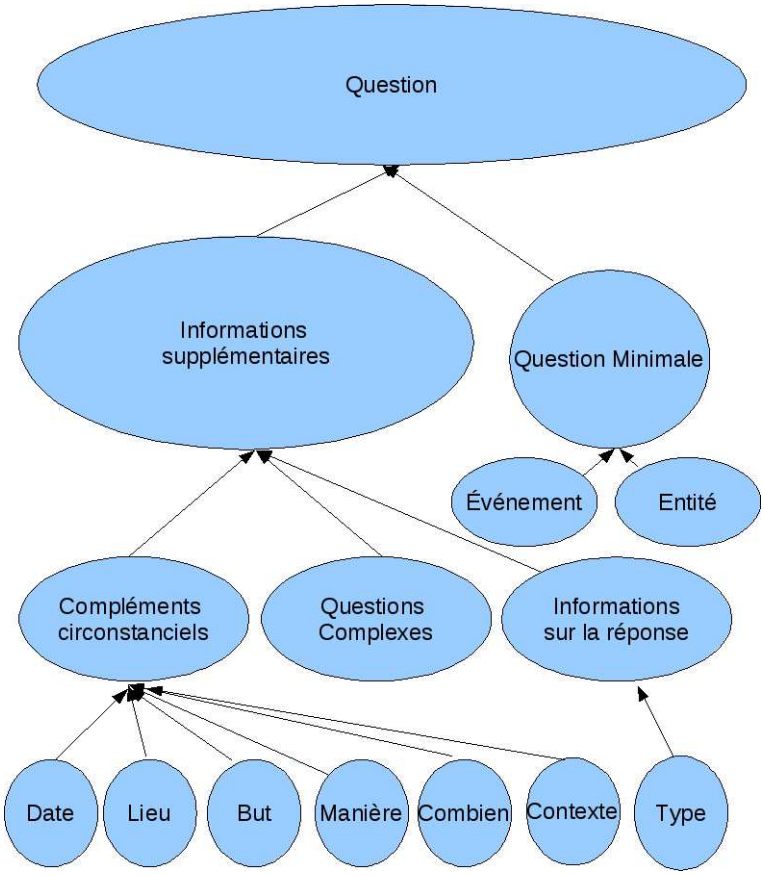

FIG. 3.2 – Décomposition d'une question

3.2.3 Évaluation

Après avoir établi les règles permettant de décomposer une question, deux évaluations restent à faire. La première permet de mesurer l'apparition des phénomènes afin de connaître leur importance et la seconde porte sur les règles afin d'évaluer leur robustesse.

839 questions ont servi à créer la décomposition et parmi celles-ci 296 (35 %) ne peuvent pas être simplifiées. Cela correspond entre autres aux questions de définitions (à 50 %) ou à des questions très courtes (« *Combien sont les Beatles ?* »).

Pour les 543 questions restantes 768 vérifications ont été détectées ce qui correspond à une moyenne de 1,4 phénomènes par question. Deux évaluations ont été menées : la proportion de questions contenant chaque phénomène ainsi que l'importance des différents critères qui correspond au rapport entre le nombre d'apparitions du phénomène et le nombre d'apparitions de tous les phénomènes (cf. tableau 3.1).

La vérification du type de la réponse, de la date et du lieu de la question sont trois éléments particulièrement importants pour lesquels il est nécessaire d'avoir des méthodes permettant de les

Phénomènes	Type spécifique	Date	Lieu	Contexte
Nb questions %	49 %	44 %	29 %	7,5 %
Importance	35 %	32 %	21 %	5 %
Phénomènes	Manière	Complexe	Combien	But
Nb questions %	4 %	4 %	0,55 %	0,18 %
Importance	3 %	3 %	0,39 %	0,13 %

TAB. 3.1 – Importance des différentes informations

traiter. Les autres phénomènes restent quant à eux plus anecdotiques.

Comme la création de règles de décomposition s'appuie sur une étude de corpus, il est nécessaire d'évaluer la robustesse des règles en analysant d'autres questions. Dans ce but, nous avons utilisé un nouvel ensemble de questions provenant de la campagne d'évaluation Quæro 2009 et contenant 509 questions. Les phénomènes contenus dans ces questions sont recherchées à l'aide de ces règles ce qui permet d'effectuer une très bonne détection puisque 93 % des phénomènes ont bien été identifiés, les problèmes restants étant notamment dus à la reconnaissance des entités nommées. Par exemple, dans la question « *Est-il possible de jouer à Dofus sans payer ?* » Dofus est vu comme un lieu et non comme un jeu.

3.2.4 Conclusion

Nous avons présenté une méthode permettant d'identifier les phénomènes dont il faut tenir compte afin de justifier une réponse. Le mécanisme consiste à décomposer les questions afin d'obtenir la question la plus réduite possible. Cette détection a permis tout d'abord de voir que de nombreuses questions (35 %) n'ont qu'une seule vérification à effectuer. Pour les autres, il est nécessaire de tenir compte du type spécifique de la réponse (présent dans 49 %), de la date (dans 44 %) et du lieu (dans 29 %).

3.3 Analyse des passages

La méthode présentée dans cette section effectue une analyse des passages permettant de vérifier que la réponse est bien justifiée par le passage. Pour cela différentes informations sont à vérifier :

- le passage traite du même sujet que la question. Pour cela, la présence dans le passage des mots de la question est recherchée aussi bien sous la forme de termes simples que de termes complexes. Ces termes peuvent aussi être sous forme de variantes ;
- il le fait de manière identique. La proximité des mots de la question dans le passage est calculée en suivant l'hypothèse que si les mots sont proches alors ils sont liés. Cette vérification ressemble ainsi à une méthode de vérification de paraphrases locales ;
- la date mentionnée dans la question se trouve également dans le passage. L'étude présentée en 3.2 a montré que la date est une information particulière de la question. Le système s'assure alors que la date contenue dans la question se trouve telle quelle dans le passage. Dans le cas contraire la réponse est vue comme non justifiée.

Notre méthode suit une approche par apprentissage comme de nombreuses autres approches présentées dans l'état de l'art. Ce type d'approche permet de combiner différents critères calculés par des analyses de différentes natures. Trois grands types de critères sont classiquement utilisés et sont similaires à ceux retenus dans ce modèle : les termes communs au passage et à la question, la proximité des termes et les vérifications propres à la validation de réponses comme la vérification du type de la réponse ou la redondance de la réponse.

Le système est ensuite évalué en utilisant les données de la campagne de validation de réponses AVE 2006 [Peñas et al. 2006]. Dans cette campagne les participants devaient reconnaître si un triplet réponse est valide en renvoyant OUI le cas échéant et NON sinon. Le corpus d'évaluation est plus particulièrement présenté en 3.3.4. Afin d'étudier l'intérêt des différents critères, ils sont évalués grâce à une partie de ces données servant à tester globalement la méthode, ce qui correspond à 578 triplets réponses parmi lesquels 416 (71 %) ne sont pas valides. Les critères sont évalués indépendamment et de manière combinée en tenant compte de la précision, du rappel et de la f-mesure des réponses valides [Grappy, et al. 2008 ; Ligozat, et al. 2007b ; Ligozat, et al. 2007a]. Nous avons aussi participé à la campagne d'évaluation AVE 2008 en combinant la méthode avec l'approche suivie dans FIDJI [Moriceau, et al. 2008b].

3.3.1 Présence des termes dans le passage

3.3.1.1 Présence globale

Un premier critère porte sur le taux de mots de la question présents dans le passage. Elle permet d'évaluer pour quelle part les informations contenues dans la question se retrouvent dans le passage justificatif. Pour ce faire, une mesure globale de similarité entre la question et le passage est fondée sur le taux de recouvrement lexical. De manière générale, quand on parle de termes ou de mots de la question présents dans le passage, cela sous-entend le mot sous sa forme initiale mais aussi sous forme de variantes. En effet, il a été constaté à travers le système de questions réponses QALC [de Chalendar, et al. 2002] qu'il fallait mieux tenir compte des variantes.

Ce critère est un des plus communément utilisé pour la validation de réponses. Breck [2009] s'était intéressé au problème dans le cadre de l'implication textuelle en vérifiant que tous les mots non vides de l'hypothèse se trouvent dans le texte. Ainsi, si un mot non vide de l'hypothèse ne se trouve pas dans le passage, alors il n'y a pas implication. Cette vérification permet d'obtenir 61 % de bons résultats.

Dans notre travail, les mots et les termes complexes de la question sont recherchés et reconnus dans le passage par FASTR [Jacquemin 1996] ce qui permet de reconnaître les différentes variations morphologiques d'un terme ainsi que certains de ses synonymes. Bien évidemment, un certain nombre de mots ne sont pas porteurs de sens et donc leur présence dans le passage justificatif n'indique rien. Ce sont les déterminants, les prépositions et les adverbes. Ces termes ne sont pas considérés dans le calcul. Le critère a comme valeur la proportion de termes simples communs.

Afin d'évaluer le critère, la réponse est vue comme valide si la proportion de termes de la question présents dans le passage est supérieure à une valeur seuil, autour de 0,5, et non valide sinon. Afin d'avoir une idée des résultats, ils sont comparés à une baseline qui consiste à reconnaître tous les triplets réponses comme valides (cf. tableau 3.2).

Méthode	Précision	Rappel	F-mesure
Baseline	0,29	1	0,44
Proportion	0,50	0,71	0,59

TAB. 3.2 – Résultats de la validation des réponses, en considérant les termes communs au passage et à la question

Les résultats montrent que l'utilisation de ce système permet effectivement d'améliorer les résultats, f-mesure 0,59 contre 0,44 pour la baseline. Cette augmentation est due à une meilleure détection des réponses correctes. Toutefois nous pouvons aussi remarquer que seul un triplet réponse sur deux vu comme valide l'est réellement (précision 0,50). Ainsi il ne suffit pas que le passage contienne beaucoup de mots de la question pour que la réponse soit effectivement valide. Cette méthode se retrouve dans la plupart des systèmes et est dite « sac de mots ». Nous la reprendrons ainsi comme baseline.

3.3.1.2 Importance des termes selon leur catégorie

Le critère précédent vérifiait que les mots pleins de la question se trouvaient bien dans le passage sans en privilégier certains. Afin d'affiner le processus, une nouvelle hypothèse peut être formulée : l'importance de la présence des mots dépend de leur catégorie morphosyntaxique. Par exemple, un nom propre absent du passage semble plus pénalisant que l'absence d'un adjectif.

Un ensemble de critères est utilisé à cet effet : la proportion de noms communs, de noms propres, de verbes et d'adjectifs communs au passage et à la question. Pour calculer la proportion de verbes, les auxiliaires ne sont pas pris en compte. Nous pouvons remarquer que le système MLENT [Kozareva et al. 2006] utilise de tels critères.

De manière pratique, une valeur de pertinence est donnée à chaque catégorie conservée. Cette valeur correspond à la proportion de mots de la catégorie de la question présents dans le passage.

Une autre vérification effectuée porte sur les bitermes. Un biterme est un ensemble de deux mots l'un à la suite de l'autre reconnus comme étant liés, comme « prix Nobel ». Si un biterme de la question se trouve dans le passage cela signifie souvent que les mots du biterme présents dans le passage sont utilisés dans le même sens que dans la question. Il est souvent plus pertinent de trouver un biterme que de trouver chacun des termes séparés. Le critère correspondant est égal à la proportion de bitermes de la question se trouvant dans le passage justificatif. Là encore toutes les questions ne contiennent pas un biterme (cf. tableau 3.3).

3.3.1.3 Termes importants de la question

D'autres critères plus sémantiques peuvent être considérés. Ces critères prennent en compte le rôle d'un mot dans la question. Dans les systèmes QALC [de Chalendar et al. 2002] et FRASQUES [Ferret et al. 2002] un certain nombre de termes sont ainsi reconnus lors de l'analyse des questions et marqués comme des termes importants. Ils permettent notamment d'identifier les passages pertinents et d'extraire la réponse. Ces termes sont :

Critère	Taux de présence	Précision	Rappel	F-mesure
Noms communs	46 %	0,34	0,86	0,49
Noms propres	53 %	0,33	0,88	0,48
Nombres	20 %	0,31	0,85	0,45
Verbes	26 %	0,29	0,75	0,41
Adjectifs	25 %	0,52	0,23	0,32
Bitermes	75 %	0,42	0,42	0,42
Focus	67 %	0,5	0,7	0,59
Type attendu	55 %	0,29	0,66	0,4
Verbe principal	62 %	0,32	0,54	0,41

TAB. 3.3 – Évaluation des mots importants de la question

- **le focus :** par définition, cet élément devrait être repris dans le passage pour exprimer la réponse. Ce critère semble donc particulièrement pertinent. Au moment de cette étude, il correspondait à un groupe nominal, sujet ou complément du verbe principal si la question mentionne un événement ;
- **le verbe principal :** c'est le verbe présent dans la question et ayant un rôle important dans la formulation de la réponse quand il introduit un fait ou une action. Il permet de formuler l'événement ou l'action mentionné dans la question. Il n'est donc ni un auxiliaire, ni un verbe modal et ne sert pas non plus à poser la question de manière indirecte. La présence de ce terme dans le passage montre que l'action de la question se trouve également dans le passage justificatif. Malheureusement, le verbe est soumis à de nombreuses variations difficiles à reconnaître comme un synonyme ou une nomalisation ;
- **le type spécifique :** n'est pas toujours présent dans le passage, mais quand il l'est il vient souvent indiquer que la réponse correspond au type. Ainsi dans la question « *Quel sport pratique Zinédine Zidane ?* » , le focus est « Zinédine Zidane », le type spécifique « sport » et le verbe principal « pratiquer ».

La section 3.2 a montré qu'une question pouvait se réduire en une hypothèse minimale à laquelle des informations supplémentaires sont ajoutées. La vérification de cette hypothèse se rencontre en pratique par le test de la présence du focus (l'élément dominant de la question) et du verbe principal dans le passage.

L'évaluation des termes importants de la question (cf. tableau 3.3) montre que le focus est un critère pertinent puisqu'il obtient une f-mesure égale à celle traitant de la proportion de termes de la question présents dans le document (0,59). Ce résultat pouvait être attendu car c'est le terme dominant de la question qui doit obligatoirement se trouver dans le passage. Le verbe principal semble malheureusement peu pertinent car il est souvent absent du passage ou présent sous forme de variation non reconnue. Le type de réponse attendu obtient un bon rappel, mais a une précision très faible car il est absent d'un grand nombre de questions, ce qui ne le rend pas très discriminant.

L'évaluation des critères portant sur la catégorie morphosyntaxique des termes montre que les mots autres que les adjectifs obtiennent un fort rappel et une faible précision ce qui indique que les réponses correctes sont souvent reconnues mais aussi que de nombreuses réponses sont considérées

Méthode	Précision	Rappel	F-mesure
Baseline	0,50	0,71	0,59
Combinaison critères morphosyntaxique	0,42	0,68	0,52
Combinaison termes importants	0,43	0,84	0,57
Tous	0,50	0,80	0,62

TAB. 3.4 – Combinaison des critères étudiant la présence des mots de la question dans le passage justificatif

comme correctes à tort. Par ailleurs, il n'y a pas une catégorie qui ressort par rapport aux autres bien que les proportions de noms communs et de noms propres soient supérieures aux autres.

L'étape suivante consiste à effectuer différentes combinaisons des critères afin de voir l'apport à la simple proportion des termes de la question dans le passage ; ce qui est fait par apprentissage grâce à la combinaison d'arbres de décision par la méthode bagging de Weka[1] . La justification du choix de ce classifieur est donnée en 3.3.4. Trois combinaisons sont effectuées : les différentes catégories morphosyntaxiques, les termes prédominants de la question et l'ensemble de tous les critères portant sur la présence des termes de la question dans le passage (cf. tableau 3.4).

Ce tableau montre que les différentes distinctions prises séparément ne permettent pas d'égaler ni de dépasser les résultats de la baseline. Cela s'explique car la vérification de la présence des termes prédominants ne permet pas de vérifier toutes les informations voulues. De plus, les termes sont souvent peu présents. En revanche, ces critères combinés avec la vérification globale des termes de la question permettent d'améliorer la baseline et plus précisément d'augmenter le nombre de réponses valides reconnues comme telles (le rappel).

3.3.2 Vérification de la date

L'événement contenu dans un certain nombre de questions est souvent complété par sa date. Par exemple, à la question « *Quel pays l'Irak a-t-il envahi en 1990 ?* » la date « 1990 » vient compléter l'action qui sans elle pourrait faire mention d'un tout autre événement. Ce phénomène a été présenté en 3.2 comme l'une des informations complémentaires les plus fréquentes puisqu'elle correspond à 32 % des phénomènes externes. Pour vérifier cette information, nous avons considéré que la date devait être précisée dans le passage justificatif sans quoi la réponse ne sera pas considérée comme justifiée.

Bien souvent, une même date se retrouve sous la même forme dans le passage et dans la question. La première vérification consiste à rechercher la date de la question dans le passage sous la même forme. Aucune variation de cette date n'est autorisée à une exception près. Quand la date de la question correspond à la date de création du document il arrive très fréquemment qu'elle ne soit pas contenue dans le passage et soit implicitement celle du document. Ainsi nous pouvons trouver « le 17 mai » et rarement « 2011 » pour qualifier la date de la sortie de « Tree of Life » au cinéma. Pour ces cas, nous considérons une date comme justifiée si elle correspond à la date de création du document dont la réponse a été extraite. L'application de cette vérification consiste à reconnaître une réponse

[1]WEKA : http ://www.cs.waikato.ac.nz/ml/weka

comme non valide s'il n'y a pas de correspondance de date. En pratique, sur 3 000 triplets réponses provenant de la campagne d'évaluation AVE 2006, 177 triplets sont vus comme ne correspondant pas au niveau de la date et parmi eux 172 (97 %) triplets réponses sont effectivement non valides.

Une vérification plus complexe a été effectuée par Barbier [2009] dans le cadre du projet CONIQUE. Elle permet de vérifier, quand la date se trouve dans le document, que celle-ci vient effectivement qualifier l'action contenue dans la question.

Le système recherche l'élément que la date qualifie dans la question. Celui-ci peut être soit le verbe de la question soit l'un de ses mots. Pour cette détection, il s'appuie sur le verbe ; si c'est un verbe support (par exemple « avoir lieu ») c'est forcément le focus qui est porteur. Le mot servant à introduire la date est également utilisé. Par exemple, si le mot est « de », l'élément porteur est le focus, tandis que si c'est « depuis », le terme est le verbe. Une fois la date et l'élément concerné identifiés, il est vérifié que la date quantifie bien cet élément dans le passage, en pratique que les deux termes sont dans la même phrase suffisamment proches. Cette vérification ne permettant pas d'améliorer significativement les résultats, elle n'a pas été reprise.

3.3.3 Plus Longue Chaîne Commune (LCC)

Le critère suivant traite de la proximité des termes de la question et de la réponse entre eux, ce qui permet de considérer des paraphrases locales. L'idée étant que les termes proches sont vraisemblablement en relation et donc le passage a davantage de chances de traiter des informations demandées par la question que si les termes étaient éloignés.

La chaîne cherchée, l'hypothèse, est une reformulation de la question sous forme déclarative à laquelle la réponse a été ajoutée. Pour valider pleinement la réponse, le passage devrait contenir l'hypothèse telle quelle ou sous une forme de paraphrase avec des variations sous phrastiques pouvant porter sur les termes de la question. Afin d'estimer si cette hypothèse se trouve dans le passage, la plus grande chaîne de mots consécutifs de l'hypothèse est recherchée sans que les mots soient forcément dans le même ordre. L'exemple suivant sera utilisé dans la suite de notre explication :

Question : Qui est le père de la reine Elisabeth 2 ?
Passage : **Georges VI**, le **père** d' **Elisabeth 2**, l'actuelle **reine** d'Angleterre ...
Réponse : Georges VI
Hypothèse : *Georges VI est le père de la reine Elisabeth 2*.

Tous les systèmes de questions réponses tiennent compte de la proximité des termes dans le passage. Par exemple Gillard, et al. [2006] calculent une mesure de compacité plus complexe tenant compte des mots de la question et de la réponse dans le cas où la réponse attendue est une entité nommée pour sélectionner la réponse.

L'utilisation de la proximité des termes dans un système de validation de réponses est donc aussi un critère assez classique pour la validation de réponse. Ainsi, certains systèmes calculent la proximité de la réponse par rapport aux mots de la question et d'autres [Kozareva et al. 2006 ; Pakray et al. 2009] calculent la plus longue chaîne commune au passage et à l'hypothèse. Deux différences principales existent entre les différents systèmes. Pour certains, il est nécessaire que les mots se trouvent dans le même ordre dans le passage et dans l'hypothèse afin d'être davantage certain que le sens est le même dans les deux cas. L'autre considération vient de l'utilisation de mots « bonus ». Ces mots

permettent de reconnaître deux mots comme liés même s'ils sont séparés par d'autres mots. Pakray et al. [2009] autorisent de tels mots mais, en revanche, s'assurent que les termes se trouvent dans le même ordre. Kozareva et al. [2006] n'autorisent aucun mot bonus mais ne se préoccupent pas de l'ordre d'apparition des termes.

La méthode proposée ici ne tient pas compte de l'ordre des différents mots car dans de nombreux cas l'information ne se trouve pas dans le même ordre dans le passage et la question. Dans le passage de l'exemple précédent « reine » se trouve après « Elisabeth 2 » contrairement à la question. Deux mots sont également considérés comme liés s'ils sont séparés par des mots bonus. Les mots vides ainsi qu'un certain nombre d'items sont autorisés. Pour chaque paire de mots de la question, le système autorise aussi qu'un seul mot bonus éventuel sépare ces deux mots. Cela permet l'ajout de modifieurs de nom comme, par exemple, l'adjectif « actuelle » dans l'exemple précédent.

Au final, l'algorithme (cf. algorithme 1) recherche la plus longue chaîne de mots consécutifs mais non ordonnés présents dans le passage et l'hypothèse.

Afin de faciliter le rapprochement entre l'hypothèse et le passage, les mots sont normalisés : les variantes du passage sont ramenées au terme de l'hypothèse. L'algorithme identifie ensuite dans le passage tous les mots présents dans l'hypothèse ainsi que leur emplacement ce qui permet d'ordonner ces mots selon leur apparition dans le passage. L'emplacement des différents items autorisés, les mots vides de sens et les signes de ponctuation, est aussi collecté.

L'algorithme consiste alors à parcourir l'ensemble des emplacements en créant à chaque fois une nouvelle chaîne de mots. Cette chaîne est agrandie itérativement si un nouveau mot est soit directement adjacent soit séparé de la chaîne en cours par un ensemble d'items autorisés et par un éventuel unique mot bonus. Ces traitements permettent d'obtenir différentes chaînes et seule la plus longue est conservée. L'algorithme s'effectue en un temps linéaire en fonction de la taille du passage. En effet, la première étape consistant à identifier les termes du passage est linéaire en fonction de la taille du passage. L'étape consistant à calculer les chaînes communes est polynomiale en fonction des mots trouvés et a donc une complexité inférieure à la première étape.

La chaîne obtenue pour l'exemple est donc : « *Georges VI, le père de Elisabeth 2 reine* ».

La formule suivante définit la chaîne calculée de manière plus formelle :
chaîne $= m_1...m_n | \forall m_i, m_i \in$ *hypothèse et* $m_i n^* l n^* m_{i+1} \in passage$ *avec n* un ensemble de mots vides de sens et l un seul éventuel mot bonus.*

Pour avoir une idée de la validité de la réponse, un poids est associé à la chaîne obtenue. Ce poids correspond au rapport entre le nombre de mots significatifs de la chaîne calculée et le nombre de mots significatifs de l'hypothèse. Le poids associé à l'exemple est de $\frac{8}{10}$, car la chaîne calculée contient 8 mots, tandis que l'hypothèse en contenait 10.

Afin de détecter un seuil permettant de dissocier les réponses valides de celles non valides un module par apprentissage a été appliqué et a permis de découvrir une frontière de 0,54. Il faut donc qu'au moins la moitié des mots de l'hypothèse se trouvent dans le passage assez proches les uns des autres pour que la réponse soit validée.

Le tableau 3.5 montre que de bons résultats sont obtenus puisque la f-mesure (0,64) est supérieure à celle obtenue par la combinaison des critères traitant de la présence des termes de la question dans

Algorithm 1 LCC (String[] hypothese, String[] passage) : double

(String mot,int place) [] L1 ; # Liste des mots de l'hypothèse présents dans le passage et leurs emplacements
int [] L2 ; # les emplacements dans le passage des items autorisés
for $i = 1 \rightarrow taille(passage)$ **do**
  $motencours \leftarrow passage[i]$;# remplissage de L1 et L2
  if $motencours \in hypothese$ **then**
    ajouter (motencours,i) à L1 ;
  end if
  if motencours est autorisé **then**
    ajouter i à L2 ;
  end if
end for
int taillemax=0 ; int i=1 ;
while $i < taille(L1)$ **do**
  $String\ chaine \leftarrow L1[i].mots$; # chaque chaîne est initialisée
  $int\ place \leftarrow L1[i].place$; # sa place
  $tailleencours \leftarrow 1$;
  $boolean\ fini \leftarrow false$;
  $boolean\ motbonusutilise \leftarrow false$;
  $int\ j \leftarrow i + 1$; # indice du mot suivant
  while $j < taille(L1)$ and not (fini) **do**
    $String\ chainesuivante \leftarrow L1[j].mots$; # pour tous les mots suivants
    $int\ placesuivante \leftarrow L1[j].place$; # trouve sa place et sa valeur
    if $chainesuivante \in chaine$ **then**
      $fini \leftarrow true$; # si le mot est déjà dans la chaine on ne le rajoute pas
    else
      $int\ k \leftarrow place + 1$;
      # on vérifie qu'on peut rajouter le nouveau mot à la chaine existante
      while $k < placesuivante$ and not (fini) **do**
        if not ($passage[k] \in L2$) **then**
          if $motbonusutilise = false$ **then**
            $motbonusutilise \leftarrow true$; # si le mot n'est pas dans L2 c est le mot bonus
          else
            $fini \leftarrow true$; # si le mot bonus est déjà utilisé on n'ajoute pas le mot
          end if
        end if
        $k \leftarrow k + 1$;
      end while
    end if
    if not(fini) **then**
      $j \leftarrow j + 1$; # aucune erreur n'a été vue ; le mot peut être ajouté
      $chaine \leftarrow chaine\ chainesuivante$;
      $motbonusutilise \leftarrow false$;
      tailleencours++ ;
      i++ ; # pour que le même mot ne soit pas comptabilisé deux fois
    end if
  end while
  if $tailleencours > taillemax$ **then**
    $taillemax \leftarrow tailleencours$;
  end if
  i++ ;
end while
return taillemax ;

Précision	Rappel	F-mesure
0,53	0,80	0,64

TAB. 3.5 – Plus longue chaîne commune

Méthode	Précision	Rappel	F-mesure
Toujours OUI	0,22	1	0,36
50 % de OUI	0,23	0,5	0,31

TAB. 3.6 – Tests de base

le passage (0,62). Cela est notamment dû à un rappel particulièrement élevé (0,80) qui indique qu'un grand nombre de réponses valides sont reconnues comme telles.

Afin de mieux évaluer ces résultats, nous avons créé une méthode de base qui calcule la plus longue chaîne commune au passage et à l'hypothèse en considérant les termes dans le même ordre et sans aucun mot bonus. Elle obtient une f-mesure de 0,55 ce qui témoigne de l'intérêt de notre définition car ses résultats sont bien meilleurs. Cela montre également que souvent les phrases sont formulées en employant les mots dans un ordre différent.

3.3.4 Validation de réponses produites par des systèmes de questions réponses : le système AVAL (Answer VALidation)

Afin de se rendre compte de l'intérêt des différents critères, leur combinaison a été évaluée dans un cadre de validation de réponses. Les données proviennent de la campagne d'évaluation des systèmes de validation de réponses AVE 2006 [Peñas et al. 2006]. Lors de cette campagne les systèmes recevaient en entrée un ensemble de triplets réponses constitués d'une question, d'un passage justificatif et d'une réponse et devaient détecter si le triplet est valide en renvoyant OUI le cas échéant et NON sinon. Les données provenant de cette campagne ont été produites par différents systèmes de questions réponses qui devaient fournir plusieurs réponses accompagnées d'un passage justificatif pour les différentes questions. Le fait de s'évaluer sur ces données permet de s'assurer de la fiabilité du système de validation de réponses indépendamment du système ayant servi à obtenir les données.

2 987 triplets réponses pour 190 questions différentes servent à évaluer notre système. Parmi ceux-ci 2358 (79 %) ne sont pas valides et seulement 629 sont valides.

Afin de mieux appréhender les données, des baselines ont été créées. La première reconnaît tous les triplets réponses comme valides et la seconde détecte aléatoirement un triplet sur deux comme valide (cf. tableau 3.6). Ces deux mesures correspondent à celles notamment présentés par les organisateurs de la campagne AVE 2007 [Peñas et al. 2007].

Comme le système a été évalué dans ce cadre, des vérifications supplémentaires liées aux données fournies, ont été apportées au système, comme une étape de prétraitement.

Méthode	Précision	Rappel	F-mesure
Toujours OUI sur base complète	0,21	1	0,35
Toujours OUI sur base réduite	0,30	1	0,46
OUI sauf sur exemples reconnus	0,30	0,95	0,45

TAB. 3.7 – Tests de base

3.3.4.1 Prétraitement du corpus

L'ensemble de test contient beaucoup plus de réponses non valides que de réponses valides. Or, un certain de nombre de ces réponses sont non valides pour une raison simple et peuvent donc être reconnues facilement. Ces cas sont les suivants :

– le passage ne contient pas la réponse. La réponse ne peut clairement pas être justifiée par le passage puisqu'il n'en parle pas. Nous pouvons toutefois nous demander comment il se fait qu'un système de questions réponses ait obtenu cette réponse ? Vraisemblablement, ce cas correspond à une erreur du système ;

– la réponse est contenue entièrement dans la question. Dans ce cas, aucune information supplémentaire n'est donnée par la réponse et cela correspond généralement à une réponse incorrecte. C'est le cas par exemple de la réponse « Tim Burton » à la question « *Citer un film de Tim Burton.* ». Dans certaines situations très rares, des réponses peuvent néanmoins être correctes comme « Banque de France » à la question « *Quel est la plus grande banque de France ?* ». ;

– le passage est entièrement contenu dans la réponse. Cela est dû à une erreur du système de questions réponses qui n'a pas pu extraire la réponse ou correctement former le passage ;

– un certain nombre de questions attendent une réponse d'un type d'entité nommée particulier. Ainsi une question comme « *Où est situé le Taj Mahal ?* » attend un lieu en retour que ce soit une ville ou un pays. Le module de vérification utilise l'analyse des questions du système FRASQUES afin de sélectionner le type d'entité nommée attendu en réponse. Le passage justificatif est également annoté en entité nommée afin de reconnaître toutes les entités qu'il contient et ainsi de détecter le type d'entité nommée correspondant à la réponse. Une comparaison entre les deux types est ensuite possible. Les réponses dont le type d'entité nommée ne correspond pas à celui attendu par la question sont alors rejetées. Cela permet par exemple de rejeter des dates pour les questions attendant un lieu en réponse.

Cet ensemble de vérifications permet de rejeter 818 triplets réponses soit près de 33 % des triplets contenus dans la base totale. Parmi ceux-ci 793 (97 %) sont effectivement non valides et donc seuls 25 sont valides. Ce filtre permet de doubler la proportion de triplets réponses valides qui passe de 21% à 43 %. Parmi les différentes vérifications, la plus poussée porte sur la vérification du type d'entité nommée de la réponse. Cette vérification peut poser problème si l'entité nommée de la question ou de la réponse est mal détectée. Toutefois le filtre effectue 95 % de bonnes décisions.

Afin d'observer l'importance de ce filtre, la baseline qui consiste à reconnaître toutes les réponses comme valides a été calculée sur la base complète sans aucun filtre, la base réduite en ne considérant pas les réponses filtrées et la base complète sur laquelle les réponses filtrées ont été marquées comme non valides (cf. tableau 3.7).

Nous voyons donc que l'étape de filtre améliore les résultats puisque la f-mesure passe de 0,35 à 0,45. Par la suite nous effectuerons nos calculs sur la base réduite car les résultats sont similaires à ceux obtenus en rejetant les réponses filtrées. De plus, cela permet de mieux comprendre les résultats obtenus.

La validation des réponses est détectée à l'aide d'un certain nombre de critères combinés par apprentissage. Dans ce but, il est nécessaire de définir la base d'apprentissage et la base de test. La base d'apprentissage correspond à $\frac{3}{4}$ des données et la base de test à $\frac{1}{4}$. Elles sont toutes les deux extraites de la base réduite. Afin que la base d'apprentissage ne contienne pas trop de triplets réponses non valides et qu'elle contienne $\frac{2}{3}$ de réponses non valides, certains de ces triplets ne sont pas présents dans cette base. La base de test contient quant à elle 578 triplets réponses parmi lesquelles 416 (71 %) ne sont pas valides. Les résultats présentés dans les sections précédentes ont été calculés sur cette base de test.

3.3.4.2 Critère spécifique : extraction de la réponse

Les différentes vérifications présentées jusqu'à présent analysent les passages justificatifs afin de détecter s'ils sont susceptibles de contenir la réponse. Toutefois rien n'indique que la réponse extraite soit celle attendue.

Ce critère utilise le système de questions réponses FRASQUES afin d'effectuer une recherche d'une réponse dans le passage justificatif. L'idée étant que si FRASQUES reconnaît la réponse proposée, cela signifie que deux systèmes de questions réponses sont capables d'extraire cette réponse, ce qui permet d'augmenter les chances que la réponse soit correcte. Bien sûr cette vérification ne peut s'appliquer que pour vérifier des réponses provenant d'autres systèmes de questions réponses et ne pourrait par exemple pas prendre place dans l'intégration de la méthode de validation de réponses au système FRASQUES.

Dans FRASQUES, l'extraction de la réponse dépend du type de réponse attendu. Si la question attend en réponse une entité nommée, le système extraiera l'entité du bon type la plus proche des mots de la question. Sinon, la réponse est recherchée grâce à des patrons d'extraction. Ces patrons sont articulés autour du focus, du verbe principal ou du type général et correspondent à des règles syntaxiques locales. Dans le premier cas, si la réponse à valider est également extraite par FRASQUES alors les termes de la question sont assez proches de la réponse ce qui peut indiquer une bonne correspondance. Les patrons d'extraction étant des règles indiquant l'existence de la relation attendue entre la réponse et le focus (ou le verbe principal), leur application permet de voir la réponse comme liée à ces termes dominants.

D'un point de vue pratique, le système FRASQUES est appliqué sur les passages de la base de test puis la réponse obtenue est comparée à la réponse à valider en comptabilisant la proportion de termes de la réponse à juger présents dans la réponse extraite. Cette combinaison permet de rassembler des réponses différentes mais ayant une partie en commun. Afin d'effectuer des tests, un seuil de 0,5 est utilisé. Si la valeur est supérieure à ce seuil alors la réponse est considérée comme étant la même.

Les résultats ne sont pas très bons (cf. tableau 3.8), avec une f-mesure de 0,49 alors que celui de la baseline est de 0,45 ce qui est dû à une faible précision (0,39). En revanche de nombreuses réponses valides sont reconnues, le rappel est élevé (0,66). La faible précision peut s'expliquer par le fait que

Précision	Rappel	F-mesure
0,39	0,66	0,49

TAB. 3.8 – Comparaison des réponses

les mécanismes d'extraction de réponse sont proches de celui développé par le système FRASQUES ainsi la réponse extraite par FRASQUES est la même que celle proposée dans 48 % des cas.

3.3.4.3 Résultats

Nous avons donc établi un ensemble de critères participant à la validation de la réponse. Les critères traités sont donc les suivants :

– la proportion de termes simples et complexes de la question présents dans le passage tels quels ou sous forme de variation ;
– la répartition des termes par catégorie morphosyntaxique ;
– la présence des mots clés de la question dans le passage ;
– le calcul de la plus longue chaîne commune au passage et à l'hypothèse ;
– la comparaison de la réponse avec celle extraite par le système FRASQUES.

L'étape suivante consiste donc à les combiner. Le classifieur choisi pour cette approche correspond à une combinaison d'arbres de décision grâce à la méthode bagging. Ce classifieur a été choisi de manière empirique car c'est celui obtenant les meilleurs résultats. De plus, de par son approche qui consiste à sélectionner récursivement le critère répartissant le mieux les données, il est possible d'analyser l'intérêt des différents critères. La figure 3.3 montre les différents traitements effectués pour détecter la validité d'un triplet réponse. Elle montre ainsi l'ordre des différents modules avec d'abord les modules de FRASQUES puis l'étape de filtre et la recherche de traits pour finir avec la combinaison des critères permettant de détecter la validité des réponses.

Afin d'évaluer les résultats obtenus, il est possible de les comparer avec ceux des autres systèmes ayant participé à AVE. Précisons que notre travail est postérieur à l'évaluation et s'est effectué sur une partie de son corpus. Le système présenté est comparé avec celui obtenant les meilleurs résultats sur le français (MLENT) [Kozareva et al. 2006] et celui obtenant les meilleurs résultats toutes langues confondues (COGEX) [Tatu et al. 2006] (cf. tableau 3.9). Notons qu'une comparaison complète n'est pas possible car les systèmes ne sont pas évalués sur les mêmes données. La comparaison est donc effectuée à un niveau global.

Le système MLENT est une combinaison d'une approche par apprentissage et d'une approche logique. L'approche par apprentissage a comme critères la proportion de termes communs, la plus grande chaîne de mots consécutifs présents dans le passage et l'hypothèse et la proportion de mots différents du passage à l'hypothèse en suivant le même ordre. Le système COGEX suit une approche logique.

D'assez bons résultats sont obtenus, supérieurs à ceux des autres systèmes. Nous pouvons remarquer qu'il y a un rappel très élevé (0,82) et une précision assez basse (0,57) ce qui montre que la plupart des triplets valides sont reconnus mais aussi que de nombreux triplets sont vus comme valides à tort. Afin d'approfondir les résultats, nous pouvons étudier la matrice de confusion présente dans le

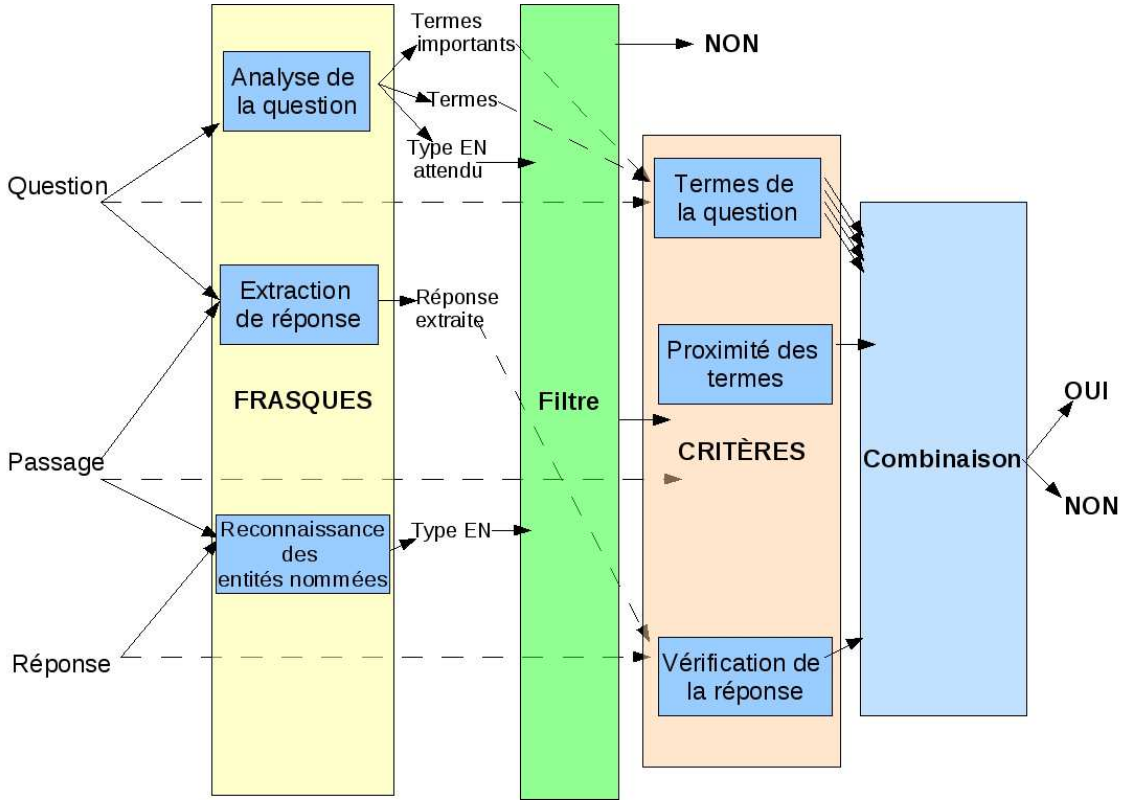

FIG. 3.3 – Mécanisme de validation de réponses : le système AVAL

tableau 3.10 dans laquelle on peut voir que 77 % des données sont correctement classées. De plus, très peu d'erreurs sont effectuées quand une réponse est déclarée non valide, car elle l'est effectivement dans 92 % des cas.

3.3.4.4 Participation à la campagne AVE 2008

Afin d'évaluer complètement notre système, nous avons participé à la campagne AVE 2008 [Rodrigo et al. 2009]. Cela permet de s'assurer que l'approche s'adapte effectivement à d'autres données et que nos hypothèses étaient bien fondées.

Système	Précision	Rappel	F-mesure
MLENT	0.34	0.73	0,57
COGEX	0.53	0.78	0,63
AVAL	**0,57**	**0,82**	**0,67**

TAB. 3.9 – Etude du passage

Valeur détectée/attendue	Valide	Non valide
Valide	57 %	43 %
Non valide	8 %	92 %

TAB. 3.10 – Matrice de confusion sur les données d'AVE 2006

Méthode	Précision	Rappel	F-mesure
AVAL sur AVE 2006	*0,57*	*0,82*	*0,67*
AVAL	0,67	0,60	0,63
FIDJI	0,88	0,42	0,55
AVAL + FIDJI	0,75	0,52	0,61

TAB. 3.11 – Utilisation du système de questions réponses FIDJI pour la campagne AVE 2008

Afin de participer à la campagne d'évaluation une combinaison de notre système de validation de réponses et du système de questions réponses FIDJI [Moriceau et al. 2009] a aussi été menée [Moriceau et al. 2008b]. FIDJI est un système de questions réponses s'appuyant sur une analyse syntaxique des passages. Ce système vérifie ainsi que les relations syntaxiques se trouvant dans la question sont également dans le passage justificatif et permet d'ajouter des critères syntaxiques à notre système. FIDJI fournit différents critères qui sont ajoutés à ceux déjà présentés afin d'effectuer une combinaison globale par apprentissage. Trois critères sont ainsi ajoutés :

– la proportion de dépendances syntaxiques de la question absentes du passage justificatif ;
– le fait que la réponse corresponde au type d'entité nommée attendu par la question ;
– une combinaison des deux critères précédents et de l'utilisation du module d'extraction de la réponse afin de vérifier que la réponse proposée peut être extraite par FIDJI. La combinaison détecte si la réponse est valide en renvoyant une valeur booléenne. Pour qu'une réponse soit valide il faut qu'au moins 30 % des liens de la question se trouvent dans le passage.

Le corpus AVE 2008 français comporte 199 triplets réponses parmi lesquels seuls 52 sont valides. Pour cette approche nous avons également intégré une vérification simple du type spécifique. Cette vérification est l'un des critères présenté lors de la vérification du type de la réponse, en 3.4.4 qui consiste à chercher le type dans la page Wikipédia associée à la réponse. Le tableau 3.11 présente les résultats sur cette campagne en utilisant ou non les critères venant de FIDJI.

Méthode	Précision	Rappel	F-mesure
LINA	0,56	0,46	0,51
ITQA	0,54	0,78	0,64
AVAL + FIDJI	0,75	0,52	0,61

TAB. 3.12 – Résultats pour AVE 2008

Le tableau montre que les résultats sur la précision et le rappel sont très différents en fonction des données utilisées (AVE 2006 ou AVE 2008). Toutefois, l'ensemble 2008 est peu représentatif car il contient trop peu de triplets valides ce qui entraîne qu'une erreur "coûte" 2 %.

Nous pouvons également voir que le système AVAL obtient des résultats meilleurs sans tenir

compte des critères venant du système FIDJI (f-mesure 0,63 vs 0,61) ce qui nous amène à penser qu'il faudrait utiliser autrement les critères syntaxiques. Toutefois ces critères permettent d'augmenter la précision car FIDJI obtient une précision très élevée (0,88) et un rappel plus bas (0,42).

Les résultats sont comparés à ceux obtenus par les autres systèmes participant à cette tâche. Trois systèmes ont participé à cette campagne sur le français ce qui correspond à 5 évaluations différentes (cf. tableau 3.12). Dans notre cas les évaluations correspondent à AVAL + FIDJI et à FIDJI seul. AVAL + FIDJI est le système ayant obtenu les meilleurs résultats, le second système, LINA, présenté par Jacquin, et al. [2008], obtient une f-mesure de 0,51. Toutes langues confondues, le système le plus performant, ITQA [Wang & Neumann 2009], obtient une f-mesure de 0,64 sur l'anglais. Ce système suit aussi une approche par apprentissage avec des connaissances syntaxiques et lexicales. Ces informations montrent que notre système est effectivement performant mais doit être amélioré notamment au niveau de la précision. Pour cela, il faudrait ajouter des critères permettant de détecter les réponses non valides. Ces vérifications devraient se faire en dehors du passage car les critères analysant le passage sont assez complets si ce n'est la prise en compte des relations syntaxiques. C'est pourquoi nous avons élaboré une solution permettant d'affiner la vérification du type de la réponse candidate.

3.4 Vérification du type

Cette section porte sur le type de la réponse. Comme nous l'avons vu, un grand nombre de questions attendent une réponse d'un type particulier. Si une réponse ne correspond pas au type spécifique attendu par la question alors le triplet réponse est forcément non valide quel que soit le passage. Cette vérification peut donc s'appliquer séparément de l'analyse du passage. Elle porte sur la réponse en elle-même et ne peut donc pas être utilisée dans un cadre d'implication textuelle. Elle peut cependant aussi être utilisée dans un cadre plus large de recherche d'entité nommée.

Il existe une infinité de questions possibles et donc un très grand nombre de types possibles qu'il n'est pas possible de connaître à l'avance. Un type peut, en effet, être composé de plusieurs mots avec, par exemple, un adjectif venant compléter un nom comme « grande banque ». Une taxonomie contenant tous les cas devrait couvrir tous les groupes nominaux que l'on peut trouver dans des questions ce qui n'est clairement pas possible. C'est pour cela que l'utilisation de base de connaissances ne peut correspondre qu'à un type partiel de vérification qu'il faudra compléter.

Le problème peut donc se formaliser ainsi : pour un type spécifique attendu par la question et une réponse proposée, est-ce que la réponse est compatible avec le type ? Ce problème peut être considéré comme une tâche de classification en deux ensembles : soit la réponse correspond au type soit elle n'y correspond pas. Il est alors possible de résoudre ce type de problèmes en suivant une approche par apprentissage supervisé [Grappy & Grau 2010a ; Grappy & Grau 2010b ; Grappy & Grau 2011].

3.4.1 Types de réponses

Avant de présenter la solution adoptée, il est nécessaire de définir ce que sont les types de réponses. Deux types de réponses peuvent être identifiés dans une question : le type spécifique et le type d'entité nommée.

Le type spécifique peut être vu comme la dénomination du concept dont la réponse est une instance (type « *acteur* » à la question « *Quel acteur joue dans Danse avec les loups ?* ») ou un hyponyme (type « *oiseau* » à la question « *Quel est l'oiseau le plus rapide ?* »). Le type est présent dans la question et dépend de ce que l'utilisateur précise. Il ne fait donc référence à aucune classification préexistante car un très grand nombre de types sont possibles. On dira que la réponse est compatible avec le type (ou est du type) attendu si elle est une instance ou un hyponyme du concept de ce type comme « Kevin Costner » pour le type « acteur » ou « autruche » pour le type « oiseau ». Nous pouvons noter que le type spécifique est constitué d'un à plusieurs mots (acteur ou premier ministre).

Le type d'entité nommée est le nom de la classe d'entité nommée à laquelle la réponse appartient. La question « *Quel acteur joue dans Danse avec les loups ?* » peut attendre ainsi une PERSONNE en retour. Le type d'entité nommée est généralement moins précis que le type spécifique comme dans l'exemple précédent avec le type spécifique acteur. Ce type d'entité ne suit pas la formulation exacte de la question et les types d'entités nommées sont généralement en nombre fixe, structurés suivant les cinq types classiques : personne, lieu, organisation, date et entités numériques (une description plus précise est donnée section 3.4.3.1).

Plusieurs rapprochements entre ces deux types peuvent être effectués selon les différentes questions :

– la question n'attend ni type spécifique ni type d'entité nommée, c'est par exemple le cas des questions booléennes ou des questions de définition (« *Qu'est ce que le Nasdaq ?* ») ;

– la question précise un type d'entité nommée en retour sans qu'il y ait un type spécifique associé, comme (« *Où est Paris ?* ») ;

– la question permet d'identifier un type spécifique qui n'a pas de type d'entité nommée associé comme « *Quel film remporta la palme d'Or en 1994 ?* » pour laquelle « film » est le type spécifique ;

– la question précise un type spécifique associé à un type d'entité nommée. Dans ce cas le type spécifique est très souvent plus précis que le type d'entité nommée. Par exemple la question « *Quel acteur a joué dans Danse avec les Loups ?* » attend une réponse de type d'entité nommée PERSONNE et de type spécifique « acteur ». Dans un certain nombre de cas, il arrive que les deux types soient égaux. Ainsi, la question « *Dans quel lieu des massacres de musulmans ont-ils été commis ?* » a « lieu » comme type spécifique et type d'entité nommée.

D'un point de vue pratique, l'extraction de ces deux types est effectuée par le module d'analyse des questions du système de questions réponses FRASQUES [Grau, et al. 2005]. Il utilise pour cela des règles portant sur des critères syntaxico-sémantiques, le type de pronom interrogatif ainsi que la forme syntaxique des questions. Le type spécifique correspond ainsi souvent au premier groupe nominal présent après le pronom interrogatif. Il est composé du nom principal du groupe accompagné éventuellement d'un adjectif.

3.4.2 État de l'art

Notre but consiste donc à détecter si la réponse correspond bien au type spécifique attendu par la question. Les systèmes de questions réponses effectuent très souvent des vérifications portant sur le type de la réponse mais se placent plus particulièrement au niveau des entités nommées. En effet, plus un système sait reconnaître de types d'entités différentes meilleures sont ses performances. Ainsi, le

système conçu par Sekine, et al. [2002] tient compte de 200 types d'entités nommées. Celui créé par Hovy, et al. [2001] en comporte 122. L'approche présentée par Harabagiu, et al. [2000] en utilise encore plus puisqu'il se fonde sur WordNet. Bien que ces approches soient efficaces, elles reposent sur une liste de types définis a priori et ne peuvent donc pas couvrir tous les cas possibles. Même WordNet ne peut contenir tous les types car ceux-ci diffèrent selon l'ajout de modifieurs tels que des adjectifs. Il est donc intéressant de procéder à des vérifications portant sur les types spécifiques.

De nombreuses méthodes recherchent des relations d'hyperonymie ou plutôt des couples (hyperonymes, hyponymes). Elles ont pour point d'origine les travaux de Hearst [1992] qui cherchent des patrons d'extraction en corpus. Dans un premier temps, le système part d'un ensemble de patrons, puis cherche, dans des documents, les phrases dans lesquelles ils apparaissent, ce qui permet d'obtenir un premier ensemble de couples hyperonymes-hyponymes. Puis les relations liant ces termes sont recherchées afin d'obtenir de nouveaux patrons de relations. Puis de nouveaux exemples sont recherchés à partir de ces règles et ce mécanisme se poursuit ainsi itérativement. Un travail similaire a été effectué par Morin [1998] sur le français avec une étape visant à rechercher des patrons depuis des exemples, suivi d'une validation des patrons par un expert et la recherche de nouveaux exemples grâce aux patrons obtenus.

Bien qu'il soit impossible de connaître tous les couples (hyperonyme, hyponyme) possibles, cette approche peut toutefois inspirer notre recherche. En effet au lieu de collecter tous les couples, on peut chercher à reconnaître une telle structure de phrase dans un document. Ainsi, la découverte dans un passage d'une phrase comme « *Robert De Niro est un acteur* » indique une relation de type entre « Robert de Niro » et « acteur ».

Les travaux menés par Magnini et al. [2002b] ont montré que la cooccurrence de la réponse et des termes de la question dans des documents est un critère pertinent permettant d'ordonner les différentes réponses. Tout d'abord les documents contenant la réponse, ceux contenant les mots de la question et ceux contenant à la fois ces termes et la réponse sont recherchés. À partir de ces informations, des valeurs permettant de détecter une confiance pour les différentes réponses sont déduites grâce à des critères statistiques. Bien que ce travail concerne l'ensemble des mots de la question il pourrait être intéressant de l'appliquer afin de marquer un lien entre le type et la réponse.

Schlobach, et al. [2004] présentent une vérification du type dans les cas où la réponse est un lieu géographique et pour cela partent d'un ensemble de 500 questions qui correspondent à 40 types de réponses différents. La vérification est effectuée grâce à la combinaison d'une approche à base de connaissances externes et d'une approche fondée sur la redondance. La première méthode porte sur des connaissances ontologiques vérifiant que le type de la réponse correspond bien au type attendu par la question ainsi que sur des recherches dans le réseau lexical WordNet. La seconde, plus statistique, part du principe que la cooccurrence de mots dans un grand corpus peut indiquer que les mots sont reliés et montre ainsi que la cooccurrence du type et de la réponse est un critère pertinent. Ce système est évalué en notant l'amélioration qu'apporte cette vérification à un système de questions réponses existant et montre une amélioration du MRR de 0,33 à 0,38.

Schlobach, et al. [2007] poursuivent ce travail en l'appliquant en domaine ouvert. Les approches sont assez semblables et combinent des informations fournies par WordNet et des informations statistiques. La vérification grâce à WordNet s'effectue en cherchant un lien possible entre la réponse et le type. La méthode statistique traite entre autre de la fréquence d'apparition de la réponse, du type, du

couple (type,réponse), de la probabilité conditionnelle d'apparition commune du type et de la réponse en fonction de l'apparition du type ou de la réponse. Une autre mesure tient compte de l'apparition de la règle « Réponse est un Type » avec un ou deux mots pouvant séparer la réponse et le type dans un ensemble de documents. L'évaluation permet d'obtenir le même type d'amélioration en domaine fermé qu'en domaine ouvert.

Ces travaux traitent effectivement de notre problématique. Toutefois notre travail est effectué sur le français, et il n'existe pas de base de données lexicale aussi complète et structurée que WordNet. Des tests ont été effectués sur EuroWordNet, une version de WordNet pour de nombreuses langues dont le français. Ce réseau lexical n'étant pas aussi complet que WordNet, de nombreux termes y sont absents et cette ressource n'est pas réellement utilisable pour la validation du type de la réponse.

Un autre axe de recherche peut s'apparenter à la vérification du type : la découverte de relations entre entités, évaluée par la tâche TREC Entity [Balog, et al. 2010].

Dans cette tâche, les systèmes participants reçoivent en entrée une entité ainsi qu'un type de relation donnée en langue naturelle et doivent renvoyer les entités reliées à l'entité donnée selon la relation souhaitée. Le type d'entité nommée des réponses à trouver est également donné.

Par exemple l'entité peut être « Kingston trio », la relation, ici une question, « *Quelle maison de production sort les disques de Kingston trio ?* » et la réponse sera une « organisation ». Les systèmes recherchent l'entité liée dans un ensemble de documents. En sortie, ils fournissent la réponse ainsi que la page Web la concernant.

Généralement, [Wang, et al. 2010 ; Bonnefoy, et al. 2011], les approches se décomposent en quatre temps :

– la recherche de documents pertinents en utilisant un moteur de recherche ;
– l'extraction des entités du type attendu en retour contenues dans ces documents effectuée grâce à un module de reconnaissance d'entités nommées ;
– l'ordonnancement des entités ;
– la détection des pages Web correspondant aux meilleures entités réponses.

L'étape la plus intéressante pour ce travail correspond à l'ordonnancement des entités. De nombreuses entités sont extraites depuis les documents et il faut alors retenir celles qui sont pertinentes. Pour cela, certaines approches utilisent des mesures probabilistes [Fang, et al. 2010] indiquant par exemple le lien entre l'entité et la requête ou celui entre l'entité et le type voulu en réponse. Tandis que d'autres cherchent des patrons indiquant le lien avec le type souhaité [Vechtomova 2010].

Le système présenté par Wang et al. [2010] effectue une détection des documents en créant une requête à partir de la relation et de l'entité cible puis fouille les différents documents afin de reconnaître les entités du type attendu en retour. L'ordonnancement est effectué en tenant compte de différents critères tels que la fréquence des entités réponses, le rang des documents duquel l'entité a été extraite ou le lien entre les deux entités, celle obtenue et celle donnée en entrée en tenant compte de la cooccurrence des deux termes dans différent documents.

Bonnefoy et al. [2011] considèrent cette tâche comme un problème de validation de réponses. L'approche cherche une correspondance entre les différentes entités cible et le type spécifique afin de retenir les entités correspondant le mieux au type. Pour ce faire, ils utilisent la distribution des mots et le fait qu'un terme a tendance à apparaître accompagné de certains mots particuliers. Les mots reliés à l'entité à classer, ceux reliés au type précis sont recherchés puis la distribution des mots de l'entité

et du type est calculée. La mesure obtenue permet alors de classer les entités. De plus, la proximité de la réponse par rapport aux mots de la question est considérée ainsi que le poids IDF de l'entité.

Pour notre part, nous nous sommes inspirés de la méthode de Schlobach et al. [2007], en l'adaptant à notre contexte et en rajoutant des critères, conservant ainsi une méthode reposant sur des critères de granularité et de robustesse différentes. Ainsi, certains s'appliquent sur beaucoup de réponses alors que d'autres plus pertinents ne s'appliquent que sur quelques unes. Ces différents critères sont donc complémentaires les uns des autres et peuvent être utilisés ensemble. Les critères peuvent être répartis en trois grandes catégories : l'utilisation des systèmes de reconnaissance des entités nommées, l'exploitation de l'encyclopédie Wikipédia et des mesures statistiques de cooccurrence. Ces critères nous servent à répondre à notre problématique qui consiste à déterminer si dans un triplet (réponse, type spécifique, passage) la réponse correspond au type.

3.4.3 Utilisation de systèmes de reconnaissance d'entités nommées

La première stratégie se place au niveau des entités nommées pour vérifier que la réponse correspond à l'entité nommée attendue par la question. Si le type d'entité nommée de la réponse ne correspond pas à celui attendu par la question alors il est plus que probable que le type spécifique ne soit pas vérifié non plus, puisque le type d'entité nommée en est une généralisation.

Les connaissances portant sur les entités nommées peuvent aussi servir à montrer qu'il y a correspondance au niveau du type spécifique. Les deux vérifications sont ainsi effectuées.

3.4.3.1 Filtrer les réponses

La première utilisation des entités nommées se fait de manière globale afin de vérifier que la réponse est compatible avec le type d'entité nommée attendu par la question. Si ce n'est pas le cas alors la réponse ne correspond probablement pas non plus au type spécifique. Par exemple, la question « *En quelle année eut lieu la révolution russe ?* » attend une date en réponse. L'utilisation de ce module permettra de rejeter la réponse « Alexandre Issaievitch Soljenitsyne » qui est de type PERSONNE donc n'est pas une date et encore moins une année mais admettra 1927. De ce fait, la reconnaissance du type d'entité nommée attendu par la question est en relation avec le système de reconnaissance d'entités nommées puisqu'ils partagent la même taxonomie.

Les passages justificatifs, et donc la réponse proposée qui en est extraite, sont analysés par un module qui annote les entités nommées. Le système permet de reconnaître une vingtaine d'entités nommées structurées suivant les cinq types classiques (personne, lieu, organisation, date, entités numériques) (cf. tableau 3.13). Par exemple, le type LIEU regroupe les types VILLE et PAYS. Les entités numériques permettent de typer les expressions numériques (longueur, vitesse, etc.). Le type NomPropre étiquette tous les termes contenant un nom propre non étiqueté plus finement cela afin de pallier l'absence de reconnaissance d'entités nommées. Ces entités nommées sont donc assez générales ce qui permet d'effectuer un filtrage de haut niveau. En effet, il y a moins de cas de mauvais typage d'entité nommée, comme par exemple prendre un lieu à la place d'une personne, ce qui serait plus fréquent avec un typage plus précis.

Le type de l'entité nommée de la réponse est recherché et comparé avec celui attendu par la question. Quatre cas sont alors possibles :

Noms Propres	Lieu	Ville
		Pays
	Personne	
	Organisation	
Date	Date relative	Date absolue
	Période/Durée	Âge
Expressions numériques	Longueur	Vitesse
	Volume	Température
	Poids	

TAB. 3.13 – Différents types d'entités nommées

– la question n'attend pas en réponse une entité nommée. C'est par exemple le cas de la question « *Quel oiseau est le plus rapide d'Afrique ?* ». Aucune information ne peut être fournie par ce critère et la valeur INCONNU sera donnée au couple réponse/type.

– la question attend une entité nommée et la réponse n'en est pas une. Par exemple la réponse « bateau » pour une question attendant une personne. Dans ce cas la réponse est vue comme mauvaise et la valeur NON est renvoyée.

– la question attend une entité nommée en réponse, la réponse est bien une entité nommée mais les types ne sont pas compatibles. Par exemple la réponse « 300 » de type nombre pour une question attendant un LIEU. Dans ce cas, la réponse sera considérée incompatible avec le type spécifique attendu et la valeur NON est renvoyée.

– la question attend une entité nommée en réponse et celle-ci est d'un type compatible avec le type EN attendu. Une entité nommée est vue comme compatible avec l'entité attendue en réponse si elles sont exactement du même type ou si elles sont d'un type similaire : les différents lieux sont vus comme compatibles entre eux tout comme les différentes expressions temporelles ou numériques. De plus les entités reconnues comme étant des lieux, des personnes ou des organisations sont compatibles avec une entité marquée uniquement comme un nom propre sans correspondre à une de ces catégories. Cette vérification ne permet d'avoir qu'une idée globale de la validité de la réponse et « Michel Rocard » est par exemple vu comme un « président ». Toutefois elle permet de rejeter des réponses ne correspondant pas au type de manière assez sûre.

3.4.3.2 Valider des réponses

Les entités contenues dans des textes peuvent être répertoriées dans des listes correspondant à chaque type que l'on sait reconnaître. Il semble donc intéressant de tester la présence ou l'absence de la réponse dans les listes correspondant aux types spécifiques cherchés. Si la réponse est trouvée cela permettrait de valider qu'elle correspond à ce type et si elle en est absente qu'elle n'y correspond pas. Bien sûr afin de pouvoir obtenir les résultats les plus pertinents possibles il est nécessaire de disposer d'un grand nombre de types d'entités nommées différents et de nombreuses valeurs pour ces types.

Le module d'entités nommées du système de questions réponses RITEL [Rosset 2008] permet de reconnaître un ensemble de types pouvant être assez précis comme « religion » ou « fleuve » et

actionnaire, adresse, affection, âge, album, ambassadeur, année, association, auteur, avocat, bénéfice, biscuiterie, bisquine, capitale, chef, chercheur, cheval, chorégraphe, compositeur, date, département, deux atouts, district, éditeur, endroit, état, équipe, événement, évêque, façon, film, financement, fonction, forum, frère, grade, grands parcs, groupe, heure, île, journal, lauréat, lieu, livre, maladie, maire, manière, médecin, meeting, métier, meurtrier, ministre, mode, monnaie, motif, moyen, musée, nationalité, nombre, numéro, occasion, organisation, origine, patron, parti, pays, PDG, peine, père, période, personnalité, personne, pièce, porte-parole, prénom, président, présidente, prix, procédé, profession, producteur, province, rapport, réalisateur, record, région, revue, roi, score, secte, somme, superficie, surnom, titre, traitement, tribunal, université, ville.

TAB. 3.14 – Ensemble des types de la base d'apprentissage

reconnaît 70 types spécifiques. Une liste des entités correspondant à chaque type connu a été créée en appliquant le système sur un grand corpus de textes. Dans chaque liste se trouvent l'ensemble des entités correspondant au type ce qui correspond à près de 240 000 expressions différentes. Certains types possèdent beaucoup plus de valeurs que d'autre. Par exemple, « livre » en contient 559 alors que « point cardinal » n'en contient que 31. L'annexe 1 énumère l'ensemble des types considérés. Comme le nombre de types d'entités nommées est limité, certain types spécifiques ne seront pas connus et pour ceux-ci ce module ne sera pas d'une grande utilité. En revanche, il est pertinent pour les autres.

Dans cette méthode, la réponse est cherchée dans la liste associée au type spécifique. Trois cas sont alors possibles :

– il n'y a pas de liste associée au type. Le module ne peut fournir aucune information sur le couple type/réponse et la valeur INCONNU est renvoyée.

– la réponse ne se trouve pas dans la liste du type cherché. La réponse semble ne pas correspondre au type et la valeur NON est retournée.

– la réponse se trouve dans la liste correspondant au type cherché. La réponse est vue comme compatible avec le type attendu et la valeur OUI est renvoyée. Cette vérification permet notamment de détecter que « Pulp Fiction » est un film.

3.4.3.3 Évaluation

Une méthode par apprentissage a été suivie avec une base d'apprentissage extraite depuis les réponses venant de la campagne de questions réponses EQueR. Cette campagne comporte 500 questions et parmi celles-ci 198 (40 %) sont utilisées car elles mentionnent un type spécifique. Comme certaines questions contiennent le même type spécifique, la base contient 98 types spécifiques différents (cf. tableau 3.14).

Ces types peuvent être très larges comme « lieu » et « parc » ou très précis comme « bisquine » (sorte de bateau de pêche à voile). Ils sont généralement formés d'un seul mot.

Les réponses pour lesquelles une correspondance avec le type est cherchée sont celles fournies par les systèmes de questions réponses ayant participé à la campagne EQueR. Ces systèmes pouvaient fournir jusqu'à cinq réponses. Ces données nous ont donc permis de construire la base d'apprentissage

Système	Décision	#	# Réponses du type spécifique	# Réponses n'étant pas du type spécifique
Filtre	OUI	1 411	**63 % (885)**	37 % (526)
	NON	457	29 % (132)	**71 % (325)**
	INCONNU	852	40 % (344)	60 % (508)
Validation de type	OUI	656	**77 % (506)**	23 % (150)
	NON	515	27% (138)	**73% (377)**
	INCONNU	1 549	46 % (716)	54 % (833)

TAB. 3.15 – Matrice de confusion des méthodes utilisant les entités nommées

qui contient 2 720 couples réponse/type spécifique dont la moitié (1 360) est valide, *c.-a-d.* que la réponse est du type attendu. Afin de décider de la validité des couples, une étude manuelle a été effectuée.

Ces données vont également servir à évaluer les différents critères. Pour cela, deux évaluations sont effectuées :

– la première, précise, correspond à la matrice de confusion et indique la proportion de réponses OUI et NON données correctement ;

– la seconde, globale, mesure la précision (proportion de bonnes réponses parmi celles données), le rappel (proportion de réponses correctes obtenues parmi celles attendues) et la f-mesure.

Évaluation de la méthode par filtre Le tableau 3.15 montre les résultats de l'évaluation de cette première méthode de manière précise. La première ligne indique, que lorsque la réponse est considérée comme étant du type spécifique attendu par la question (valeur « OUI »), c'est effectivement le cas dans 63% des cas. Il y a donc un taux d'erreurs de 37%, c'est-à-dire que des réponses sont déclarées comme compatibles avec le type spécifique à tort. La deuxième ligne montre que lorsque la réponse est vue comme incompatible avec le type d'entité nommée (valeur « NON »), alors elle ne correspond pas non plus au type spécifique attendu par la question à 71%, alors que cette décision est erronée pour 29% d'entre elles. La troisième ligne traite du cas où la détection ne peut être faite (31%). Ces résultats confortent l'idée que la détection des réponses ne correspondant pas au type est bien plus efficace que la détection des réponses valides avec ce critère.

La précision de la méthode est de 0,65 et le rappel de 0,45. Il est effectivement plus bas car 31% des couples (réponse, type) ne sont pas évalués.

Cette détection est mise en œuvre par la plupart des systèmes de questions réponses comme par exemple le système FRASQUES qui l'utilise lors de l'extraction des réponses. Les résultats obtenus par cette méthode peuvent être considérés comme les résultats de base qu'il s'agit d'améliorer.

Évaluation de la méthode validant certaines réponses Lorsqu'une liste correspond au type alors la valeur détectée est très souvent la bonne (75 %) que la réponse corresponde au type ou non (cf. tableau 3.15). Cela permet d'avoir une précision très élevée de 0,75. En revanche, comme peu de données (43%) sont évaluées, le rappel est bas, 0,32.

Les différentes évaluations obtenus permettent de voir que les méthodes sont complémentaires. En effet d'assez bons résultats, 75 % sont obtenus lorsque le type est connu par le système RITEL mais ne s'applique qu'à peu de données. En revanche l'approche par filtre s'applique sur beaucoup plus d'exemples, car la granularité des types est moins fine mais obtient de moins bons résultats notamment quand la réponse est vue comme correspondant au type spécifique (63 % de bons résultats contre 77 %).

L'utilisation des entités nommées peut donc permettre de valider le type de la réponse en s'appuyant sur plusieurs phénomènes : tout d'abord si le type correspond à une liste d'entités nommées collectée par le système de RITEL alors le fait que la réponse se trouve dans la liste associé au type donne une bonne indication sur la justification. Dans le cas contraire, quand le type ne correspond pas à une liste, les entités nommées peuvent aussi être utilisées afin de détecter certaines réponses ne correspondant pas au type. Toutefois des couples (réponse, type) ne sont toujours pas traités et ces critères, s'ils sont corrects, ne fournissent pas des décisions très fiables, d'où l'ajout de critères supplémentaires.

3.4.4 Recherche de définitions en corpus

Les critères précédents permettent de détecter si une réponse correspond à son type spécifique. Toutefois, ils ne peuvent pas s'appliquer pour les types spécifiques n'ayant pas un type d'entité nommée associée. Une méthode possible consisterait alors à s'appuyer sur la définition des termes à typer car généralement elle mentionne le type du terme défini. La ressource dans laquelle les recherches de définitions sont faites est l'encyclopédie Wikipédia[2]. Chacune de ses pages permet de définir son titre et de fournir un grand nombre d'informations le concernant. Par exemple, la page consacrée à « Stanley Kubrick » indique sa bibliographie, son parcours cinématographique, survole ses différents films et fournit notamment des informations relatives à son style... Plusieurs traitements peuvent être effectués pour déterminer que la réponse correspond au type spécifique : le premier concerne la structure des pages en elle-même tandis que le second recherche des patrons d'extraction dans l'ensemble des pages.

3.4.4.1 Recherche dans des pages particulières

De part sa structure encyclopédique, chacune des pages Wikipédia définit l'élément qui constitue son titre. Cela permet de formuler l'hypothèse suivante : si le type spécifique est trouvé dans la page Wikipédia dont le titre est associé à la réponse, cette dernière a de fortes chances d'être reliée à ce type et peut ainsi en être une instance ou un hyponyme.

Par exemple, la page consacrée à « Steven Spielberg » contient les mots « producteur » et « réalisateur ». Bien sûr, cette présence ne signifie pas que la réponse est du type mais si le couple testé est cohérent alors il a des chances d'être validé.

Pour ce travail nous avons utilisé la version de Wikipédia de novembre 2006 retenue pour la campagne de questions réponses CLEF 2008 [Forner, et al. 2009].

Le type, pris sous sa forme textuelle, est recherché dans les pages Wikipédia correspondant à la réponse c'est-à-dire les pages ayant pour titre la réponse ou dont le titre contient la réponse. Trois cas

[2]Wikipédia : http ://fr.wikipedia.org

sont alors possibles :

– aucun titre de page ne peut être associé à la réponse. Dans ce cas, rien ne peut être déduit et la valeur INCONNU est renvoyée. Ces cas correspondent par exemple à des entités sans grande importance comme par exemple des personnes ayant eu un rôle très ponctuel dans le temps comme le meurtrier « Alfred Henninger ».

– la page correspondant à la réponse contient bien le type. Cela implique que la réponse a une forte probabilité d'être du type cherché et la valeur OUI est renvoyée. Ainsi « une autruche » est un « oiseau ».

– la page ne contient pas le type. Dans ce cas la réponse ne correspond très probablement pas à ce type. La valeur NON est donc renvoyée. C'est par exemple le cas de « Bethléem » pour le type « planète ».

3.4.4.2 Utilisation de patrons d'extraction

Se limiter à certaines pages réduit la portée d'utilisation de Wikipédia car l'information peut aussi figurer dans d'autres pages.

Le critère présenté dans cette section s'appuie sur des patrons d'extractions, similaires à ceux présentés par Hearst [1992] et Morin [1998]. De tels patrons ont pour but d'expliciter que la réponse est une sorte de type comme « RÉPONSE est un TYPE » (*une autruche* est un *oiseau*). De manière globale, de tels patrons sont recherchés dans les pages Wikipédia et si l'un d'entre eux s'applique alors on supposera que la réponse est compatible avec le type. Cela peut donc se voir comme une utilisation du principe de Hearst [1992] avec ceci de différent que les patrons servent à valider des couples (hyponymes, hyperonymes) et non à en découvrir de nouveaux. De plus, ces patrons ne sont pas collectés mais seuls quelques uns, issus d'une analyse du corpus, sont utilisés. Le choix de Wikipédia comme corpus tient compte de sa spécificité. En effet, Wikipédia étant une encyclopédie, elle est plus à même de contenir des phrases de définitions qu'un corpus de journaux par exemple. Les patrons ayant servi pour cette étude sont les suivants :

1. ***RÉPONSE* être déterminant *TYPE***, avec de nombreuses variantes du verbe être et du déterminant, (*exemple Albert Einstein est un physicien*) ;
2. ***TYPE RÉPONSE*** (*physicien Albert Einstein*) ;
3. ***RÉPONSE*, déterminant *TYPE*** (*Albert Einstein, un physicien*) ;
4. ***RÉPONSE* (déterminant *TYPE*...)** (*Albert Einstein (le physicien ...)*) ;
5. ***RÉPONSE* : déterminant *TYPE*** (*Albert Einstein : le physicien*) ;

Afin de savoir si la réponse correspond au type attendu, les variables *TYPE* et *RÉPONSE* sont tout d'abord instanciées par leurs valeurs. Ainsi, pour vérifier que Johnny Depp est un acteur, *TYPE* prend la valeur acteur et *RÉPONSE* la valeur Johnny Depp. Cela est fait pour chacune des règles, ce qui amène à des phrases comme « *Johnny Depp est un acteur* », « *Johnny Depp sera l'acteur* », etc.

Ces phrases sont ensuite cherchées telles quelles dans les pages Wikipédia grâce au moteur de recherche Lucene [Hatcher & Gospodnetic 2004]. Si au moins un document est renvoyé, c'est qu'il contient l'expression et dans ce cas le couple (réponse, type) est validé et la valeur OUI renvoyée, sinon la réponse est considérée comme incompatible avec le type attendu et la valeur NON est donnée.

Système	Décision	#	# Réponses du type spécifique	# Réponses n'étant pas du type spécifique
Recherche dans des pages spécifiques	OUI	661	**74 % (491)**	26 % (170)
	NON	589	39 % (228)	**61 % (361)**
	INCONNU	1470	43 % (641)	57% (829)
Utilisation de Patrons d'extraction	OUI	974	**73 % (713)**	26 % (261)
	NON	1746	37 % (647)	**63 % (1099)**

TAB. 3.16 – Matrice de confusion des méthodes utilisant Wikipédia

3.4.4.3 Évaluation

Évaluation de la recherche dans des pages particulières Lorsque la réponse est considérée comme correspondant au type, c'est souvent le cas (74%) (cf. tableau 3.16). En revanche, davantage d'erreurs sont commises quand la réponse est perçue comme ne correspondant pas au type (seules 61% des décisions sont correctes). Cela peut s'expliquer par le fait que le type peut être remplacé dans la page de la réponse par un synonyme ou un terme faisant référence à un type plus général ou plus spécifique. Par exemple, Albert Einstein n'est pas reconnu comme un chercheur car le terme est remplacé par physicien.

Les résultats montrent également que peu de réponses sont évaluées, seules 46 % des réponses ont une page Wikipédia qui leur est dédiée. D'un point de vue global, la décision est correcte dans 68% des cas. Le nombre élevé de réponses non évaluées entraîne un rappel plus faible, de 0,32.

Évaluation de la méthode utilisant des patrons d'extraction Le tableau 3.16 montre tout d'abord que toutes les valeurs peuvent être évaluées par cette méthode, contrairement à la précédente. Ainsi le rappel aura une valeur égale à la précision. De plus, les résultats sont plutôt bons quand la réponse OUI est renvoyée (73% des réponses correspondent bien au type). Les cas d'erreurs restantes correspondent aux cas où une règle s'applique sans signifier une relation d'hyponymie. Cela peut être par exemple le cas des règles 2 ou 3 qui peuvent s'appliquer dans des structures de listes comme : *le président Nicolas Sarkozy, le premier ministre François Fillon et la ministre Roseline Bachelot* qui peut indiquer que Nicolas Sarkozy est un premier ministre ce qui est faux.

Quand la réponse NON est donnée de moins bons résultats sont obtenus puisque seuls 63 % des résultats renvoyés sont corrects. De manière globale 66 % des résultats sont corrects ce qui se traduit par une précision et un rappel de 0,66.

Une combinaison des deux approches aurait pu être de rechercher de tels patrons dans les pages consacrées à la réponse mais sa portée serait plus restreinte puisqu'elle ne s'appliquerait que sur très peu de couples (réponse, type). Une autre possibilité d'amélioration pourrait considérer le nombre de documents renvoyés par la seconde méthode afin de décider du OUI.

3.4.5 Recherche en corpus

Le troisième type de critère est d'ordre statistique. Il se place dans un cadre plus général et s'intéresse à l'apparition de la réponse et du type dans un ensemble de documents, quel que soit le document ou la relation les liant. On suppose que si le type et la réponse apparaissent souvent dans les mêmes

documents, alors ils sont liés. Cette idée s'inspire beaucoup de celle de Magnini et al. [2002b] qui l'utilisaient afin d'effectuer de l'ordonnancement de réponses.

Afin de mettre en évidence les relations entre la fréquence d'apparition du type et de la réponse et le fait que la réponse soit du type attendu, un système par apprentissage a été créé. Ce système reprend un ensemble de critères décrits par Magnini et al. et repris dans Schlobach et al. [2004] et Schlobach et al. [2007]. Ils correspondent à des mesures de cooccurrence en corpus et sont les suivants :

– **les proportions d'apparition** : le rapport entre le nombre de documents contenant le type et la réponse et le nombre de documents contenant la réponse ou le nombre de documents contenant le type. Ce critère permet de détecter les cas où la réponse apparaît fréquemment accompagnée du type. Si les mots apparaissent souvent dans les mêmes documents alors ils sont probablement liés.

$$C1 = \frac{nb\ documents(réponse + type)}{nb\ documents(réponse)} \tag{3.1}$$

$$C2 = \frac{nb\ documents(réponse + type)}{nb\ documents(type)} \tag{3.2}$$

– **les fréquences d'apparition** du type, de la réponse et de l'ensemble type+réponse. Ces critères permettent de savoir si les termes apparaissent souvent et notamment si le type ou la réponse est un terme souvent utilisé. Ces informations permettent de renforcer les proportions d'apparition en différenciant les cas où les deux termes sont fréquents et sont toujours ensemble des cas où l'occurrence d'un terme est rare. Dans un tel cas la probabilité d'apparition commune peut être forte mais fortuite.

$$C3 = \frac{nb\ documents(réponse)}{nb\ documents} \tag{3.3}$$

$$C4 = \frac{nb\ documents(type)}{nb\ documents} \tag{3.4}$$

$$C5 = \frac{nb\ documents(réponse + type)}{nb\ documents} \tag{3.5}$$

– **la mesure PMI** (*Pointwise Mutual Information*), mesure classique de statistique qui permet de mesurer la force d'association de deux termes. Elle correspond au rapport entre la fréquence d'apparition commune et le produit des fréquences d'apparition du type et de la réponse. Cela constitue une combinaison des informations obtenues par les critères précédents.

$$\begin{aligned} C6 &= \frac{Fréquence(réponse + type)}{Fréquence(réponse) * Fréquence(type)} \\ &= \frac{C5}{C4 * C3} \\ &= \frac{nb\ documents(réponse + type) * nb\ documents\ total}{nb\ documents(réponse) * nb\ documents(type)} \end{aligned} \tag{3.6}$$

Collection	Décision	#	# Réponses du type spécifique	# Réponses n'étant pas du type spécifique
Wikipédia	OUI	822	**66 % (542)**	34 % (280)
	NON	725	31 % (220)	**69 % (505)**
Le Monde	OUI	634	**76 % (479)**	24 % (155)
	NON	913	31 % (283)	**69 % (630)**

TAB. 3.17 – Matrice de confusion de la méthode utilisant des mesures statistiques sur Le Monde

Deux corpus de documents ont été utilisés afin d'effectuer les calculs statistiques. Le premier est l'encyclopédie Wikipédia et le second les articles du journal « Le Monde » de 1992 à 2000. Ces articles sont utilisés car les réponses ont été extraites de ces journaux qui contiennent donc toutes les réponses et notamment celles n'apparaissant pas dans Wikipédia telles que les réponses ayant eu une brève importance historique quand les questions portent sur un fait divers. De plus, ces articles sont plus nombreux, ce qui semble plus intéressant dans une étude statistique. Des documents de nature différente entraînent des formulations différentes dans lesquelles l'utilisation des mots n'est pas la même. Il semble donc intéressant de tester les deux corpus.

Ici nous n'avons pas affaire à un critère isolé mais à un ensemble de critères visant à marquer le même type de phénomène. Une combinaison a été créée afin de calculer le critère global. Les résultats fournis par chacun d'eux sont donc combinés grâce à une combinaison d'arbres de décision par le bagging fourni par le système WEKA. Cette combinaison a aussi été choisie afin de combiner l'ensemble des critères permettant de vérifier le type spécifique de la réponse, en 3.4.6. Comme les différents critères sont complémentaires elle permet de distinguer l'utilisation de chacun d'eux en fonction des cas.

L'évaluation de la méthode ne peut se faire sans base de test. Celle-ci est décrite dans la section suivante. Les deux méthodes obtiennent de bons résultats avec 68 % et 72 % de résultats corrects (cf. tableau 3.17). Dans ces cas nous avons donc une précision, un rappel et une f-mesure égaux puisqu'aucune restriction n'est faite sur les réponses et qu'elles sont toutes évaluées.

De meilleurs résultats sont obtenus en examinant les articles de journaux qu'en interrogeant Wikipédia. La différence a lieu lorsque la réponse est vue comme correspondant au type. Une bonne détection est observée dans 66 % des cas sur les articles Wikipédia et dans 76 % pour les articles de journaux. Cette différence est probablement due à la différence de taille des deux corpus ; comme le corpus Le Monde est plus important que Wikipédia, la présence commune fréquente de certains mots est plus significative.

3.4.6 Combinaison des critères

Après avoir décrit l'ensemble des critères, l'étape finale consiste à les combiner. La combinaison d'arbres de décision grâce au bagging a été choisie pour cette tâche.

Les critères utilisés sont tous ceux présentés précédemment :

1. le filtre sur les entités nommées ;

2. la validation du type grâce aux entités nommées ;
3. la présence du type dans la page Wikipédia de la réponse ;
4. l'application de patrons d'extraction dans les pages Wikipédia ;
5. l' ensemble des critères statistiques calculés sur Wikipédia :
   – le rapport entre le nombre d'apparitions de la réponse et du type ensemble et le nombre d'apparitions du type ou de la réponse ;
   – la fréquence d'apparition du type, de la réponse et de l'ensemble type+réponse ;
   – la mesure PMI ;
6. l'ensemble des critères statistiques calculés sur les articles du journal Le Monde.

Parmi les critères, l'ensemble des critères statistiques a été gardé sans faire de combinaison au préalable afin de permettre au classifieur d'effectuer la meilleure combinaison possible des différents critères. De plus, ils ont été calculés à la fois sur les documents Wikipédia et sur les articles de journaux car ces deux bases permettent d'avoir des informations différentes pouvant être combinées.

3.4.7 Résultats

Après voir vu l'ensemble des critères ainsi que la manière de les combiner il reste à évaluer la solution proposée.

Comme la combinaison des critères est effectuée par apprentissage il est nécessaire d'avoir une base de test et une base d'apprentissage. La base d'apprentissage est celle construite à partir de l'évaluation EQueR [Grau, et al. 2008].

La base de test est construite à partir des données fournies par la campagne de validation de réponses AVE 2006 [Peñas et al. 2006]. Les données de cette campagne permettent d'utiliser 1 547 paires réponse/type spécifique dont la moitié (762) est valide, i.e. la réponse est du type spécifique. Ces paires sont issues de 90 questions et correspondent à 47 types spécifiques différents. Rappelons que les types spécifiques sont reconnus lors de l'analyse des questions du système FRASQUES qui les reconnaît en examinant l'analyse en constituants de la question. Les passages justificatifs sont également analysés par FRASQUES afin de relever les entités nommées et notamment celle correspondant à la réponse. La figure 3.4 présente plus en détail le traitement effectué sur les exemples de la base de test. Il montre ainsi les prétraitements effectués par FRASQUES afin d'obtenir le type spécifique, le type d'entité nommée et les entités nommées contenus dans le passage. Il montre également les différents critères ainsi que leur combinaison.

Deux types d'évaluation sont effectués :
– l'évaluation des critères pris séparément sur la base de test ;
– l'évaluation de la combinaison des critères ;

Les résultats de chacun des critères obtenus sur la base de test permettent d'apprécier l'importance des différents critères et notamment de détecter ceux qui sont les plus prometteurs. Cela permet également de s'assurer que les remarques effectuées sur la base d'apprentissage s'appliquent également sur une autre base (cf. tableau 3.18).

Les méthodes sont assez robustes puisque les résultats obtenus par chacune d'elles sur la base de test sont assez semblables à ceux obtenus sur la base précédente. Nous pouvons remarquer que la

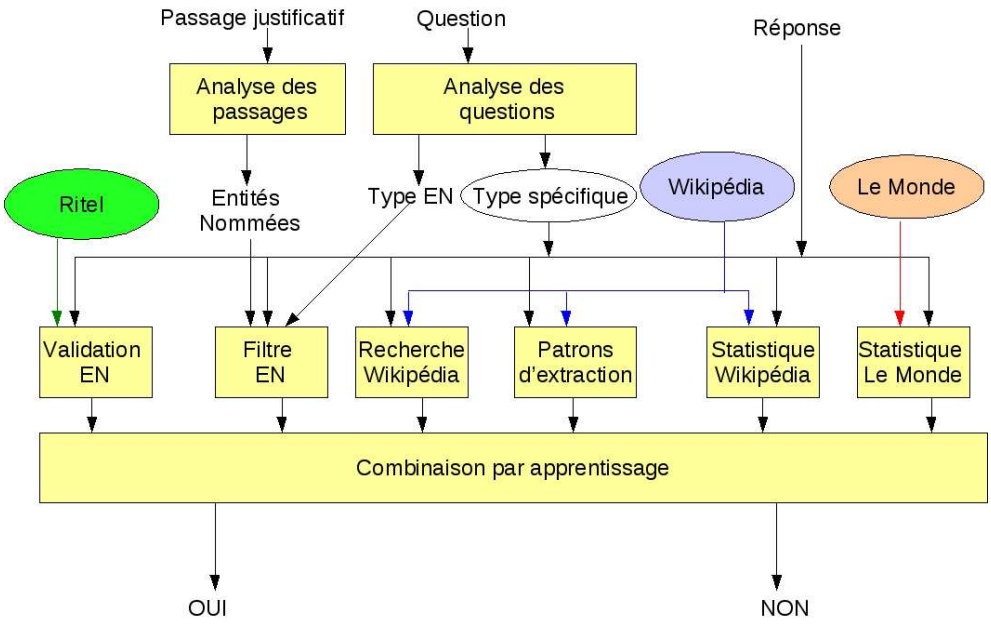

FIG. 3.4 – Le système de vérification du type

vérification de la présence du type dans la page Wikipédia associée à la réponse obtient de meilleurs résultats sur la base de test. Le rappel passe notamment de 0,32 à 0,46 ce qui indique que proportionnellement davantage de réponses ont pu être évaluées par cette méthode. Cela s'explique par la nature des données. En effet, les données de la base de test contiennent plus souvent des noms de personnes pour lesquelles il existe des pages Wikipédia associées.

La seconde information pouvant être extraite du tableau concerne l'importance des différents critères. Tout d'abord, nous pouvons voir que la validation des réponses en utilisant les listes d'entités nommées est la méthode la plus sûre (précision 0,80). Ce qui revient bien à notre idée originelle qui est que même si ce critère ne peut pas être utilisé pour tous les cas quand une information est trouvée elle est très souvent bonne car ses performances sont liées aux bonnes performances du système

Critère	Précision	Rappel	F-mesure
1) Filtre sur les entités nommées générales	0,69	0,54	0,60
2) Validation grâce aux listes d'entités nommées	**0,80**	0,32	0,45
3) Recherche dans la page Wikipédia de la réponse	0,72	0,46	0,57
4) Utilisation de patrons syntaxiques	0,70	0,70	**0,70**
5) Critères statistiques sur Wikipédia	0,68	0,68	**0,68**
6) Critères statistiques sur Le Monde	0,72	**0,72**	**0,72**
Combinaison	**0,80**	**0,80**	**0,80**

TAB. 3.18 – Résultats des critères sur la base de test et leur combinaison

Décision	# Réponses du type spécifique	# Réponses n'étant pas du type spécifique
OUI	**80 % (603)**	20 % (149)
NON	20 % (159)	**80 % (636)**

TAB. 3.19 – Matrice de confusion de la méthode globale

Base de test	Proportion de couples correctement classés
Avec EN (1205)	82%
Sans EN (342)	74%

TAB. 3.20 – Vérification du type en fonction des entités nommées

d'entités nommées utilisé. La recherche du type dans la page Wikipédia associée à la réponse obtient également de bons résultats quand elle peut être appliquée (précision 0,72).

De manière plus globale, les méthodes utilisant les patrons syntaxiques et celle étudiant la co-occurrence dans les textes obtiennent de bons résultats, précision 0,70 et 0,72. De plus comme elles s'appliquent à toutes les réponses, elles obtiennent une précision, un rappel et une f-mesure égaux.

Comme les méthodes sont différentes elles peuvent être complémentaires. Pour certains cas elles ont ainsi des décisions différentes et si une méthode classe mal un exemple, les autres n'effectueront pas cette erreur et la décision finale sera la bonne. Ce qui montre l'intérêt de la combinaison qui revient à donner une note à chaque exemple en tenant compte des observations effectuées par les différentes méthodes.

Pour évaluer la combinaison, il faut comparer ses résultats à ceux obtenus par d'autres méthodes. La plupart des systèmes de questions réponses utilisant la détection du type d'entité nommée comme filtre, les résultats obtenus par cette méthode servent de baseline.

La méthode complète permet de classer correctement 80 % des données (cf. tableau 3.18). Ce pourcentage élevé montre que la méthode choisie est efficace. Les résultats obtenus par la combinaison de méthodes sont très nettement supérieurs à ceux reposant uniquement sur un module de reconnaissance des entités nommées. La f-mesure la surpasse ainsi de 0,20 puisque les résultats sont bien supérieurs au niveau de la précision (0,80 vs 0,69). De plus la méthode est applicable à tout type de questions alors que la baseline ne s'applique qu'à 54 % des réponses. Les résultats obtenus par la méthode sont également supérieurs à ceux obtenus par tous les critères pris séparément ce qui montre bien l'intérêt de l'utilisation de la combinaison.

Afin d'étudier plus en détail les résultats, la matrice de confusion de la méthode a été calculée (cf. tableau 3.19). Elle montre que l'approche obtient des résultats similaires quelle que soit la valeur retournée (OUI (80%), NON (80%)).

Afin de mieux comprendre les résultats, une étude distinguant les cas où la réponse est associée à un type d'entité nommée des cas où elle n'en n'a pas, a été menée dans deux buts. Tout d'abord cela permet de voir l'intérêt de la vérification du type d'entité nommée. Cela permet également de s'assurer que la méthode peut être utilisée dans les cas où la question n'attend pas de type d'entité nommée. Le tableau 3.20 indique qu'une meilleure décision est prise pour les réponses ayant un type

d'entité nommée associé. Cela peut s'expliquer par le fait que la plupart des réponses de la base d'apprentissage sont associées à un type d'entité nommée (78%) et témoigne de l'utilité des critères portant sur les entités nommées. Nous pouvons aussi remarquer que de bons résultats sont obtenus quand la question attend une réponse n'étant pas d'un type d'entité nommée particulier ce qui fournit une solution sans avoir recours à des bases de connaissances dans tous les cas. Ce second résultat montre que cette vérification pourrait effectivement permettre d'améliorer les systèmes de questions réponses puisqu'une vérification pourrait être effectuée sur davantage de questions.

Des tests ont montré que lorsque le type était composé de plus d'un mot le système a davantage de mal à reconnaître les réponses correspondant au type puisqu'il n'effectue que 46 % de bonnes détections. Les cas correctement traités correspondent à des termes souvent utilisés ensemble comme « premier ministre ». Pour les autres cas comme « grande entreprise », seules 5 % des réponses correctes sont détectées. Ces informations montrent qu'il vaut mieux garder un type spécifique court et que pour une question comme « *Dans quel film de Kevin Reynolds Kevin Costner a-t-il joué* » le type « film » est à privilégier à « film de Kevin Reynolds ».

Avant de clore cette section, voyons quelques résultats obtenus :
– Hosni Moubarak est un président ;
– Krypton est une planète ;
– Bethléem n'est pas une planète ;
– Barings n'est malheureusement pas une « grande banque ». Cela doit être dû à la présence de l'adjectif ;
– le Parti Socialiste est, à tort, un président pour une question comme « *Quel président succéda à Jacques Chirac ?* ». Cela est sûrement dû au rapport fréquent entre ces deux termes. Par exemple le critère portant sur l'étude en corpus montre que ces deux termes sont souvent reliés. De plus le type se trouve dans la page Wikipédia reliée à la réponse avec des phrases comme « Le président Mitterrand ». Notons que cette erreur aurait pu être évitée en utilisant davantage la vérification des entités nommées.

Les résultats peuvent aussi être rapprochés de ceux des systèmes de recherche d'entités nommées cherchant toutes les entités nommées présentes dans un corpus de texte. La campagne MUC, présentée par Grishman & Sundheim [1995], a permis d'évaluer ces systèmes en se focalisant sur la recherche des types d'entités nommées personne, lieu, organisation, date, expressions de temps, pourcentage et unité monétaire. Le système ayant eu les meilleurs résultats à cette campagne obtient une f-mesure de 0,93. Ces résultats sont supérieurs à ceux obtenus par notre système. Toutefois cette campagne s'intéresse seulement à sept types d'entités, d'un niveau de granularité supérieur à celui de notre système, qui peut de plus détecter tout nouveau type apparaissant dans une question. Cela introduit une différence de résultats car il semble plus facile de montrer qu'un nom correspond à une personne que de montrer que c'est un acteur.

Lors de la campagne TREC Entity, les systèmes étaient évalués à l'aide de différentes mesures dont la précision aux différentes positions. Le meilleur système obtient une précision de 0,3075 ce qui revient à une bonne découverte de relation pour moins d'$\frac{1}{3}$ des cas. Il se fonde sur un formalisme logique et obtient des résultats bien inférieurs aux nôtres ce qui peut s'expliquer par la différence de la tâche (vérification vs découverte). De plus, les réponses sont évaluées selon l'exactitude du lien donné.

3.4.8 Intérêt pour la validation de réponses

Nous avons évalué notre système de manière indépendante c'est-à-dire en regardant la proportion de détections correctes. Comme le cadre général de cette thèse est la validation de réponses, nous l'avons étudié en le replaçant dans cette thématique.

La validation de réponses, comme nous l'avons vu précédemment, consiste, à partir de triplets réponses constitués d'une question, d'une réponse et d'un passage justificatif, à détecter si la réponse est effectivement valide. Bien sûr quand la réponse ne correspond pas au type spécifique attendu par la question alors elle a très peu de chances d'être valide.

Une première étude traite donc du rapport entre la validité du type de la réponse et la validité de la réponse. Remarquons que certaines approches utilisent ce critère afin d'ordonner les réponses. Par exemple Huang et al. [2009] effectuent de l'ordonnancement de passages en vérifiant notamment qu'une instance du type spécifique se trouve dans le passage.

Dans notre cas, l'étude porte sur la proportion de réponses correspondant au type qui sont effectivement valides ainsi que sur la proportion de réponses n'y correspondant pas qui ne le sont pas.

Le tableau 3.21 présente une correspondance entre la vérification du type et la validité de la réponse pour les réponses de la base de test. Notons que cette base contient beaucoup plus de réponses non valides (80%) que de réponses valides ce qui est dû à la nature même de la tâche puisque les systèmes proposent cinq réponses dont au mieux une est valide. Comme certaines réponses n'ont pas été évaluées par les organisateurs en termes de validité, la base contient 1 457 triplets au lieu de 1 547.

Réponse du type	# Réponses valides	# Réponses non valides
OUI (702)	**34 % (236)**	66 % (466)
NON (755)	7 % (53)	**93 % (702)**
TOTAL (1457)	20 % (289)	80 % (1168)

TAB. 3.21 – Rapport entre la vérification du type et la validité des réponses

Si la réponse est vue comme ne correspondant pas au type cherché alors elle est très souvent non valide (93 %). Toutefois nous pouvons remarquer que lorsque la réponse est valide la vérification du type effectue 18 % d'erreurs ($\frac{53}{289}$).

En revanche rien ne peut être dit quand la réponse est considérée comme étant du type attendu puisque les réponses reconnues comme étant du type attendu sont le plus souvent incorrectes (66 %). Ce faible résultat montre que la méthode est nettement insuffisante pour détecter la validité des réponses. En effet la validité dépend également du contenu des passages permettant de vérifier qu'ils justifient effectivement la réponse et malheureusement les passages sélectionnés par les systèmes de questions réponses ne sont pas tous pertinents. Ainsi de nombreuses réponses sont incorrectes tout en correspondant au type. Par exemple « Tim Burton » est bien un réalisateur mais ne répond pas à la question « *Quel est le réalisateur du retour du Jedi ?* ». Ces cas se présentent notamment quand plusieurs réponses du même type se trouvent dans le même passage comme par exemple « *Mars Attack de Tim Burton, Le retour du Jedi de Georges Lucas* ».

Ces différentes informations permettent de proposer un premier système de validation qui consiste à filtrer les réponses ne correspondant pas au type spécifique. En évaluant le système à l'aide des mesures utilisées pour la validation de réponses (la précision, le rappel et la f-mesure sur les réponses OUI), une précision de 0,34, un rappel de 0,81 et donc une f-mesure de 0,48 sont obtenus alors que le cas le plus simple consistant à dire toujours OUI n'a une précision que de 0,20 et un rappel de 0,33 ce qui montre que de bien meilleurs résultats sont quand même obtenus.

3.4.9 Conclusion

Nous avons proposé une méthode permettant de vérifier que la réponse correspond bien au type attendu par la question en combinant différents critères par apprentissage. Les premiers utilisent les entités nommées afin de rejeter des réponses non compatibles ou d'en valider d'autres en utilisant les entités nommées comme une base de connaissances. Les seconds traitent des particularités des pages Wikipédia en cherchant le type dans la page associée à la réponse ou en tenant compte de certaines structures de phrases spécifiques. Les derniers sont d'ordre statistique et observent l'apparition commune du type et de la réponse dans un ensemble de documents.

Les évaluations ont montré que de bons résultats sont obtenus, 80 % de bonnes détections. Cette méthode fournit une première information sur la validité d'une réponse en se focalisant plus particulièrement sur la problématique consistant à savoir si elle correspond à une réponse possible à la question. Toutefois, cette seule information ne suffit pas à valider la réponse car d'autres informations sont à vérifier comme l'action de la question. C'est pourquoi la section 3.5 intègre les deux types de vérification.

La vérification du type de la réponse peut également être utilisée par les systèmes de questions réponses lors de l'extraction des réponses qui s'appuie bien souvent sur l'extraction des groupes nominaux correspondant au type d'entité nommée attendu en retour par la question. Ainsi si la question attend une réponse n'étant pas d'un type d'entité nommée connue mais possède un type spécifique, la vérification du type pourrait permettre de restreindre l'ensemble candidat à des réponses plus pertinentes. Dans ce cadre, la reconnaissance des entités nommées est une tâche clé puisque plus un système sait reconnaître de types d'entités nommées, meilleures sont ses performances.

3.5 Intégration de la vérification du type dans le système de validation de réponse

Nous avons donc vu dans les sections précédentes différents critères contribuant à vérifier que la réponse est valide.

– l'analyse des passages qui vérifie que le passage contient bien les informations demandées par la question ;

– la vérification du type de la réponse qui permet de rejeter certaines réponses ne correspondant pas à la question.

L'étape suivante consiste donc à les combiner. Pour ce faire, deux méthodes peuvent être envisagées :

– la première place la vérification du type comme un filtre. Les réponses ne correspondant pas au type sont reconnues comme fausses. La validité des autres est détectée par application de la

Méthode	Précision	Rappel	F-mesure
Baseline	0,5	0,62	0,55
Comme filtre	0,59	0,56	0,57
Comme critère	0,51	0,62	0,56

TAB. 3.22 – Intégration de la vérification du type

méthode de validation. En effet, nous avons vu que ce filtre permet de détecter correctement les réponses non valides ;

– la seconde considère la vérification du type comme un critère fourni au classifieur ; l'idée étant que la classification détectera la meilleure utilisation de cette vérification.

Comme seules les questions ayant un type spécifique sont concernées pour cette partie, les bases d'apprentissage et de test sont extraites des ensembles utilisées pour évaluer le système AVAL et seules les questions mentionnant explicitement un type de réponse attendu sont retenues. La base d'apprentissage contient 488 exemples dont 186 sont valides. La base de test contient 302 exemples dont 80 sont valides.

Le tableau 3.22 présente les résultats des deux intégrations possibles de la vérification du type à la validation de réponses. Les résultats à améliorer sont ceux obtenus par le système AVAL (méthode baseline).

Ce tableau montre que l'ajout du type comme critère n'améliore que peu les résultats. En effet la méthode obtient les mêmes résultats avec ce critère que sans (f-mesure 0,56 vs 0,55). Cela n'est pas surprenant car cette vérification ne constitue pas le critère principal pour justifier une réponse et de ce fait le classifieur ne le considère pas comme un critère fortement discriminant. Ainsi, le classifieur privilégie la proximité des termes de l'hypothèse dans le passage et la présence dans le passage de l'ensemble des termes de la question.

L'étape de filtre, avant l'apprentissage, reconnaît 111 réponses comme n'étant pas du type attendu. Parmi elles 11 (10%) sont malheureusement valides.

Le traitement de la vérification du type de la réponse comme filtre permet d'augmenter la précision (0,5 vs 0,59). La proportion de réponses correctement vues comme valides est donc plus élevée. Toutefois le rappel diminue lui aussi ce qui est dû aux 10% d'erreurs de la vérification du type (0,56 vs 0,62) ce qui entraîne une faible augmentation de la f-mesure.

Une étude plus fine des résultats montre que sans considérer cette vérification 49 faux positifs sont détectés (la réponse est vue comme valide à tort). Ce sont ces faux positifs que la vérification du type pourrait faire diminuer. Or, sur ces 49 fausses réponses seules 17 (35%) peuvent effectivement être résolues en utilisant la vérification du type et sur ces 17 cas, 14 (82%) sont effectivement bien résolus. Le faible nombre de réponses pouvant être améliorées explique que le critère est peu utilisé dans un processus de validation par apprentissage et donc les résultats globaux restent analogues. Des tests similaires ont été effectués avec des bases d'apprentissage et de tests différents, notamment en considérant également des questions sans type spécifique pour augmenter les exemples permettant de valider une réponse, mais ces modifications ne changent pas les résultats. Afin de mieux évaluer les résultats, la taille de la base de test devrait être augmentée.

3.6 Conclusion et perspectives

Nous avons vu dans ce chapitre une méthode détectant la validité des réponses. Elle part du principe que plusieurs vérifications sont à effectuer afin de s'assurer de la validité de la réponse.

Parmi ces vérifications, celles de la première catégorie étudient le passage justificatif afin de vérifier qu'il contient bien les informations nécessaires à valider une réponse. La seconde s'applique quant à elle pour les questions attendant un type spécifique en réponse et a pour but de vérifier que la réponse proposée correspond bien au type attendu.

Comme une approche par apprentissage est suivie, il est possible d'ajouter d'autres critères. Parmi ces différents ajouts, présentés plus en détail lors des perspectives finales en 5.1, une vérification pourrait porter sur le lieu de la question qui a été marqué comme un élément important mais pour lequel aucune vérification particulière n'a été effectuée. Le second grand type de perspectives traite des relations entre les différents termes de la question dans le passage. Le système AVAL considère deux termes comme reliés s'ils sont proches dans le document. Une vérification plus poussée pourrait venir reconnaître plus finement la relation en tenant compte des relations syntaxiques ou de l'utilisation de règles de paraphrases ce qui permettrait de mieux reconnaître deux phrases de même sens mais formulées différemment.

Le système AVAL considère la validation de réponses comme un élément externe aux systèmes de questions réponses et venant vérifier les réponses renvoyées. Cette place est intéressante s'il s'agit de valider les sorties de plusieurs systèmes, qui ont chacun leurs propres critères de décision. Elle n'offre aucun intérêt à la suite d'un SQR, puisque ce dernier aurait déjà opéré des choix. Ou alors il faut considérer de très nombreuses réponses, et cela nous ramène à une problématique de réordonnancement.

Chapitre 4

Utilisation de la validation de réponses dans un système de questions réponses : le système QAVAL

Sommaire

Dans les chapitres précédents, nous avons présenté notre système de validation de réponses. Le système a été évalué à l'aide de données faites des résultats de systèmes de questions réponses en détectant les réponses valides. Mais nous avons vu que cette approche pouvait aussi être intégrée dans un système de questions réponses pour choisir les réponses à proposer.

La plupart des systèmes de questions réponses recherchent la réponse à une question dans des collections constituées d'articles de journaux car ces corpus ont été construits pour les campagnes d'évaluation. Mais ils peuvent aussi être appliqués sur des documents provenant du Web. Soit ils passent par des moteurs existant sur le Web, soit ils interrogent des collections crawlées sur Internet.

Toutefois, cette recherche se heurte à deux problèmes : la nature des documents, qui contiennent des images, des listes ... pour lesquels il est nécessaire de faire un prétraitement et le fait que les documents peuvent comporter des fautes d'orthographe et de syntaxe.

Les systèmes de questions réponses participant à l'évaluation créée dans le cadre du projet Quæro [Quintard et al. 2010] recherchent des réponses aux questions dans des documents issus du Web et rencontrent davantage de difficultés sur ce type de documents que lors de recherches dans des articles de journaux. En effet, en 2008, sur les 167 questions factuelles, les systèmes obtiennent une précision comprise entre 15,9 % et 38,6 %. En 2009, sur les 295 questions factuelles, les résultats sont un peu meilleurs puisqu'ils se situent entre 27,5 % et 50,2 %. Les résultats obtenus par le meilleur système sont bien moins bons que ceux qu'il a obtenus sur la langue française lors de la campagne CLEF 2006 (68,95 %). En ce qui concerne FRASQUES [Ferret et al. 2002], notre système de référence, une analyse des résultats a montré que les nouveaux problèmes venaient essentiellement de la sélection des documents et du fait que les documents contiennent des phrases incomplètes et mal segmentées ce qui posait problème lors de l'extraction de réponses à partir de passages composés d'une seule phrase.

Notre système de validation de réponses peut s'appliquer à tout type de documents. En effet, il ne traite principalement que de la syntaxe locale qui se traduit par des critères portant sur la présence des mots de la question dans le passage ou leur proximité dans ce dernier. D'autres critères recherchent une information en dehors des documents. La création d'un système de questions réponses robuste à tout type de documents peut donc s'appuyer sur ce module et utiliser une notion de passages justificatifs centrés sur un nombre de mots et non sur la notion de phrase.

Afin de pouvoir ordonner les différentes réponses, une méthode classique consiste à pondérer chacune d'elles suivant un score de confiance. De nombreux systèmes d'ordonnancement de réponses [Suzuki et al. 2002 ; Martin et al. 2001] utilisent une approche assez similaire avec un ensemble de critères tels que la présence des termes de la question dans le passage avec des possibles variations, la présence des entités nommées de la question ou leur proximité dans le passage. D'autres systèmes, [Higashinaka & Isozaki 2008 ; Huang et al. 2009], effectuent des vérifications sur la réponse afin de reconnaître qu'elle correspond à la question comme la correspondance d'entités nommées ou des vérifications plus ou moins précises du type spécifique. Nous utilisons des critères analogues avec la présence des mots de la question dans le passage et la vérification du type spécifique auxquels nous intégrerons des critères propres à notre système de questions réponses.

Nous nous intéresserons exclusivement aux questions factuelles car ce sont celles qui relèvent de notre méthode de validation, contrairement aux questions complexes, et, à l'inverse des questions de définition, l'extraction d'une réponse ne s'appuie pas uniquement sur des patrons d'extraction. Ce type de questions porte sur la recherche d'une caractéristique d'une entité ou d'un événement telles que sa date (« *Quand le pont de Normandie a-t-il été inauguré ?* »), son lieu (« *Où ont eu lieu les jeux olympiques de 1992 ?* ») ou une autre information (« *Dans quel film de Kevin Reynolds Kevin Costner a-t-il joué ?* »).

Le système de questions réponses doit pouvoir analyser tout type de documents et pour cela suit une approche un peu particulière et différente de FRASQUES. En effet, des passages de textes de taille fixe sont tout d'abord recherchés. Ainsi les réponses à valider sont extraites d'un passage de

texte et non d'une phrase. Les passages sont extraits grâce au moteur de recherche. Afin de pouvoir s'adapter aux documents Web, un prétraitement des documents a aussi été effectué.

La seconde innovation consiste à extraire un grand nombre de réponses candidates depuis ces passages afin de ne pas omettre la bonne réponse quitte à avoir de nombreuses réponses candidates. Il est donc nécessaire d'effectuer un bon ordonnancement de celles-ci afin de pouvoir faire ressortir la réponse correcte.

4.1 Architecture du système

La stratégie du système QAVAL (Question Answering by VALidation) s'appuie sur un processus en six temps (cf. figure 4.1) :

- l'analyse des questions ;
- la recherche de passages de petite taille ;
- la sélection des passages les plus pertinents parmi ceux renvoyés ;
- l'annotation des passages afin de marquer les informations provenant de la question ;
- l'extraction des réponses candidates depuis les différents passages ;
- l'ordonnancement des réponses par application de la méthode de validation de réponses.

Le système QAVAL a été élaboré par différentes personnes. Je me suis particulièrement intéressé à l'extraction et à l'ordonnancement des réponses. En effet, ces deux tâches correspondent à utiliser la validation de réponses dans un système de questions réponses. Il est toutefois nécessaire de présenter plus en détail les étapes précédentes.

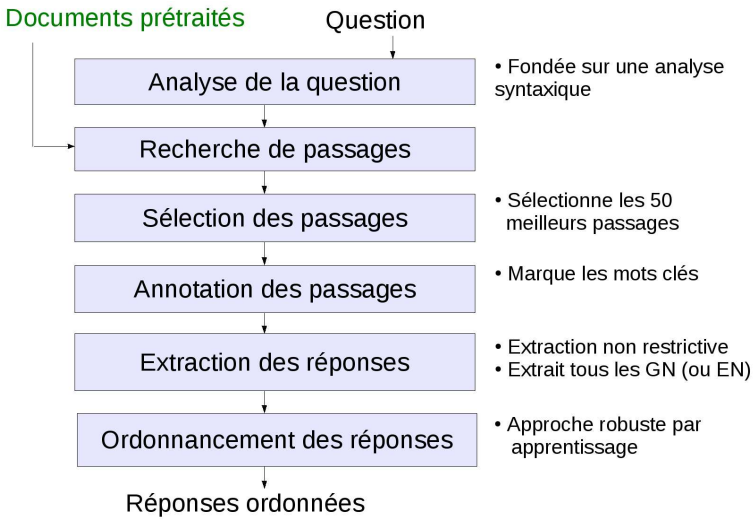

FIG. 4.1 – Le système QAVAL

FIG. 4.2 – Exemple de passage Web

4.1.1 Prétraitement des documents

Les documents Web sont dans un format permettant un rendu visuel par les navigateurs. Afin de comprendre les spécificités posées par des documents Web, considérons la question « *Citer une chute d'eau en Islande.* ». La page Web présentée en figure 4.2 est un document duquel la réponse « Dettifoss » peut être extraite et validée. Dans ce document il est bien indiqué que Dettifoss est une chute d'eau islandaise. Mais les informations de la question sont réparties dans le titre et le texte accompagnant les différentes images. Une partie de la structure HTML du document est présentée dans le tableau 4.1. Afin de pouvoir traiter les documents dans le cadre de notre système de questions réponses, il est nécessaire de les transformer dans un format textuel adapté à nos processus d'analyse. Ce travail a été réalisé par Falco [2010b] et correspond à la création du système Kitten[1].

Notre système de questions réponses nécessite donc que les documents soient au format textuel où toutes les balises sont supprimées. Pour cela, les documents sont tout d'abord transformés en format XHTML en appliquant *HTMLCleaner*[2] et *jTIDY*[3]. Le contenu textuel des documents est alors obtenu depuis ces structures à l'aide d'un ensemble de filtres portant sur les balises XHTML puisque certaines indiquent les paragraphes ou les titres alors que d'autres pointent les scripts qu'il ne faut pas

[1] Kitten Is A Textual Treatment for Extraction and Normalization
[2] http ://htmlcleaner.sourceforge.net/
[3] http ://jtidy.sourceforge.net/

```
<head>
<title>Islande Dettifoss Hveravellir</title>
</head>
<td width="10"><font color="white">
<tr><td valign="top">
<a href="../Photos/Islande/detifoos1.jpg">
<img border="0" vspace="10" hspace="10" src="../photogallery/photo12604/detifsos1.jpg"
width="100" height="54" title="Dettifoss" /></a>
</td></tr>
<tr valign="top"><td>
<font face="Tahoma"><FONT SIZE="1"></FONT><P>
<FONT SIZE="1">Dettifoss</FONT></P><FONT SIZE="1">
</FONT><FONT FACE="Times New Roman"></FONT></font></td></tr>
<tr valign="top"><td>
<font face="Tahoma"><FONT SIZE="2"></FONT><P><FONT SIZE="2">
La chute d'eau de Dettifoss est la plus puissante d'Europe</FONT></P><FONT SIZE="2">
</FONT><FONT FACE="Times New Roman"></FONT></font></td>
</tr>
</table>
</td>
```

TAB. 4.1 – Structure HTML du document

considérer.

Une extraction linéaire de ce document produit : « *Islande Dettifoss- Hveravellir Dettifoss La chute d'eau de Dettifoss est la plus puissante d'Europe...* ». Cette structure comporte de nombreux termes non reliés les uns aux autres et sur lesquels une analyse syntaxique produit un résultat difficilement exploitable par notre système de questions réponses.

L'organisation visuelle des documents peut indiquer certaines relations. Ainsi certains groupes de mots peuvent être reconnus comme une phrase même s'ils ne se terminent pas par un point ; c'est par exemple le cas classique d'un titre. L'utilisation d'expressions régulières a permis de reconnaître de tels cas et ainsi de marquer les fins de phrases.

Avec ces traitements, le document de l'exemple se transforme en :

« *Islande. Dettifoss- Hveravellir. Dettifoss. La chute d'eau de Dettifoss est la plus puissante d'Europe ...* ». Dans ce texte il est alors plus facile de reconnaître que Dettifoss est une réponse pertinente, à condition de considérer tout le passage pour valider les informations contenues dans la question et pas seulement la phrase la plus pertinente.

Des structures spéciales telles que les tableaux et les listes nécessitent des traitements particuliers. Dans le tableau 4.2, un rapprochement linéaire éloignerait « zone C » et « 13-04 29-04 ». Il serait donc difficile de connaître les dates des vacances de printemps de cette zone puisque la structure « *Vacances Zone A Zone B Zone C Hiver 9-02 25-02 2-02 18-02 16-02 4-03 Printemps 6-04 22-04 30-03 15-04 13-04-29-04* » serait considérée. La méthode mise en œuvre a consisté à reconnaître ce type de tableau

Vacances	**Zone A**	**Zone B**	**Zone C**
Hiver	9-02 25-02	2-02 18-02	16-02 4-03
Printemps	6-04 22-04	30-03 15-04	13-04 29-04

TAB. 4.2 – Exemple de tableau

afin de rapprocher les informations liées, ce qui permet d'obtenir la structure suivante : *Vacances ; Hiver / Zone A / 9-02 25-02. Vacances ; Hiver / Zone B / 2-02-18-02 (...) Vacances ; Printemps / Zone A / 6-04-22-04. Vacances ; Printemps / Zone B / 30-03-15-04 (...)*. Dans cette structure des mots sont répétés de manière à ce que chaque croisement corresponde à une phrase et ainsi la réponse 13-04 29-04 peut être obtenue en appliquant des critères d'extraction usuels.

Les documents venant du Web peuvent correspondre à différents types de formats. Les plus courants sont bien sûr les documents HTML mais il y a aussi des fichiers XML, DOC ou PDF. Afin de rendre ces documents utilisables par les systèmes de questions réponses, ils sont mis sous forme textuelle. Ainsi les fichiers XML sont transformés de manière linéaire et le contenu des autres fichiers est obtenu en utilisant leur structure. Les documents obtenus sont ensuite transformés par le système Kitten.

4.1.2 Analyse des questions

L'analyse des questions permet d'obtenir les informations utiles aux processus de recherche et de sélection des réponses. Ces informations sont celles reconnues par le système FRASQUES [Ferret et al. 2002], présenté en 1.1, intégrant une reformulation du focus effectuée par El Ayari [2009]. L'analyse des questions par décomposition, présentée en 3.2 n'a pas été réutilisée en l'état, mais l'analyse qui a été réécrite met en œuvre une notion de focus analogue à la notion d'hypothèse minimale, et une extraction des entités nommées est effectuée. Les informations retenues correspondent aux éléments suivants :

– les mots clés de la question utiles afin de rechercher les passages ;
– le focus : l'élément à propos duquel on demande une information. Avec la nouvelle formulation il correspond ainsi soit à l'entité sur laquelle porte la question soit à l'événement qu'elle contient. A la question « Quel est le poids de la tour Eiffel » le focus est « tour Eiffel ». A la question « *En quelle année le pont de Normandie a -t-il été inauguré ?* » le focus est « inaugurer ». La distinction est effectuée en tenant compte du verbe présent dans la question. S'il s'agit d'un verbe support, appartenant à une liste conçue manuellement, alors le focus est une entité, sinon il s'agit d'un événement. Les arguments du verbe présents dans la question sont également extraits en tant que modifieurs (pont de Normandie à la question précédente), ainsi que les modifieurs de l'entité ;
– le type d'entité nommée attendu (DATE à la question précédente) ;
– le type spécifique attendu, année à la question précédente.

A partir de ces informations, la catégorie de la question a été déduite. Elle représente le type de relations qui devra exister entre la réponse et le focus ou le type dans les documents : modifieur du nom, complément du verbe, complément circonstanciel... Comme le système s'applique uniquement sur les questions factuelles cinq catégories sont possibles :

– **instance** définit les questions ayant pour but de nommer une instance d'un certain type. Ces

questions sont très souvent sous forme affirmative (*Nommer un film de Steven Spielberg.*). Pour ce type de questions, la vérification du type spécifique peut être un critère très pertinent ;

– **combien** caractérise les questions attendant une expression numérique en retour (*Combien de fois Lance Armstrong a-t-il gagné le Tour de France ?*). Pour ce type de questions, une réponse peut être recherchée à l'aide de patrons d'extraction s'articulant autour de l'unité précisée dans la question ;

– **argument** définit les questions contenant un événement qu'il est nécessaire de vérifier. La réponse est un argument du verbe, le sujet ou l'objet. Ces questions peuvent attendre une entité nommée en réponse (*Qui chante "on n'a pas tous les jours 20 ans" ?*) ou non (*Quel vin boire avec des fruits de mer ?*) ;

– **complément circonstanciel** définit les questions attendant un lieu ou une date en réponse (*Où se situe la station Saint Michel ?*). Pour ce type de questions, la réponse est un complément circonstanciel ;

– **modifieur du nom** reconnaît les autres types de questions (*Quelle est la norme de hauteur sous plafond ?*).

L'analyse de la question de FRASQUES a été réécrite par Fleifel [2010] pour tenir compte de la redéfinition du focus et s'appuie sur une analyse syntaxique produite par XIP [Aït-Mokhtar, et al. 2002] et un ensemble de règles.

Par exemple à la question « *Quel vin boire avec des fruits de mer ?* », la réponse attend un vin en réponse, le focus et le verbe principal est « boire », la catégorie est donc « argument » et les mots clés « fruit », « mer », « fruit de mer », « vin » et « boire » sont reconnus.

4.1.3 Recherche de passages

Le module suivant recherche les passages censés contenir la réponse. Le système FRASQUES commençait par interroger un ensemble de documents préalablement découpés en passages puis analysait les passages pour en sélectionner les phrases les plus à même de contenir la réponse. Cela permettait d'obtenir de bons résultats sur des documents tels que des articles de journaux. Toutefois si on veut appliquer ce raisonnement sur des documents Web, on se heurte à la nature des phrases. En effet, vue leur définition dans les documents Web, davantage d'informations sont réparties sur plusieurs phrases. Une stratégie différente a donc été établie.

Cette nouvelle approche cherche directement les courts passages de texte desquels la réponse sera extraite sans passer par le niveau de phrases. Dans ce but, le moteur de recherche Lucene a été utilisé lors de l'indexation des documents et la recherche de passages ce qui permet de paramétrer la taille des documents renvoyés. Des études empiriques ont permis de voir que la taille optimale est de 300 caractères [Falco 2010a]. Le fait d'extraire des passages de taille fixe a comme inconvénient de tronquer les phrases. Pour contrer cela, les passages détectés sont agrandis en complétant la première et la dernière phrase. Afin de prendre en compte une plus grande diversité de documents, les mots sont tronqués à leur racine ce qui permet de trouver davantage de variantes d'un même mot. Dans le même but, une expansion de la requête a été faite en rajoutant certains synonymes des mots la constituant.

4.1.4 Sélection des passages

Les passages retournés par Lucene sont analysés par Fastr [Jacquemin 1996]. Il repère les termes de la question présents dans le passage ainsi que leurs variantes. Les termes peuvent être constitués d'un à plusieurs mots. Les variations peuvent être morphologiques, syntaxiques ou sémantiques. À chaque variante est associé un poids, d'autant plus fort que la variation est fiable. Ainsi une correspondance exacte aura un poids plus élevé qu'une dérivation morphologique, dont le poids est supérieur à celui obtenu par un synonyme. Les passages peuvent donc être pondérés et ordonnés et le poids du passage correspond à la somme des poids de l'ensemble des mots de la question. Cet ordonnancement de passage est celui effectué par les systèmes FRASQUES et QALC et présenté dans [Ferret, et al. 2001]. Une fois les passages ordonnés, ceux qui sont le plus à même de contenir la réponse correcte doivent se trouver dans les premières positions. Une étude a montré que les réponses se trouvaient très souvent parmi les 50 premiers passages et par conséquent seuls ceux-ci sont retenus. Cette sélection permet de n'effectuer l'extraction des réponses que sur les passages les plus pertinents, ce qui réduit le temps de traitement du système.

4.1.5 Annotation des passages

Afin d'identifier les informations portées par la question, les passages sont analysés et annotés. Dans un premier temps les entités nommées numériques sont marquées par le système de reconnaissance de FRASQUES. Puis l'analyseur XIP [Aït-Mokhtar et al. 2002] construit la représentation syntaxique de chaque phrase des différents passages et calcule ainsi ses relations syntaxiques ainsi que les syntagmes contenus dans la phrase. Comme le système est appliqué sur des documents provenant du Web où les relations syntaxiques trouvées sont peu fiables, seuls les syntagmes sont conservés. L'analyseur permet également de détecter les entités nommées non numériques contenues dans le passage.

Dans un second temps, les termes importants de la question sont marqués : le type spécifique, le focus et le verbe principal. Les entités nommées du type attendu par la réponse sont aussi indiquées. L'analyseur WMatch [Rosset, et al. 2008] est appliqué dans ce but. Il permet d'appliquer des expressions régulières sur des arbres. Voyons un exemple.

Question : Quelle était la date de sortie de Ratatouille au cinéma ?
Réponse : 1er août 2007
Passage : Ratatouille . Date de sortie du film au cinéma : le 1er août 2007.
Passage annoté : <GN> <FOCUS> Ratatouille|NN|ratatouille </FOCUS>
</GN> .|SENT|. <SURGN> <GN> <NOUN> Date|NN|date </NOUN> </GN>
<PREP> de|IN|de </PREP> <GN> <NOUN> sortie|NN|sortie </NOUN>
</GN> <PREP> du|IN|de </PREP> <GN> <NOUN> film|NN|film </NOUN>
</GN> <PREP> au|IN|à </PREP> <GN> <NOUN> cinéma|NN|cinéma
</NOUN> </GN> </SURGN> :|PUNCT| : <DET> le|DT|le </DET>
<EN-Réponse> <DATE> <ADJ> 1er|JJ|1er </ADJ> <NOUN> août|NN|août
</NOUN> <NUM> 2007|CD|2007 </NUM> </DATE> </EN-Réponse>

Dans cet exemple, le focus est marqué, ainsi que la date et la catégorie morphosyntaxique de chaque terme. Nous voyons également que les groupes nominaux sont rassemblés ce qui permet

de créer le SURGN qui correspond à des groupes nominaux complexes. La réponse candidate est indiquée par la balise EN-Réponse.

4.2 Extraction de réponses

4.2.1 Extraction de réponses candidates

A cette étape nous disposons d'un ensemble de passages dont certains contiennent vraisemblablement la bonne réponse. Il reste encore à l'extraire et la placer en première position. Le système QAVAL traite de questions factuelles. Ce type de questions attend en réponse soit une entité nommée, par exemple calendaire, soit un groupe nominal. Le module d'analyse des questions permet de distinguer ces deux cas ainsi que le type d'entité nommée attendu le cas échéant.

L'extraction des réponses a généralement pour but d'extraire les réponses les plus appropriées. Pour cela, les différents systèmes s'appuient sur des critères de type proximité des termes de la question, des critères syntaxiques (patrons d'extraction ou relations) et reconnaissance du type d'entité nommée. Ainsi le système FRASQUES utilise des patrons syntaxiques pour certaines questions tandis que pour les autres il extrait la réponse du bon type d'entité nommée la plus proche des mots de la question. Ces traitements sont efficaces sur des articles de journaux, toutefois certaines réponses correctes peuvent ne pas être détectées. De plus, le système QAVAL doit pouvoir s'appliquer également sur les documents Web avec un découpage en phrases particulier, sur lesquels les patrons d'extraction s'appliquent moins bien que sur les articles de journaux.

Notre approche, dans QAVAL, est volontairement laxiste afin de laisser la décision au système de validation. L'extraction des réponses dépend du fait que la réponse attende une entité nommée en réponse ou non.

– Si la question attend une réponse d'un type particulier (« *Qui est le président des États Unis ?* » attend une personne en réponse) alors toutes les entités nommées du type attendu contenues dans le passage sont extraites ainsi que les noms propres puisque ce sont les seules réponses pouvant être pertinentes. Ces entités sont indiquées lors de l'annotation des passages et donc il ne reste plus qu'à les extraire. Dans l'exemple concernant le film Ratatouille, la date a été marquée comme une entité nommée du type attendu par la réponse.

– Si la question n'attend pas une réponse d'un type d'entité nommée particulier (« *Citer un succès de Michael Jackson.* ») alors la réponse est un groupe nominal. Ce groupe est présent dans les passages justificatifs. Tous les groupes nominaux contenus dans les passages sont extraits. Ces groupes sont reconnus et marqués par l'analyseur XIP lors de l'analyse des passages. Deux types de groupes nominaux sont marqués. Les groupes de petite taille qui correspondent généralement à un seul nom éventuellement accompagné d'un adjectif et les groupes plus longs qui les combinent. Les groupes nominaux : « Ratatouille », « Date de sortie du film au cinéma », « Date »,« sortie », « film » et « cinéma » pourraient être extraits du passage précédent si la question n'attendait pas un certain type d'entité nommée en réponse.

Dans les deux cas, des patrons d'extraction sont également utilisés. Ils dépendent de la catégorie de la question. Toutefois, ces patrons sont encore à améliorer et ne constituent pas à l'heure actuelle un critère très fiable en dehors de ceux qui s'appliquent pour les questions de catégorie “combien”. A

chaque réponse extraite est attribué un score qui tient compte du score de la phrase et permet d'évaluer la pertinence de la réponse.

4.2.2 Filtre des réponses

Après avoir extrait les réponses candidates, il reste à les ordonner afin que la réponse correcte se trouve en première position. Pour cela, nous appliquons une version modifiée de notre module de validation de réponses.

Ravichandran et al. [2003] montrent que la validation de réponses et l'ordonnancement de réponses sont deux tâches différentes. La validation de réponses vérifie que la réponse est valide en renvoyant une valeur booléenne et l'ordonnancement vise à classer les différentes réponses en appliquant une valeur de confiance à chacune d'elle. Ces différentes présentations sont détaillées en section 1.2.

Dans le cadre du système QAVAL, le but recherché est bien sûr de placer la bonne réponse en première position mais néanmoins il est préférable d'avoir un système qui propose la bonne réponse parmi les 3 ou 5 premières plutôt qu'un système qui se trompe plus souvent en ne sélectionnant qu'une réponse par question. Nous transformons ainsi notre système de validation de réponses en un système d'ordonnancement pour lesquelles les valeurs de confiance retournées sont comprises entre -1 et 1. Une valeur de -1 indique que la réponse est non valide. Une valeur de 1 indique que la réponse est valide.

Notre système d'ordonnancement se déroule en deux étapes. Tout d'abord un ensemble de réponses sont identifiées comme non valides puis les réponses sont ordonnées par le classifieur. L'étape de filtre vise à réduire le nombre de réponses candidates, ce qui permet d'améliorer le temps de traitement des différentes réponses. De plus, comme les filtres sont très fiables, les performances du système ne sont pas affectées (cf. section 4.3.3). Les réponses filtrées, car ne correspondant pas à la réponse de manière triviale, sont déclassées directement avec un poids de -2. Les réponses ne sont pas directement éliminées car dans un cas hypothétique où toutes les réponses seraient déclassées, la bonne réponse pourrait figurer dans cet ensemble.

La première vérification s'assure que la réponse n'apparaît pas entièrement dans la question. Ce cas est assez fréquent car pour certaines questions tous les groupes nominaux présents dans les passages sont extraits. Or, pour une question contenant un groupe nominal, un passage pertinent devra contenir ce groupe et par conséquent il est reconnu comme une réponse potentielle. Par exemple la question « *Citer un succès de Michael Jackson.* » contient le groupe nominal « Michael Jackson » qui n'est clairement pas la réponse appropriée.

L'information de la date contenue dans la question est l'un des éléments clés qui doit être vérifié pour décider de la validité de la réponse au même titre que la vérification du type spécifique ou celle de l'action de la question. Lors de la validation de réponses, la vérification de la date est effectuée en s'assurant que celle-ci est contenue dans le passage justificatif et en considérant le triplet comme non valide si ce n'est pas le cas. Lors de l'ordonnancement, cette vérification peut s'appliquer également. De plus, si la date est mentionnée d'une manière différente dans un passage alors il peut y avoir un autre passage dans lequel elle se trouvera. La même vérification a été effectuée sur toutes les entités nommées contenues dans la question. Les dates, les lieux et les personnes de la question sont

ainsi considérés. Ces vérifications servent de filtre permettant de déclasser certaines réponses et ne constituent pas un critère car elles sont fiables, ce qui permet de diminuer le nombre de réponses.

La vérification de la présence du lieu de la question dans le passage est une première indication sur le fait que l'action a effectivement lieu à l'endroit voulu, ce qui correspond à une des informations à vérifier. Ici encore il se peut que le lieu ne se trouve pas dans le passage mais vu le nombre de passages extraits, plusieurs contiennent souvent la bonne réponse, et on peut être à peu près sûr que certains le contiennent.

Les personnes présentes dans les questions sont souvent celles effectuant ou ayant subi l'action de la question donc ayant un rôle très important dans l'événement. Une personne possède peu de variantes dans les documents, à part l'oubli du prénom ou l'appel d'une personne par sa fonction comme « le président » pour « Nicolas Sarkozy » ou un pronom personnel. Comme ces termes sont généralement proches dans le document pour permettre un usage anaphorique, il doit exister un passage les contenant ce qui reflète bien l'intérêt de travailler sur les passages et non sur une phrase. Le fait de vérifier que la personne se trouve dans le passage est un premier signe indiquant que l'information principale de la question est mentionnée dans le document.

A la question « *Quel pays l'Irak a-t-il envahi en 1990 ?* » un bon passage doit contenir « Irak » et « 1990 ». En effet, s'il ne contient pas le premier terme on ne peut pas savoir quelle invasion est mentionnée. Si la date ne se trouve pas dans le passage alors il pourrait bien, par exemple, parler de l'invasion de l'an 586. En revanche, le type spécifique « pays » peut ne pas se trouver dans le passage et le verbe « envahir » peut se trouver sous une autre forme comme « annexion ».

A ce point, les réponses issues de ces vérifications sont des candidates possibles, car leur type d'entité nommée correspond à celui que la question attend en retour ce qui ne signifie pas que nous soyons sûrs d'avoir extrait la réponse correcte parmi cet ensemble. Le passage a de bonnes chances de traiter de l'entité ou de l'événement mentionné dans la question. De plus, la date et le lieu contenus dans la question se retrouvent dans le passage ce qui constitue deux vérifications importantes comme nous l'avons montré lors de la décomposition de question, en section 3.2.

4.2.3 Ordonnancement des réponses par apprentissage

Les réponses pour lesquelles aucun score de confiance n'a encore été fourni sont évaluées par notre système de validation de réponses. Il reprend les critères suivants :

– la présence des termes de la question dans le passage ;
– la proximité des mots par calcul de la plus longue sous chaîne commune au passage et à l'hypothèse (question + réponse) ;
– la correspondance entre la réponse et le type spécifique attendu par la question.

Le problème n'est plus de détecter la valeur de validité d'un ensemble de réponses extraites par différents systèmes de questions réponses, mais d'indiquer un score à chaque réponse permettant de discerner la plus appropriée. Pour ce faire, nous avons gardé les critères de validation auxquels nous en avons ajouté de nouveaux pour pallier le fait que les réponses et les passages ne sont plus issus d'un système de questions réponses, qui les avaient donc sélectionnés par l'application d'un certain nombre de critères.

4.2.3.1 Rang du passage

Une des étapes ayant servi à obtenir les passages a consisté à les ordonner afin de sélectionner les 50 meilleurs. Il est naturel de tenir compte du rang du passage obtenu comme un critère. Logiquement plus le rang du passage est élevé et plus il a de chances de contenir la réponse valide ou au moins de traiter des mêmes informations que la question. De plus, des études ont montré que lorsque la réponse est dans un passage, alors celui-ci est souvent dans les vingt premières positions. Ce critère porte donc sur les passages et permet de caractériser les plus prometteurs. Il vient en complément des critères sur les termes. Les autres critères ajoutés portent sur la réponse et visent à permettre ainsi de distinguer la bonne réponse parmi plusieurs extraites du même passage.

4.2.3.2 Mesure de densité

Le système de questions réponses FRASQUES extrait certaines réponses en sélectionnant l'entité nommée la plus proche des mots de la question. La proximité des termes de la question par rapport à la réponse est un critère classique des systèmes de questions réponses. Ce critère est calculé ici par la moyenne des distances séparant chaque mot de la question et la réponse dans le passage justificatif. La distance entre la réponse et un mot de la question est calculée en comptabilisant le nombre de mots non vides séparant ces deux termes. Si un mot de la question est absent du passage justificatif, alors la distance sera considérée comme égale à la taille, en nombre de mots non vides, du passage. Le score total associé à une réponse correspond à la moyenne des distances ainsi calculées sur l'ensemble des mots non vides de la question. Ainsi ce critère permet de tenir compte aussi bien de la proximité des mots de la question que de la proportion de termes non vides de la question absents du passage justificatif et de la proximité de la réponse par rapport aux mots de la question. La formule suivante présente plus formellement la valeur du critère pour une réponse *rep* et un passage *p* :

$$score(rep, p) = \frac{\sum_{i=1}^{|M|}(distance(rep, M_i)) + N * taille(p)}{|M+N|}$$

avec M l'ensemble des mots non vides de la question présents dans le passage et N l'ensemble des mots non vides absents

L'utilisation d'une mesure de densité est un critère assez fréquent. Par exemple Gillard et al. [2006] utilise une mesure de compacité tenant également compte de la distance des mots de la question entre eux et a ainsi montré que la densité était un critère pertinent pour la sélection de réponses et le filtrage des passages.

4.2.3.3 Redondance de la réponse

Un critère souvent considéré dans les systèmes de questions réponses consiste à comptabiliser le nombre de fois où la même réponse a été extraite. L'idée est que plus une réponse a été extraite plus elle semble pertinente. Généralement, cette notion de redondance porte sur plusieurs documents.

Dans notre cas, la réponse est extraite depuis chaque passage dans lequel elle se trouve. Ce critère indique donc la proportion de documents dans laquelle la réponse se trouve. Comme une réponse peut se trouver, dans certaines situations, plusieurs fois dans le même passage, le critère indique aussi cette redondance.

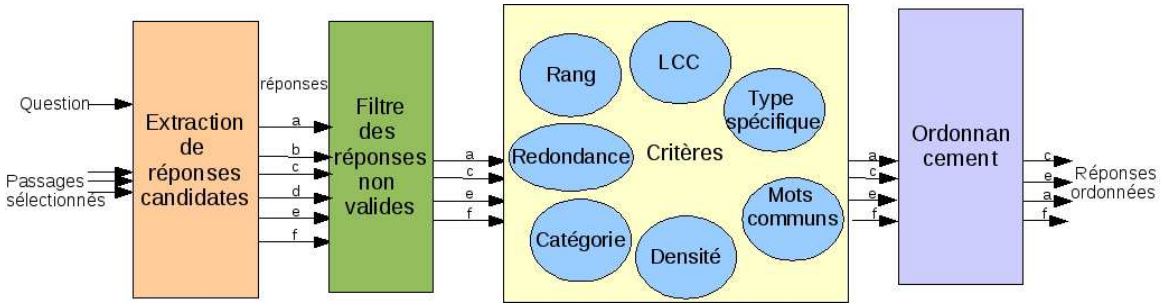

FIG. 4.3 – Extraction de réponses du système QAVAL

4.2.3.4 Catégorie de la question

La stratégie d'extraction des réponses dépend de la catégorie de la question. La catégorie de la question sert à indiquer le lien qui doit se trouver dans le document entre la réponse et le focus ou le type spécifique. Dans notre système l'ajout d'un critère tenant compte de la catégorie de la question doit permettre de différencier la combinaison des critères d'une catégorie à une autre. Les différentes catégories ont été présentées en 4.1.2.

D'autres critères auraient pu être ajoutés en liaison avec la catégorie : la distance de la réponse avec le focus ou le type. Un autre critère pourrait tenir compte du fait qu'un patron d'extraction puisse s'appliquer ainsi que son degré de confiance. Mais, comme notre étude a porté sur les documents Web, les patrons s'appliquent rarement.

4.2.3.5 Ordonnancement

Après avoir calculé l'ensemble des critères, il reste à les combiner. La figure 4.3 récapitule l'extraction des réponses du système QAVAL avec les différents modules. Le système de validation de réponses utilisait la combinaison bagging d'arbres de décision fournie par WEKA afin de classer les réponses en valides ou non. Ici cette information n'est pas utile puisqu'il s'agit d'ordonner les réponses les unes vis à vis des autres. Dans ce but, au lieu d'apprendre une valeur de validité, un score de confiance est appris. A l'aide de ces valeurs, il est alors possible d'ordonner les réponses de manière à ce que la plus fiable soit en première position. Différents tests présentés en 4.3 ont permis de voir que la combinaison d'arbres de décision était la méthode obtenant les meilleurs résultats.

Comme une approche par apprentissage est suivie, il est nécessaire d'avoir des bases d'apprentissage. La section 4.3 présente plus en détail les deux bases utilisées. Afin de créer les bases, le système recherche les réponses à différentes questions. Comme les réponses correctes à ces questions sont connues, il est possible de déterminer automatiquement les réponses découvertes qui semblent valides. Une évaluation manuelle de ces réponses permet ensuite de déterminer celles qui le sont réellement. Le fait de suivre une telle approche permet d'être au plus près des données testées. Toutefois, chaque fois qu'un changement a lieu dans le système avant l'ordonnancement des réponses, de nouvelles réponses valides peuvent être extraites. Il est donc nécessaire d'analyser à chaque changement les réponses obtenues par le système. Afin de ne pas avoir trop de travail à effectuer, les réponses déjà

évaluées sont mises en mémoire. Ainsi chaque réponse correcte extraite est comparée à une réponse évaluée puis si elle n'est pas déjà connue, elle est présentée à un des créateurs du système qui l'évalue.

4.2.3.6 Implémentation des critères

Le module d'ordonnancement des réponses consiste donc à combiner différents critères par un mécanisme d'apprentissage. Il semble intéressant de revenir sur l'implémentation du calcul des critères.

Comme l'ordonnancement s'applique sur de très nombreuses réponses, il est nécessaire que le temps passé pour connaître la valeur des différents critères soit court, ce qui est le cas pour l'ensemble des critères sauf la validation du type de la réponse. Pour les critères traitant de l'apparition des mots de la question dans le passage, un parcours de la question permet d'extraire les mots et leur catégorie puis un parcours du passage permet d'extraire la valeur des différents critères. Le critère portant sur la densité des mots de la question autour de la réponse est lui aussi assez rapide. Dans un premier temps un parcours du passage permet de connaître la position des mots de la question et la position de la réponse. Puis il suffit ensuite de comparer ces positions pour obtenir la valeur du critère. Au final ce critère se calcule encore en fonction de la taille du passage et peut être calculé en même temps que les précédents. La présentation du critère LCC, en 3.3.3, a montré qu'il était également calculé en un temps linéaire fonction de la taille du passage.

Le critère le plus long à calculer est la validation du type de la réponse car elle utilise de nombreuses recherches dans des documents par le moteur de recherche Lucene. Afin de rendre le calcul de ce critère plus rapide, une base de connaissances est créée à partir des résultats déjà obtenus, avec l'idée qu'un résultat déjà calculé peut être utile pour un autre couple réponse-type spécifique. En effet, de nombreuses questions ont le même type, par exemple « année », et peuvent avoir les mêmes réponses.

Un des critères utilisé pour la vérification du type de la réponse consiste à rechercher le type dans la page Wikipédia associée à la réponse. Afin de se passer de l'utilisation d'un moteur de recherche, une table de hachage a été construite avec comme clé le titre de la page et comme valeur l'adresse de la page.

Comme la vérification du type est utilisée comme un critère la valeur produite n'est plus une valeur booléenne mais un score indiquant la confiance qu'a le classifieur dans le fait que la réponse corresponde au type, ce qui permet d'avoir une valeur plus précise.

4.3 Expérimentation

4.3.1 Les données de travail

Le système QAVAL doit pouvoir s'appliquer aux documents provenant du Web mais en donnant aussi de bons résultats sur des articles de journaux. Il a donc été évalué sur les deux types de collections.

Commençons par définir le corpus Web. Il correspond aux données d'évaluation des systèmes de questions réponses proposé dans le cadre du projet Quæro [Quintard et al. 2010]. Le corpus a été

constitué automatiquement par Exalead[4]. Il correspond aux URLs des documents Web. Les documents ont été collectés entre mai et juin 2008 en sélectionnant les pages visités par les utilisateurs du moteur d'Exalead. Ainsi, lorsqu'un visiteur fournit une requête au moteur de recherche, celui-ci lui retourne un ensemble de documents qui ont été retenus pour construire un corpus d'à peu près deux millions de pages. L' utilisation de ce corpus est effectuée en deux temps. Tout d'abord, un ensemble de 500 000 documents a été extrait pour permettre de mettre au point les systèmes, puis, un passage à l'échelle sera effectué sur la base complète.

Lors de la campagne 2010, les questions ont été créées à partir des requêtes des utilisateurs, sans examiner les documents, et correspondent à l'information cherchée par l'utilisateur. Trois types de questions sont présentes :

– les questions écrites directement à partir des requêtes ;

– les questions orales qui correspondent à une retranscription de demandes orales et peuvent donc contenir des erreurs (« *le chemin de fer est plus rapide , mais plus rapide que quoi ?* ») venant de l'aspect spontané des questions ;

– les questions orales réécrites de manière à être davantage correctes (« *Quels sont les moyens de transport moins rapides que le chemin de fer ?* »).

Dans un premier temps, nous nous focaliserons uniquement sur les questions de la première catégorie.

Afin de tester aussi sur des articles de journaux, les collections provenant des campagnes d'évaluation QA@CLEF 2006 [Magnini et al. 2006] sont utilisées. Elles correspondent aux articles du journal Le Monde et aux dépêches de l'agence ATS. Le corpus contient en tout 150 000 documents.

Pour évaluer automatiquement notre système, il est nécessaire de disposer des réponses attendues. Les questions viennent de différentes campagnes d'évaluation ce qui permet d'être au plus près des évaluations classiques mais aussi de connaître les réponses potentielles à l'avance et d'être sûr que les réponses peuvent se trouver dans les documents. Pour l'évaluation sur les documents venant du Web, 147 questions provenant de la campagne Quæro 2010 sont utilisées. Pour celles sur les documents journalistiques, 126 questions provenant de la campagne EQueR [Grau 2005] sont considérées.

Comme l'ordonnancement de réponses suit une approche par apprentissage, il est nécessaire de constituer des bases d'apprentissage. Deux bases ont été créées. La première concerne les articles de journaux et la seconde les documents Web. Les corpus sont créés de manière semi automatique en utilisant QAVAL. Un ensemble de questions provenant des campagnes QA@CLEF2005 et QA@CLEF2006 est fourni au système qui en extrait des réponses dans les différents types de documents.

Au final deux corpus ont été créés. Le premier, pour les documents Web, contient 1 047 réponses parmi lesquelles 349 sont valides. Le second est composé de 2 850 réponses dont 950 sont valides. Notons que le nombre de réponses non valides a été diminué afin de ne pas en avoir trop et correspond ainsi à $\frac{2}{3}$ des données. Sans cela il y aurait 53 781 réponses non valides pour la recherche dans les documents journalistiques ce qui est beaucoup trop élevé (98,26 %).

Le MRR (*Mean Reciprocal Rank*) sur les cinq premiers rangs est calculé afin d'évaluer le système ainsi que la proportion de réponses correctes se trouvant soit au premier rang soit dans les cinq pre-

[4]http ://www.exalead.com/search/

	MRR	Premier rang %(#)	Cinq premiers rangs %(#)
Quæro	**0,43**	**35% (51)**	**56% (82)**
Baseline Quæro	0,30	22% (32)	43,5% (64)
EQueR	**0,47**	**39% (49)**	**60% (76)**
Baseline EQueR	034	27% (34)	47% (59)

TAB. 4.3 – résultats de QAVAL

miers. Afin de détecter les réponses valides, la réponse proposée est comparée à une réponse connue comme correcte. Puis, une évaluation manuelle permet de valider les réponses.

4.3.2 Évaluation globale de QAVAL

Afin d'évaluer notre système d'ordonnancement, nous avons comparé les résultats de QAVAL avec ceux obtenus en utilisant un ordonnancement des réponses simple. La réponse la plus proche des mots de la question est extraite depuis les cinq premiers passages pouvant la contenir. Les réponses sont ordonnées suivant l'ordre des passages. Cet ordonnancement correspond à celui effectué par le système FRASQUES dans le cas où une entité nommée est attendue en réponse.

Le tableau 4.3 montre que des résultats assez satisfaisants sont obtenus puisque le MRR augmente de 48 % (0,43 vs 0,30) pour la recherche sur les documents Web et de 38 % pour la recherche dans les autres documents. Nous pouvons aussi voir que la méthode est robuste puisqu'elle obtient des résultats proches pour les deux recherches (MRR 0,43 et 0,47) avec des résultats légèrement supérieurs pour la recherche dans les articles de journaux.

Afin de mieux évaluer les résultats portant sur les articles issus du Web, nous pouvons les comparer à ceux obtenus par les trois autres systèmes ayant participé à la campagne d'évaluation QUAERO 2010. Lors de cette campagne, le MRR est calculé sur les trois meilleures réponses. Sur ces questions, le meilleur résultat est de 0,71, le second de 0,46 et le dernier de 0,40. Le MRR de notre méthode est de 0,41, ce qui est inférieur aux résultats du meilleur système mais comparable aux résultats des autres.

Les résultats sont donc plutôt satisfaisants mais peuvent être améliorés puisque seules 56 % des questions ont une réponse correcte parmi les cinq premières renvoyées. Il est donc nécessaire de savoir ce qui a entraîné ces erreurs. Dans ce but, la figure 4.4 évalue chaque étape (extraction des documents, sélection des documents, extraction des réponses, validation des réponses). Pour chacune d'elle, la figure présente le nombre de questions pour lesquelles une réponse peut être fournie avant cette étape ainsi que le nombre d'erreurs dues à ce module. Par exemple pour l'extraction des réponses, nous pouvons voir que sur les 147 questions de départ 125 ont au moins un passage pertinent et parmi celles-ci seules 108 ont une réponse correcte extraite.

Cela montre qu'une assez bonne détection des documents a été effectuée puisque 85 % des questions possèdent un passage contenant la bonne réponse parmi les passages sélectionnés mais il y a encore des possibilités d'améliorations. Le module le moins efficace est sans doute l'ordonnancement de réponses qui effectue seulement 76 % de bons ordonnancements. Ces résultats peuvent s'expliquer

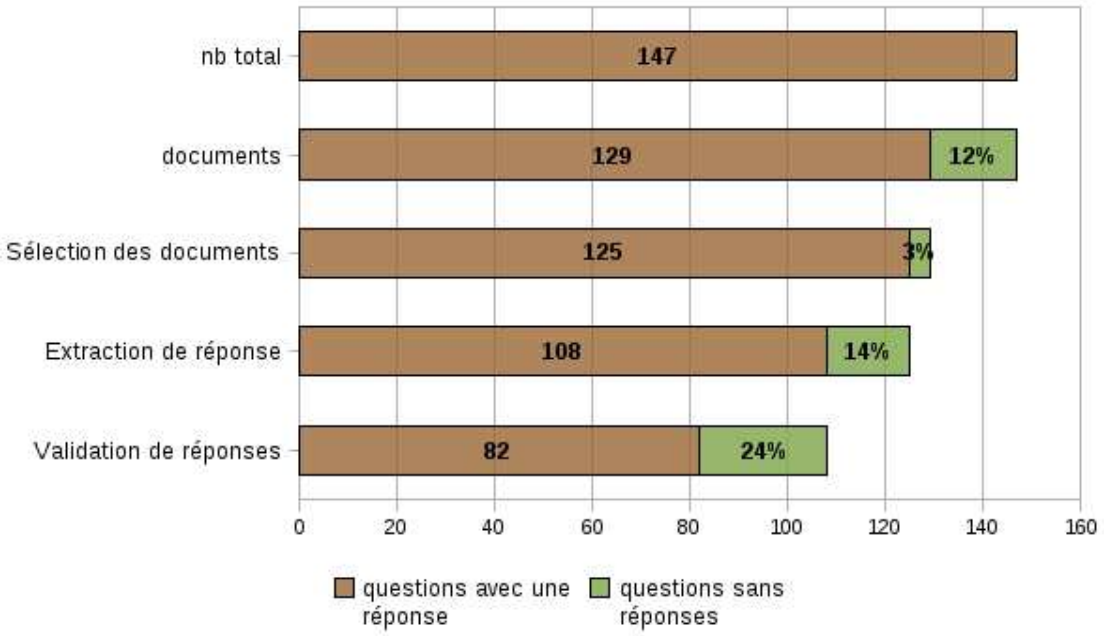

FIG. 4.4 – Répartition des erreurs

de par le grand nombre de réponses à évaluer. L'évaluation est assez laxiste au niveau de l'ensemble des étapes sauf de l'ordonnancement de réponses puisqu'une réponse est considérée comme valide si elle est correcte alors que pour l'ordonnancement il est aussi nécessaire qu'elle soit justifiée. Par exemple, un document est considéré comme valide s'il contient la réponse et ce quel que soit son contenu. L'évaluation de l'ordonnancement est présentée plus en détail en 4.3.3.

Dans [Grappy, et al. 2011] nous avons présenté en détail une évaluation du système de prétraitement de documents. Ces résultats sont donnés ici à titre indicatif. Kitten est comparé à une baseline qui extrait directement le contenu textuel en utilisant des tags ainsi qu'au logiciel BoilerPipe [Kohlscëtter, et al. 2010]. Ces trois systèmes sont évalués à l'aide de la proportion de passages sélectionnés contenant une bonne réponse et du MRR calculé sur les réponses finales (cf. tableau 4.4)

Système	# Passage correct	MRR
Baseline	114 (77 %)	0,28
BoilerPipe	121 (82 %)	0,32
Kitten	**130 (88 %)**	**0,43**

TAB. 4.4 – Évaluation du prétraitement des documents

Les résultats obtenus par Kitten sont meilleurs que ceux obtenus par la baseline (+11 % pour la

détection de passages). Cela indique que les prétraitements spécifiques sont pertinents. La différence est particulièrement marquante au niveau du MRR avec une amélioration de 53 % ce qui montre l'intérêt d'effectuer ces prétraitements pour les différents modules. De plus, de meilleurs résultats sont obtenus par notre méthode que par le logiciel BoilerPipe ce qui peut s'expliquer par le fait que ce dernier ait été créé dans un but différent, la classification de documents.

Afin de vérifier que la taille du passage est suffisamment grande, des évaluations ont été effectuées sur des passages de 600 caractères et de 1 000 caractères. Ces évaluations n'ont pas permis de voir de changements significatifs puisque sur l'ensemble le plus grand, seules 3 questions ont une réponse supplémentaire contenue dans les documents détectés. De plus, en utilisant une taille de passages plus élevée, davantage de réponses sont extraites, ce qui entraîne plus de temps de traitement, et risque de détériorer les résultats de l'ordonnancement puisque davantage de réponses incorrectes sont extraites.

Les évaluations précédentes, provenant de la campagne QUAERO ont donc été effectuées sur les questions correctement écrites, afin de tester la robustesse de notre système, des tests ont été réalisés sur les 42 questions factuelles présentes sur l'oral et les 43 questions réécrites (cf. tableau 4.5). Le système rencontre davantage de problèmes sur les questions orales que sur les questions correctement écrites puisque le MRR est deux fois moins élevé. Toutefois, tous les systèmes participant à la campagne accusent une baisse significative, le meilleur système n'a un MRR que de 0,36 et le moins bon qui s'appuie essentiellement sur la syntaxe des questions obtient 0,09. Les questions réécrites sont également plus difficiles que les questions dites classiques car leur formulation est plus éloignée de celle présente dans les documents.

	MRR	Premier rang	Cinq premiers rangs
Questions orales	0,24	20%	25%
Questions réécrites	0,30	23 %	40 %
Questions classiques	0,43	35 %	57 %

TAB. 4.5 – Résultats de QAVAL sur les questions orales et réécrites

4.3.3 Évaluation de l'ordonnancement

Le système QAVAL obtient donc d'assez bons résultats globalement, mais l'ordonnancement de réponses effectue 24 % d'erreurs, ce sont les questions dont la bonne réponse a été extraite mais n'est pas placée parmi les cinq premières. Cette sous section s'intéresse plus particulièrement à cet ordonnancement. Tout d'abord il est à noter que de très nombreuses réponses sont à ordonner. Ainsi, sur les données Quæro, 11 405 réponses sont candidates ce qui correspond à une moyenne de 77 réponses par question.

Avant l'ordonnancement de réponses, une erreur a été commise pour 26 % des questions et il n'est donc plus possible d'obtenir une réponse valide pour celles-ci. Une nouvelle évaluation porte sur les questions pour lesquelles une réponse correcte peut être ordonnée. Ainsi, cet évaluation est effectuée que sur 108 questions provenant de la campagne du projet Quæro au lieu de 147. Comme pour l'évaluation précédente nous avons évalué l'ordonnancement et la baseline sur les deux collections (cf.

tableau 4.6) Ainsi, de meilleurs résultats sont obtenus par notre méthode que par l'extraction simple avec une amélioration de 45 % de MRR sur les questions Quæro et de 36 % pour EQueR.

	MRR	Premier rang	Cinq premiers rangs
Quæro (108 questions)	**0,58**	**46%**	**76%**
Baseline Quæro	0,40	30%	59%
EQueR (96 questions)	**0,61**	**50%**	**78%**
Baseline EQueR	0,45	35%	61%

TAB. 4.6 – Évaluation de l'ordonnancement des réponses

	MRR	Premier rang	Cinq premiers rangs
Avant	0,38	28 %	54 %
Après	0,53	46 %	62 %

TAB. 4.7 – Évaluation de l'ordonnancement des passages

Afin d'évaluer complètement le module d'ordonnancement de réponses, des mesures ont été effectués sur les passages puisqu'en ordonnant les réponses le système ordonne également les passages. La première mesure étudie l'impact de cet ordonnancement sur les questions Quæro en calculant les différentes mesures avant et après l'ordonnancement des réponses (cf. tableau 4.7). Dans ce cas l'évaluation n'est plus faite sur la réponse mais sur les passages et plus particulièrement sur le fait qu'il contienne la bonne réponse.

Les résultats sur les passages sont effectivement bien meilleurs après l'ordonnancement de réponses avec une amélioration du MRR de près de 40 % ce qui est dû à la forte proportion de passages corrects placés en première position, 46 %. Ce nombre est à comparer avec la proportion de bonnes réponses en première position détectées par QAVAL, 34 %. La différence permet de constater qu'un grand nombre de passages en première position possède la réponse correcte, mais, qu'une autre réponse contenue dans ce passage a été considérée comme valide à tort. Il faudrait donc rajouter des vérifications portant sur la réponse. Ces valeurs nous permettent également d'évaluer le rang du passage. Ainsi, dans 54 % des cas la réponse se trouve parmi les cinq premiers passages. Dans 64 % elle se trouve parmi les dix premiers passages et 71 % se trouvent dans les 20 premiers.

4.3.3.1 Évaluation de l'extraction des réponses

La figure 4.4 a montré que l'étape d'extraction des réponses effectue 14 % d'erreurs. Les erreurs se rencontrent que la question attende une réponse d'un certain type d'entité nommée ou non. Un des premiers problèmes vient de la reconnaissance des entités nommées soit lors de l'analyse des passages soit lors de l'analyse des questions, par exemple pour « *Quelle est la norme de hauteur sous plafond ?* » une hauteur est attendue mais le système ne le reconnaît pas. Quand la question n'attend pas de type d'entité nommée particulier, le système effectue 23 % d'erreurs qui sont dues au fait que la réponse ne correspond pas à un groupe nominal et peut soit être plus grand « *Quelle est l'adresse*

de la clinique de la sauvegarde à Lyon ? » soit être plus court « *Par quelle lettre commence le nom des chiens nés en 2008 ?* ».

L'extraction de réponses dépend du type de question à savoir si elle attend ou non une entité nommée en réponse. Cette étude porte sur cette distinction. Les questions attendant une entité nommée en réponse obtiennent un MRR de 0,52 alors que les autres ne sont qu'à 0,25. Cela peut s'expliquer car pour les questions n'attendant pas d'entité nommée en retour beaucoup plus de réponses sont extraites. Il faudrait donc un moyen permettant de limiter le nombre de réponses extraites ou de mieux caractériser la réponse en tenant par exemple compte de la distance au focus ou des relations syntaxiques avec le focus.

Un test similaire s'applique pour les questions attendant une entité nommée en réponse et correspondant à un groupe nominal : les personnes et les organisations. Pour ces questions, ce ne sont plus les entités nommées mais les groupes nominaux présents dans les passages qui sont extraits. Le test a été effectué sur les questions de la campagne EQueR et a montré que le MRR originellement de 0,47 descend à 0,39.

Avant d'être ordonnées, des réponses jugées incorrectes sont déclassées. Sans ce filtre, le MRR des questions EQueR est alors de 0,39 alors qu'il est de 0,47 en l'utilisant ce qui témoigne de son intérêt. Sur les 15 434 réponses correspondant à cette campagne, 4 514 (29 %) ont été filtrées et parmi elles seules 15 (0,33 %) étaient valides ce qui témoigne de l'efficacité du filtre.

4.3.3.2 Étude du classifieur

Afin de s'assurer que la combinaison d'arbre de décision est bien le meilleur classifieur, différents tests ont été menés. Les premiers consistent à utiliser différents classifieurs pour l'ordonnancement des réponses Quæro. Les classifieurs étudiés sont : un arbre de décision seul, un réseau de neurones, un SVM, un système bayésien et un réseau bayésien. Tous ces classifieurs sont proposés par WEKA (cf. tableau 4.8).

	MRR	Premier rang	Cinq premier rangs
Bagging	**0,43**	**34%**	**56%**
Arbre de décision (J48)	0,35	25 %	49 %
Réseau de neurones	0,35	27 %	48 %
SVM	0,25	17 %	36 %
Bayésien naïf	0,37	25 %	49 %
Réseau bayésien	**0,43**	**34%**	**57%**

TAB. 4.8 – Différents classifieurs

Le bagging et le réseau bayésien sont les systèmes obtenant les meilleurs résultats sans doute grâce à leur combinaison de différents classifieurs ce qui permet d'avoir des scores de confiance fins. Le SVM par exemple ne fournit pour la plupart des cas que deux valeurs (la réponse est justifiée et la réponse n'est pas justifiée) ce qui ne permet pas de classer les réponses et entraîne de plus faibles résultats.

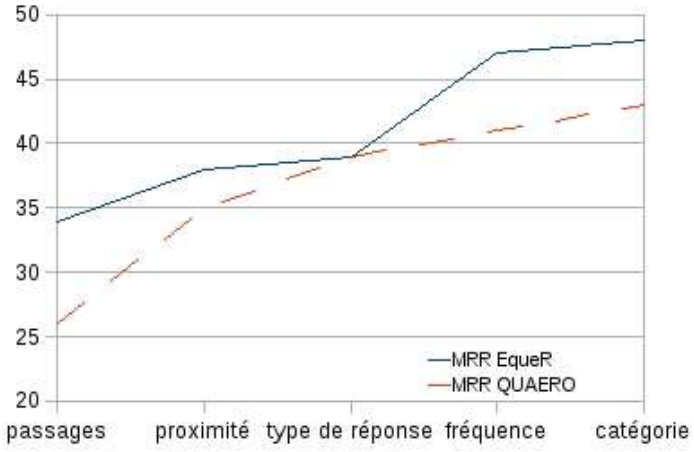

FIG. 4.5 – Importance des critères

Un test additionel a consisté à utiliser le système SVMRANK [Joachims 2009]. Celui-ci utilise un SVM afin d'ordonner des données. Un SVM consiste à créer une marge visant à séparer les données. Pour cela les données de la base d'apprentissage sont placées dans un espace composé de nombreuses dimensions. Le test a été effectué sur les données QUAERO et a permis d'obtenir des résultats moyens (MRR 0,34) et inférieurs à ceux obtenus par le bagging. Cela est probablement dû au fait que la plupart des critères ne s'appliquent pas à toutes les questions et qu'une question peut donner lieu à peu de critères.

4.3.3.3 Importance des critères

L'ordonnancement des réponses est effectué par une combinaison de critères par apprentissage. Il est donc intéressant d'étudier de plus près cette combinaison. La combinaison étant effectuée par un ensemble d'arbres de décision, il est possible de distinguer l'ordre des critères. Il est à noter que tous les critères sont exploités mais que les plus importants sont : la fréquence de la réponse, le rang du passage et la proximité des mots de la question.

La figure 4.5 présente plus en détail l'importance des différents critères puisqu'elle montre comment chaque critère améliore les résultats. Pour cela, les deux ensembles Quæro et EQueR sont étudiés. Nous pouvons voir que pour EQueR la fréquence de la réponse et la proximité des termes sont deux critères prédominants mais qu'en revanche la validation du type de la réponse ne fournit aucune information. A l'inverse pour Quæro cette vérification est importante mais la fréquence de la réponse ne l'est pas. Ces informations montrent bien que les corpus sont différents.

L'apport de la vérification du type de la réponse a été mesuré sur les données de la collection Quæro en retirant ce critère de l'ensemble de départ. Le MRR passe alors de 0,43 à 0,41 ce qui montre que ce critère est bien pertinent d'autant plus qu'il ne s'applique pas à toutes les questions, seulement à 54 %.

Une étude plus spécifique de ce critère a été menée. Tout d'abord, les questions sont décomposées

selon qu'elles précisent un type spécifique ou non. Les questions sans type spécifique sont souvent plus simple « *Où est le restaurant El Bario ?* » ce qui entraîne de meilleurs résultats (MRR 0,46 contre 0,40). Pour les questions avec un type spécifique, ce critère est effectivement pertinent puisque sans lui le MRR descend à 0,37.

La redondance de la réponse est une information particulièrement discriminante pour les questions venant de la campagne EQueR. Cela peut s'expliquer car sur cette collection la moyenne des redondances sur toutes les réponses est de 1,7 alors que celle des réponses correctes est de 9,6. Ce qui montre une vraie distinction entre ces deux valeurs. Cette différence ne se retrouve pas sur la collection Quæro.

4.3.3.4 Comparaison avec des systèmes existants

Avant de terminer les évaluations, nous pouvons comparer notre module d'ordonnancement de réponses avec ceux utilisés par d'autres systèmes d'ordonnancement. Il est à noter que ces systèmes s'appliquent sur l'anglais où des ressources supplémentaires, comme WordNet, sont disponibles. Le système présenté par Cui et al. [2005] s'applique sur des articles de journaux et ordonne les passages en se fondant sur les relations syntaxiques du passage et de la question. L'évaluation est effectuée sur un ensemble de 413 questions factuelles venant de la campagne TREC-12. Les résultats que nous avons obtenus sont supérieurs aux leurs puisqu'ils obtiennent un MRR de 0,48 là où les nôtres sont de 0,50 et 0,53.

Le système présenté dans [Suzuki et al. 2002] combine différents critères portant sur la présence des termes de la question dans le passage, la présence des entités nommées et la présence des mots importants de la question dans le passage. Le système est évalué sur 1 358 questions ayant au moins une réponse correcte associée. L'évaluation obtient un MRR de 0,70 alors que la baseline n'en avait un que de 0,47. Il y a donc une amélioration de 49 %. Nos résultats sont un peu inférieurs car nous n'avons une amélioration que de 0,40 à 0,58 soit 45 % pour les questions pour lesquelles il existe une réponse correcte associée.

Harabagiu & Hickl [2006] s'intéressent à l'intégration de la détection de l'implication textuelle dans un système de questions réponses. La détection est effectuée par une combinaison de critères lexicaux, de vérifications syntaxiques et de vérifications de paraphrases. Quatre intégrations de ce module dans un système de questions réponses sont effectuées : l'ordonnancement de réponse, de passage, une réécriture de questions et une méthode hybride. L'évaluation est effectuée sur un ensemble de 500 questions et la baseline qui consiste à ne pas utiliser l'implication textuelle obtient un MRR de 0,30. La première utilisation, la plus proche de la notre obtient un MRR de 0,41 pour les questions avec un type spécifique et 0,38 pour les questions sans, ce qui correspond à une amélioration respectivement de 37 % et 27 %. La meilleure utilisation obtient un MRR de 0,56 et 0,40 ce qui correspond à une amélioration respectivement de 87 % et 37 %. Notre méthode est donc plus performante que leur ordonnancement de réponses simple mais moins que leur méthode hybride qui nécessite des règles de transformations de questions. De plus la détection d'implication tient compte de ressources telles que les règles de paraphrases qui ne sont disponibles que sur l'anglais.

4.4 Conclusion

Alors que la plupart des systèmes de questions réponses rencontrent des difficultés lorsqu'ils s'appliquent sur les documents Web, le système QAVAL apporte une solution robuste. Pour cela il s'appuie sur un ordonnancement de réponses utilisant en grande partie notre module de validation de réponses. De nombreuses réponses sont extraites depuis de courts passages de texte, de trois phrases environ, puis le module est appliqué pour les ordonner. Le système obtient des résultats similaires en effectuant une recherche sur des articles de journaux que sur les documents Web. De plus les résultats sont bons puisqu'ils surpassent la baseline de près de 48 %. Ces informations permettent de voir que notre module de validation de réponses peut effectivement s'appliquer correctement au sein d'un système de questions réponses.

Même si les résultats sont bons il est toujours possible de les améliorer. Dans notre cas nous avons montré que l'extraction des réponses serait meilleure en utilisant des conditions permettant de limiter le nombre de réponses à ordonner par exemple en reconnaissant davantage d'entités nommées. Une autre amélioration possible s'applique quand plusieurs réponses sont issues d'un même passage et auraient pour but d'en dégager la meilleure. Une possibilité pourrait être de tenir compte de la syntaxe en recherchant les relations de la question dans le passage ou en vérifiant que les relations portant sur le focus se trouvent effectivement dans le passage.

Jusqu'à présent, le système ne s'applique qu'aux questions factuelles. Afin d'être le plus exhaustif possible il faudrait qu'il puisse s'appliquer à d'autres types de questions et entre autre aux questions booléennes. Dans ce type de questions la réponse OUI ou NON est à renvoyer en vérifiant qu'il existe un document pouvant contenir l'information. Les traitements à effectuer pour ce type de questions sont présentés plus en détail dans les perspectives finales en 5.1.1.

Chapitre 5

Conclusion et perspectives

Pour trouver la réponse à une question, les systèmes de questions réponses doivent retrouver des passages contenant l'information demandée et les analyser afin d'en extraire la réponse courte. Cette information est formulée différemment dans les documents tant au niveau lexical que syntaxique et dans certains cas elle ne peut être qu'inférée du passage. De plus, certains types de documents, comme ceux issus du Web, suivent des styles très spécifiques ce qui peut poser problème pour certaines analyses.

La validation de réponses a pour but de détecter si une réponse est valide en s'appuyant sur le passage de texte l'accompagnant. Elle peut ainsi permettre d'améliorer les systèmes de questions réponses en ne renvoyant que les réponses valides à l'utilisateur. Ce manuscrit a traité tout particulièrement de cette tâche et a répondu aux questions suivantes :

– Qu'est-ce qu'une réponse valide ?
– Comment mettre en pratique la notion de validation de réponses pour détecter automatiquement les réponses valides ?

En étudiant les différentes campagnes d'évaluation des systèmes de validation de réponses, nous avons vu qu'une réponse est valide si elle est compatible avec la question et si les différentes informations demandées ou présentes dans la question se retrouvent dans le passage justificatif.

Une étude de corpus a permis d'étudier les différents phénomènes à traiter afin de reconnaître ces éléments, ainsi que leur présence et absence possible du document. Cette étude a permis de constater que les réponses partiellement justifiées étaient trois fois plus présentes que les réponses clairement justifiées ce qui montre la difficulté de la tâche et la nécessité de tenir compte de ces phénomènes. Parmi ceux-ci, le plus présent est la reformulation d'un terme de la question suivi par l'absence du type de la réponse du document.

Tout cela complexifie un appariement global entre la question et le texte. Ainsi nous avons proposé et développé une stratégie qui décompose les éléments à retrouver. Cela permet d'intégrer différents types de vérification de ces éléments dans les textes.

La décomposition de question a pour but de réduire la question en une structure courte du type sujet-verbe-objet, similaire à la notion de focus déjà présent dans les systèmes de questions réponses FRASQUES et QALC, et un ensemble de vérifications annexes. Pour mesurer le taux de présence de ces vérifications, une étude d'un corpus de questions a été réalisée. Elle a permis de faire ressortir

les trois vérifications les plus présentes : la vérification du type spécifique, le type précis de réponse attendu par la question (présent dans 49 % des questions ayant au moins une vérification secondaire), de la date (44 %) et du lieu (29 %).

Au final, une réponse est valide si :
– elle est compatible avec la question notamment au niveau du type spécifique attendu ;
– les autres informations contenues dans la question se retrouvent dans le passage justificatif sachant qu'elles peuvent se trouver sous une forme différente. Parmi celles-ci, il est nécessaire de vérifier que le passage mentionne l'action de la question, sa date et son lieu.

A l'aide de cette définition, la validation peut être effectuée par des processus distincts : l'analyse des passages pour vérifier qu'ils contiennent bien les informations demandées par la question et la vérification du type spécifique de la question.

L'analyse des passages justificatifs porte sur les différentes informations de la question et s'appuie sur différentes vérifications. La première détecte que le passage mentionne bien les mêmes éléments que la question en effectuant une comparaison de termes communs. Comme tous les termes n'ont pas la même importance, différentes comparaisons ont été effectuées portant notamment sur la catégorie morphosyntaxique des termes ou leur importance dans la question. Dans ce cadre, un traitement tout particulier a été fait pour les dates. Comme une date possède peu de variations, si la date de la question ne se retrouve pas dans le passage alors la réponse sera considérée comme non valide.

Le deuxième grand type de critères vérifie que les termes sont employés de la même façon dans le passage que dans la question en se fiant à l'idée que s'ils sont proches dans le texte alors ils sont probablement liés et expriment de ce fait la même information que la question. La mise en pratique de cette idée a consisté à calculer la plus longue chaîne de mots de la question et de la réponse présents dans le passage. Dans cette chaîne les mots doivent être suffisamment proches les uns des autres sans tenir compte de leur ordre.

Ce système a été évalué sur sa capacité à reconnaître les réponses valides. Pour ce faire, il a notamment participé à la campagne d'évaluation AVE 2008. Dans celle-ci, les données sont les sorties de différents systèmes de questions réponses, ce qui devait permettre de ne pas être dépendant d'un système particulier. Le système de validation s'appliquait après une étape de filtre visant à supprimer des réponses non valides reconnaissables de manière triviale. Une autre amélioration tenait compte de la comparaison de la réponse à évaluer et de celle extraite du passage justificatif par le système de questions réponses FRASQUES. Sur le français, le système a obtenu les meilleurs résultats et se place parmi les meilleurs toutes langues confondues ce qui témoigne de son efficacité.

L'une des difficultés d'un système de questions réponses est de typer finement les entités des textes. Ainsi, plus un système sait reconnaître de types différents, meilleurs sont ces performances. Ainsi nous avons étudié ce problème afin de vérifier que la réponse est compatible avec la question posée pour les questions explicitant un type spécifique attendu. Par exemple « *Quel président succéda à Jacques Chirac ?* » attend un président en réponse. Très souvent cette information n'est pas contenue dans le passage justificatif et par conséquent différents critères tenant compte de connaissances et de documents externes ont été considérés :

– des utilisations d'entités nommées soit pour valider certaines réponses à l'aide d'un système de reconnaissance à grain fin soit pour en rejeter d'autres grâce à une taxonomie plus large ;

– des recherches dans l'encyclopédie Wikipédia en suivant l'idée que le contenu d'une page vise à expliquer son titre. Ainsi le type est cherché dans la page associée à la réponse ;

– des recherches de structures de phrases particulières indiquant un lien entre le type et la réponse telles que « *Réponse est un Type* » ;

– des fréquences d'apparition en corpus du type et de la réponse en supposant que s'ils apparaissent souvent dans les mêmes documents alors ils sont liés.

Cette méthode est efficace puisqu'elle effectue 80 % de bonnes détections. Malheureusement, elle n'a pas permis d'améliorer considérablement la détection des réponses valides quand elle est appliquée sur les sorties des systèmes.

Afin de répondre à notre troisième problématique, à savoir « Comment utiliser la validation de réponses dans un système de questions réponses ? » nous avons intégré la validation de réponses dans un SQR pour ordonner différentes réponses, après leur extraction. Dans ce but, un score de confiance est calculé pour chaque réponse.

D'un point de vue pratique, notre module de validation a été placé au sein du système de questions réponses QAVAL qui a été évalué sur des questions factuelles. De très nombreuses réponses sont extraites depuis de courts passages de textes et ordonnées par le module de validation mettant en œuvre des critères supplémentaires. Ainsi, certains critères se retrouvent dans la plupart des SQR et d'autres sont spécifiques à notre système ; outre la vérification du type spécifique, il s'agit des caractéristiques déduites de l'analyse des questions (focus, catégorie des questions).

Le premier des critères porte sur la densité des mots de la question autour des différentes réponses. Ce critère suppose que plus la réponse est proche des mots de la question plus elle leur est liée et correspond à une des méthodes classiques d'extraction de réponses. Un autre critère correspond au rang du passage duquel la réponse a été extraite. Ce rang est issu d'une étape d'ordonnancement de passages, fondé sur les termes du passage, visant à reconnaître les passages les plus à mêmes de contenir la réponse correcte. Le dernier critère porte sur la redondance de la réponse. En effet, si la même réponse a pu être extraite depuis différents passages alors elle est souvent liée aux mots de la question et a donc plus de chances d'être correcte qu'une réponse extraite une seule fois.

Le système QAVAL a été évalué sur deux types de collections distinctes : les articles de journaux et les documents issus du Web. De par sa structure spécifique, cette dernière catégorie est plus délicate à traiter par un système de questions réponses. Le système est cependant efficace dans les deux cas puisqu'il obtient des résultats comparables sur les deux collections, proches de ceux obtenus par les autres systèmes traitant de la même collection issue du Web. L'étape d'ordonnancement de réponses est plutôt efficace puisque, quand cela est possible, elle place la réponse correcte en première position dans 50 % des cas et parmi les cinq meilleurs pour plus de 75 %. De plus, les résultats surpassent de 45 % ceux obtenus par une méthode de base visant à sélectionner la réponse la plus proche des mots de la question en suivant l'ordre des passages.

Bien sûr ce n'était qu'une manière de répondre à ces problématiques et d'autres systèmes, présentés lors de l'état de l'art, par exemple [Zanzotto & Moschitti 2006] ou [Tatu et al. 2006], se fondent sur l'utilisation d'une structure particulière du passage et de la forme déclarative de la question contenant la réponse en s'appuyant sur des analyses syntaxique profondes. L'évaluation du système QAVAL a montré que le système pouvait s'adapter à différents types de documents comme les documents issus

du Web. Les analyseurs syntaxiques n'étant pas robustes à ces documents, une telle évaluation de la validité d'une réponse n'aurait pas été possible.

5.1 Perspectives

Nous avons donc vu dans ce manuscrit la création d'un système de validation de réponses ainsi que son intégration dans un système de questions réponses. L'approche obtient de bons résultats néanmoins des améliorations peuvent être envisagées. Trois grands types de perspectives sont présentés ici :

– l'amélioration du système QAVAL ;
– la vérification du lieu de l'événement précisé dans la question ;
– la reconnaissance de paraphrases permettant de vérifier que les relations entre les termes de la question se trouvent dans le passage.

5.1.1 Amélioration du système QAVAL

Les perspectives à court terme portent sur l'amélioration du système QAVAL. La première amélioration possible porte sur les patrons d'extractions. Comme nous l'avons vu, ces patrons permettent de marquer un lien entre la réponse et le focus ou le type. Une plus grande prise en compte de tels patrons semble donc un critère pertinent qui permettra de caractériser les réponses valides. Dans le cadre du système FRASQUES des patrons existaient mais, malheureusement, ils s'appliquent plutôt sur des documents correctement rédigés et s'appliquent peu sur les documents issus du Web. Il serait intéressant de mener une étude sur les formulations des réponses. Si elles possèdent des régularités, on pourrait alors envisager d'apprendre ces patrons.

L'analyse des résultats a montré que le système QAVAL s'applique moins bien quand de nombreuses réponses sont extraites. Une amélioration possible pourrait consister à réduire le nombre de réponses extraites tout en restant assez large pour être sûr de reconnaître la bonne réponse. Des possibilités de traiter ce problème pourraient porter sur la distance entre la réponse et le focus en rejetant des réponses trop éloignées ou sur les liens syntaxiques liant ces deux termes.

La dernière amélioration possible concerne l'utilisation du système QAVAL pour traiter d'autres catégories de questions. En effet, jusqu'à présent il a été développé pour répondre à des questions factuelles, ce qui restreint ses possibilités d'usage.

La première catégorie non traitée concerne les questions de définition qui attendent en réponse la définition d'une entité comme par exemple « *Qu'est ce qu'un airbus ?* ». Pour ces questions la méthode la plus courante consiste à s'appuyer sur des patrons d'extraction tels que « Focus est un réponse » qui se rencontrent notamment dans le passage « *un airbus est un avion.* ». Pour valider une réponse il faudrait vérifier qu'un patron s'applique bien. L'utilisation de la vérification du type de la réponse peut également être considérée. En effet, au lieu de vérifier que la réponse est une instance du type spécifique, il faudrait reconnaître que le focus est du type désigné par la réponse (que l'airbus est un avion) dans des autres documents.

La validation de réponses peut également permettre de résoudre les questions booléennes. Pour ces questions une affirmation est proposée par l'utilisateur comme « *Jean-Baptiste Poquelin était -il*

Molière ? ». Pour ces questions le mécanisme consiste à extraire l'ensemble des passages correspondant à une requête contenant les mots clés de la question et à vérifier si l'un d'eux explique la forme déclarative de la question. A la question précédente, un tel passage peut par exemple être « *Molière est aussi connu sous le nom de Jean-Baptiste Poquelin.* ». Si c'est le cas alors le système renvoie la valeur OUI. Pour répondre NON il faut détecter un passage qui contredit l'hypothèse. Par exemple, à la question « *Le brésil a-t-il gagné la coupe du monde de football en 1998 ?* » il faut répondre NON en détectant un passage de texte comme « *La France a remporté la coupe du monde de football en 1998.* ». Comme deux pays ne peuvent pas remporter cette compétition, le texte contredit donc l'hypothèse.

On peut remarquer que ce traitement est similaire à celui effectué lors de la détection d'implication textuelle. Comme la validation de réponses est proche de cette tâche nous pouvons utiliser une version modifiée de notre système. Ainsi la présence des mots de la question dans le passage peut être considérée ainsi que le calcul de la plus longue chaîne commune. En revanche les autres critères tels que la vérification du type ou la densité des mots de la question autour de la réponse semblent peu pertinents.

Pour pouvoir traiter ces questions, il faudrait ajouter des critères spécifiques par exemple un critère marquant la polarité du passage en comptant le nombre de négations ce qui permettrait de détecter les cas où la réponse NON est à fournir. Ce trait pourrait également être utile à un système de questions réponses mais ces phénomènes sont peu rencontrés sur des questions factuelles. Il faudrait également tenir compte des relations sémantiques entre les mots afin de détecter les cas d'inférence.

5.1.2 Vérification du lieu

La décomposition des questions a montré que l'information de lieu était présente dans de nombreuses questions et constitue le troisième grand type de vérification le plus présent après la vérification du type spécifique et de la date de la question. La vérification consisterait alors à rechercher le lieu contenu dans la question dans le passage. Par exemple pour la question « *Dans quelle ville eut lieu la finale de la Coupe du monde de football aux USA en 1994 ?* » il faut vérifier que l'événement a lieu aux États Unis. Contrairement aux dates, un lieu peut souvent se trouver sous une forme différente, par exemple être remplacé par un de ses méronymes. Ainsi un pays peut être remplacé par une ville ou une région. A la question précédente, la réponse « Pasadena » est validée grâce au passage « *La finale de la coupe du monde de football se déroule le 17 juillet 1994 au Rose Bowl de Pasadena* ». Il reste alors à montrer que Passadena est une ville américaine. Le problème peut se formaliser de la manière suivante : étant donné un lieu présent dans le passage justificatif, correspond-il à celui présent dans la question ou est-il l'un de ses méronymes ?

Pour ce faire, nous pourrions vérifier cette relation pour chaque nouvel exemple. Le mécanisme pourrait se rapprocher de celui vérifiant la correspondance entre la réponse et le type spécifique attendu par la question qui était effectué à l'aide d'un système de reconnaissance d'entités nommées et de méthodes de fouille de différents textes. Tout d'abord, la page Wikipédia d'une ville ou d'une région mentionne le pays dans lequel elle se trouve. Ainsi « France » est présente dans la page associée à « Paris ». Ensuite certaines structures de phrases comme « *Lieu1 est en Lieu2* » (Paris est en France) indiquent la relation de méronymie. A ce sujet on peut noter que certains systèmes de recherche de relations d'hyperonymie, par exemple [Morin 1998], s'appliquent également à ce type de recherches.

Pour finir, la fréquence en corpus pourrait également être un signe de liaison puisque le pays se trouve souvent dans les documents relatifs à la ville. Par exemple, Google nous indique que Marseille se trouve dans 159 000 000 documents et que 103 000 000 (65 %) contiennent également le mot « France ».

Pour effectuer ce travail, il resterait alors à créer un ensemble de données suffisamment grand pour l'apprentissage et définir les règles exprimant les relations cherchées.

La vérification du type de la réponse pourrait également être appliquée pour gérer la coréférence. La vérification du type peut s'appliquer plus spécifiquement quand le coréférent n'est pas porté par un pronom mais par un groupe nominal. Ce groupe correspond souvent au « type » de l'antécédent. Par exemple, dans le passage « *Steven Spielberg persuade George Lucas qu'« Indy » est un personnage taillé pour Harrison Ford. L'acteur et le réalisateur sont mis en relation par George Lucas* » il faut détecter si le terme « l'acteur » fait référence à Steven Spielberg ou à Harison Ford. Comme ce dernier est le seul acteur parmi les deux, c'est à lui qu'il est fait référence. Dans le cadre de la validation de réponses, les antécédents intéressants seraient restreints à certaines entités de la question.

5.1.3 Utilisation de la syntaxe et détection de paraphrases

La dernière perspective concerne notre module de validation de réponses. Comme il suit une approche par apprentissage, il est toujours possible de rajouter de nouveaux critères. Ainsi le traitement des informations spatiales pourrait constituer un nouveau critère.

Un autre grand type de vérifications pourrait montrer que les mots présents dans le passage sont utilisés dans le sens demandé par la question et pour cela vérifier qu'ils sont reliés de manière globale ou apparaissent sous d'autres formes. L'approche présentée dans ce manuscrit utilise pour cela une notion de paraphrase locale grâce à la proximité des mots de la question dans le passage ainsi que les variations de termes reconnues par FASTR. Malheureusement, cela ne suffit pas toujours à s'assurer que les termes sont effectivement reliés.

Différentes améliorations pourraient porter sur les liens entre les termes. Tout d'abord, de nombreux systèmes [Rus 2006 ; Marneffe, et al. 2006 ; Moriceau & Tanier 2009] vérifient que les relations syntaxiques présentes dans la question se trouvent également dans le passage. Dans le chapitre 3.3, notre module d'analyse des passages a été complété par une vérification portant sur la présence des relations syntaxiques de la question dans le passage. Pour cela, la proportion de relations syntaxiques de la question contenues dans le passage était calculée. Les résultats ont montré que le système était moins performant en intégrant cette vérification. Il faudrait donc trouver un moyen plus subtil de tenir compte de ces relations. La section 3.2 a montré que chaque question pouvait se décomposer en une phrase minimale et un ensemble d'informations secondaires, la relation principale étant souvent de la forme sujet-verbe-objet. Pour valider une réponse, il semble pertinent de montrer que le sens de la phrase minimale peut être inféré du passage justificatif. Bien souvent le verbe de cette phrase ne se retrouve pas dans le passage justificatif alors que les deux autres composants sont présents.

Si la relation verbale ne se trouve pas dans le passage cela ne veut pas forcément signifier que l'information portée en est absente. En effet, une paraphrase de cette relation peut être présente et c'est un cas assez souvent rencontré lors de l'étude de corpus que nous avons menée. Cette paraphrase correspond alors à un chemin dans la représentation syntaxique de la question reliant le sujet et l'objet

du verbe dont on cherche une paraphrase. Le nouveau problème peut donc se formaliser ainsi : étant donné un chemin présent dans le passage et reliant deux termes de la question ou un terme de la question et la réponse, correspond-il à la relation verbale reliant les deux mêmes termes dans la question ? Une telle vérification permettrait par exemple de reconnaître « poignarder à mort » comme paraphrase de « assassiner » à la question « *Qui a assassiné Henri IV ?* » et le passage « Ravaillac poignarda à mort Henri IV ».

Une solution possible pourrait être d'effectuer des recherches dans des documents. La plupart des systèmes de paraphrases ont pour but de rechercher la paraphrase d'une phrase donnée et partent de l'hypothèse présentée dans [Lin & Pantel 2001] : si deux formulations ont tendance à avoir des contextes en commun alors elles sont liées et se trouvent en situation de paraphrase.

Dans les systèmes de paraphrases, les phrases sont représentées sous forme de chemins dans un arbre syntaxique et les contextes correspondent aux extrémités de ces chemins. Dans le cas de l'exemple précédent, les contextes sont « Henri IV » et « Ravaillac » et les chemins « assassiner » et « poignarda à mort ».

Une approche possible pourrait se rapprocher de celle présentée dans [Pantel, et al. 2007] qui recherche les différents chemins dans des documents Web. De ces chemins sont issues des extrémités différentes. Dans notre exemple, l'extrémité gauche correspond à un grand nombre d'assassins et l'extrémité droite à leurs victimes. Ces extrémités sont recherchées pour les deux chemins et comparées ce qui permet d'avoir un score de confiance sur le fait que les phrases soient paraphrases.

Il reste alors de nombreux travaux à mener afin d'appliquer cette approche, notamment pour définir la comparaison des contextes à effectuer en tenant compte du contexte global de la question. Le contexte global peut permettre de réduire les problèmes de polysémie d'un verbe. Par exemple l'expression « jouer à » porte sur une personne et soit un jeu (« Marie joue à la Marelle ») soit une salle de spectacle (« Dany Boon joue à l'Olympia ») soit une ville (« Steve Mandanda joue à Marseille »).

Annexe A

Liste des entités nommées du système RITEL

- acronyme ;
- adresse ;
- année ;
- animal ;
- astre ;
- célèbre cométe ;
- continent ;
- compétition ;
- couleur ;
- équipe ;
- film ;
- fleuve ;
- fonction ;
- fonction religieuse ;
- fonction publique ;
- grade militaire ;
- groupe de musique ;
- heure ;
- jour de la semaine ;
- journal ;
- langue ;
- lieu ;
- lien familial ;
- livre ;
- ministère ;
- monnaie ;
- montagne ;
- monument ;

- mois ;
- mouvement (politique) ;
- musée ;
- œuvre musicale ;
- origine ;
- parti (politique) ;
- pays ;
- peuple ;
- personne ;
- phénomène météorologique ;
- point cardinal ;
- prix ;
- prénom ;
- province ;
- région ;
- religion ;
- route ;
- sport ;
- société de production ;
- titre ;
- transport ;
- type d'animal ;
- type de film ;
- unité de fréquence ;
- unité de distance ;
- unité de masse ;
- unité de mémoire ;
- unité de superficie ;
- unité de taille de fichier ;
- unité de temps ;
- unité de vitesse de transmission ;
- unité de vitesse ;
- unité de volume ;
- université ;
- ville ;
- voie ;

Bibliographie

R. Adams (2006). 'textual entailment through extended lexical overlap'. In *Proceedings of RTE-2 Workshop*.

S. Aït-Mokhtar, J.-P. Chanod et C. Roux (2002). 'Robustness beyond shallowness : incremental deep parsing'. *Natural Language Engineering* **8**.

C. Ayache, B. Grau et A. Vilnat (2006). 'EQueR : the French Evaluation Campaign of Questions Answering Systems'. In 5^{th} *Conference on Language Resources and Evaluation (LREC 2006)*.

J. Y. Bahadorreza Ofoghi (2009). 'UB.dmirg : A Syntactic Lexical System for Recognizing Textual Entailments'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

K. Balog, P. Serdyukov et A. P. de Vries (2010). 'Overview of the TREC 2010 Entity Track'. In *TREC 2010 Working Notes*.

R. Bar-Haim, I. Dagan, I. Greental, I. Szpektor et M. Friedman (2007). 'Semantic inference at the lexical-syntactic level for textual entailment recognition'. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

R. Bar-Haim, I. Szpektor et O. Glickman (2005). 'Definition and Analysis of Intermediate Entailment Levels'. In *Proceedings of ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*.

V. Barbier (2009). *Utilisation de connaissances sémantiques pour l'analyse de justifications de réponses à des questions*. Ph.D. thesis, Université Paris Sud.

D. Battistelli, J. Couto, J.-L. Minel et S. R. Schwer (2008). 'Représentation algébrique des expressions calendaires et vue calendaire d'un texte'. In *Actes de TALN/JEP 08*, pp. 1–10.

P. Bédaride et C. Gardent (2010). 'Syntactic Testsuites and Textual Entailment Recognition'. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

L. Bentivogli, E. Cabrio, I. Dagan, D. Giampiccolo, M. L. Leggio et B. Magnini (2010a). 'Building Textual Entailment Specialized Data Sets : a Methodology for Isolating Linguistic Phenomena Relevant to Inference'. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

L. Bentivogli, P. Clark, I. Dagan, H. Dang et D. Giampiccolo (2010b). 'The Sixth PASCAL Recognizing Textual Entailment Challenge'. In *Proceedings of TAC'2010*.

L. Bentivogli, I. Dagan, H. T. Dang, D. Giampiccolo et B. Magnini (2009). 'The Fifth PASCAL Recognizing Textual Entailment Challenge'. In *Proceedings of TAC'2009*.

G. Bernard (2011). *Réordonnancement d'hypothèses dans un systèmes de questions réponses*. Ph.D. thesis, Université Paris Sud.

D. Bernhard, B. Cartoni et D. Tribout (2011). 'Évaluer la pertinence de la morphologie constructionnelle dans les systèmes de Question-Réponse'. In *Actes de TALN 2011*.

E. Boldrini, M. Puchol-Blasco, B. Navarro, P. M. Martínez-Barco et C. Vargas-Sierra (2009). 'AQA : a multilingual Anaphora annotation scheme for Question Answering'. *Procesamiento del lenguaje natural* **42** :97–104.

L. Bonnefoy, P. Bellot et M. Benoit (2011). 'Une approche non supervisée pour le typage et la validation d'une réponse à une question en langage naturel : application à la tâche Entity de TREC 2010'. In *COnférence en Recherche d'Infomations et Applications*.

J. Bos et K. Markert (2006). 'When logical inference helps determining textual entailment (and when it doesn't)'. In *The Second PASCAL Recognising Textual Entailment Challenge. Proceedings of the Challenges Workshop*, pp. 98–103.

W. Bosma et C. Callison-Bursh (2006). 'Paraphrase substitution for Recognizing textual entailment'. In *Working Notes for the CLEF 2006 Workshop (AVE)*.

E. Breck (2009). 'A simple system for detecting non-entailment'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

E. Cabrio et B. Magnini (2011). 'Towards component-based textual entailment'. In *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, pp. 320–324.

J. J. Castillo (2009). 'Sagan in TAC2009 : Using Support Vector Machines in Recognizing Textual Entailment and TE Search Pilot task'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

T. Chklovski et P. Pantel (2004). 'VerbOcean : Mining the Web for Fine-Grained Semantic Verb Relations'. In D. Lin & D. Wu (eds.), *Proceedings of EMNLP 2004*, pp. 33–40.

C. L. A. Clarke, G. V. Cormack et T. R. Lynam (2001). 'Exploiting redundancy in question answering'. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pp. 358–365, New York, NY, USA. ACM.

D. Clarke (2006). 'Meaning as Context and Subsequence Analysis for Entailment'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

J. Cohen (1960). 'A Coefficient of Agreement for Nominal Scales'. *Educational and Psychological Measurement* **20**(1) :37.

H. Cui, R. Sun, K. Li, M. yen Kan et T. seng Chua (2005). 'Question answering passage retrieval using dependency relations'. In *SIGIR 2005*, pp. 400–407.

I. Dagan et O. Glickman (2004). 'Probabilistic Textual Entailment : Generic Applied Modeling of Language Variability'. In *Proceedings of the PASCAL Workshop on Learning Methods for Text Understanding and Mining*.

I. Dagan, O. Glickman et B. Magnini (2005). 'The PASCAL Recognising Textual Entailment Challenge'. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop*.

G. de Chalendar, T. Dalmas, F. Elkateb-Gara, O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz, L. Monceaux, I. Robba et A. Vilnat (2002). 'The Question Answering System QALC at LIMSI, Experiments in Using Web and WordNet'. In *Proceedings of TREC11*.

R. De Salvo Braz, R. Girju, V. Punyakanok, D. Roth et M. Sammons (2005). 'An inference model for semantic entailment in natural language'. In *Proceedings of the 20th national conference on Artificial intelligence - Volume 3*, pp. 1043–1049.

A. Echihabi, U. Hermjakob, E. Hovy, D. Marcu, E. Melz et D. Ravichandran (2004). 'How to Select an Answer String ?'. In T. Strzalkowski & S. Harabagiu (eds.), *Advances in Textual Question Answering*. Kluwer.

S. El Ayari (2009). *Évaluation transparente du traitement des éléments de réponse à une question factuelle*. Ph.D. thesis, Université Paris Sud.

M.-H. Falco (2009). *Analyse des questions dans un système de question-réponse*. Mémoire de master, Université Paris Diderot - Paris 7.

M.-H. Falco (2010a). 'Quartely Progress Report Numéro 11'. Tech. rep., QUAERO.

M.-H. Falco (2010b). 'Quartely Progress Report Numéro 9'. Tech. rep., QUAERO.

L. Fang, N. Si, S. Somasundaram, Z. Al-Ansari, a Yu et Y. Xian (2010). 'Purdue at TREC 2010 Entity Track : a Probabilistic Framework for matching types between Candidates and Target Entities'. In *The Nineteenth Text REtrieval Conference (TREC 2010) Proceedings*.

C. Fellbaum et G. Miller (1998). *WordNet : An Electronic Lexical Database*. Christiane Fellbaum.

O. Ferrandez, D. Micol, R. Munoz et M. Palomar (2007). 'The contribution of the University of Alicante to AVE 2007'. In *Working Notes of CLEF*.

O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz et C. Jacquemin (2001). 'Document selection refinement based on linguistic features for QALC, a Question Answering system.'. In *3rd International Conference on Recent Advances in Natural Language Processing (RANLP 2001)*.

O. Ferret, B. Grau, M. Hurault-Plantet, G. Illouz et C. Jacquemin (2002). 'Quand la réponse se trouve dans un grand corpus.'. *Ingénierie des Systèmes d'Information* **7**(1-2) :95–123.

O. Ferrández, R. Muñoz et M. Palomar (2008). 'A lexical-semantic approach to AVE'. In *CLEF*.

O. Ferrández, R. Muñoz et M. Palomar (2009). 'Alicante University at TAC 2009 : Experiments in RTE'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

L. Fleifel (2010). 'Analyse des questions d'un système de questions-réponses'. Tech. rep., ENSIIE.

P. Forner, D. Giampiccolo, B. Magnini, A. Peñas, Álvaro Rodrigo et R. Sutcliffe (2010). 'Evaluating Multilingual Question Answering Systems at CLEF'. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

P. Forner, A. Peñas, E. Agirre, Eneko, I. Alegria, F. Corina, M. Nicolas, O. Petya, P. Prokopis, R. Paulo, S. Bogdan, S. Richard, S. Erik et T. K. Sang (2009). 'Overview of the Clef 2008 multilingual question answering track'. In *CLEF'08 : Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*.

A. Garcia-Fernandez (2010). *Génération de réponses en langue naturelle orales et écrites pour les systèmes de question-réponse en domaine ouvert*. Ph.D. thesis, Université Paris Sud.

D. Giampiccolo, B. Magnini, I. Dagan et B. Dolan (2007). 'The third PASCAL recognizing textual entailment challenge'. In *Proceedings of the ACLPASCAL Workshop on Textual Entailment and*.

L. Gillard, P. Bellot et M. El-Bèze (2006). 'Influence de mesures de densité pour la recherche de passages et l'extraction de réponses dans un système de questions-réponses.'. In *CORIA'06*, pp. 193–204.

O. Glickman (2006). *Applied Textual Entailment*. Ph.D. thesis, Senate ofBar Ilan University.

I. Glöckner, S. Hartrumpf et J. Leveling (2007). 'Logical validation, answer merging and witness selection a study in multi-stream question answering'. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, pp. 758–777.

I. Glöckner (2006). 'University of Hagen at QA@CLEF 2006 : Answer Validation Exercise'. In *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop*.

I. Glöckner (2008). 'University of Hagen at QA@CLEF 2008 : Answer Validation Exercise '. In *Results of the CLEF 2008 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2008 Workshop*.

A. Grappy et B. Grau (2010a). 'Answer type validation in question answering systems'. In *RIAO*, pp. 9–15.

A. Grappy et B. Grau (2010b). 'Validation du type de la réponse dans un système de questions réponses'. In *COnférence en Recherche d'Infomations et Applications - CORIA*, pp. 131–146.

A. Grappy et B. Grau (2011). 'Validation du type de la réponse dans un système de questions réponses'. *Document Numérique volume 14 n°=2* .

A. Grappy, B. Grau, M.-H. Falco, A.-L. Ligozat, I. Robba et A. Vilnat (2011). 'Selecting answers to questions from Web documents by a robust validation process'. In *The 2011 IEEE/WIC/ACM International Conference on Web Intelligence*.

A. Grappy, B. Grau, O. Ferret, C. Grouin, V. Moriceau, I. Robba, X. Tannier, A. Vilnat et V. Barbier (2010). 'A Corpus for Studying Full Answer Justification'. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

A. Grappy, A.-L. Ligozat et B. Grau (2008). 'Evaluation de la réponse d'un système de question-réponse et de sa justification'. In *COnférence en Recherche d'Infomations et Applications - CORIA*.

B. Grau (2005). 'EQueR, une campagne d'évaluation des systèmes de question/réponse'. *Journée Technolangue/Technovision (ASTI'2005)* .

B. Grau, G. Illouz, L. Monceaux, P. Paroubek, O. Pons, I. Robba et A. Vilnat (2005). 'FRASQUES, le système du groupe LIR, LIMSI'. In *Atelier EQueR, Conférence (TALN'05)*.

B. Grau, A.-L. Ligozat, I. Robba, A. Vilnat et L. Monceau (2006). 'FRASQUES : A Question Answering system in the EQueR evaluation campaign'. In *5^{th} Conference on Language Resources and Evaluation (LREC 2006)*.

B. Grau, A. Vilnat et C. Ayache (2008). 'EQueR : évaluation de systèmes de question-réponse'. *L'évaluation des technologies de traitement de la langue : les campagnes Technolangue Traité IC2, série Cognition et traitement de l'information* .

R. Grishman et B. Sundheim (1995). 'Design of the MUC-6 evaluation'. In *MUC*, pp. 1–11.

S. Harabagiu et A. Hickl (2006). 'Methods for using textual entailment in open-domain question answering'. In *Proceedings of the 44th annual meeting of the Association for Computational Linguistics*.

S. M. Harabagiu, M. A. Paşca et S. J. Maiorano (2000). 'Experiments with open-domain textual Question Answering'. In *Proceedings of the 18th conference on Computational linguistics*, pp. 292–298.

S. Hartrumpf (2008). 'Semantic Decomposition for Question Answering'. In *ECAI*, pp. 313–317.

E. Hatcher et O. Gospodnetic (2004). *Lucene in Action (In Action series)*. Manning Publications Co.

M. A. Hearst (1992). 'Automatic Acquisition of Hyponyms from Large Text Corpora'. In *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539–545.

J. Herrera, A. Rodrigo, A. Peñas et F. Verdejo (2006). 'UNED Submission to AVE 2006'. In *Workshop CLEF 2006*, Alicante, Spain.

A. Hickl, J. Williams, J. Bensley, K. Roberts, B. Rink et Y. Shi (2006). 'Recognizing Textual Entailment with LCC's GROUNDHOG System'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

R. Higashinaka et H. Isozaki (2008). 'Automatically Acquiring Causal Expression Patterns from Relation-annotated Corpora to Improve Question Answering for why-Questions'. *ACM Transactions on Asian Language Information Processing (TALIP)* **7** :6 :1–6 :29.

E. Hovy, L. Gerber, U. Hermjakob, C.-Y. Lin et D. Ravichandran (2001). 'Toward semantics-based answer pinpointing'. In *HLT '01 : Proceedings of the first international conference on Human language technology research*, pp. 1–7.

Z. Huang, M. Thint et A. Celikyilmaz (2009). 'Investigation of question classifier in question answering'. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 2 - Volume 2*, EMNLP '09, pp. 543–550, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Inkpen, D. Kipp et V. Nastase (2006). 'Machine Learning Experiments for Textual Entailment'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

C. Jacquemin (1996). 'A symbolic and surgical acquisition of terms through variation'. *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing* pp. 425–438.

C. Jacquin, L. Monceaux et E. Desmontils (2008). 'The Answer Validation System ProdicosAV Dedicated to French'. In *CLEF*, pp. 452–459.

O. Jahna et R. Dragomir (2004). 'Comparing Semantically Related Sentences : The Case of Paraphrase Versus Subsumption'. In *Proceedings of Coling 2004*, pp. 1265–1268.

J. J.Castillo (2008). 'The contribution of FaMAF at QA@CLEF 2008.Answer ValidationExercise'. In *CLEF*.

T. Joachims (2009). 'Training Linear SVMs in Linear Time'. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

C. Kohlscëtter, P. Fankhauser et W. Nejdl (2010). 'Boilerplate detection using shallow text features.'. In B. D. Davison, T. Suel, N. Craswell, & B. Liu (eds.), *WSDM*, pp. 441–450. ACM.

M. Kouylekov et M. Negri (2010). 'An Open-Source Package for Recognizing Textual Entailment'. In *ACL (System Demonstrations)*.

M. Kouylekov, M. Negri, B. Magnini et B. Coppola (2006). 'Towards Entailment-based Question Answering : ITC-irst at CLEF 2006'. In *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*.

Z. Kozareva et A. Montoyo (2006). 'MLEnt : The Machine Learning Entailment System of the. University of Alicante'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

Z. Kozareva, S. vasquez et A. Montoyo (2006). 'Adaptation of a machine-learning textual entailment system to a multilingual answer validation exercise'. In *Evaluation of Multilingual and Multimodal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006*.

D. Laurent, S. Nègre et P. Séguéla (2005). 'QRISTAL, le QR à l'épreuve du public'. *Traitement Automatique des langues* .

D. Laurent, P. Séguéla et S. Nègre (2010). 'Cross lingual question answering using qristal for clef 2006'. *Evaluation of Multilingual and Multi-modal Information Retrieval* pp. 339–350.

A.-L. Ligozat (2006). *Exploitation et fusion de connaissances locales pour la recherche d'informations précises*. Ph.D. thesis, Université Paris Sud.

A.-L. Ligozat, B. Grau, A. Vilnat, I. Robba et A. Grappy (2007a). 'Lexical validation of answers in Question Answering'. In *Web Intelligence*, pp. 330–333.

A.-L. Ligozat, B. Grau, A. Vilnat, I. Robba et A. Grappy (2007b). 'Towards an Automatic Validation of Answers in Question Answering'. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007) (2)*, pp. 444–447.

D. Lin (1998a). 'Dependency-based Evaluation of MINIPAR'. In *Proc. Workshop on the Evaluation of Parsing Systems*.

D. Lin (1998b). 'An Information-Theoretic Definition of Similarity'. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304. Morgan Kaufmann.

D. Lin et P. Pantel (2001). 'Discovery of inference rules for question-answering'. *Nat. Lang. Eng.* **7**(4) :343–360.

C. E. Lipscomb (2000). 'Medical Subject Headings (MeSH)'. *Bull Med Libr Assoc.* .

M.A.Garcia-Cumbreras, J. Perea-Ortega, F. M. Santiago et L. Urena-Lopez (2007). 'SINAI at QA@CLEF2007.Answer Validation Exercise'. In *CLEF*.

B. Magnini, D. Giampiccolo, P. Forner, C. Ayache, V. Jijkoun, P. Osenova, A. Peñas, P. Rocha, B. Sacaleanu et R. F. E. Sutcliffe (2006). 'Overview of the CLEF 2006 Multilingual Question Answering Track'. In *Working Notes for the CLEF 2006 Workshop*.

B. Magnini, M. Negri, R. Prevete et H. Tanev (2002a). 'Comparing Statistical and Content-Based Techniques for Answer Validation on the Web'. In *Proceedings of the VIII Convegno AI*IA*.

B. Magnini, M. Negri, R. Prevete et H. Tanev (2002b). 'Is It the Right Answer ? Exploiting Web Redundancy for Answer Validation'. In *proceedings of the 40th annual meeting of the association for computational linguistics*, pp. 425–432.

P. Malakasiotis (2009). 'AUEB at TAC 2009'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

M.-C. D. Marneffe, B. Maccartney, T. Grenager, D. Cer, A. Rafferty et C. D. Manning (2006). 'Learning to distinguish valid textual entailments'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

E. Marsi, E. Krahmer, W. Bosma et M. Theune (2006). 'Normalized alignement ofdependecy trees for detecting textual entailment'. In *CLEF*.

A. I. Martin, M. Franz et S. Roukos (2001). 'IBM's Statistical Question Answering System-TREC-10'. In *Proceedings of TREC10*.

R. Mihalcea et D. I. Moldovan (2001). 'eXtended WordNet : progress report'. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pp. 95–100.

D. Moldovan, S. Harabagiu, A. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu et O. Bolohan (2002). 'LCC Tools for Question Answering'. In *Proceedings of the 11th Text REtrieval Conference (TREC-2002)*.

V. Moriceau, B. Grau, O. Ferret et J.-L. Minel (2008a). 'Conique : Inférences en contexte pour trouver, justifier et présenter des réponses à des questions en domaine ouvert'. In *15ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2008)*.

V. Moriceau et X. Tanier (2009). 'Apport de la syntaxe dans un système de question-réponse : étude du système FIDJI'. *Traitement Automatique du Langage Naturel (TALN)* .

V. Moriceau, X. Tannier, A. Grappy et B. Grau (2008b). 'Justification of Answers by Verification of Dependency Relations - The French AVE Task'. In *Working Notes of CLEF Workshop*.

V. Moriceau, X. Tannier et B. Grau (2009). 'Utilisation de la syntaxe pour valider les réponses à des questions par plusieurs documents'. In *CORIA*.

E. Morin (1998). 'PROMÉTHÉE un outil d'aide à l'acquisition de relations sémantiques entre termes'. In *Traitement automatique des langues naturelles*.

A. Moschitti, S. Quarteroni, R. Basili et S. Manandhar (2007). 'Exploiting syntactic and shallow semantic kernels for question answer classification'. In *Proceedings of ACL-07*, pp. 776–783.

R. Nairn, C. Condoravdi et L. Karttunen (2006). 'Computing relative polarity for textual inference'. In *Proceedings of ICoS-5 (Inference in Computational Semantics)*.

E. Newman, J. Dunnion et J. Carthy (2006). 'Constructing a Decision Tree Classifier using Lexical and Syntactic Feature'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

J. Nicholson, N. Stokes et T. Baldwin (2006). 'Detecting Entailment Using an Extended Implementation of the Basic Elements Overlap Metrics'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

P. Pakray, S. Bandyopadhyay et A. Gelbukh (2009). 'Lexical based two-way RTE System at RTE-5'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

P. Pantel, R. Bhagat, T. Chklovski et E. Hovy (2007). 'ISP : Learning inferential selectional preferences'. In *Proceedings of NAACL 2007*.

K. Papineni, S. Roukos, T. Ward et W.-J. Zhu (2002). 'BLEU : a method for automatic evaluation of machine translation'. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

A. Peñas, Á. Rodrigo, , V. Sama et F. Verdejo (2006). 'Overview of the Answer Validation Exercise 2006'. In *CLEF*.

A. Peñas, Á. Rodrigo et F. Verdejo (2007). 'Overview of the Answer Validation Exercise 2007'. In *CLEF*, pp. 237–248.

D. Perez, E. Alfonsecaia et P. Rodríguez (2005). 'Application of the Blue algorithm for recognising textual entailments'. In *Proceedings of the Recognising Textual Entailment Pascal Challenge*.

A. Perini (2009). 'Detecting Textual Entailment with Conditions on Directional Text Relatedness Scores'. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

L. Quintard, O. Galibert, G. Adda, B. Grau, D. Laurent, V. Moriceau, S. Rosset, X. Tannier et A. Vilnat (2010). 'Question Answering on web data : the QA evaluation in Quæro'. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

D. Ravichandran, E. Hovy, , F. J. Och et F. J. Och (2003). 'Statistical QA - Classifier vs. Re-ranker : What's the difference ?'. In *Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering*.

A. Rodrigo, A. Peñas et F. Verdejo (2009). 'Overview of the answer validation exercise 2008'. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF'08, pp. 296–313, Berlin, Heidelberg. Springer-Verlag.

Á. Rodrigo, A. Peñas, J. Herrera et F. Verdejo (2006). 'The Effect of Entity Recognition on Answer Validation'. In *CLEF*, pp. 483–489.

A. Rodriguo, A. Peñas et F. Verdejo (2007). 'UNED at Answer Validation Exercise 2007'. In *CLEF*.

S. Rosset (2008). *Systèmes de dialogue (oral) homme machine : du domaine limité au domaine ouvert*. Document d'habilitation à diriger des recherches, Université Paris Sud.

S. Rosset, O. Galibert, G. Bernard, E. Bilinski et G. Adda (2008). 'The LIMSI participation to the QAst track'. In *Working Notes of CLEF 2008 Workshop*.

S. Rosset et S. Petel (2006). 'The Ritel Corpus - An annotated Human-Machine open-domain question answering spoken dialog corpus'. In *5^{th} Conference on Language Resources and Evaluation (LREC 2006)*.

V. Rus (2006). 'Two Related Lexico-Syntactic Approaches to Entailment'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

E. Saquete, P. Martínez-barco, R. Muñoz et J. L. Vicedo (2004). 'Splitting complex temporal questions for question answering systems'. In *Proceedings of ACL'04*, pp. 566–573.

F. Schilder et B. Thomson McInnnes (2006). 'Word and tree-based similarities for textual entailment'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

N. Schlaefer, J. Ko, J. Betteridge, G. Sautter, M. Pathak et E. Nyberg (2007). 'Semantic Extensions of the Ephyra QA System for TREC 2007'. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC)*.

S. Schlobach, D. Ahn, M. de Rijke et V. Jijkoun (2007). 'Data-driven Type Checking in Open Domain Question Answering'. *Journal of Applied Logic* **5**(1) :121–143.

S. Schlobach, M. Olsthoorn et M. D. Rijke (2004). 'Type Checking in Open-Domain Question Answering'. In *Proceedings of European Conference on Artificial Intelligence*, pp. 398–402. IOS Press.

S. Sekine, K. Sudo et C. Nobata (2002). 'Extended Named Entity Hierarchy'. In *Proceedings of* 3^{rd} *International Conference on Language Resources and Evaluation (LREC'02)*, pp. 1818–1824.

D. Shen et D. Klakow (2006). 'Exploring correlation of dependency relation paths for answer extraction'. In *Proceedings of ACL2006*, pp. 889–896.

J. Suzuki, Y. Sasaki et E. Maeda (2002). 'SVM answer selection for open-domain question answering'. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, pp. 1–7.

K. Séjourné (2009). *Question réponse et interaction*. Ph.D. thesis, Université Paris Sud.

M. Tatu, B. Iles et D. Moldovan (2006). 'Automatic Answer Validation Using COGEX'. In *Working Notes for the CLEF 2006 Workshop*. Springer.

A. Téllez-Valero, M. M. y Gómez, L. V. Pineda et A. Peñas (2010). 'Towards Multi-Stream Question Answering Using Answer Validation'. *Informatica (Slovenia)* **34**(1) :45–54.

M. Tonoike, T. Utsuro et S. Sato (2004). 'Answer validation by keyword association'. In *Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data*, ROMAND '04, pp. 95–103.

A. Téllez-Valero, A. Juárez-González, M. M. y Gómez et L. Villaseñor-Pineda (2008). 'INAOE at QA@clef 2008 : evaluating answer validation in spanish question answering'. In *CLEF*.

A. Téllez-Valero, M. M. y Gomez et L. Villasenor-Pineda (2007). 'INAOE at AVE 2007 :Experiments in Spanish Answer Validation'. In *CLEF*.

L. Vanderwende et W. B. Dolan (2006). 'What Syntax Can Contribute in the Entailment Task'. In *Machine Learning Challenges, Evaluating Predictive Uncertainty, Visual Object Classification and Recognizing Textual Entailment, First PASCAL Machine Learning Challenges Workshop, MLCW 2005*.

L. Vanderwende, A. Menezes et R. Snow (2006). 'Microsoft Research at RTE-2 : Syntactic Contributions in the Entailment Task : an implementation'. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

O. Vechtomova (2010). 'Related Entity Finding : University of Waterloo at TREC 2010 Entity Track.'. In *The Nineteenth Text REtrieval Conference (TREC 2010) Proceedings*.

A. Volokh et G. Neumann (2010). ‘Comparing the benefit of different dependency parsers for textual entailment using syntactic constraints only’. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval ’10, pp. 308–312.

E. M. Voorhees (2002). ‘Overview of the TREC 2002 Question Answering Track’. In *Proceedings of the Eleventh Text REtrieval Conference (TREC)*, pp. 115–123.

E. M. Voorhees et D. M. Tice (1999). ‘The TREC-8 Question Answering Track Evaluation’. In *Text Retrieval Conference TREC-8*, pp. 83–105.

D. Wang, Q. Wu, H. Chen et J. Niu (2010). ‘A Multiple-Stage Framework for Related Entity Finding : FDWIM at TREC 2010 Entity Track’. In *The Nineteenth Text REtrieval Conference (TREC 2010) Proceedings*.

R. Wang et G. Neumann (2007). ‘DFKI-LT at AVE 2007 :Using Regonizing Textual Entailment for Answer Validation’. In *CLEF*.

R. Wang et G. Neumann (2009). ‘Information synthesis for answer validation’. In *Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, CLEF’08, pp. 472–475.

M. Yashar, M. Negri, E. Cabrio, M. Kouylekov et B. Magnini (2009). ‘Using Lexical Resources in a Distance-Based Approach to RTE’. In *Proceedings of the TAC 2009 Workshop on Textual Entailment*.

F. Zanzotto, A. Moschitti et M. Pazienza (2006). ‘Learning textual entailment from examples’. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

F. M. Zanzotto et A. Moschitti (2006). ‘Experimenting a general purpose textual entailment learner in AVE’. In *Working Notes of the Answer Validation Exercise at the Cross Language Evaluation Forum*.