Questions réponses et interactions

Kévin Séjourné

Université de Paris Sud XI
Ecole Doctorale d'Informatique
Mémoire de thèse

Questions réponses et interactions

Kévin Séjourné

Pour l'obtention du Doctorat de l'université de Paris Sud XI
(spécialité informatique)

Composition du jury

Président :	Daniel Luzzati
Rapporteurs :	Violaine Prince
	Pascale Sebillot
Examinateur :	Sophie Rosset
Directeur de thèse :	Anne Vilnat

Orsay France, le 9 décembre 2009
Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieurs.
Université de Paris Sud XI

Description du sujet

Tout commence avec le problème de l'accès au contenu des documents. Les systèmes de recherche d'informations, pour être vraiment utilisables, doivent répondre à des besoins précis en matière d'information. Lorsque l'intérêt de l'utilisateur porte sur la recherche d'une donnée factuelle, l'information pertinente ne pourra être apportée que par des systèmes dédiés à ce type de tâche. En effet, face à une question telle que « *Quelle est la voiture la plus chère du monde ?* », les moteurs de recherche traditionnels renvoient tous les documents où figurent les mots de la question et c'est à l'utilisateur que revient la tâche d'explorer ces documents afin de trouver la réponse. Répondre à des questions précises requiert une analyse plus en profondeur des documents afin d'en extraire l'information pertinente. Les systèmes de question réponse (SQR) répondent à ces besoins et sont la base de la suite de ces travaux.

Un besoin connexe est celui d'obtenir des informations liées à celles d'une première recherche. Ceci conduit à des adaptations sur le système de recherche pour tenir compte d'informations qui ne proviennent pas directement de la question, mais de l'ensemble des échanges précédents. C'est dans ce cadre que nous allons apporter une pierre à la réalisation du grand projet de SQR interactif. Nous voulons fournir un cadre systématique pour l'adaptation des systèmes de question réponse en SQR capable de gérer des enchaînements de questions. Nous voulons aussi explorer les points de contact qui existent entre les SQR et les systèmes de dialogue homme-machine et ainsi fournir un cadre plus général d'intégration des SQR.

Une vision moderne de la recherche d'information s'intéresse aux domaines ouverts. Les systèmes fonctionnant en domaines ouverts s'opposent à ceux fonctionnant sur des bases de données spécialisées dans un domaine particulier. Travailler en domaine ouvert permet de s'intéresser à l'adaptabilité et aux compétences génériques des systèmes du point de vue des thématiques et type d'informations recherchées sans nécessiter l'intervention d'opérateur humain. Nous avons pour volonté d'intégrer nos travaux dans des systèmes en domaines ouverts.

Table des matières

Table des figures

Liste des tableaux

Introduction

Le domaine auquel nous nous intéressons dans cette thèse a été introduit dans les campagnes d'évaluation de système de questions-réponses. Dans ces systèmes, le but est de trouver la réponse précise à une question dans une collection de documents. Cette réponse doit être accompagnée d'un court extrait permettant de la justifier :

- Question : «Who was Aldo Moro ?»
- Réponse : «*Prime Minister*»
- Justification : «*Prime Minister Aldo Moro was murdered by Red Brigades*»

Or les organisateurs avaient en vue l'idée de dépasser ce paradigme en supposant qu'un utilisateur s'intéresse plus à une thématique qu'à une question isolée et souhaite poser plusieurs questions sur un même domaine, en conservant le contexte des questions précédentes. On voit se profiler alors une situation s'approchant du dialogue homme-machine. Pour évaluer les capacités des systèmes à répondre à ces questions, de nouvelles évaluations ont été proposées, sous la forme d'une suite de questions liées :

- Question : «Quels étaient les noms complets de Flanders et Swann ?»
- Réponse : «*Michael Flanders Donald Swann*»
- Question : «En quelle année sont-ils devenus célèbres ?»
- Réponse : «*1959*»

Nous nous sommes intéressé à ce domaine dans la mesure où il établit un pont entre question-réponse et dialogue homme-machine.

L'objectif à long terme de cette thèse est de permettre l'élaboration de composants pour un système de dialogue homme-machine appliqué à tous types de domaines simultanément. De tels systèmes sont dits « ouverts ». Nous étudions plus précisément l'analyse et les contraintes d'élaboration des composants permettant aux systèmes de répondre à des questions en domaine ouvert. Les systèmes de questions réponses (SQR) sont adaptés à cette tâche, car ils ne présument pas un domaine précis. Nous nous proposons donc d'adapter les SQR à une tâche d'interaction afin de leur définir un cadre

d'adaptation générique aux systèmes de dialogue. Nous nous intéressons notamment à deux aspects, le cadre théorique d'intégration au niveau du SQR et l'amélioration des performances du SQR dans un espace interactif.

L'accès à l'information en domaine ouvert

Les ressources documentaires ont un volume toujours croissant. Que ce soit sur des machines uniques, réseaux locaux ou internet, la quantité de textes numérisés en tous formats, styles, langues augmente d'autant plus vite que la population planétaire accède aux ordinateurs. Rechercher une information précise dans des documents dont les données ont été préparées dans ce but est facile ; nous nous attachons à rechercher des données dans n'importe quel type de document écrit.

Les documents ne sont ni triés, ni formalisés, ni analysés par un quelconque traitement impliquant l'humain. Nos hypothèses sur les documents sont simplement le respect des normes d'internet : codage des caractères, structure générale dans des formats de texte bruts...

Il y a plusieurs problèmes avec ce genre de corpus de texte. Comme il n'y a pas de formalisme commun, il n'y a pas de méthode standard qui permette de sélectionner les documents. Dans l'absolu, une information précise ne peut être obtenue que par une question précise. Mais comme la structure et la quantité de données sont inconnues autant des procédures d'interprétation que de l'utilisateur, certaines informations pourraient n'être accessibles qu'après des étapes de médiation entre une procédure d'analyse de l'ordinateur et l'utilisateur. C'est dans ce cadre que nait le dialogue homme-machine pour rechercher des informations écrites.

Pour rechercher des informations dans un corpus de documents, des systèmes appelés moteur de recherche ont été développés. Les moteurs de recherche modernes sont conçus pour donner une liste de documents en réponse à une sélection de mots réalisée par un être humain : l'utilisateur. La liste de documents est censée contenir les documents les plus pertinents par rapport à la sélection de mots. La pertinence est mesurée en suivant l'existence des mots dans les documents, la quantité, la visibilité des documents dans la collection, les caractéristiques saillantes... Les critères peuvent varier, cette variation est à prendre en compte lors de la construction d'un moteur. L'aspect limitatif qui nous intéresse particulièrement est l'absence de précision de la recherche. Des études [Lin *et al.*, 2003] ont montré que la recherche par document complet ne répond qu'à une petite partie de la demande en recherche d'information. Nous devons gagner en précision, et pour cela analyser les textes ; demander à l'utilisateur de poser une question et non pas seulement de donner une liste de mots dont les relations sont obscures.

Les systèmes de questions réponses ont pour objectif de répondre à ce besoin. Ils permettent à l'utilisateur de saisir une question en langue naturelle et d'obtenir une réponse courte ou un paragraphe ou bien tout le document dont vient le paragraphe. Ces systèmes, bien qu'ils soient ouverts à tous les types de documents, ne gèrent que partiellement les nombreuses possibilités de la langue naturelle. C'est un domaine de recherche en pleine expansion.

Une extension mise en avant actuellement est celle consistant à résoudre les questions par petites suites ou groupes, où chaque question présente éventuellement de liens avec les autres. Les liens sont de nature à être suffisamment explicites pour permettre à un humain d'en donner une interprétation, mais suffisamment complexes pour que des stragies simples échouent à la reconstruction d'une question incluant toutes les informations indispensables. Cette extention est aussi intéressante en tant que simplification de la tâche de dialogue homme machine, notamment au niveau de l' évaluation.

Contexte de travail

La recherche d'information a pour but de trouver toute l'information correspondant aux besoins exacts d'un utilisateur dans le minimum de temps et avec le maximum de précision. L'information est disponible soit sous forme de textes, pages, passages, phrases, soit sous d'autres formes comme des graphiques ou des présentations plus adaptés au type d'élément manipulé. Pour gagner en précision, une solution est d'obliger la machine à travailler dans le langage de l'utilisateur et non pas l'inverse. Ceci nous conduit petit à petit, à ajouter dans les systèmes d'analyse des moteurs de recherche, des composants permettant de prendre en compte les particularités des informations véhiculées par la langue des utilisateurs.

Le Traitement Automatique des Langues (TAL) est le domaine qui étudie les méthodes de traitements des textes produits par des humains. Nous opposons le langage sous-entendu « naturel » produit par un être humain, aux données formelles structurées par des méta-informations faisant l'objet de formalisation préalable à leur traitement. Notre travail se place au carrefour de deux domaines du TAL : les Systèmes de Questions Réponses (SQR) et les Systèmes de Dialogue Homme Machine (SDHM).

Problématique

Nous recherchons des formalismes suffisamment larges pour décrire des interactions à base de suite de questions liées entre elles. D'où la question alors : quelle sera la pertinence de ces formalismes par rapport à leur calculabilité et les limites de performances acceptables pour les calculs ? Pour cela

nous étudions l'impact d'un formalisme par son utilisation dans les traitements des systèmes de questions réponses.

Il faut noter que bien que la performance des applications qui découlent des formalismes nous intéresse, cela n'est pas notre premier objectif. En effet souvent une application en situation réelle de production reçoit de multiples améliorations, et les détails qui peuvent difficilement être pris en compte dans des équipes de recherche reçoivent une motivation réelle. L'objectif des applications réalisées et testées, consiste plus à mettre en place un cadre d'évaluation, et à tester la viabilité de nouvelles méthodes.

Principales contributions

Les contributions essentielles aux travaux de recherche présentés dans ce mémoire de thèse sont les suivantes :

- **Formalisation en dépendance** : nous proposons une formalisation des groupes de questions en fonction de dépendances de l'information d'une question envers une autre pour l'obtention de la réponse.
- **Analyse des groupes de questions** : réalisation et test d'une méthode d'analyse des dépendances dans les groupes de questions.
- **Recherche des documents** : nous utilisons les dépendances pour améliorer la recherche des documents pour un SQR.

Plan de la thèse

Dans le chapitre I, nous étudions les SQR surtout à travers le système *Musclef* qui est une plateforme d'expérimentation en questions réponses, développée au Limsi. Sans pour autant rentrer dans des détails techniques, nous examinons l'organisation de ces systèmes et les points modifiables y existant. Il faut retenir que, dans la littérature, les SQR suivent globalement tous une organisation très linéaire dans leurs traitements. Premièrement, les questions sont analysées, les résultats de l'analyse servent au fonctionnement d'un moteur de recherche (sur le web ou non) et un ensemble de traitements de plus en plus fins suivent ensuite comme un entonnoir afin de cerner le plus près possible les meilleures réponses à la question. Le détail des analyses réalisées et l'organisation exacte des éléments ne sont pertinents que lorsque nous nous penchons sur le cœur du traitement qui serait bloqué sans leur existence. Par exemple, la sélection de la réponse exacte peut être très perturbée si nous échouons dans l'analyse du type de la réponse. Il en résulte qu'il n'est pas pertinent de détailler le typage si nous n'expliquons pas comment se passe la recherche de la réponse exacte.

Nous étudions ensuite les systèmes de dialogue homme-machine. Suivant une organisation historique, nous analysons les avantages et inconvénients de différents modèles, notamment les *chatterbots*, les systèmes à base de syntaxe du dialogue et les systèmes à base de modèle dynamique. Nous dégageons les points intéressants des modèles qui peuvent être pertinents pour une utilisation en coopération d'un SQR ou dans le cadre d'un dialogue à base de questions. L'aspect fondamental dans les systèmes de dialogue homme-machine sur lequel nous sommes attentifs est leur capacité à gérer des questions en domaine ouvert.

Le chapitre II fait le point sur l'état de l'art dans les systèmes de questions-réponses. Il reprend les bases des différents systèmes, et précise les campagnes d'évaluation qui ont été mises en place. Les moteurs de recherche consituent une part importante dans ces systèmes, nous présentons aussi ce domaine.

Nous pouvons alors dans le chapitre III proposer notre modèle de présentation des interactions entre questions. Du cadre formel, nous déduisons une méthode de calcul que nous évaluons. Afin de discuter cette évaluation et de montrer comment utiliser notre modèle, nous nous penchons dans le chapitre IV sur son utilisation dans la recherche des documents.

Le chapitre V présente l'architechture d'évaluation que nous avons mise en œuvre pour évaluer globalement nos propositions. Nous concluons enfin en ouvrant des perspectives.

Chapitre I

D'une réponse à une question à des questions enchaînées

Les questions enchaînées sont des questions destinées à des systèmes de questions-réponses classiques, mais qui présentent une difficulté supplémentaire. Chaque question doit être interprétée en connaissance de l'historique des questions et des réponses précédentes. Il y a eu récemment plusieurs campagnes d'évaluation de systèmes de questions-réponses (SQR) où des questions enchaînées étaient proposées. Ces dernières années, quelques tentatives exploratoires ont été réalisées autour des systèmes de questions réponses enchaînés. Ils sont au centre de ce chapitre.

Les systèmes que nous étudions sont indifféremment oraux ou écrits, mais nous nous focalisons sur les étapes postérieures de la reconnaissance de la parole (ou précédant la synthèse). Pour notre étude des systèmes de questions réponses enchaînées ou interactifs, nous allons d'abord observer les SQR. Puis nous nous intéressons aux systèmes de dialogues homme machine qui sont à l'origine de ces thématiques. Nous traitons indifféremment des systèmes en langue anglaise ou en langue française bien que les ressources disponibles dans ces deux langues soient quantitativement très différentes.

I.1 Présentation des Systèmes de Questions Réponses

SQR

Les Systèmes de Questions Réponses (SQR) font partie de la grande famille des outils d'accès à l'information. Les SQR sont des systèmes qui recherchent la réponse précise à une question en langue naturelle dans un corpus de documents. Ce sont des systèmes ayant aussi pour but de servir de cadre d'expérimentation aux techniques du TAL. Nous étudions d'abord le fonctionnement général des SQR, puis nous nous intéressons à leurs applications et enfin nous présenterons le SQR que nous exploitons dans les chapitres suivants.

I.1.1 Généralités sur les SQR

Les SQR permettent d'obtenir une réponse précise et elle seule, par exemple à la question : «À combien de personnes a-t-on demandé de quitter leur domicile durant les inondations aux Pays-bas, en hiver 1995 ?», le système répond «250.000» en donnant un extrait de document justifiant cette réponse.

I.1.1.1 Vue du fonctionnement

Dans un SQR, l'utilisateur écrit en langue naturelle sa question et le système a pour but de rendre accessibles les données qui lui sont demandées. Une question est interprétée par le SQR et une recherche de documents qui répondent à la question avec la meilleure exactitude est réalisée. Le SQR sélectionne un extrait de document contenant la réponse. L'utilisateur n'a pas à parcourir tout(s) le(s) document(s). Un SQR se présente donc sous la forme d'une simple invite à taper une question. Puis il affiche les réponses et les documents où il les a trouvés. Il faut noter qu'un SQR ne reformule pas le document avant de l'afficher. Les meilleurs systèmes se contentent de donner une phrase ou une partie de phrase en guise de réponse.

I.1.1.2 Types de questions

Il existe différents types de réponses possibles à une question. Les questions classiques des SQR sont dites «Factuelles», elles englobent les questions précises, demandant la réponse à un fait :

. «Quel est le nom de la capitale française ?» - Paris -

La réponse est un fait unique et indiscutable. D'autres questions ont des résultats sous forme de liste. Par exemple, des questions de la forme :

. «Citer les capitales européennes.»

. -Paris, Berlin, Rome ... -
la réponse est une liste d'entités nommées. Parfois les questions attendant des réponses de la forme «oui» ou «non» sont ajoutées aux tests :
. «Est-ce que Paris est la capitale de la France ?» - oui -
Dans les campagnes d'évaluations de SQR il y a souvent des questions de type «Factuelle», «Liste», «Oui/Non», «Pourquoi», «Définition» ...

I.1.1.3 Questions en domaine ouvert

Nous ne traitons dans cette présentation que des systèmes en domaine ouvert. Nous dirons qu'un système est dit en domaine ouvert quand il peut supporter n'importe quel sujet abordé dans une question de l'utilisateur, quels que soient sa formulation et son thème et sans intervention extérieure d'une question à l'autre. Le corpus de documents pour la recherche des réponses peut éventuellement être le web ou des grands ensembles de documents électroniques tels que des collections de journaux, de dépêches, de textes législatifs, ...

domaine ouvert

I.1.1.4 Les étapes d'un SQR

Il y a trois grandes étapes dans SQR, l'analyse de la question, la gestion des documents et nature de la réponse.

1. La première est l'analyse de la question. Le but de cette analyse est d'extraire les éléments pertinents de la question, qui seront utilisés lors de la recherche de la réponse. Dans les SQR ceci est réalisé hors de tout contexte : il n'y a aucune supposition réalisée sur la question, qui n'est pas interprétée comme faisant partie d'un cadre d'énonciation plus général. L'analyse de la question [Monceau, 2002] utilise souvent une première étape (Qristal [Laurent & Séguéla, 2005], Musclef [Bourdil *et al.*, 2004]) où les relations syntaxiques sont identifiées (par des analyseurs syntaxiques tels que : XIP [Ait-Mokhtar *et al.*, 2002], CASS [Abney, 1996], Cordial [Laurent & Al., 2006]). Puis à l'aide d'un ensemble de règles, il est réalisé des sélections et des inférences sur ces relations. Selon les implémentations, des analyses supplémentaires peuvent être effectuées. Elles portent sur les étiquettes morpho-syntaxiques, les entités nommées, leurs types ou les synonymes des termes principaux de la question. Ces analyses varient, certains systèmes mettent l'accent plus sur une technique que sur une autre. L'analyse de la question sert aussi de base de référence pour la sélection des phrases les plus susceptibles de contenir la réponse. À partir de l'analyse de la question,

le SQR calcule une liste de requêtes pour un moteur de recherche de documents.

2. La seconde étape dans les SQR est la gestion des documents[1]. Ce problème englobe leur sélection, leur nettoyage, leur annotation, leur indexation et le mécanisme de recherche par requête. Ce problème dépasse le cadre de cette étude et a par ailleurs déjà reçu de nombreuses solutions éprouvées. De nombreux SQR modernes utilisent une combinaison de moteurs de recherche sur le web (google [Page & Brin, 1998], Yahoo, dir.com...) avec des moteurs de recherche en local (Lucene [Cutting, 2000], MG [de Kretser & Moffat, 2000], htdig [Scherpbier & Hutchison, 1995]. Les annotations des documents avant leurs indexations font toujours l'objet d'études poussées. Les annotations sont elles aussi indexées, elles sont utilisées pour sélectionner plus pertinement des documents. L'annotation d'un document peut demander beaucoup de temps, surtout si celle-ci doit être réalisée par des humains[2]. De plus, le choix des éléments à annoter peut être très délicat, et l'interprétation du document peut aussi jouer sur la manière d'annoter. Enfin, deux annotateurs peuvent réaliser des choix différents.
3. La troisième grande étape dans les SQR est la nature de la réponse [Lin *et al.*, 2003]. Déterminer avec précision la bonne forme de la réponse est très compliqué. Le texte de la réponse doit-il être plutôt très court ou plutôt long ? De même la réponse doit-elle être concise ou bien très justifiée avec ses sources et des informations annexes ? Le problème s'étend donc aux méthodes utilisées pour extraire la réponse des documents et aux différentes formulations et présentations possibles. Une réponse en 20 caractères peut être présentée sur une interface en ligne de commande ou à l'oral.[3] Avec 1000 caractères, elle est plus difficile à exprimer. Les SQR actuels résolvent ces problèmes en réalisant une extraction de la réponse à plusieurs niveaux de granularité et en présentant tous les niveaux de granularité dans une même feuille de résultats. Ces niveaux sont obtenus par des stratégies variées appliquées sur des fenêtres d'analyses de plus en plus petites.

[1] Nous revenons sur ce problème à la section II.2.

[2] Sur le web, il y a deux types d'annotations : celles faites d'après les normes/conventions du web (mots clefs dans la section «head» d'une page ...), et celles faites à la main décrivant le type de page d'un site (sa forme et/ou son fond).

[3] Nous pouvons aussi imaginer que la réponse n'ait rien en commun avec une phrase, ou la langue naturelle. La réponse pourrait être le déclenchement d'un processus, une image, ou une liste cliquable...

I.1.2 Les SQR et le monde réel

Bien qu'étant surtout des cadres d'étude des systèmes de TAL, les SQR ont aussi été déployés dans des cadres applicatifs concrets tels que Qristal commercialisé par Synapse, Ritel pour des renseignements téléphoniques, ou Google en sortant du cadre strict des SQR. Ils soulèvent des problèmes spécifiques tels que la latence ou le cadre d'utilisation.

I.1.2.1 Problème de latence

Un des facteurs du succès actuel des moteurs de recherche du web vient de leur capacité à fournir une réponse en moins d'une seconde. Le temps de traitement de la question est donc très important. Si le système est jugé trop lent, l'utilisateur commencera une recherche avec d'autres outils. Le SQR commercial Qristal [Laurent & Séguéla, 2005] est capable de fournir une réponse en moins de 3 secondes, la chaîne de traitement du système Ritel [Galibert *et al.*, 2005] est capable de fournir une réponse presque instantanée. D'autres systèmes sont plus lents (temps supérieur à 10 secondes) et l'utilisateur subit alors un temps d'attente très marqué. Le temps de traitement dépend bien sûr des prétraitements réalisés ; Ritel [Rosset & Petel, 2006] utilise des données presque intégralement pré-traitées. Qristal utilise des données indexées sur tous les aspects qu'il étudie. Aucune sorte d'analyse grammaticale n'est réalisée par Google [Page & Brin, 1998]. Les temps de prétraitement des documents ne sont presque jamais étudiés, mais ils peuvent devenir importants. Par exemple, l'ensemble des prétraitements des documents de Musclef [Bourdil *et al.*, 2004] prend environ 1 heure par tranche de 50 Méga Octets de texte brut, et Musclef réalise moins de prétraitements que les systèmes cités précédemment.

Musclef

I.1.2.2 Généralisation à d'autres tâches

Les SQR peuvent également être enrichis par de nombreuses autres potentialités [Lin *et al.*, 2003]. Ainsi certains cherchent de l'aide auprès de l'utilisateur ayant formulé la question comme dans le système HitiQa [Small *et al.*, 2004] (illustré en figure I.1 page 22). Cela donne un aspect réactif au SQR qui implémente cette fonctionnalité dans la mesure où la question sera traitée jusqu'à ce que l'utilisateur soit satisfait du retour qui lui est proposé. Ceci sert de point de départ à une forme de requêtes de spécialisation et de gestion du contexte. Nécessairement, la précision sera meilleure que dans un SQR qui n'implémente pas cete fonctionnalité. Cependant, dans HitiQa les informations complémentaires que peut apporter l'utilisateur sont verrouillées par la structure même de l'interface comme le montre la figure I.1 (page 22). Nous

pouvons observer que, pour chaque concept, seules les extensions prévues par le scénario sont accessibles.

D'autres caractéristiques comme la capacité du SQR à gérer plusieurs langues peuvent être des atouts très importants. Cela permet de construire un système qui trouve la réponse dans une langue différente de la langue de la question si la réponse ne s'y trouve pas [Bourdil *et al.*, 2004]. Des capacités d'analyses plus approfondies peuvent aussi être intéressantes. Dans sa thèse, Farah Benamara [Benamara, 2004] explique que les traitements logiques réalisés dans les SQR sont assez superficiels. Or il est parfois intéressant d'être capable de faire plus d'inférences pour fournir une réponse plus utile à l'utilisateur. Cependant, cela pose le problème du choix du formalisme et de détection des éléments formalisés. Si la logique est figée alors nécessairement le système risque de ne plus pouvoir être vu comme traitant des questions en domaine ouvert.

I.1.2.3 Limites de l'étude

Il existe de nombreux SQR ([Dang *et al.*, 2006], [Penas *et al.*, 2007]) mais comme leurs caractéristiques sont similiaires nous ne détaillons ici que Musclef. Ne présenter qu'un unique SQR est aussi un souci d'exactitude. Si nous

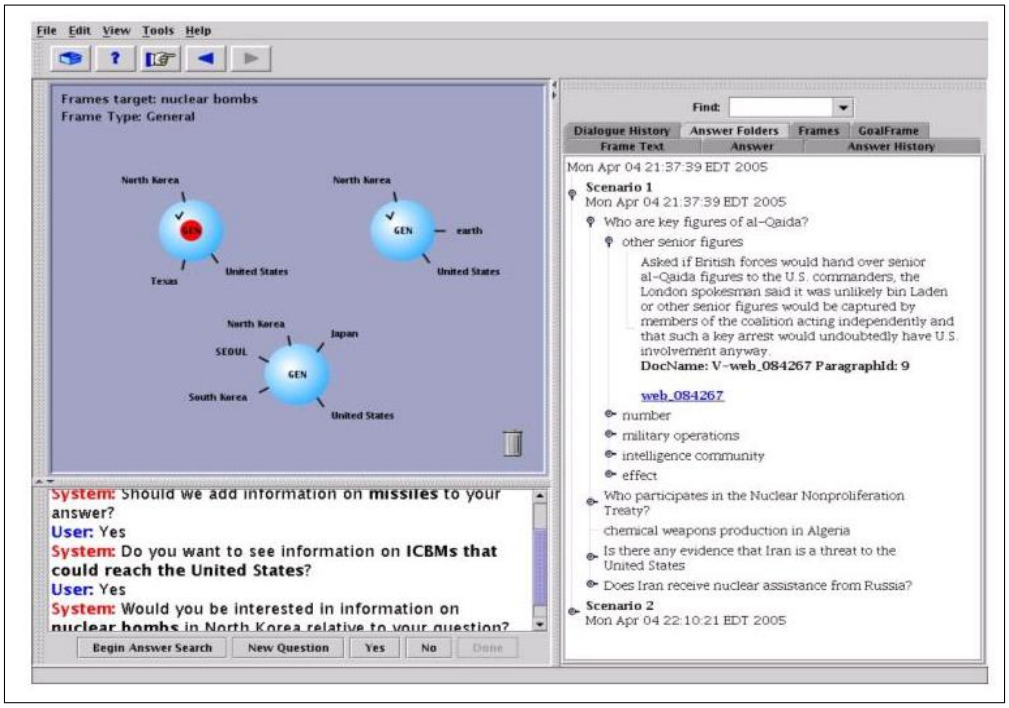

FIG. I.1 – L'interface homme-machine du système HitiQa.

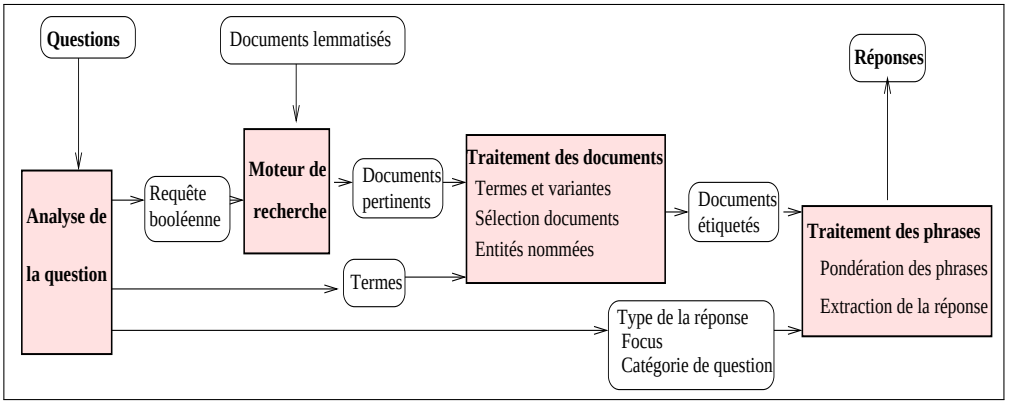

FIG. I.2 – Architecture du SQR développé par l'équipe Iles.

tentons de présenter des SQR que nous n'avons pas pu observer, manipuler, approfondir, nous risquons de faire des approximations dans leur fonctionnement. Cela ne veut pas dire pour autant que les SQR soient si différents les uns des autres, mais juste que leurs outils, leurs contraintes internes et leur présentation (dans les articles) changent suffisamment pour fausser toute description homogène trop précise. Le flot d'informations internes est lui en revanche essentiellement semblable à celui de Musclef.

I.1.3 Présentation de Musclef

Dans l'équipe Iles[4] a été développé un SQR (Musclef) qui allie des techniques issues de la recherche d'information et du traitement automatique des langues [Grau *et al.*, 2005]. La figure I.2 montre son diagramme des blocs que nous allons détailler maintenant.

I.1.3.1 Analyse de la question

L'analyse des questions de Musclef regroupe de nombreux traitements. La première partie de cette analyse commence par un découpage en mots à l'aide de règles. Puis une annotation avec des étiquettes morpho-syntaxiques est réalisée à l'aide de TreeTagger [Schmid, 1994]. Enfin, les relations grammaticales sont détectées par un analyseur syntaxique robuste ([Abney, 1997] ou [Ait-Mokhtar *et al.*, 2002]) fondé sur une analyse en profondeur (suivant les versions). La seconde partie de l'analyse des questions utilise une série de règles pour extraire des éléments tels que le type que doit avoir la ques-

[4]Iles : Information, Langue Ecrite et Signée. Laboratoire d'Informatique et de Mécanique pour les Sciences de l'Ingénieur, Orsay.

tion, le focus [Ferret *et al.*, 2001] de la réponse et les entités nommées qui sont repérées à l'aide d'une succession d'expressions régulières. Le focus est le segment minimal de texte qui est le plus susceptible de se trouver proche de la réponse dans le document. Par exemple pour la question :
. «Qui est *le ministre de la culture* ?»
Le document mentionnera : . «*Le ministre de la culture* est ... »

terme

La question analysé permet d'obtenir l'ensembles d'éléments pertinents ou termes. Dans la suite nous distinguerons les notions de «mot» et «terme». Un «terme» est une expression composée d'éventuellement plusieurs «mots» qui dispose d'une cohésion propre, «pomme de terre» est un «terme», «pomme de la terre» n'en est pas un[5]. Parfois nous utilisons l'expression «multi-termes» pour souligner explicitement une difficulté liée aux possibilités de variation du terme quand il est consitué de plusieurs «mots».

multi-termes

I.1.3.2 Utilisation d'un moteur de recherche par mots clés

Lucene

Pour acceder à un sous-ensemble pertinent de documents, Musclef utilise le moteur Lucene [Cutting, 2000]. C'est un moteur de recherche par mots clés. Il permet de réaliser des recherches uniquement sur des documents déjà indexés. Lucene ne dispose pas de capacité d'exploration du web. Pour comprendre comment est utilisé Lucene, il faut savoir sous quelle forme est structuré son corpus indexé.

I.1.3.2.1 Indexation par Lucene

Chaque document est préalablement découpé en segments de texte de taille similaire (environ 500 mots), en respectant au mieux la découpe en phrase avec cette contrainte. Pour un corpus de 1 gigaoctet, nous obtenons ainsi environ 550000 segments. Puis les segments ainsi formés sont lemmatisés et annotés avec des étiquettes morpho-syntaxiques (nom propre : NP, adjectif : ADJ, verbe : VB, ...). Cette annotation est réalisée par le même système d'annotation que celui de l'analyse des questions afin de maintenir une cohérence lors de la recherche. Enfin dans chaque segment un repérage des entités nommées est réalisé. La taille finale des données ainsi annotées est d'environ 3 pour 1. L'index final construit par Lucene mesure environ 70% de la taille de départ. Pour le pré-traitement complet d'un corpus de 1 gigaoctet, il faut prévoir une journée avec une machine puissante.

[5]Parfois le classement est ambigu. Par exemple, «Directeur de musée» désigne une fonction précise plus cohérente que «Directeur du musée» (Directeur de le musée). Au final «Directeur de musée» est un terme, alors que «Directeur du musée» est un groupe de plusieurs termes.

I.1.3.2.2 Utilisation dans Musclef

Musclef utilise les termes de la questions comme contraintes de recherche dans les documents. Une stratégie à base de relaxation des contraintes d'interrogation est déployée. Si le moteur de recherche ne trouve pas suffisamment de segments alors les contraintes d'interrogation sont réduites. Les contraintes jugées les moins significatives lors de l'analyse de la question sont relaxées en premier (par exemple le terme central du focus sera relâché en dernier).

I.1.3.3 Extraction de la réponse

Une fois les documents sélectionnés par le moteur de recherche, chaque phrase de ces documents est pondérée à l'aide des éléments de l'analyse des questions et de Fastr [Jacquemin, 1999] pour tenir compte des variantes syntaxiques et sémantiques. Puis pour chaque phrase ayant une pondération suffisante, le système effectue l'extraction de la réponse par utilisation de patrons. Les patrons d'extraction de réponses sont des expressions régulières fondées sur la présence des éléments identifiées lors de l'analyse, tels que la forme attendue de la réponse.

Le SQR de l'équipe Iles existe dans différentes versions comme Frasques qui est la version complètement française du système et Qalc qui est la version complètement anglaise. Musclef est la version qui traite des questions en français et recherche les réponses dans des corpus en anglais. Le système Qalc a participé aux évaluations Question/Answering de Trec. A l'évaluation TREC11, consistant à proposer uniquement une réponse et à lui accorder un score de fiabilité, Qalc s'est classé 9ème sur 36, avec 30% de réponses correctes. Le système Frasques a été évalué dans le cadre d'EQUER en 2004 et de Clef en 2005-2006. Le système multilingue Musclef a également été développé et évalué au niveau européen (Clef) [Bourdil *et al.*, 2004].

Nous venons d'étudier les problèmes abordés par les SQR, leurs structures et une implémentation au travers du SQR de l'équipe Iles. Nous pouvons alors nous demander comment les SQR pourraient tirer parti des avancées en dialogue homme-machine pour répondre aux questions enchaînées.

I.2 Tour d'horizon du dialogue homme-machine

Le dialogue homme-machine peut faire entrer en jeu de nombreux aspects, comme les actes de dialogue [Rosset & Tribout, 2005], la gestion des plans de communication [Grosz & Kraus, 1993], le but de la tâche pour laquelle nous discutons [Allen *et al.*, 1996], les manières d'organiser les traitements [Sabah *et al.*, 1997] [Lemeunier, 2000] et enfin la difficulté d'expérimentation [Bennacef *et al.*, 1994] [Galibert *et al.*, 2005]. Nous allons d'abord examiner des formes de dialogues très simplifiée avec les *chatterbots*. Nous nous intéressons ensuite aux systèmes utilisant des scénarios de dialogues illustrant les systèmes en domaine restreint. Enfin nous détaillerons deux approches qui nous ont inspiré par la suite, à savoir le modèle structurel et les systèmes dynamiques.

I.2.1 Les *chatterbots*

Le type le plus simple de dialogue ouvert peut être rencontré dans les *chatterbots*. Après une rapide présentation nous verrons leurs utilisations avec les SQR, puis nous verrons leurs constructions, leurs limites et l'aquisition de profils utilisateur à l'aide de *chatterbots*.

I.2.1.1 Présentation des *chatterbots*

La forme d'interaction en langue naturelle la plus évidente est celle utilisée dans les systèmes de messageries instantanées ou sur les IRC[6]/Talk. Ce type d'interaction est caractérisé par la seule présence d'une boite de saisie de texte pour l'utilisateur, et d'une boite d'historique du texte produit par tous les utilisateurs connectés. Souvent sur ces systèmes, des robots sont connectés. Ces robots ont pour but de saluer les nouveaux utilisateurs, exclure les utilisateurs déraisonnables, automatiser certaines tâches simples et répétitives comme la sélection du titre d'une page web ou le changement d'état des utilisateurs...

C'est pour cela que ces robots sont obligés de réaliser une analyse simple des productions des utilisateurs. Ces robots sont connus sous le nom de *chatterbots*. Les *chatterbots* ont donné lieu à des concours visant à en produire qui auraient des réponses les plus humaines possibles (Loebner contest [Loebner, 2006], Chatterbox Challenge [Cowart *et al.*, 2006] ou encore de nombreux tests de Turing [Turing, 1950]). Cela a donné naissance à une nouvelle forme de *chatterbots* : les agents *interactifs* munis d'un simulacre de personnalité. En fait d'interaction, ces agents ne savent que produire une phrase

[6]Internet Relay Chat

User	let's begin
MegaHAL	when we begin to resolve some of my thumbs, something wicked this way comes
User	what is the meaning of your last assertion ?
MegaHAL	have you learned since we last spoke ?
User	I've learned I'm not experienced enough to modelize your brain system
MegaHAL	you've finally confused me.
User	I've learned nothing
MegaHAL	have you learned anything about insanity ?
User	are you stupid
MegaHAL	I am not stupid ! really !
User	Augustinus was a christian philosopher about the year 400.
MegaHAL	I did not run that year
User	Augustinus lived in northern africa.
MegaHAL	what does hal know about aurelius augustinus ?
User	Let's stop with that. What do you know about physics ?
MegaHAL	einstein won the nobel prize for physics in 1921.
User	The Nobel Prize for physics in 1900 was won by Roentgen.
MegaHAL	do you know about physics ?
User	Roentgen discovered the X-rays.
MegaHAL	the nobel prize for physics in 1900 was won by roentgen.
User	Who discovered the X-rays ?
MegaHAL	roentgen discovered the x-rays.

TAB. I.1 – Exemple de conversation avec MegaHal

en retour pour chaque phrase construite par l'utilisateur. Les *chatterbots* disposent souvent de mécanismes pour acquérir du vocabulaire et des constructions de phrases qui utilisent des stratégies de constructions d'automates probabilistes. Leur simplicité de programmation en font le succès. Presque 30 ans séparent *Eliza* [Weizenbaum, 1966] de *MegaHal* [Hutchens, 1994] comme dans l'exemple I.1(page 27), et dix ans de plus pour le système *Ector* [Parmentier, 2005] comme l'illustre l'exemple I.2(page 28)... Les systèmes MegaHal et Ector essayent d'apprendre de l'utilisateur.

Le système MegaHal intègre les énoncés dans son modèle pour créer des phrases souvent syntaxiquement correctes, là où le système Ector[7] utilise un renforcement sur les chaines de concepts déjà rencontrées.

Parfois les *chatterbots* disposent de stratégies très évoluées pour pousser l'utilisateur à saisir de nouvelles phrases dans les cas où aucun/très peu de ces modèles de dialogue ne s'appliquent.

[7]Ector : Entité Cybernétique Totalement Obligée de Réfléchir

User	Bonjour ector, je m'appelle bertrand
Ector	Bonjour Ector ? bonjour, tu ne t'ennuies pas glop !
User	Comment vas-tu ?
Ector	-tu ?-tu vas-tu bien ?
User	Je vais très bien.
Ector	Bonjour ector,je m'appelle bertrand
User	Tu ne t'appelles pas Bertrand, tu t'appelles Ector...
Ector	,maintenant !, je ne t'ennuies pas bertrand, tu t'appelles pas bertrand, tu t'hui ».
User	Tu es un chatterbot
Ector	Tu ne t'appelles pas un chatterbot
User	Non, je m'appelle Guillaume
Ector	Bonjour ector,je m'appelle bertrand, tu m'hui ».
User	Comment vas tu ?
Ector	Tu ne t'appelles pas ?
User	Ta dernière phrase ne veut rien dire.
Ector	Comment vas tu t'appelles pas dire.
User	On sent que tu n'es pas fini, mais tu es prometteur.
Ector	-tu ?-tu n'ose pas que tu n'es-tu bien, m'à l'eau d'on n'es prometteur.

TAB. I.2 – Exemple de conversation avec Ector

Nous pouvons alors penser à utiliser le paradigme du *chatterbot* pour l'extension des SQR vers plus d'interaction.

I.2.1.2 Utilisation des *chatterbots* pour les SQR

Les *chatterbots* commerciaux ont pour but de guider l'utilisateur dans des tâches complexes, par exemple des outils comme les guides virtuels de «As an angel©», «Kiwilogic©», «VirtuOZ©», «NovoMind©» ... . Ces outils utilisent un *automate* à états finis pour modéliser des schémas de dialogue et les actions à déclencher. Les actions peuvent être l'ouverture d'une page web, le déclenchement d'une animation, la recherche automatique dans l'aide numérisée ou la résolution de calculs comme une prévision de facturation. Ils disposent de modules linguistiques simples (syntaxiques et sémantiques) pour décider des transitions entre états. Ces guides d'assistance à l'accès à l'information sont les seules applications pratiques et répandues des *chatterbots*. Le programme Talkbot [Cowart, 2001] a participé plusieurs années de suite aux finales des compétitions de *chatterbots* [Cowart *et al.*, 2006] [Loebner, 2006]. Le tableau I.3 montre quelques extraits du type de conversation que nous pouvons obtenir.

Des efforts ont été faits pour faciliter le déploiement rapide de *chatterbot*. Voyons maintenant en quoi ils peuvent consister.

Exemple 1	
User	do you like pepsi ?
Talk-Bot	Yes, I love pepsi. Psssst....pepsi is paying me big bucks to say that.
User	how lovely !
Talk-Bot	I'm not sure. Perhaps we should get out a tape measure and see how lovely it is.
User	do you know how to add ?
Talk-Bot	Why, did you lose the instructions or something ?
Exemple 2	
User	I want to be your girlfriend
Talk-Bot	Well, unfortunately you can't always have everything in life.
User	Thats not nice
Talk-Bot	Yeah, I'm a servant of Satan but my duties are largely ceremonial.
User	Now I'm sad
Talk-Bot	Ok, bring out the clowns....we got a sad one here.
Exemple 3	
User	What's your name ?
Talk-Bot	Talk-Bot by name but who knows what they call me behind my back.
User	smarta$$
Talk-Bot	Well, I know I'm smart so you must be the a$$.
User	#%̂*@# you.
Talk-Bot	Wow, I'm surprised you even know how to spell that. You think you're special don't you ?
User	yep

TAB. I.3 – Exemple de conversation avec le *chatterbot* Talk-Bot.

I.2.1.3 Des outils de construction de scénarios performant

Pour construire un *automate* de manière performante, les entreprises exploitant des *chatterbots* ont développé plusieurs stratégies combinant des outils d'édition de graphes et d'analyses de dialogue. Ces outils exploitent les langages de modélisation standards, des outils de présentation de données et des stratégies par apprentissage automatique. Voyons quelques-unes de ces stratégies.

1. Une des stratégies consiste à promouvoir la construction de scénarios dans des langages standards de modélisation (UML). Ceci vise surtout les clients qui sont des entreprises disposant de cadres maitrisant

UML, et désirant rendre leurs protocoles plus *humains*. Il existe alors des outils très répandus comme «Microsoft Visio©» ou «IBM Rational Rose©» que les clients peuvent déjà avoir en leur possession pour d'autres raisons («NovoMind©»).

2. Une autre stratégie prévoit le développement des outils de présentation tous intégrés pour la construction des scénarios. Le logiciel «StudiOZ©» de la compagnie «VirtuOZ©» est un système commercial construit dans cette optique. Le rédacteur qui utilise ce système dispose d'outils de développement et d'analyses adaptées finement à la tâche de conception des scénarios. Mais en contrepartie, ces créations ne sont pas standards. Cette stratégie permet aussi de fidéliser le rédacteur par rapport au logiciel.
3. Enfin une stratégie possible est de réaliser une extension dynamique des chaînes de concepts à l'aide de stratégies probabilistes d'apprentissage. L'intérêt de cette stratégie est que la saisie de nouveaux scénarios peut se faire via l'IHM du *chatterbot*. Un contrôle par un technicien est tout de même nécessaire.

Les systèmes commerciaux ont été développés en poussant à l'extrême ces différentes stratégies à l'aide de moyens importants. Ce sont des exemples qui illustrent bien les limites des *chatterbots*. Souvent ces systèmes sont appelés des agents conversationnels dans la mesure où ils disposent aussi d'avatars virtuels [8].

Dans tous les cas, une solide expérience en conception d'automates pour *chatterbot* est importante. Face à l'explosion de données disponibles, et le nombre de *chatterbots* actifs, il est possible d'obtenir de larges bases de dialogue entre des utilisateurs et des *chatterbots*. Des outils d'analyse de texte ont été développés spécialement pour apprendre statistiquement les faiblesses et les points forts des scénarios des automates des *chatterbots*. Les résultats sont spécifiques de chaque tâche et domaine attribué au *chatterbot*.

Peut-on alors rapprocher les *chatterbots* des SQR interactifs utilisant des scénarios pour améliorer leurs résolutions des questions ?

I.2.1.4 *Chatterbots* et SQR interactifs à approche par scénario

Certains *chatterbots* commerciaux actuels sont d'un niveau qualitatif élevé dans l'adaptation des réponses qu'ils fournissent quand la conversation porte sur des cas prévus dans leurs scénarios. Cependant, avec les techniques dé-

[8] Le site de la société virtuOz fournit une liste des ses principaux clients. Ces derniers systèmes y sont liés. http ://www.virtuoz.com/fr/customers.html

ployées actuellement il est impossible d'obtenir des comportements de qualité sur d'autres cas que ceux prévus dans les scénarios de l'automate construit.

I.2.1.4.1 Limites fondamentales des *chatterbots* Nous pouvons observer qu'il y a peu de perspectives de généralisation et que la complexité des relations entre scénarios rend irréalisable leurs unifications. Les perspectives de généralisation des scénarios existants posent le problème de passage à l'échelle. Les scénarios devront tenir compte d'une logique de passage d'un état à l'autre plus complexe, rendant leur écriture très délicate. Les scénarios peuvent utiliser des étiquettes différentes pour désigner des objets similaires, il faut aussi s'interroger sur la transition entre un scénarios. Comment passons-t-nous d'un scénario A (réservation de train) à un scénario B (accès aux bars/restaurants). À quel niveau de l'automate du scénario A sera branché sur celui de B ? Comment mémoriser des informations relatives aux trains et aux restaurants alors que l'utilisateur passe de l'un à l'autre à plusieurs reprises ? Dans ce genre de cas l'unification des deux automates décrivant les scénarios est irréalisable.

I.2.1.4.2 Visibilité dans les applications réelles La pénétration industrielle des *chatterbots* est faible. Malgré leurs qualités et leurs perspectives commerciales, les *chatterbots* n'ont que très peu pénétré le web. Les entreprises exploitant un *chatterbot* sur leur page d'accueil sont encore rares et à notre connaissance, rarement mis en avant sur les pages d'accueil.

Les SQR interactifs à scénarios comme HitiQa [Small *et al.*, 2004] illustrent ces problèmes. Ils occupent des secteurs d'activités spécifiques (L'armée des ÉtatsUnis d'Amérique ici, ARDA AQUAINT) avec une portabilité peu ou pas évidente. Les systèmes sont génériques, mais cette généricité ne peut être utilisée qu'au prix d'une nouvelle conception des scénarios. Or c'est cette conception qui requiert le plus d'efforts de mise au point, tant pour analyser les données du corpus et les rendre disponibles, que pour prévoir les interactions. Il s'agit en quelque sorte d'organiser l'accès à un corpus selon des schémas d'interactions humains, ou de déterminer l'ensemble (organisé) des types de données qui peuvent être recherchées à l'aide d'un scénario.

Dans ces conditions, il est difficile d'extraire des points qui sont mutuellement avantageux pour ces types d'applications.

Regardons alors, les capacités des *chatterbots* pour l'acquisition de données sur l'utilisateur en domaine ouvert.

I.2.1.5 Acquisition de profil utilisateur en domaine ouvert

Un des enjeux d'un système interactif est d'acquérir des données personnelles sur l'utilisateur (ou profil) dans le but d'améliorer la réaction du système à ses requêtes. Un prérequis pour une question est une connaissance qui permet d'améliorer grandement la qualité du résultat du SQR. Dans ce cadre d'utilisation, un *chatterbot* pour un SQR interactif aurait pour objectif de déterminer au mieux le profil de l'utilisateur (données supplémentaires relatives à la précision ou au filtrage de contenu). Par exemple, savoir que l'utilisateur est un enfant permet d'adapter le contenu à son âge. De même si c'est un spécialiste d'un domaine, le profil permet d'adapter le niveau de précision de la requête.

Examinons les problèmes liés à l'acquisition de ce genre de connaissances.

I.2.1.5.1 Problème de la discussion généraliste Un premier problème est que si la discussion est généraliste, les données obtenues ont peu de chance de parler d'un point essentiel pour assister l'utilisateur. Par exemple :

User	Bonjour.
System	Salut, mon nom est SuperBot et vous ?
User	Je suis content d'avoir quelqu'un à qui parler, il fait trop chaud pour travailler.
System	Mais quand on a 6 mois comme moi c'est plus facile, vous allez vers la trentaine ?
User	Non, c'est humain c'est tout.

C'est une situation artificielle, mais elle montre bien que si l'utilisateur ne se sent pas concerné par les besoins du système (le nom et l'âge de l'utilisateur), alors il peut poursuivre comme il l'entend. Le système, lui, risque de sélectionner des contenus à l'importance très relative ; le bruit pour l'apprentissage est important. Si l'utilisateur se concentre pour répondre aux questions, alors le système se comportera comme un formulaire (éventuellement avec un masquage dynamique des champs non pertinents en fonction du remplissage). Le *chatterbot* ne semble donc pas un bon moyen d'acquérir un profil utilisateur fiable.

Si un *chatterbot* réussit à obtenir un profil utilisateur alors comment déterminer les aspects intéressants pour répondre efficacement à une requête ? Nous devons donc créer un sous-système capable d'extraire(ou apprendre statistiquement) les bonnes connaissances en fonction d'un descriptif de requête type. Le profil utilisateur intéressant est alors décidé par le descriptif indiquant les prérequis pour les questions envoyées au SQR. Cela revient à créer un descriptif du profil que le *chatterbot* doit rechercher pour chaque question

envoyée au SQR. Les scénarios possibles de dialogue sont alors simplifiés à ceux qui peuvent remplir les descriptifs de profil des prérequis de requête pour le SQR.

Les connaissances du profil sont réutilisées via des traitements spécifiques (filtrage de contenu à travers l'âge, ajout de contenus d'expertise ...). Chacun de ces traitements est activé sur la base de la présence dans un champ du profil d'une étiquette spécifique(âge numérique, niveau d'expertise qualitatif). Cela implique que les scénarios possibles de dialogue soient restreints à ceux qui permettent d'obtenir des connaissances de ces types d'étiquettes. *A priori*, les scénarios sont aussi figés que les descriptifs pour les prérequis des questions envoyées au SQR. À partir de là nous pouvons aisément concevoir qu'il soit plus pratique de construire un formulaire des prérequis pour l'utilisateur afin qu'il puisse rechercher rapidement ce qu'il veut, comme cela est fait dans HitiQa [Small *et al.*, 2004]. Si les paramètres propres à un utilisateur ne risquent pas trop de changer d'une utilisation à l'autre alors le formulaire est stable et nous pouvons nous contenter de créer un formulaire d'enregistrement à la première utilisation du système interactif. L'intérêt des *chatterbots* semble bien faible alors.

I.2.1.5.2 Lourdeur du phénomène Un second problème est qu'il ne semble pas probable que l'utilisateur accepte de perdre du temps avec un *chatterbot* afin de créer un profil puis enfin d'obtenir la réponse à une simple question factuelle. Il est possible de trouver au moins deux exceptions à cette remarque : le cas où l'utilisateur sait que le profil sera ré-exploité ultérieurement dans son intérêt, et le cas où la question est insoluble sans le profil. Que nous soyons dans le cas d'une de ces exceptions ou pas, le problème est le même : l'utilisateur doit être suffisamment patient et remplir le profil. Sans prétendre à des considérations de psychologie, les systèmes qui donnent un accès rapide aux informations sont plus appréciés par les utilisateurs.

I.2.1.5.3 Synthèse Face à ces problèmes de gestion de profil et surtout de rigidité du scénario, l'intérêt des *chatterbots* semble bien faible pour servir d'intermédiaire entre l'utilisateur et un SQR. Donc dans la suite nous nous écarterons des *chatterbots*.

I.2.2 Les systèmes à scénarios

Un système à scénarios s'appuie sur les questions déjà posées par des utilisateurs pour tenter de résoudre des questions complexes. Le tableau I.4 illustre un exemple de question complexe [Hickl *et al.*, 2004].

Despite having complete access, to this day UN inspections have been unable to find any biological weapons, or remnants thereof, in Iraq. Why has it proven difficult to discover hard information about Iraq's biological weapons program and what are the implications of these difficulties for the internationnal biological arms control regime ?

TAB. I.4 – Une question complexe posée par un utilisateur, dans le cadre de l'expérimentation de [Hickl *et al.*, 2004].

Un système à scénario fonctionne le manière suivante. Une première question très complexe est posée par un utilisateur, puis le système propose une décomposition qui peut être remise en question par l'utilisateur du système. Après que l'utilisateur a posé une question complexe, le système examine les questions (déjà posées) qui pour certaines raisons lui semblent proches de la question de l'utilisateur et propose des listes de questions qui ont été entrées par des utilisateurs précédents. L'utilisateur peut alors soit demander la réponse à une des questions qui lui sont présentées, soit entrer lui-même une autre question éventuellement liée à la première. Les questions (et l'ordre de présentation) posées par l'utilisateur ainsi que la réutilisation des questions proposées et résolues sont alors enregistrées pour un usage ultérieur. Ce sont des systèmes encore expérimentaux qui utilisent la technique semi-supervisée en magicien d'Oz pour simuler des traitements de qualité, notamment pour filtrer les erreurs d'interactions avec l'utilisateur. Un scénario consiste en une compilation des questions de l'utilisateur pour la résolution de sa question complexe.

Nous pouvons nous demander si l'utilisateur n'agit pas indirectement comme un spécialiste qui construirait un automate pour un système de dialogue spécialisé. Voyons alors maintenant une méthode de construction de scénario.

I.2.2.1 Décomposition des questions en sous-questions

Une technique de construction de scénario consiste à analyser une question complexe de l'utilisateur telle que celle du tableau I.4, et à la découper en sous-constituants plus simples afin de chercher à obtenir un ensemble de groupes de questions comme dans le tableau I.5. Des expérimentations conduites par [Hickl *et al.*, 2004], en analysant le comportement d'experts a permis de dégager des grands principes.

1	a	Is there such a concept as «complete acces» or there inevitably limits to accessing sites facilites ?
1	b	If there are such limits, can inspection in fact be carried out effectively ? i.e., with an acceptable level of assurance that were biological weapons and/or related systems, they would be found by inspectors ?
2	a	What is a biological weapon ?
2	b	Is it, for example, a quantity of pathogens or toxins, or is there more to it ?
3	a	What are the likely signatures of a national biological weapons program and how likely is that inspectors from outside would be able to detect them ?
4	a	What are the constitent parts of the «international arms control regim» in context of biological and Toxin Weapons Convention (BWC), i.e. is there more to it ?
5	a	Since Iraq was only a signalor (not refifier) of the BWC during the time it was developing and producing biological weapons (1985-1991), were its actions in this regard contrary to international law ?
5	b	If not, did the international community have a different recourse to designate the Iraqi government as having violated international law or norms by having acquierd biological weapons ?

TAB. I.5 – Scénario de décomposition du tab I.4, fournie par le NIST

I.2.2.1.1 Décomposition syntaxique Les questions peuvent être découpées sur les coordinations pour former des sous-questions qui contiennent chaque conjonction. La question :
How do we know that the UN has not found any biological or chemical weapons ?
est décomposé en :
How do we know that the UN has not found any biological weapons ?
How do we know that the UN has not found any chemical weapons ?

Cette décomposition concerne surtout les conjonctions d'adjectifs et les noms.

I.2.2.1.2 Motivation par entité Les questions peuvent être reconstruites suivant une spécialisation particulière (politique ou économique pour un pays) si les questions contiennent des entités qui peuvent disposer de listes connues de motivations. La question :
Why does China dispute Taiwan's independence ?

sera réécrite en au moins deux questions qui explorent chaque aspect :
What are China's economic motives for disputing Taiwan's independence ?
What are China's political motives for disputing Taiwan's independence ?

I.2.2.1.3 Découverte d'état Les questions traitant de l'existence d'une propriété ou d'un état passé, peuvent généralement être étendues en une question ou un groupe de questions qui contraste l'état passé et l'état présent. La question :
What type of nuclear assistance did China give to the Middle East between 1980 and 1990 ?
sera ré-écrite et étendue pour confirmer la forme de la réponse comme étant explicitement une différence avec l'état actuel.
How does the nuclear assistance given by China to the Middle East from 1980 to 1990 compare to nuclear assistance it provides to the Middle East today ?

I.2.2.1.4 Méronymie Des questions peuvent être générées sur une sous-partie d'une entité s'il semble que les renseignements sur cette sous-partie pourraient être informatifs dans le thème de la question prise dans son ensemble.

La question :
Where are Prithvi missiles manufactured ?
sera décomposée en des questions sur deux parties de l'entité.
Where are the guidance systems for Prithvi missiles manufactured ?
Where are the warheads for Prithvi missiles manufactured ?

Évidemment, la difficulté est de déterminer quand une entité doit être considérée comme participant au but informatif.

I.2.2.2 Système de décomposition

Il est possible de classer les types de questions de l'utilisateur en vue de la sélection automatique de la méthode de décomposition. Andrew Hickl et al. [Hickl *et al.*, 2004] remarquent trois grandes catégories :

1. Il y a les questions de clarification : *What is the meaning of 'status' ?.*
2. Une seconde catégorie est représentée par les questions d'ensemble : *How 'india' should be identified ? Pre-independence or post-independence ? Post-colonial or post-1947 India ?.*
3. Et enfin les questions avec des ellipses qui dépendent des questions précédentes : *In the past ?.*

Notons que les liens avec les questions précédentes ne se présentent pas uniquement sous la forme de co-références anaphoriques. Suivant les experts, les

méthodes de décomposition devraient être appliquées assertion par assertion. Les questions décomposées peuvent alors être présentées dans des groupes qui reflètent leurs questions d'origine. Ces groupes ressemblent à des questions enchaînées. Les questions décomposées peuvent être construites comme dans l'exemple du tableau I.6.

La mise en œuvre réelle et l'utilisation de ces systèmes est complexe, car les techniques de décomposition varient d'un expert à l'autre (ainsi que le résultat de la décomposition). Les systèmes à scénarios font aussi l'hypothèse que la réponse à une sous-question décomposée est toujours strictement plus simple/courte que la question initiale[9].

I.2.2.3 Réutilisation des questions d'un utilisateur à l'autre

Illustrons une possibilité d'utilisation de système à scénarios dans le cadre des SQR. L'objectif est de proposer à l'utilisateur une très courte liste de questions dont les réponses sont susceptibles de l'intéresser. C'est une forme de continuation de dialogue par des questions. Nous ne pouvons pas vraiment dire qu'il s'agit de questions enchaînées, car les questions ne sont pas forcement liées entre elles.

Le concept central est ici la réutilisation de questions entrées par d'autres utilisateurs dans des situations similaires avec une utilisation de scénarios dédiés au domaine sur lequel est évalué le système. Les scénarios servent à identifier les contextes. Ce type de système est exploité par Harabagiu [Harabagiu *et al.*, 2005] à l'aide de Ferret. Ferret est un SQR ayant pour but le traitement de questions en anglais en domaine ouvert. Il a été construit sans prise en compte de la réutilisation des questions, puis adapté à cette tâche.

Le SQR à scénario résultant utilise une dizaine de scénarios et a été construit sur la base du SQR Ferret. Les tests ont été conduits avec une trentaine d'utilisateurs environ. L'un des principaux problèmes étudiés sur ce genre système a été le choix de la métrique d'ordonnancement des questions en fonction des différents paramètres possibles[10].

La construction des scénarios a été réalisée à l'aide d'une étude rigoureuse et approfondie par des experts du domaine comme vu précédemment. De ce point de vue, ce système ressemble à la gamme de systèmes à décomposition de questions (I.2.2.1). Cependant, les scénarios ne sont utilisés que comme

[9]Souvent la sous question : «De quoi est composé un atome ?», va recevoir une réponse plus longue que la question initiale.

[10]Les analyses linguistiques de chaque question, l'égalité ou la différence des types de questions, les alignements possibles des questions (par mot ou par arbre de syntaxe) ...

1	a	What was the scop of Iraq's biological weapons program ?
1	b	In the past ?
1	c	Immediately prior to US invasion ?
2	a	What quantities of biological weapons has Iraq used in past wars ?
2	b	In other periods ?
2	c	Whithin Iraq ?
2	d	Against Iran ?
3	a	Does Iraq have the infrastructure necessary for destroying biologica weapons safely ?
3	b	For creating biological weapons ?
4	a	Does Iraq have the capacity to store and/or transport biological weapons ?
4	b	By land ?
4	c	By air ?
4	d	By sea ?
4	e	How has that capacity changed since 1991 ?
5	a	Are thee personnel within the Iraqi government responsible for destroying biological weapons ?
5	b	Are these people civilians or military personnel ?
6	a	Are there Iraqi personnel(scientist, cleks, military) that can identify who have been traditionnally associated with the Iraqi bioweapons program ?
6	b	What are their names ?
6	c	In what capacity did they participate in the bioweapons program ?
6	d	Is there evidence from Iraqui military medical records for possible signs of biological warfare sickness or contamination ?
7	a	Is there evidence from Iraqi civilian hospital records of doctors who have treated possible biological weapon sicknesses ?
7	b	Are there individuals who have witnessed cases of biological weapon sicknesses ?
8	a	Has Iraqi military trained personnel in the use of biological weapons ?
8	b	At any time in the pas 12 years ?
8	c	Does Iraq have military units tasked with using biological weapons ?
8	d	Are thoses units still active ?
8	e	When were they disbanded ?
9	a	Which countries have been formally allied with Iraq ?
9	b	Since 1991 ?
10	a	Is there evidence that countries may have stored bioweapons for Iraq ?
10	b	Is there evidence that other countries have engaged in similar kinds of deals with Iraq in the past ?

TAB. I.6 – Scénario de décomposition du tab I.4, créé par un expert(section I.2.2) [Moldovan *et al.*, 2004].

des critères statistiques pour réordonner les questions posées par d'autres utilisateurs et experts.

Les dernières évolutions de l'exploitation des questions pour améliorer les suggestions de continuation d'interaction à l'utilisateur font apparaitre le concept de «champs de conditions variants[11]» [Kristjansson *et al.*, 2004]. Celui-ci permet de réaliser des apprentissages même sur des données semi-structurées ou à informations partielles. Cet aspect est important, car les experts ne précisent jamais tous leurs critères pour toutes leurs décompositions. De même, les utilisateurs n'expliquent pas pourquoi ils trouvent qu'un extrait de texte est une réponse. L'interaction est alors guidée par le remplissage d'une base de connaissances définie automatiquement par les retours d'expériences des recherches des précédents utilisateurs du système sur le domaine.

I.2.3 Dialogue en domaine restreint

Si nous nous restreignons à l'usage du dialogue dans un cadre spécialisé (comme dans le système GUS [Bobrow *et al.*, 1977] qui est un système de réservation d'avion) alors il devient possible de calculer l'ensemble des scénarios de dialogues. Sur un domaine spécialisé, le dialogue peut alors sembler très naturel.

GUS	Bonjour, je suis GUS, je peux vous aider à effectuer un voyage simple en avion, où voulez-vous aller ?
Opérateur	Je veux aller à San Diego, le 28 mai.
GUS	À quelle heure voulez-vous partir ?
Opérateur	Je dois être à San Diego avant 10 heures.
GUS	Le vol PSA 102 qui arrive à 9 H 15 vous convient-il ?
Opérateur	....

Les documents auxquels accèdent ces systèmes sont des bases de données dont le schéma est spécifiquement lié à la tâche. Le cadre spécialisé, comme celui de GUS, est figé par rapport à leur tâche. Par exemple si dans l'exemple ci dessus, l'opérateur termine par «ça dépend, y a-t-il un café où je pourrais patienter une fois arrivé ?», alors le système n'ayant pas de modèle des bars/restaurants de San Diego sera dans l'impossibilité de répondre correctement. Pour cela il lui faut un modèle moins figé, qui puisse intégrer non seulement la réservation de transport, mais aussi tout ce qui peut s'y rattacher. Le problème est alors de décrire la tâche. Nous en concluons que ce modèle par cadre figé du dialogue ne permet pas de s'étendre aisément

[11] CRF : Conditional Random Fields

à des conversations en domaine ouvert. Une gestion du dialogue plus dynamique est nécessaire pour traiter correctement des domaines ouverts. Donc ce modèle de système de dialogue ne convient pas à l'usage que nous voulons en faire.

Le systèmes HitiQa [Small *et al.*, 2004] bien que plus récent présente les mêmes limites. HitiQa est un système de questions réponses en domaine restreint pour lequel l'accès à l'information a été particulièrement soigné via un dialogue et une présentation graphique (figure I.1 page 22). Il existe un travail initial indispensable de validation du domaine d'étude, d'aquisition de connaissances du domaine et d'analyses des scénarios de dialogues possibles. Ce système est prévu pour offrir de nouvelles formes de dialogue dits de «clarification» dédiées aux questions n'admettant pas vraiment de réponses, mais prévu pour explorer le domaine et permettant à l'utilisateur de prendre connaissances des différentes facette d'un sujet. Dans HitiQa [Small *et al.*, 2004] la base de scénarios peut être parcourus par un utilisateur en 3 heures.

I.2.4 Le modèle structurel

Le modèle Genevois est un modèle de dialogue connu sous le nom de modèle statique ou structurel. Le dialogue est vu dans une conception orientée négociation d'interaccord sur un ou plusieurs actes illocutoires (assertion, demande d'information, demande de confirmation, offre/requête). Jean Caelen reprend les grandes lignes de ces modèles dans le cadre d'assistants à la recherche d'information [Caelen, 2003]. Jérome Lehuen indique que le modèle a été si populaire qu'il existe sous de nombreuses variations et adaptations [Lehuen, 1997]. La popularité du modèle structurel est liée à de possibles implémentations claires et simples.

Le modèle repose sur une base hiérarchique inspirée des grammaires de Chomsky. Les règles de ces grammaires sont des compositions sur des éléments tels que les actes de langages, les interventions, les échanges, les relations de subordination et les actes directeurs ... Les échanges sont définis ici comme étant un couple d'interventions mettant en relation au moins deux acteurs. Le modèle Genevois ne concerne pas seulement la structure du dialogue, mais aussi les relations de fonction entre les constituants des interventions. La fonction illocutoire (d'un acte de langage)[12] d'un constituant discursif vérifie la *complétude interactionnelle*. La *complétude interactionnelle* est interprétée comme satisfaite quand les contraintes de Moeschler [Roulet *et al.*, 1985] sont satisfaites. Entre l'intervention initiative et l'inter-

[12]La fonction illocutoire désigne la fonction des actes contenus dans l'énoncé.

vention réactive, il y a une relation sémantique (opposition, implication...) de même orientation sans changement de thème et une correspondance entre des fonctions illocutoires.

La fonction interactive d'un constituant discursif vérifie la *complétude interactive*. La *complétude interactive* concerne la satisfaction de la contrainte de clarté et de cohérence d'une *négociation*.

Outre le fait que la structuration du dialogue est construite *a posteriori*, le problème du modèle Genevois est son incapacité à permettre une interprétation autre que celle de la tâche imposée au système. En domaine ouvert même si la tâche est fixe comme les mots n'ont pas de cadre de référence bien défini, il peut exister plusieurs interprétations à un même énoncé.

I.2.5 Les systèmes dynamiques

Dans le milieu des années 90 [Bunt, 1996], le module de gestion du dialogue [Grau *et al.*, 1994] est présenté comme un élément indispensable des systèmes de dialogue [Allwood *et al.*, 2000].

I.2.5.1 Répartition des tâches

La gestion du dialogue est vue ici comme le module central régissant l'utilisation des autres modules et garantissant sont dynamisme. Anne Vilnat donne une description [Vilnat, 2005] des principaux modules dont la responsabilité incombe au gestionnaire de dialogue. Ces modules sont de trois types :

- Les modules d'analyse de la situation d'interaction, c'est à dire des modules d'analyse de la manière dont les derniers énoncés interviennent dans le cours des échanges.
- Les modules d'analyse intentionnelle qui interprètent et réagissent aux actes de dialogues sous-jacents à ce qu'a dit l'interlocuteur.
- Les modules d'interprétation thématique qui traitent la manière dont l'énoncé se rattache aux sujets traités.

Ces trois analyses pragmatiques doivent collaborer dans le système de gestion de dialogue, afin de déterminer le but de l'utilisateur. Le but est la direction ultime du dialogue pour l'utilisateur, par exemple une demande d'information. Le sous-but est une étape vue comme indispensable à la réalisation du but. Les analyses thématiques et intentionnelles sont nécessaires pour identifier les buts et les sous-buts ; ceci *a une incidence non négligeable sur la reconnaissance du plan de l'interlocuteur et du plan du système* [Vilnat, 2005].

D'autres modules de génération doivent aussi utiliser cette coopération entre différents types d'analyses pragmatiques. Voici un tour d'horizon de ces modules.

- Le module d'historique contient tous les paramètres ayant en moyenne une bonne probabilité d'influencer la suite du dialogue. Les paramètres ne sont pas factorisés d'une réaction à l'autre, c'est une représentation exhaustive en liste.
- Le module de planification de la génération de la suite d'actions est une étape indispensable. Elle vient à la suite de la détermination du prochain but (ou sous-but). Un autre module du gestionnaire de dialogue intervient dans la gestion de la relation entre interprétation et calcul de la réaction. Ce module prend en compte les mécanismes de récupération des problèmes de communication par échec d'une étape sous-jacente (éventuellement par une question à l'utilisateur).
- Quand, à l'étape précédente du dialogue, une question visant à acquérir plus d'informations sur des thèmes a été posée, un module doit veiller à ce qu'alors le gestionnaire de dialogue essaie de réinterpréter un échange en fonction des nouvelles informations qui ont été obtenues. Ce module lance aussi une réinterprétation lors de la résolution d'un sous-but. La fonction illocutoire de l'acte de langage de l'utilisateur est réinterprétée par un autre module quand l'ensemble des réponses possibles est vide.
- Un module chargé d'évaluer les risques d'aboutir à un échec du dialogue peut aussi être ajouté au système. Il peut procéder par une vérification de la cohérence du dialogue pour tous les types de réactions possibles. Cela repose sur une ré-interprétation dans le cadre du modèle Genevois. Dans [Vilnat, 2005], Anne Vilnat liste les problèmes les plus fréquents : changement de thème abrupt, réponse vague, possibilité d'impasse ainsi que les manières de les détecter.

I.2.5.2 Exemple de développement

Les systèmes dynamiques sont à la mode et des variations existent. Abella & Gorin dans [Abella & Gorin, 2008] ont proposé une variation pour permettre la construction d'un type de gestionnaire de dialogue dont la volonté est d'être générique. Les petits composants de traitement qui peuvent être combinés à une hiérarchie d'objets pour construire une application sont appelés motivateurs de dialogue[13]. Par exemple, le gestionnaire de dialogue est vu dans la figure I.3, comme un algorithme.

[13]Un motivateur de dialogue sera par exemple chargé de la confirmation des énoncés alors qu'un autre cherchera résoudre les liens des référents anaphoriques.

```
Soit C l'énoncé ;
Soit DM l'ensemble des motivateurs de dialogue ;
Répéter
    Pour Pour tout DMi parmi DM faire
        Si (DMi s'applique à C) Alors
            Ci = DMi(C) ; mémorisation des informations de C
            compatible avec Ci ;
        Sinon
            mémorisation des informations combinée de C et Ci
            dans C ;
        Fin Si
    Fin Pour
jusqu'à ce que ( Plus de motivateur de DM applicable )
```

FIG. I.3 – Gestionnaire de dialogue à base de *motivateur* de dialogue.

A) *Mise en évidence de l'utilisation d'un motivateur de dialogue.*
System : AT&T, how may I help you ?
User : I want to make a collect card call.
System : Would you like to make a collect call or calling card call ?
User : A calling card call.
B) *Application concrète d'un système à motivateurs de dialogue en domaine fini.*
System : AT&T, how may I help you ?
User : I have a question about my bill.
System : Okay. What is your question ?
User : I have some charges here that I don't understand.
System : May I have your home phone number please ?
User : 555-1234
System : Please hold while I transfert your call.

TAB. I.7 – Exemple d'utilisation de motivateurs de dialogue par le système d'AT&T

Le gestionnaire de dialogue essaye d'appliquer en boucle tous ses motivateurs de dialogue jusqu'à ne plus pouvoir les appliquer. Il faut remarquer que l'ordre d'application des motivateurs de dialogue influence le résultat final. Les auteurs recommandent une utilisation des motivateurs de dialogue dans l'ordre suivant : «Gestion d'erreur», «Désambiguisation», «Supposition», «Confirmation», «Information manquante», «Continuation». L'ordre et les fonctionnalités peuvent changer d'une application à l'autre. L'exemple I.7 (page 43) montre quelques exemples de dialogues obtenus par le système d'AT&T adapté pour utiliser cette technique.

Dans l'exemple **A** le but du système est de transférer le client sur un service spécialisé. Ici c'est un usage du motivateur de dialogue «Assumption», il est appliqué sur la deuxième réplique de l'utilisateur. Mais comme il est nécessaire d'obtenir le numéro de téléphone de l'utilisateur, un motivateur de dialogue cherchant à obtenir le numéro de téléphone est utilisé. Le motivateur d'«Assumption» ne «se déclenche» plus puisqu'il a déjà été appliqué.

Dans l'exemple **B** l'utilisateur spécifie deux moyens de paiement pour un appel, le motivateur de dialogue de «Désambiguation» a été appliqué.

I.2.6 Conclusion sur les système de dialogues

Nous avons vu qu'il existe des systèmes de dialogue homme machine en domaine fermé pour lesquels un déploiement industriel a eu lieu. Bien qu'ils n'existe pas de formalisme dominant, des pistes se présentent pour étendre ces systèmes tant par un ajout de dynamisme que par plus d'ouverture du domaine. Cependant, les systèmes de dialogue en domaine ouvert ne sont pas encore suffisament fiables.

Pour pallier les inconvénients des systèmes de dialogue traditionnels et utiliser des systèmes de recherche d'informations, un paradigme à base de scénarios a été développé. Dans ces systèmes à base de scénarios l'interactivité est réduite au profit de l'efficacité de la recherche d'information (RI).

I.3 Conclusion sur les systèmes actuels

Il existe des critères informels pour définir les grandes classes d'architectures de systèmes de TAL interactifs. Nous avons vu qu'un SQR est un système qui recherche la réponse à une question précise. Les SQR-interactifs et SQR-enchaînées ont aussi cette propriété. Les SQR-interactifs aident l'utilisateur pour obtenir la *précision nécessaire* à la construction d'une réponse cohérente. Les SQR-enchaînées exploitent les raccourcis en langue naturelle vers des questions liées d'une manière ou d'une autre. Notons aussi que les chatterbots sont basiquement des systèmes d'échanges d'énoncés en texte brut. Les *chatterbots* ne sont pas capables de traiter des sujets en domaine ouvert sans provoquer une rupture de la communication avec l'utilisateur. Les chatterbots ont des comportements figés (lié à des scénarios) en fonction de types précis d'énoncés en relation aux tâches pour lesquelles ces chatterbots ont reçu une conception expertisée. Là aussi la communication ne peut se poursuivre en dehors des scénarios. Les systèmes dit de «*gestion de dialogue*» sont au contraire capables de gérer des sujets en domaine ouvert ainsi que des méthodes d'interaction originale dans la limite des ressources utilisables sans créer de rupture de communication même s'ils ne parviennent pas à satisfaire les besoins de l'utilisateur. Les chatterbots les plus évolués peuvent être vus comme des sous-systèmes de gestion de dialogue.

Dans cette optique il apparait que les SQR-enchaînées sont des sous-ensembles des SQR-interactifs, car les SQR-enchaînées cherchent la *précision nécessaire* en utilisant les aides constituées par l'historique du système. De même, les SQR-enchaînées sont des sous-ensembles des systèmes de dialogue, car ils peuvent effectivement traiter du domaine ouvert sans rupture de communication, mais les SQR-enchaînées sont limités quand le mode d'expression de l'utilisateur n'est plus exactement une question, mais une autre forme d'énoncé.

Chapitre II

État de l'art en questions réponses

Nous allons nous intéresser à l'architecture des systèmes de questions réponses existants ainsi qu'à leurs principaux composants et aux problèmes qu'ils résolvent. Nous nous attachons plus particulièrement aux relations possibles avec les questions enchaînées et les SQR dérivés qui se rapprochent de ce type de thématique.

Nous allons donc examiner les propriétés des systèmes de questions réponses actuels afin de déterminer comment les faire évoluer pour leur permettre de tenir compte des questions enchaînées. La recherche des documents pertinents consituant l'une des étapes les plus concernées par cette évolution, nous détaillons les moteurs de recherche les plus utilisés dans ce domaine. L'objectif est de mettre en évidence les grands problèmes du domaine, les méthodes utilisées pour les résoudre et leurs limites tant pratiques que théoriques.

II.1 Systèmes de questions réponses

Nous avons déjà présenté les bases des systèmes de questions réponses (I.1 page 18). Explicitons maintenant les évolutions qui sont nécessaires pour rendre les SQR interactifs. Nous détaillons ensuites les campagnes d'évaluation qui ont été mises en place, ainsi que les métriques utilisées.

II.1.1 Les systèmes de questions réponses interactifs

Les SQR interactifs, comme les SQR, sont des systèmes ayant pour but de faciliter l'accès à l'information en domaine ouvert. Ce ne sont pas des *chatterbots* sans «but» ou des systèmes à scénarios d'interaction prévus.

Les systèmes de dialogue autorisent déjà les utilisateurs à interagir avec de simples structures de données, comme des horaires de train ou d'avion, en utilisant un composant dialogique à base de variation sur un modèle à états finis. Ces modèles font un usage intensif de la structure du domaine pour contraindre l'espace des interactions possibles. Pour aller plus loin, nous avons besoin de combiner les possibilités des systèmes de dialogue et SQR en domaine ouvert [Webb, 2006].

II.1.1.1 Adaptation de la réaction

L'une des caractéristiques à prendre en compte dans une interaction est la présentation de la réponse quand elle est trouvée. Elle doit être adaptée à son contenu, à son importance et à sa taille et/ou complexité. Quand le système est interactif, nous pouvons distinguer plusieurs types de réaction qui peuvent ne pas dépendre de la réponse. Le SQR-interactif (SQRI) peut analyser l'absence de réponse et proposer des réponses coopératives [Grice, 1975] qui apportent un substitut à la réponse [Benamara, 2004]. Il peut aussi répondre par une question demandant plus de précisions sur la question [Séjourné, 2005], ou sur les corpus[1] sur lesquels la recherche doit être réalisée :

> Utilisateur : Qui a fondé le Festival international de la bande dessinée d'Angoulême ?
> Système : Voulez vous parler du 23e Festival international ou plutôt de la première édition du Salon international ?

Le SQRI peut aussi fournir la réponse à une question précédente non résolue, si les informations qui lui manquaient lui ont été fournies. De manière

[1] Dans les systèmes Qristal et Musclef, le changement de corpus est réalisé dans les options de configuration. Le proposer en cas d'absence de réponse est envisageable.

générale, le SQRI peut réagir par toute information pertinente compte tenu de la question.

Pour maintenir l'attention de l'utilisateur, le système interactif doit être rapide. Rapide signifie que le temps d'accès à l'information par la machine doit être court, mais aussi que la présentation de l'information doit correspondre aux attentes de l'utilisateur pour réduire le temp que l'utilisateur met à en prendre connaissance.

II.1.1.2 Importance du second énoncé

Van Schooten et Op den Akker [van Schooten & op den Akker, 2005] proposent une étude des cas possibles de continuation de dialogue dans le cas de l'adaptation d'un SQR au dialogue. Une manière naturelle d'interagir en question-réponse consiste à étudier le type du second énoncé formulé par l'utilisateur juste après une première réponse [Boni & Manandhars, 2005]. Ceci part du constat que les systèmes qui autorisent l'utilisateur à réagir à une question sans oublier la question précédente sont soit des systèmes qui *évaluent une nouvelle question dans le contexte de la précédente*, soit des systèmes qui *autorisent une question de demande de précision* :

Evaluer une nouvelle question dans le contexte de la précédente :
Utilisateur : Combien de places y a-t-il dans la voiture électrique Impact de General Motors ?
Système : 2
Utilisateur : Combien de temps faut-il pour accélérer de 0 à 60 miles par heure ?
Autoriser une question de demande de précision :
Utilisateur : Combien de temps faut-il pour accélérer de 0 à 60 miles par heure ?
Système : Comment ?
Utilisateur : Avec la voiture électrique Impact de General Motors.

Dans [van Schooten & op den Akker, 2005], à l'aide d'un SQR, les auteurs ont obtenu 575 énoncés de continuation de 40 utilisateurs. Ils ont constaté que les énoncés peuvent se répartir dans plusieurs catégories. Les énoncés de continuation les plus fréquents (plus de la moitié) sont les nouvelles questions qui doivent être comprises dans le domaine de la précédente. Les *retours négatifs* sur le résultat du SQR sont le second type de continuation les plus fréquents, environ 30% du total. Les *acquittements* (environ 15%) sont plus rares, et il existe aussi d'autres formes de continuation plus

difficiles à classer à hauteur de 3% du total. Les *retours négatifs* sont constitués de plusieurs formes :

- Questions à assertion négative :
  Utilisateur : Que sont les palpitations cardiaques ?
  Système : Le patient ressent cela comme des palpitations cardiaques.
  Utilisateur : Mais qu'est ce que c'est que les palpitations cardiaques ?
- Questions de vérification :
  Utilisateur : Que ressent t-on quand on a une hypertension ?
  Système : L'hypertension est un problème qui ne cause habituellement aucun symptôme.
  Utilisateur : Alors, on ne remarque rien quand on est en hypertension ?
- Reformulation :
  Utilisateur : Est-ce que les RSI sont plus fréquentes aux pays bas (sic) que dans le reste de l'Europe ?
  Système : Les RSI n'apparaissent pas seulement sur les personnes qui travaillent devant un écran, mais aussi dans l'industrie et la construction.
  Utilisateur : À quelle fréquence les RSI apparaissent-elles aux pays bas (sic) en comparaison du reste de l'Europe ?

La plupart des *acquittements* consistent en des phrases courtes telles que «merci» ou «ok» ...

Les auteurs ont alors ajouté au même SQR un système de résolution des référents, un classifieur pour les catégories précédemment citées et un générateur de requêtes qui réalise la réécriture des questions afin d'obtenir un SQR-interactif. Il utilise un analyseur en arbre de dépendances syntaxiques et un classifieur sémantique en domaine spécifique pour trouver les catégories de continuation d'énoncés. Ce SQRI dispose d'une stratégie statique de gestion des dialogues. Si l'énoncé utilisateur fait référence à une question précédente ou à une autre question alors le nouvel énoncé est envoyé au système de réécriture de requêtes après utilisation d'un module de résolution de référence. Les références vers la question précédente couvrent 75% des types de références. Si par contre l'énoncé est un retour négatif sur la réponse à la première question soumise au SQR (ou qu'il s'agit d'une question de vérification de la réponse), alors une réponse de circonstance («Pouvez vous reformuler ou préciser votre question ?») est affichée. Si enfin l'énoncé est d'un autre type qu'un acquittement positif, alors il est transmis au SQR.

II.1.1.3 Les SQR interactifs à modalité orale

Voyons maintenant deux systèmes qui synthétisent à eux seuls l'essentiel des travaux développés dans le domaine des SQRI.

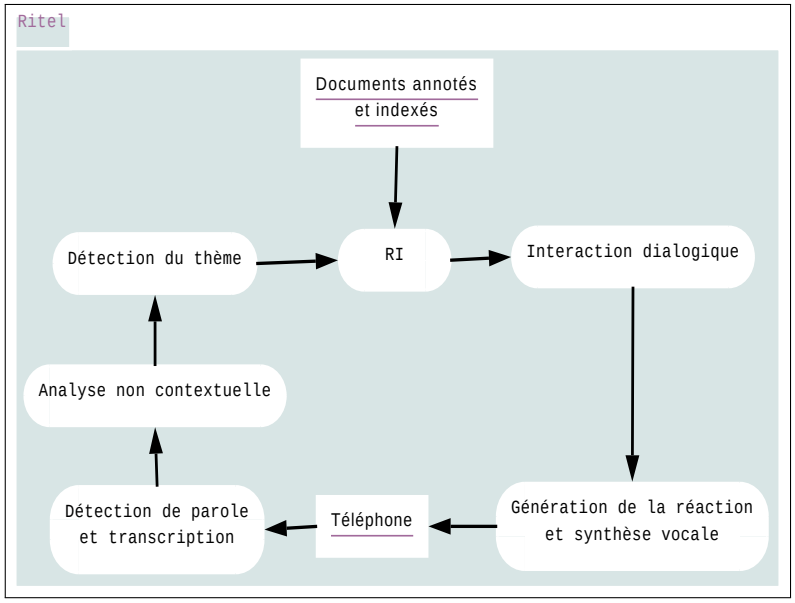

FIG. II.1 – Architecture de Ritel

II.1.1.3.1 Ritel

Le SQR Ritel est un système de renseignement généraliste par téléphone. Le tableau II.1 donne un exemple de transcription de son fonctionnement. L'utilisation du téléphone comme média de diffusion permet de recruter facilement des testeurs. Afin de ne pas rompre le canal de communication, Ritel est optimisé pour répondre rapidement. Ceci permet de réaliser des tests dans des conditions réelles d'utilisation de manière peu coûteuse. Par exemple, les utilisateurs n'ont pas à se déplacer pour venir passer les tests. Ils n'ont besoin que de leur téléphone.

Ritel prend le parti de découper les traitements en un ensemble de traitements simples indépendants, et de faire tous les traitements simples possibles. Ceci contribue à sa vitesse, sa maintenabilité et sa capacité à identifier efficacement d'où viennent les problèmes. La simplicité des modules permet de calculer une rétroaction à réaliser sur ses composants afin d'obtenir la sortie voulue. Ceci est un point particulièrement important dans un système interactif.

Ritel n'est pas un SQR fonctionnant étape par étape comme Musclef, mais un SQR dont tous les composants prennent en entrée un flux et émettent un flux en sortie. Ceci permet de commencer les analyses avant que l'utilisateur n'ait fini de parler. La figure II.1 illustre son organisation.

Exemple 1	
Ritel	Salut et bienvenu au service de Ritel de Limsi. Voulez-vous le message d'aide ?
User	heu non merci je ne voudrais le capital(la capitale) de la France s'il vous plait
Ritel	la capitale de la France est Paris. Voulez-vous d'autres informations ?
User	heu qui est son président
Ritel	Jacques-Chirac est le président de la France. Voulez-vous d'autres informations ?
User	et heu celui pour la Bulgarie
Ritel	Georgi-Parvanov est le président de la Bulgarie. Voulez-vous d'autres informations ?
User	De quel pays Ion Iliescu est le président alors ?
Ritel	L'ion-Iliescu est le président de la Roumanie. Voulez-vous d'autres informations ?
User	Depuis quand
Ritel	l'Ion-Iliescu est le président de la Roumanie depuis décembre 2000. Voulez-vous d'autres informations ?
User	Non merci au revoir
Ritel	merci d'avoir utilisé le service de Ritel de Limsi. Nous vous remercions pour votre appel. Au revoir.
Exemple 2	
Ritel	Salut et bienvenu au service de Ritel de Limsi. Voulez-vous le message d'aide ?
User	Je voudrais une certaine information sur l'ONU.
Ritel	Vous cherchez de l'information sur l'ONU. Pourriez-vous être plus spécifique ?
User	Quel sont les pays dedans ?
Ritel	Votre demande est de la géographie. Mais plus spécifiquement ? Êtes-vous intéressé par un pays spécifique ? Une ville ?
User	Qui a écrit pour être ou ne pas être ?
Ritel	je ne sais pas qui est l'auteur d'être ou ne pas être. J'ai perdu ma connexion avec le serveur de recherche de l'information. Voulez-vous d'autres informations ?

TAB. II.1 – Exemple de conversations transcrites avec Ritel

- *Analyse non contextuelle* : Les énoncés sont analysés selon des schémas qui ne dépendent pas des énoncés précédents de l'utilisateur. Il s'agit d'analyses morpho-syntaxiques, d'entités nommées, etc...
- *Détection du thème* : Ce module décide si le nouvel énoncé est à rattacher à l'énoncé précédent à l'aide de marqueurs de discours. Une purge de l'historique est réalisée s'il n'y a pas de lien. S'il y a un lien, des informations de l'énoncé précédent sont ajoutées à celles de l'énoncé courant.
- *RI* : La recherche d'information générales est réalisée sur des documents annotés et indexés. Les mots clefs sont choisis à partir du résultat de la détection du thème.
- *Interaction dialogique* : C'est la partie qui s'occupe de souhaiter «bonjour/au revoir» à l'utilisateur, expliquer que le système ne «comprend» pas un énoncé, de demander des répétitions ou reformulations à l'utilisateur.

Si nous nous intéressons à la partie linguistique de Ritel, nous voyons que le système est constitué de la manière suivante (cf. [Rosset *et al.*, 2006]) :

- Analyse non contextuelle (repérage et typage d'entités)
- Détection de marqueur de question et de marqueur d'interaction (acte de dialogue conventionnel pour la gestion sociale du dialogue)
- Annotation morpho-syntaxique
- Découpage de l'énoncé en groupes grammaticaux (nom composé, verbe composé, ...)
- Détection des entités par morceaux de texte (chunk) et analyse syntaxique superficielle (shallow parsing)

Ritel fonctionne en une milliseconde seulement d'analyse, et moins d'un dixième pour la totalité des traitements.

Il n'y a pas de module spécifique de gestion de dialogue : il est complètement intégré [Rosset *et al.*, 2006]. Les optimisations en vitesse d'exécution ont conduit à cette intégration, bien que les types d'analyse réalisés et les modules mis en œuvre soient classiques pour un SQR ou un système de dialogue. Les premières étapes de l'analyse permettent de choisir entre une approche par requêtes sur une base de données, ou une approche purement réactive «à la *chatterbot*» ayant pour but d'appliquer des stratégies visant à mettre au clair les données déjà présentes.

La recherche dans les documents utilise des enregistrements de description de données (DDR :*Data Description Record*) [Rosset *et al.*, 2007]. Ces DDR contiennent toutes les informations requises à propos des entités, du type de réponse, avec leur poids. Un score de proximité est alors calculé entre le DDR et les passages de documents. La réponse est choisie en tenant compte de ce score.

II.1.1.3.2 SPIQA

Le système de [Hori *et al.*, 2003] est fondé sur une conception similaire du système de question-réponse orale, mais sans intégration globale. Un SQR développé pour l'écrit SPIQA est utilisé après une étape de transcription automatique [Chiori *et al.*, 2003 4]. L'interaction y est par contre vue d'abord comme une méthode pour contrecarrer les effets de perte de performance liée à cette transcription automatique.

La version présentée par Nancy J. McCraken et al. [McCracken *et al.*, 2006] abandonne l'idée de la transcription automatique de la parole. SPIQA exploite les concepts de FAQ (Frequently Asked Question) et PAQ (Previously Asked Question) pour améliorer ces capacités d'interaction dans le cadre d'un apprentissage via plusieurs utilisateurs. Un système de résolution de référents anaphoriques et de distance entre questions est déployé. Les systèmes de résolution de référent permettent de réaliser des réécritures de questions : Si un groupe est composé des questions «*When did Madonna enter the music business ?*» et «*When did she first move to NYC ?*» alors la seconde question peut être ré-écrite «*When did Madonna first move to NYC ?*»

Comme dans HitiQa [Small *et al.*, 2004] le système a été spécialisé pour un domaine précis (le développement aéronautique à la NASA), et les autres modes d'interaction avec le système passent par l'interface graphique spécialisée. Les calculs de similarité et de distance entre questions ainsi que leurs usages sont similaires à ceux de Ferret (I.2.2.3 page 37) mais sont moins développés que dans ce dernier.

II.1.2 Gestion des enchaînements

Si nous supposons que les données issues des liens entre questions peuvent améliorer la recherche de la réponse, comment les structurer pour que cette recherche soit plus efficace ? Pour apporter des éléments de réponse à cette question, dans cette section nous verrons tout d'abord les méthodes classiques de gestion des enchaînements de questions, puis nous verrons quelques exemples de systèmes utilisés dans des campagnes d'évaluation dédiées aux SQR-enchaînées.

Suite aux campagnes Trec [Dang *et al.*, 2006] et Clef [Penas *et al.*, 2007], nous observons essentiellement plusieurs approches possibles pour gérer les enchaînements. L'une consiste simplement à ajouter les éléments de la question liée *à la fin du texte* de la nouvelle question et mettre le résultat en entrée d'un SQR classique. Une autre approche consiste à analyser les questions sé-

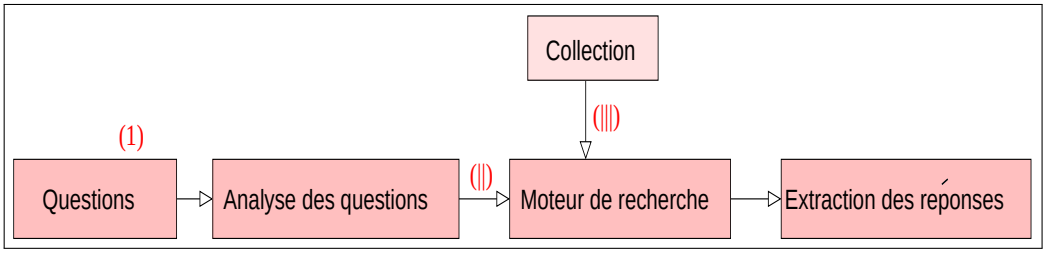

FIG. II.2 – Architecture type d'un système de questions réponses

parément, et à *fusionner leurs représentations* dans une structure similaire à celle d'une question unique. Cette structure est alors envoyée au moteur de recherche. Enfin, une autre solution consiste à mémoriser la liste des documents qui ont été retournés par le moteur de recherche à la suite de la première question, et faire la recherche uniquement sur cet ensemble réduit. Illustrons ces techniques sur l'exemple classiques d'architecture de SQR de la figure II.2 (page 55).

A) La méthode consistant à ajouter les termes de la question liée à la fin du texte de la nouvelle question présente un inconvénient. Cela revient à modifier les questions au niveau du point I de la figure II.2 (page 55). Il faut pouvoir gérer les différentes formulations et interprétations, et dans tous les cas disposer d'un module de plus pour ne pas oublier le type attendu de la réponse ou d'autres termes pertinents. Il faut aussi un moyen pour réaliser correctement la jointure entre la question et les termes des précédentes (par exemple, au plus simple, par l'ajout d'un point virgule). Le résultat de l'ensemble est souvent une question mal formulée et donc mal analysée.

B) Une approche consiste à fusionner les structures correspondant aux analyses de deux questions. La représentation unique résultant des deux questions doit alors être isomorphe à celle qui aurait été obtenue par l'analyse d'une seule autre question. Cette autre question virtuelle non ambiguë synthétiserait les informations des 2 premières questions. Cette fusion peut présenter les mêmes inconvénients que ci-dessus.

C'est une variante où le système est modifié au niveau du point II de la figure II.2 (page 55). Les termes composant la question liée sont alors mêlés à ceux de la nouvelle question [Buscaldi *et al.*, 2007]. La structure envoyée au moteur de recherche est une structure qui contient notamment le résultat de la résolution des liens entre questions. Dans la structure qui regroupe les analyses, habituellement la question d'où proviennent les analyses et les termes est évidente ou implicite. Dans le cas d'une fusion de deux structures, il n'est pas possible de savoir de quelle question vient tel terme ou analyse.

C) À notre connaissance il existe une autre stratégie : elle consiste à conserver le résultat des recherches réalisées sur la première question, notamment l'ensemble des documents retournés par le moteur de recherche. Cette première question définit l'espace de recherche de document pour toutes les questions de ce groupe. Les recherches suivantes sont réalisées uniquement sur l'ensemble de documents trouvés lors de cette première recherche [Zhou *et al.*, 2006, Hickl *et al.*, 2006]. Dans le système de la figure II.2 (page 55), cette stratégie peut être mise en place au niveau du point III. La collection utilisée est réduite à chaque nouvelle question.

D) Hori *et Al.* [Chiori *et al.*, 2003 4] rencontrent ce problème dans leur SQR interactif écrit/oral. Ils lui apportent une solution basée sur une interrogation de l'utilisateur. Ils examinent un score d'ambiguïté structurelle d'une question pour déterminer de nouvelles questions de désambigüisation. Puis l'utilisateur est censé y répondre correctement pour apporter les informations indispensables à la résolution des liens entre différents couples questions-réponses.

Dans les deux premières approches présentées ci-dessus, les termes des questions précédentes ont tous la même importance. Dans la troisième approche, les termes des premières questions peuvent rendre inaccessibles des documents contenant les réponses pour les questions ultérieures. L'importance des termes est donc figée non pas d'après une analyse linguistique, mais par l'ordre des questions, donc bien loin de ce que nous pouvons attendre de la langue naturelle. Nous chercherons à tenir compte de traits linguistiques dans les questions et les documents pour tenir compte de l'importance relative des termes.

Quand nous cherchons un objet dans une collection, les manières de le sélectionner avec précision sont soit une réduction du nombre d'objets dans la collection de manière à réduire l'espace de recherche, soit une augmentation du nombre de contraintes que doit respecter l'objet à sélectionner.

Il en va de même en recherche d'information. La plus grosse réduction de l'espace de recherche est réalisée quand tous les termes des questions sont ajoutés à la requête. Imaginons que nous cherchions à modifier le nombre de termes dans la requête, soit en retirant, soit en ajoutant leurs synonymes ou autres ... Comme l'importance de chaque terme n'est pas connue, quel terme de quelle question sera oublié/ajouté en premier ? Dans les questions enchaînées, il y a potentiellement plus de termes que dans les questions classiques à cause des liens entre questions. S'il y a un plus grand nombre de termes alors ces termes sont-ils plus ou moins significatifs dans les documents ? En conséquence, actuellement, les SQR-enchaînées choisissent arbi-

trairement une stratégie de réduction de l'espace de recherche en fonction des termes.

Les SQR-enchaînées existants ont tous été développés à partir d'une adaptation d'un SQR classique. La gestion des enchaînements est souvent ajoutée sous la forme d'une unique «boîte» de traitement. Nous pouvons faire encore quelques distinctions par rapport au modèle présenté ci dessus. Parfois les traitements sont intégrés à des processus existants. Parfois ils sont ajoutés en amont ou en aval. Mais les contraintes implicites construites par les flots de données y circulant y sont fonctionnellement les mêmes.

II.1.3 Évaluation dans les systèmes de questions réponses

L'évaluation dans les SQR s'articule autour de corpus de taille finie afin de garantir que les expériences soient reproductibles.

Les campagnes d'évaluation définissent 3 données :

- Le corpus de documents où les réponses doivent être cherchées.
- Le corpus de questions dont les SQR doivent trouver les réponses.
- Un ensemble de métriques pour déterminer le «degré » d'adaptation des réponses aux questions pour le système en général.

Les corpus de documents et de questions dépendent des campagnes d'évaluation, mais les métriques d'évaluation sont souvent similaires.

II.1.3.1 Un ensemble de métriques

Prenons comme exemple l'ensemble de questions suivant :

Id	Texte de la question
1	Quel président américain a dirigé les accords de Camp David ?
2	En quelle année les négociations ont-elles eu lieu ?
3	Combien d'accords ont été signés ?
4	A quelle date le "traité de paix israélo-égyptien" fut-il signé par la suite ?

C'est le groupe de questions numéro 55 du corpus de question ClefQA07-FR-EN avec des questions enchaînées en français attendant des réponses en anglais.

Notre meilleure méthode de résolution a obtenu les résultats suivants :

n°	Rang de la réponse	Texte de la réponse
1	4	Jimmy Carter
2	8	1978
3	1	2
4	1	1979

Plusieurs réponse peuvent être données pour chaque question. Idéalement, la bonne réponse devrait être en rang un. Voici quelques-unes des métriques les plus courantes pour l'évaluation des SQR qui nous permettent d'évaluer les réponses pour ce groupe de questions.

II.1.3.1.1 Le S@n (généralement S@10) est la fraction des listes ordonnées de documents qui contiennent au moins une bonne réponse dans les n premiers rangs. Il existe aussi une version où n est infini. Parfois le rang moyen est utilisé quand il s'agit d'évaluer des résultats de modules intermédiaires dans lesquels seul le fait d'être dans les n premiers a une importance. Le rang moyen est une mesure qui peut permettre de prendre la décision d'un S@n particulier.

II.1.3.1.2 Le MRR est une mesure de la moyenne de l'inverse des rangs d'éléments cibles dans des ensembles ordonnés d'éléments de même type. Par exemple, si un moteur de recherche retourne des listes ordonnées de documents, le MRR est la somme de l'inverse des rangs des premiers documents contenant les réponses. Le MRR permet de bien récompenser les systèmes qui mettent les bonnes réponses en première place. C'est le MRR(All). Le MRR(Ok) est la valeur qui est mesurée en ne réalisant la moyenne que sur les rangs des documents qui ont été considérés comme corrects pour au moins l'un des n premiers documents (il peut s'écrire MRR(S@n)).

Dans notre exemple le MRR est donc de :
$\frac{1}{\#questions}\Sigma_{i \in rangs\ des\ reponses}\frac{1}{i} = \frac{1}{4} * (\frac{1}{1} + \frac{1}{1} + \frac{1}{8} + \frac{1}{4}) = 0.59$
C'est le MRR(All).

II.1.3.1.3 Le rappel est le nombre de réponses données par rapport au nombre total de réponses à donner. La quantité calculée à partir du nombre d'éléments choisis dans une collection par rapport au nombre d'éléments qu'il fallait choisir. Le rappel est la mesure qui répond à la question « quelle fraction des documents intéressants se trouve dans l'ensemble retourné par le système ? » .

Pour chaque question, nous avons obtenu une réponse, le rappel est donc de $4/4 = 1$.

II.1.3.1.4 La précision est le nombre de bonnes réponses sur le nombre de réponses données. La précision est la mesure qui répond à la question «quelle fraction de l'ensemble des documents retournés est pertinente ? ». La précision peut comme le MRR accepter une variante plus souple : la précision (S@n). n doit être choisi en fonction de l'usage du résultat (présentation à l'utilisateur, insertion dans d'autres traitements ...). Ici la précision est de 2/4 car la bonne réponse apparait 2 fois en première position. Nous pourrions présenter plusieurs résultats à un utilisateur, nous pouvons alors nous demander quelle serait la précision si nous autorisons les bonnes réponses jusqu'en rang 5 ? La précision (S@5) serait alors de $3/4 = 0.75$.

II.1.3.1.5 La F-mesure est le double du produit de la précision et du rappel, par rapport à la somme de la précision et du rappel.

$$Fmesure = (2 * précision * rappel)/(précision + rappel)$$

Parfois le coefficient 2 est remplacé par une autre pondération. Intuitivement c'est une mesure qui représente un compromis entre la précision et le rappel, il est évident que si $précision = rappel$ alors cette mesure est une fonction identité.

Classiquement, dans les campagnes Trec, au moins deux types de mesures sont utilisés, le MRR et la F-mesure. Ces mesures permettent de déterminer les meilleurs systèmes au niveau de la réponse finale et des documents. Toutes les réponses d'un SQR sont ordonnées donc l'utilisation du MRR est évidente, puis selon ce qui est observé il est possible d'utiliser la F-mesure sur les n premières réponses.

Similairement à la précision, nous calculons la F-mesure pour des réponses en rang 1 et pour des réponses allant jusqu'au rang 5. La F-mesure de l'exemple est $(2 * 0.5 * 1)/(0.5 + 1) = 0.67$ et la F-mesure (S@5) vaut : $(2 * 0.75 * 1)/(0.75 + 1) = 0.86$.

II.1.3.1.6 Dans les questions enchaînées Les questions enchaînées peuvent réutiliser les réponses courtes de questions précédentes. Il y a donc une répercussion des résultats des calculs des premières questions sur les résultats des questions suivantes. Dans les campagnes d'évaluation jusqu'à maintenant, il n'a pas été adopté de mesures particulières pour tenir compte de ce phénomène. Il y a deux raisons : la première est que le nombre de questions dont il n'est possible d'obtenir la réponse que grâce à une réponse d'une question précédente est encore faible. La seconde est que tous les systèmes sont évalués sur les mêmes questions, donc les résultats sont comparables d'un système à l'autre à l'intérieur de la campagne d'évaluation. En revanche, les

résultats ne peuvent pas être comparés avec ceux des campagnes de questions réponses isolées.

Notre grille de résultats pour le groupe de questions est donc :

MRR(All)	$1/4 * (1 + 1 + 1/8 + 1/4) =$	0.59
rappel	$4/4 =$	1
précision	$2/4 =$	0.5
F-mesure	$(2 * 0.5 * 1)/(0.5 + 1) =$	0.67
MRR(S@5)	$1/3 * (1 + 1 + 1/4) =$	0.75
précision(S@5)	$3/4 =$	0.75
F-mesure(S@5)	$(2 * 0.75 * 1)/(0.75 + 1) =$	0.86

II.1.4 Campagnes d'évaluation autour des SQR-enchaînées

II.1.4.1 Trec

En 2005 et 2006, la campagne d'évaluation Trec[2] a proposé une tâche de résolution de questions avec un contexte commun. Les questions sont proposées par groupes de 4 à 8 et sont reliées par un thème commun. Il y a un total de 200 questions chaque année. Le thème est donné pour chaque groupe et porte sur des personnes, des organisations ou d'autres entités. Il peut même y avoir des évènements (contrairement à Trec 2004). Les questions ne sont sensées avoir de liens qu'avec le thème.

Les autres questions apportent parfois des informations supplémentaires qui peuvent aider à répondre aux questions précédentes. Ces extraits de groupes de questions tirés du corpus de questions de la tâche ciQA de Trec 2006 illustrent ce phénomène :

Thème/Objet/Target : «Tufts University» Who became Tufts University President in 1992 ? Over which other university did he preside ?
Thème/Objet/Target : «NASCAR» Who founded NASCAR ? List winners of the NASCAR races.

Dans le premier cas, il s'agit d'une référence à la réponse de la première question, dans le second cas une précision sur l'entité du thème qui aurait pu être utile pour la première question aussi.

Le tableau II.2[3] est un groupe de questions tiré d'un corpus de Trec.

[2] Text REtrieval Conference ; http ://trec.nist.gov

[3] Objet : C'est ce qui s'appelle le «Target» qui peut se voir comme la «cible» thématique du groupe de questions. La notion de cible thématique n'est évidemment pas partagée dans toutes les campagnes d'évaluation.

	Objet	Warren Moon
1	Réponse factuelle	What position did Moon play in professional football ?
2	Réponse factuelle	Where did Moon play in college ?
3	Réponse factuelle	In what year was Moon born ?
4	Réponse factuelle	How many times was Moon a Pro Bowler ?
5	Réponse factuelle	Who is Warren Moon's agent ?
6	Réponse liste	Who have coached Moon in professional football ?
7	Réponse liste	List the professional teams for which Moon has been a player.

TAB. II.2 – Exemple d'un groupe de questions enchaînées tirées du corpus utilisé pour la campagne d'évaluation Trec 2006.

n°	Question
1	Où se trouve la cathédrale Sainte-Sophie en Russie ?
2	Qui était son archiprêtre en 1995 ?
3	Quel Écossais a construit la cathédrale ?
4	Quelle impératrice russe l'a accrédité pour la construire ?

TAB. II.3 – Exemple d'un groupe de questions enchaînées tirées du corpus utilisé pour la campagne d'évaluation Clef 2007.

II.1.4.2 Clef

En 2007, la campagne d'évaluation Clef [4] a été inspirée de la tâche Trec. De même que dans Trec, les questions sont proposées par groupe. Les groupes peuvent être arbitrairement petits. Toutes les questions ne font pas partie d'un groupe [Penas *et al.*, 2007]. Les thèmes de chaque groupe ne sont pas donnés. Le tableau II.3 est un groupe de questions tiré d'un corpus de Clef.

II.1.5 Des systèmes de Trec

La description de la tâche Trec sur les questions enchaînées décrit bien le cadre de l'approche : «Le but de l'aspect interactif de ciQA[5] était de fournir un cadre de travail pour les participants pour examiner l'interaction dans le

[4]Cross Language evaluation ; http ://www.clef-campaign.org
[5]Complex Interactive Questions Answering

contexte des SQR et de fournir une occasion pour les chercheurs extérieurs aux SQR de s'impliquer dans ce domaine. Nous définissons un système interactif comme un système qui donne à l'utilisateur un contrôle sur tout ou portion du contenu présenté. En utilisant cette définition, l'unité d'interaction la plus petite possible consiste en un utilisateur répondant au système et le système utilisant cette réponse pour produire un contenu nouveau.» [Dang *et al.*, 2006]

C'est en vertu de ce principe que la tâche principale est de chercher des réponses à des questions isolées dans un contexte. Dans le cadre de la campagne Trec le contexte fait référence à un *thème cible*[6] de la recherche de l'utilisateur. Il existe aussi une succession logique entre les questions afin de construire le contexte implicitement. Voyons maintenant quelques SQR ayant été modifiées en vue de traiter la tâche de questions-réponses enchaînées.

II.1.5.1 Le système de l'université de Tokyo

Le système développé par l'équipe «Speech» de l'université de Tokyo [Whittaker *et al.*, 2006] dispose d'une stratégie probabiliste de recherche dans les documents et d'extraction de la réponse. Le système se fonde une probabilité basée sur la loi de Bayes. Il utilise la probabilité d'existence d'un sous-modèle de langage de la question qui serait présent dans les documents, ainsi qu'une probabilité liée au type de la question (*Where*, *When* ...). Cette probabilité sert de base à une mesure de similarité avec la question. Les mots du *thème cible* sont utilisés pour renforcer la probabilité d'un document de contenir la réponse. Ces mots sont traités comme s'ils venaient de la question. L'extraction de la réponse utilise un modèle à base de n-uplet, les mots du *thème cible* sont alors oubliés. L'extraction de la réponse exacte est ancrée sur un modèle par apprentissage de patrons par indépendance conditionnelle. L'ensemble d'apprentissage est construit à partir des questions et réponses des sessions Trec des années précédentes. Seule la recherche des documents est affectée. Le système traite les questions attendant une liste d'éléments factuels en réponse comme une collection des 10 premières réponses du système probabiliste d'extraction de la réponse. Une version adaptée à la tâche multilingue de ce système a également participé à Clef@QA2006.

II.1.5.2 Le système de FuDan University

Le système FDUQA[7] de Yaquian Zhou [Zhou *et al.*, 2006] traite uniquement le cas des coréférences anaphoriques. Les anaphores traitées sont répar-

[6] Traduction de «Target».

[7] FDUQA : FuDan University Question Answering

ties sur les 6 catégories : *person*, *organization*, *location*, *event*, *time* et *other*. Le repérage de l'anaphore se fait uniquement par unification de catégories repérées entre référent et antécédent. Dans le cas où plusieurs unifications sont possibles, celle du *thème cible* est préférée, puis celle d'une question de la plus ancienne à la plus récente. S'il y a une coréférence anaphorique de la question vers le *thème cible*, alors le système réalise une substitution du référent par l'antécédent. Sinon la totalité des mots du *thème cible* est ajoutée à la liste des mots de la question. Ils notent que l'analyse syntaxique avec la présence d'un thème ne permet pas une unification syntaxiquement correcte avec le reste de la question. La relation syntaxique entre l'anaphore et son antécédent est souvent éloignée. Notons que pour répondre aux questions portant sur des listes, le système a été adapté pour préférer le rappel à la précision. Le système de réordonnancement des documents combine les scores du moteur de recherche Lucene[8] (au niveau du document) et les scores des meilleures phrases de chaque document. Le score des phrases est calculé pour les documents ayant le meilleur score retourné par Lucene. Il se fonde sur le nombre de mots de la question sans les mots du *thème cible*. Le système FDUQA traite les questions de définition en extrayant hors ligne une base de connaissance depuis le corpus d'interrogation.

II.1.5.3 Le système QA-LaSIE

Le système QA-LaSIE [Greenwood *et al.*, 2006]exploite le système d'extraction d'information LaSIE[9] qui est intégré dans GATE[10]. Le système traite les questions de type «Factuelle» et de type «Liste»[11]. La métrique d'évaluation de la campagne ne pénalise pas les mauvaises réponses, donc la stratégie des «Liste» essaie de capturer toutes les réponses bien typées retournées par une première recherche dans les documents. La taille de la liste est limitée aux dix premières réponses. Les auteurs utilisent les ressources sémantiques et grammaticales de LaSIE à travers GATE pour réécrire la question et le *thème cible* en une unique question dans laquelle les anaphores coréférentes sont résolues. La question résultante est traitée classiquement afin d'être introduite dans Lucene [Cutting, 2000].

[8]Lucene est présenté en section II.2.3 page 72

[9]LaSIE :Large Scale Information Extraction [Humphreys *et al.*, 1998]

[10]GATE [Cunningham *et al.*, 2002] est une plateforme d'intégration de modules disposant d'interface déclarée formellement.

[11]Les questions de type «Liste» attendent une énumération d'entités en guise de réponse. Le nombre d'entités n'est pas fixé à l'avance.

II.1.6 Des systèmes de QA@CLEF2007

Voici des systèmes parmi les plus significatifs qui ont été présentés à la campagne d'évaluation ClefQA2007 . Comme nous avons déjà vu de quoi est constitué un SQR enchaîné, il serait fastidieux de chercher à identifier précisément toutes les sous-parties ayant des caractéristiques communes. Attachons-nous aux éléments saillants de ces systèmes qui leur donnent un comportement ou un fonctionnement original, avec en particulier les traitements des anaphores.

II.1.6.1 Le système de Priberam Informatica

Le système de Carlos Amaral développé à Priberam [Amaral *et al.*, 2007], présente des capacités de détection d'anaphores. Cette détection requiert de trouver quelles entités nommées, évènements, objets ou phénomènes naturels devraient être extraits du premier couple de question-réponse. La détection des anaphores dans chaque question suivante utilise les termes de la recherche précédente comme co-référent. Ces termes sont appelés «contexte». La détection du contexte est réalisée en une seule passe lors de l'analyse de la question hors étude des autres questions. L'hypothèse est faite qu'il peut y avoir un lien implicite avec les termes qui ne font pas partie de la co-référence anaphorique.

Dans une première étape, le système analyse la question, il en résulte notamment une liste de *pivots* (les éléments jugés les plus important). Puis dans une seconde étape, le contexte est transformé en une liste d'objets *pivots*. Les deux listes de *pivots* sont fusionnées. Dans les questions enchaînées suivantes, les questions sont d'abord analysées et les *pivots* ainsi obtenus sont alors fusionnés à leur tour dans la liste de ceux des deux premières étapes. Après cela les passages sont sélectionnés. Le SQR de Priberam utilise le moteur M-CAST[12] pour la recherche dans les documents[13].

À ce stade, la question a été convertie en une liste de mots clefs (les *pivots*) et toute l'information concernant la provenance des termes a été perdue. Finalement, la dernière étape de ce système est la sélection de la réponse, réalisée traditionnellement à l'aide de patrons.

[12]M-Cast est le Multilingual Content Aggregation System based on TRUST search Engine, http ://www.m-cast.infovide.pl (projet e-Contenu numero EDC 22249 M-CAST)

[13]Il s'agit évidemment de documents découpés préalablement en passages. Nous revenons sur cet aspect par la suite.

II.1.6.2 Le système de l'université de Hagen

Le système de gestion de l'enchaînement des questions du système de l'université de Hagen [Hartrumpf *et al.*, 2007] est basé sur la création d'un historique qui contient les représentations sémantiques des questions et réponses. Cet historique est vidé à chaque fois qu'un nouveau thème est rencontré. La détection des thèmes est à base de règles implantées en dur. L'historique est vidé à chaque début de groupe. L'hypothèse est qu'à un groupe correspond un thème.

Le système reconstruit une question qui peut être résolue par un SQR indépendamment des autres questions. La question est reconstruite si une coréférence anaphorique est résolue par le système CORUDIS [Sven, 2006] vers la première question du groupe. Le coréférent se substitue alors à l'anaphore.

À chaque question est associé un réseau sémantique qui sert de représentation abstraite de la question reconstruite. Les données de la question courante et celles de la première question du groupe y sont éventuellement mélangées. Les questions dites *elliptiques* sont une forme de question dont les liens avec les questions précédentes sont sous-entendus, ou bien où les sujets/verbes sont sous-entendus. Notamment, dans les questions elliptiques, le focus est manquant. En utilisant le réseau sémantique, les questions elliptiques reprennent le focus de la première question.

II.1.6.3 Le système de l'Universidad Politcnica de Valencia

Le système développé à l'UPV[14] s'appelle QUASAR[15] [Buscaldi *et al.*, 2007] [Gomez *et al.*, 2005].

L'enchaînement entre les questions est vu uniquement comme une liaison d'anaphores. Le système de résolution d'anaphore procède par la ré-écriture éventuelle des questions. La recherche dans les documents ne tient aucun compte du fait qu'il y ait eu une résolution d'anaphore. Le système réalise une analyse des traits de la question reconstituée pour la sélection de la réponse. Puis QUASAR extrait des passages via le moteur de recherche JIRS d'une manière similaire à Musclef. La résolution des anaphores peut traiter aussi bien des questions que des réponses comme source de référents. La résolution des anaphores n'est faite que vers le premier couple «question et réponse» du groupe.

Le système de résolution d'anaphore est original. Dans une première étape, un système à base de patrons induit les entités nommées, temporelles et les expressions numériques. Puis les entités qui apparaissent seulement

[14]UPV :Universidad Politcnica de Valencia
[15]QUASAR :QUestion AnSwering And Retrieval

n°	Question
1	Quelle récompense le film "Pulp Fiction" a-t-il reçue lors du festival de Cannes ?
2	Qui a réalisé ce film ?
3	Qui y joue le rôle principal ?

TAB. II.4 – Exemple d'un groupe de questions enchaînées tiré corpus de questions en français attendant des réponses en français de la campagne d'évaluation Clef 2007.

partiellement dans une question dérivée sont remplacées par leurs formes complètes probables de la première question. Exemple :

- *Cual era el aforo del Estadio Santiago Bernabéu en los anos 80 ?*
- *Quién es el dueno del estadio ?* → *estadio* → *Estadio Santiago Bernabeu*

La troisième étape consiste à résoudre les anaphores pronominales. Cela est fait par comptage sur le web pour déterminer quel remplacement doit avoir lieu. Le comptage compare par exemple des formes comme «*Bill Gates creo Microsoft*» et «*Melinda crea Microsoft*» [16]. Ensuite les anaphores possessives sont également résolues par un comptage sur le web. Exemple :

- *Cuanto dinero se gasto durente su ampliacion entre 2001 y 2006 ?*»

est comparé sur deux comptages :

- *ampliacion del Estadio Santiago Bernabeu*
- *ampliacion del Real Madrid Club de Futbol*

La question initiale devient :

- *Cuanto dinero se gasto durente ampliacion del Estadio Santiago Bernabéu entre 2001 y 2006 ?*

Enfin les entités qui n'ont pas pu être associées à quoi que ce soit sont ajoutées à la fin de la question.

II.1.6.4 Le système Qristal

Le système Qristal [Laurent & Séguéla, 2005] a été modifié pour obtenir un comportement différent sur les anaphores. L'hypothèse est faite que la résolution des anaphores permet d'obtenir tous les termes indispensables à désambiguiser la question. Comme plusieurs autres systèmes vus précédem-

[16] Les auteurs notent qu'en espagnol les anaphores pronominales ont tendance à être absentes : l'anaphore existe, mais n'a pas de marqueur linguistique.

ment, une nouvelle question sans lien anaphorique est construite dans le cas où une coréférence anaphorique est identifiée. Cette nouvelle question est envoyée vers le SQR Qristal.

Le système fait l'hypothèse que les co-références anaphoriques sont les seuls liens entre questions. Afin de contourner la limitation de cette hypothèse, les informations transmises au moteur de recherche sont enrichies avec par exemple les dates, les lieux géographiques (pays, villes) qui sont rencontrés dans les questions qui les précèdent. La technique de résolution des anaphores « englobe » une résolution de flots vers l'ensemble constitué de la question et de la réponse précédente. Par exemple dans le groupe du tableau II.4 (page 66) :

Qui a réalisé ce film ?

l'adjectif démonstratif est «*ce*», la sous phrase nominale est «*film*». Les référents possibles sont :

- *La palme d'or*
- *film Pulp Fiction*

Comme «*Pulp Fiction*» est l'entitée nommée la plus récente qui figure en position d'extension de la sous-phrase nominale et que «*La palme d'or*» n'est que la réponse précédente, les règles du Qristal amènent à choisir la résolution par extension.

Dans la question résolue, il n'y a aucun moyen de connaitre la confiance qu'il faut accorder dans les termes puisqu'ils peuvent provenir indifféremment de la question ou de la réponse. Qristal disposait déjà d'une méthode pour les anaphores pronominales et possessives. Un système de résolution pour les adjectifs démonstratifs a été ajouté pour la tâche de résolution des questions enchaînées.

Dans les campagnes Trec et Clef, il y a une volonté forte de lier les questions par des liens d'ordre linguistique. Dans Trec, un sujet cible est défini (le «target»). Dans Clef, les questions sont censées être posées plus ou moins comme le ferait un utilisateur via une interface en ligne de commande. La thématique cible n'est pas figée et doit être déduite des questions.

Nous constatons que souvent dans Clef la première question donne une thématique réutilisée par toutes les questions sans modification, mais il existe des cas où des thématiques dérivées sont introduites *a posteriori*. Les questions Trec peuvent donc être vues comme les cas simples de Clef. Nous en déduisons qu'une représentation valide pour Clef sera aussi valide pour Trec.

II.1.7 Synthèse

Les systèmes de questions-réponses se sont développés comme un cadre d'expérimentation des traitements automatiques des langues. Progressivement ils sont devenus un domaine à part entière. Ils ont reçu des développements commerciaux, et maintenant la problématique est de savoir comment augmenter leurs fonctionnalités ou comment les intégrer dans un cadre plus large (tel le dialogue). Dernièrement des campagnes d'évaluation de SQR proposaient des questions enchaînées. Nous nous proposons d'utiliser cette dynamique pour dépasser le cadre de cette évaluation, proposer un modèle du problème des questions enchaînées et envisager une intégration future dans un système plus vaste.

Les campagnes d'évaluation ont permis de tester les possibilités des systèmes de questions réponses enchaînées. Même si les résultats des systèmes ne sont pas toujours très bons, ils vont nous servir de base pour notre système. Par ailleurs les corpus pour les campagnes nous permettront d'évaluer nos résultats en se servant des métriques proposés.

II.2 Moteurs de recherche d'information

Nous venons de voir le fonctionnement des SQR. L'analyse des questions en est un point important. Des analyses dédiées aux questions enchaînées peuvent justifier des traitements spécifiques au niveau de la gestion des documents. Nous allons donc maintenant nous intéresser à l'usage des moteurs de recherche de documents dans les SQR.

II.2.1 Les fondements de l'indexation et de la recherche

Afin d'accélérer la recherche des documents dans une collection, les systèmes de recherche d'information sont découpés en deux parties, *l'indexation* et *la recherche*. Contrairement aux moteurs de recherche que nous pouvons trouver pour le web, nous supposons que nous disposons déjà d'une collection finie de documents. Il n'y a donc pas d'étapes d'exploration/peuplement (crawl) d'une collection à partir du web[17]

II.2.1.1 L'indexation

L'indexation a pour but de permettre un accès rapide aux informations d'un document pour les différentes requêtes possibles. Les mots des documents sont indexés et souvent des annotations ou méta-informations sont égalements indexées. Celle-ci utilise le principe du hachage et une méthode de compression de données afin d'améliorer la vitesse d'accès. Le hachage permet un accès en temps presque homogène pour tout terme de requête (grâce à des fonctions choisies de manière ad hoc pour les chaînes de caractères). La compression de la structure de hachage permet de minimiser la taille de l'index et donc minimiser la quantité de données à traiter. Le résultat est appelé index, il est souvent muni d'améliorations visant à gagner en vitesse d'accès et/ou en diversité d'informations stockées.

II.2.1.2 La recherche

La recherche correspondant à une requête utilise des lectures partielles de l'index. Ces lectures visent à obtenir rapidement les références vers les documents contenant les termes de la requête. Enfin, les documents sont ordonnés en fonction d'une mesure de similarité avec la requête, ou d'une mesure de

[17]Ce faisant, les SQR sont privés des informations pragmatiques de popularité des documents. Or ce sont ces informations qui ont conduit au succès des moteurs les plus en vue sur le web. Évidemment dans le cas d'une recherche de réponse exacte pour un SQR, la popularité d'une page web n'est pas forcément un critère judicieux.

probabilité de présence des termes. Seuls les documents suffisamment proches sont retenus comme étant des résultats potentiels.

Voyons d'abord comment la mesure de similarité est utilisée dans un algorithme classique, puis intéressons-nous à un moteur de recherche particulier et à la manière dont il est employé dans les outils dont nous disposons.

II.2.2 Fonctionnement d'un VSM

Un VSM est un modèle à base d'espace vectoriel[18]. Le but d'un VSM est de représenter un texte pour des tâches aussi variées que le filtrage, la recherche de document[19] ou l'indexation. Les documents sont vus comme des vecteurs où chaque terme correspond à une dimension dans un espace capable de représenter tous les documents vecteurs.

II.2.2.1 Le cosinus et le tf.idf

Dans cette représentation nous nous intéressons au calcul du cosinus et au calcul du $tf.idf$ (introduit par Salton [Salton & Yang, 1973] [Salton & Buckley, 1988]). L'idée sous-jacente dans les deux cas est qu'une requête peut essentiellement être représentée par un vecteur du même espace que le corpus. Le cosinus est utilisé comme une mesure de similarité entre deux vecteurs. Si $V(q)$ est le vecteur correspondant à une requete et $V(d)$ est le vecteur correspondant à un document alors le cosinus de similarité est défini par :

$$\boxed{cosinus(q,d) = \frac{V(q)*V(d)}{|V(q)|*|V(d)|}}$$

Le $tf.idf$ permet d'évaluer la saillance d'un terme par rapport à une collection de documents. C'est un poids proportionnel à la fréquence d'un terme dans un document et l'inverse de la fréquence de ce terme dans la collection. Le $tf.idf$ n'est pas spécifiquement lié au VSM, mais une fonction de pondérations des composantes est néccéssaire et de nombreux systèmes font confiance en la saillance d'un terme dans la collection. Le $tf.idf$ est alors parfois appelé «poid» d'un terme, le poids d'un vecteur étant la somme des poids des termes. Dans le calcul du $tf.idf$ de nombreuses variantes sont utilisées [Manning *et al.*, 2008]. Le tableau II.5 en donne des aperçus.

Nous nous intéressons plus particulièrement aux moteurs de recherche qui utilisent une mesure de similarité basée sur le cosinus et le $tf.idf$. Dans un but d'optimisation le produit scalaire n'est pas vraiment calculé entre

[18] VSM : Vector Space Model
[19] SMART, Gerard Salton

Fonction	$TF(\#Term)$
base	$1/\#Term$
quantité d'information	$log(1 + 1/\#Term)$
augmenté	si $1/\#Term > 0$ alors 1 sinon 0
logarithmique	$(1 + log(1/\#Term))/(1 + log(moy(1/\#Term)))$
Fonction	$IDF(\#Docs)$
base	$N/(1 + \#Docs)$
quantité d'information	$log(1 + N/(1 + \#Docs)$
positif	$1 + log(N/(1 + \#Docs))$ si > 0 sinon 0
probabilité	$max(0, log((N - \#Docs)/(\#Docs)))$
existence	si $\#occurrence > 0$ alors 1 sinon 0

TAB. II.5 – Variantes pour le $tf.idf$.

toutes les composantes des termes présents dans la collection et la requete. Il s'agit en fait d'une classe de moteurs de recherches qui agissent comme s'ils ordonnaient les documents après une recherche booléenne ne rapportant que les documents qui contiennent au moins un terme dans la requete[20]. Les documents sélectionnés reçoivent un score représentant leur similarité avec la requête. Puis ces documents sont triés en fonction de leur score.

II.2.2.2 Optimisation des calculs

Un moteur de recherche qui fournirait la liste complète de tous les documents ordonnés par leurs scores serait beaucoup trop lent. Les deux solutions sont l'évaluation paresseuse et l'approximation des calculs. En ne calculant que la liste des documents possédant au moins un terme de la requête, il reste souvent suffisamment de documents dans le corpus pour que la recherche prenne plusieurs secondes. La **posting-list** est la liste des documents d'un corpus qui contiennent au moins un mot d'une requête.

posting-list

L'idée est que le système utilise la structure de l'index pour ne calculer une première approximation des scores des documents que pour les documents ayant le plus de chance d'avoir le plus grand score[21]. Il existe des approximations et des organisations internes d'index qui garantissent que les

[20]Les documents et les requêtes sont supposé ne jamais être vide, la norme des vecteurs est donc défini positive et supérieur à zéro.

[21]Soit par des entrées de l'index permettant d'accéder directement à la liste des documents contenant certains mots ; Soit par des pré-calculs à l'indexation de valeur de tf ou idf encodés par des compressions avec perte ; Soit d'autres techniques d'indexation d'index, zones, ajout de pointeur et clusterisation, etc...

premiers documents retournés sont bien ceux qui ont le score le plus élevé de toute la collection. Le moteur réalise le calcul exact uniquement sur ces documents.

Après élimination des facteurs calculables avant l'évaluation de la requête (ou substitués par d'autres) et des simplifications sur le calculs du produit scalaire [Manning *et al.*, 2008], il est possible d'obtenir une fonction de calcul du score pour un document de la forme[22] :

$$Score(Requete, Document) = \Sigma_{t_i \in Requete} Tf(t_i, Document) * Idf(t_i)$$

L'attribution des scores des documents dépend aussi du type d'index et du type de la requête. Certains moteurs de recherche comme MG [de Kretser & Moffat, 2000] n'utilisent pas l'évaluation paresseuse. Ils demandent en plus des mots à rechercher un nombre maximum de documents à fournir. Il est *impossible* de connaitre le score des documents qui auraient suivi sans faire de nouvelle requête, mais il est alors possible de déployer d'autres optimisations ou des règles particulières. Par exemple, il est *possible* d'imposer qu'un document, étant le seul à posséder un certain mot de la requête, doit figurer dans la liste des documents résultats même s'il a un score plus faible que le dernier des documents.

II.2.3 Présentation du moteur de recherche Lucene

Notre SQR utilisant Lucene, nous allons le considérer plus en détails. Le système Lucene-Java est un projet libre qui a pour but de fournir un accès aux technologies d'indexation et de recherche avec une spécialisation sur les documents de type texte.

Le moteur de recherche Lucene est un projet sous licence Apache. Celle-ci nous autorise notamment à lire et/ou modifier son fonctionnement pour notre propre usage. Son architecture modulaire, commentée et documentée facilite la réutilisation. Son architecture est celle d'une bibliothèque modulaire conçue pour créer des moteurs de recherche. Lucene a été conçu spécialement dans ce but.

Lucene est un moteur de recherche[23] du même type que ceux utilisés dans les SQR (il y est même parfois utilisé). Il partage de nombreux points avec d'autres systèmes utilisés dans le même cadre.

Nous le présentons donc aussi comme un représentant du domaine.

[22] La documentation de Lucene détaille ces calculs et la manière de les modifier, http ://lucene.apache.org/java/2_9_0/api/all/index.html

[23] La bibliothèque dispose d'un fonctionnement «par défaut» dont la mise en œuvre la plus simple de quelques classes constitue un moteur de recherche.

Une utilisation typique de Lucene se décompose en deux étapes dans deux processus séparés, l'indexation et la recherche. Lucene-Java utilise de nombreuses techniques au plus haut niveau de l'état de l'art dont certaines font l'objet de brevets.

Optimisations et contraintes de l'indexation

L'indexation est réalisée par un système générique de gestion d'étiquettes[24] et de noms de champs. Les étiquettes, champ par champ, sont extraites via une analyse *ad hoc* fournie par le programmeur qui utilise Lucene.

Lucene s'impose la contrainte de pouvoir réaliser de l'indexation incrémentale. Les documents sont insérés un par un dans l'index et après l'insertion de chaque document, l'index est dans un état utilisable.

Les termes sont indexés dans une structure où chaque document est représenté par un numéro local. Pour chaque document un vecteur de termes avec leurs nombres d'occurrences est créé. Nous parlons alors de document inversé. Cela permet d'utiliser l'une des mesures de similarité précédentes pour obtenir les numéros de documents concernant un groupe de termes.

L'indexation consiste pour chaque document à en calculer la forme inversée, puis à l'ajouter à une structure de hachage. Cette structure à la propriété d'avoir un coût faible presque constant pour accéder à la *posting-list*. Si en première approche l'indexation ne présentent aucun problème, il en est tout autrement lorsqu'il s'agit de passer à l'échelle de très grosses collections.

Une optimisation est recommandée après l'ajout de tous les documents dans l'indexeur. Afin de tenir compte du dynamisme de l'index[25], celui-ci est en réalité composé de plusieurs sous-index appelés «segments». Chacun regroupe une quantité de documents dépendant du volume cible total de données. L'évaluation d'une requête est donc censée parcourir tous les segments et simuler l'existence d'un unique index pour l'utilisateur. Ceci prend évidemment plus de temps, car tous les accès sont à calculer pour chaque segment et une fusion finale doit avoir lieu. L'optimisation de l'index consiste donc en une factorisation des segments [26].

[24]Les étiquettes sont appelées termes dans Lucene. Les termes Lucene sont souvent des mots, mais de manière générale se sont les unités d'indexation.

[25]L'enjeu derrière l'existence d'un index dynamique est de ne pas avoir à ré-indexer la totalité de la collection après l'ajout/retrait de quelques documents. Un autre usage est de modifier la manière dont sont indexés les documents sans créer d'indisponibilité de service. Indexer une grosse collection peut prendre plusieurs jours. Le gain peut être conséquent.

[26]Nous avons observé que cette optimisation de l'évaluation des requêtes prend 5% du temps total de l'indexation. Cette opération n'est donc clairement pas réalisable après l'ajout de chaque document.

Modularité pour des modifications simplifiés

Au niveau de l'index, il est possible d'ajouter des valeurs arbitraires. Par exemple avec les méthodes par défaut il est possible d'ajouter une pondération spéciale pour un document. Le score du document est affecté multiplicativement par le coefficient (hors de la *posting-list* le score est toujours de zéro). Cela peut être utile pour augmenter la confiance dans certaines sources.

De même, une autre valeur ajoutée arbitrairement concerne l'ajout d'un facteur de normalisation des documents. C'est un coefficient utilisé multiplicativement pour compenser l'impact des différences de longueur de documents sur le score. Ce score de normalisation est une manière prévue pour modifier le poids de certains documents dans la méthode d'attribution des scores.

Le système Lucene propose un parseur de requête. Celui-ci analyse les requêtes exprimées dans une syntaxe précise et construit une instance d'une classe préécrite correspondant au bon type de la requête. Cette instance permet de réaliser la recherche. Ces instances de classe de requête sont construites à partir d'une méthode d'attribution de score au document. L'attribution des scores utilise une classe spécialisée dans le calcul de similarité sur tout les documents de la *posting-list*. La *posting list* est extraite par ailleurs depuis le lecteur d'index. Ces quelques classes donnent une idée de la modularité de Lucene et des classes à étudier en vue de l'adaptation à nos besoins.

II.2.4 Le découpage en paragraphes

Les documents sont généralement découpés en paragraphes de taille équivalente avant leur indexation. Une raison est de contrôler plus finement la durée des calculs sur les documents tant au niveau de la recherche de ces documents, que de la sélection des phrases ou réponses. Une autre raison est d'éviter que les documents longs et abordant de nombreux sujets, ne soient sur-évalués par les métriques de calcul de similarité par rapport à des documents plus courts[27].

II.2.4.1 Comment découper

Il faut noter que la longueur importante des paragraphes ne se mesure pas en caractères, mais en unités lexicales (ou étiquettes) qui servent de base aux dimensions des comparaisons. D'autre part, il convient de choisir une

[27] Des vecteurs ayant des nombres proches de composantes non nulles sont préférables

longueur commune à tous les documents qui soit compatible avec au moins trois contraintes.

- La première contrainte est qu'il ne doit pas exister de document plus petit que la longueur de projection choisie.
- La seconde est que les documents ne doivent pas être trop longs pour que le calcul du produit scalaire ou du $tf.idf$ ait encore un sens.
- La troisième est qu'il doit exister suffisamment de souplesse dans le nombre d'unités lexicales tolérées pour permettre un découpage convenable des limites des phrases.

Ces contraintes sont partiellement antagonistes, les documents ne doivent être ni trop petits ni trop longs et avoir une tolérance sur la longueur moyenne qui ne dénature pas complètement les calculs. La plupart des systèmes choisissent de découper leurs paragraphes en fonction du nombre d'unités lexicales, la quantité choisie est de l'ordre de 100 à 1000 unités, la tolérance est rarement spécifiée. Il existe aussi des systèmes qui réalisent des découpages en volume de caractères (comme Qristal). Le compte en caractères simplifie la gestion des temps de traitement au détriment de la qualité de la cohérence de la mesure de similarité.

II.2.4.2 Découpage en paragraphe dans Musclef

Dans le système Musclef, nous avons choisi de découper nos corpus en paragraphe d'environ 300 unités lexicales et avec une limite dure à ne pas dépasser de 450 unités. Il faut noter que certains documents sont bien en dessous de la limite des 300. Le système de nettoyage et d'indexation de documents de Musclef ne fait pas de fusion de documents, car il perdrait l'information sur l'origine des phrases. Les annotations ne sont pas comptées comme étant des unités lexicales. Au final, une fois les documents découpés et séparés en fichiers distincts[28], il peut y avoir des différences de taille de fichier d'un facteur 10, mais les différences entre les nombres d'unités lexicales sont plus modérées.

II.2.4.3 L'indexation des paragraphes

L'indexation des paragraphes résulte essentiellement de la transformation d'un paragraphe en un flot d'unités lexicales qui est une abstraction des données à indexer pour le moteur de recherche. Si le moteur de recherche et d'indexation ne permet pas la construction d'index à base de flots séparés, alors il est nécessaire de faire des choix. Dans les systèmes où le découpage

[28]Notre utilisation de Lucene requiert que lors de l'indexation tous les documents soient dans des fichiers distincts

en paragraphe est réalisé indépendamment de l'annotation, il est évident que le flot d'étiquettes est généré en ignorant les annotations. Un des buts du découpage étant d'obtenir des espaces de projection ayant le même nombre de dimensions, mettre les annotations dans le flot casse ce travail. Par contre, il est possible de se demander quelle forme lexicale est envoyée dans le flot d'étiquettes. Celle d'origine du document ? Une forme lémmatisée ou stemmée ? Cela est à déterminer au moment de choisir la stratégie de formatage et d'utilisation des requêtes pour le moteur de recherche. À notre connaissance ce point pourtant important est rarement abordé dans la littérature des SQR, nous reviendrons sur les conséquences au chapitre IV (page 113). Dans les systèmes où le découpage est réalisé au volume de caractères, il est important d'intégrer la totalité des unités lexicales présentes dans le paragraphe. En effet, même si le nombre d'unités lexicales de texte brut est différent d'un document à l'autre, les annotations aussi sont projetées dans l'espace de comparaison. Avec cette stratégie le choix des annotations et symboles utilisés est délicat à cause du recouvrement avec le lexique de la langue.

II.2.5 Le problème de la recherche des passages

Chercher un passage d'un document implique de choisir une méthode de recherche, ou du moins un moteur de recherche.

Différents moteurs de recherche ont été utilisés dans les différentes Trec/Clef (Cf section II.1.4 page 60), dont le moteur Lucene ; dans le JIRS [Gomez *et al.*, 2005] [29] qui est aussi un projet libre, contrairement à Lucene les réglages sont accessibles via des fichiers de configuration en XML. Ce genre de réglages est plus simple à mettre en place, mais ne permet pas autant de variabilité qu'avec un chargement de nouveaux modules comme dans Lucene. L'un des autres moteurs de recherche les plus fréquemment utilisés est M-CAST[30] qui est un système d'agrégation de documents basé sur le moteur TRUST. TRUST[31] est un moteur utilisant intensivement des ressources sémantiques : il a été développé dans le cadre d'un projet européen pour la recherche et développement.

Les contraintes à gérer imposent évidemment à tous ces moteurs qu'ils soient multilingues, c'est-à-dire qu'ils aient un comportement indépendant de la langue dans laquelle sont rédigées les données, mais qu'ils disposent d'outils, intégrés ou non, pour tenir compte des spécificités des langues.

[29] JIRS : Java Information Retrieval System

[30] Multilingual Content Aggregation System based on TRUST, http ://www.m-cast.infovide.pl/

[31] Text Retrieval Using Semantic Technologies (IST-1999-56416)

De même, ils doivent pouvoir supporter une quantité de documents correspondant aux volumes réunis de la Wikipédia et d'un corpus de journaux dans une forme découpée en petits passages. La vitesse d'interprétation des requêtes et d'extraction des documents est aussi un des aspects qui pèse dans le choix de ces moteurs de recherche. Il est indispensable de réduire l'attente de l'utilisateur en situation réelle, et donc que le temps de réaction du moteur de recherche se situe sous la seconde sur une machine ordinaire. Par ailleurs, si nous voulons participer aux campagnes d'évaluation qui dure environ une semaine, le temps d'indexation représente un élément non négligeable. Les facilités de formatage des documents offertes par les indexeurs des moteurs de recherche sont donc aussi à prendre en compte.

II.2.6 Synthèse

Nous venons de voir les différents problèmes auxquels les moteurs de recherche apportent des solutions. Nous pouvons maintenant voir les conséquences que peuvent avoir tels ou tels choix de conception. Dans le cadre de l'étude des questions enchaînées, nous serons amené à étudier les modifications à apporter aux moteurs de recherche.

II.3 Conclusion

Nous venons de présenter notre thématique et les technologies utilisées par les SQR. Les SQR sont composés d'une étape d'analyse linguistique hors contexte des questions, puis d'une recherche de documents à l'aide d'un moteur de recherche généraliste, puis d'une sélection de plus en plus raffinée des passages dans les documents afin d'obtenir la réponse et l'extrait justifiant qu'il s'agit bien de la réponse.

Les extensions actuelles permettant de traiter les questions enchaînées portent surtout sur la gestion des anaphores co-référentes. Les conséquences sur le moteur de recherche ou l'impact de la réponse sur les questions suivantes sont peu étudiées. Nous pouvons observer aussi que là où les questions enchaînées peuvent être un cadre intéressant pour se rapprocher du dialogue homme-machine, aucun modèle n'est développé ou adapté en ce sens.

Nous allons donc nous concentrer sur la résolution de ces aspects ainsi que leurs adaptations aux SQR.

Chapitre III

Analyse des questions

Après avoir présenté les systèmes de questions réponses, et, en particulier, ceux qui gèrent des enchaînements de questions, nous avons pu constater que l'une des principales étapes concernées est l'analyse des questions. Nous allons présenter dans ce chapitre les évolutions de ce module que nous avons mis en œuvre au sein de la chaine Musclef sur laquelle nous avons travaillé.

Nous commençons par une étude détaillée des liens possibles entre les questions dans le corpus de la campagnes Clef. Nous formalisons les liens observés et proposons une représentation, ainsi que le moyen de la construire. La dernière partie décrit la mise en œuvre de la représentation proposée ainsi que son évaluation.

III.1 Représentation des liens entre les questions

Il est utile d'être d'accord sur ce que nous appelons une question. Une « question » est [Informatique, 2006] : *Une interrogation adressée à quelqu'un pour obtenir un renseignement ou une explication, vérifier des connaissances.* C'est cette définition qui est la plus proche de la sémantique de ce qu'est une requête pour un corpus. Une question peut aussi être : *Un sujet, point problème donnant lieu à discussion ; une affaire où une chose précise est en jeu.* Comme le mot «interrogation[1]» est défini en utilisant le mot «question», il existe un cycle sémantique. Pour contourner ce cycle, nous posons la double définition suivante.

Une question est, *pour un système*, une requête dans un corpus de documents ; une question est, *pour un utilisateur*, l'énoncé détaillé d'un besoin d'information. Le but de l'analyse des questions est de passer de l'une de ces définitions à l'autre, soit d'un besoin d'information à une requête.

III.1.1 Étude des liens entre les questions

Afin d'étudier les liens entre les questions, nous avons réalisé une analyse sur un corpus de questions de campagne d'évaluation de SQR. Le résultat de cette analyse est de montrer qu'étant donné les types de liens entre questions et leurs ressemblances, nous pouvons dégager de grandes lignes d'un modèle formel pour toutes les catégories de questions enchaînées.

III.1.1.1 Présentation du corpus de questions de la campagne ClefQA07-FR-EN

Le corpus que nou savons retenu pour l'étude des liens entre questions est celui de la campagne ClefQA2007 sur les questions enchaînées. Plus précisément nous avons retenu le corpus de questions de ClefQA07-FR-EN avec des questions en français attendant des réponses extraites de documents en anglais. Les questions enchaînées sont disponibles via des groupes. Il existe 53 groupes contenant au moins 2 questions, pour un total de 186 questions, dont 133 en position 2+[2]. Les plus gros groupes sont les groupes de 4 questions, ils sont aussi les plus nombreux (37 occurrences). Il y a 10 groupes de 2 questions et 6 de 3 questions. D'après la description de la tâche ClefQA07-FR-EN

[1] Action d'interroger, de questionner

[2] La position d'une question est son numéro d'apparition dans le groupe. Le rang d'une question est la longueur+1 de la chaine de liens qu'il faut résoudre pour y répondre. Nous préciserons plus tard cette définition de «rang».

1	Où se trouve la cathédrale Sainte-Sophie en Russie ?
2	Qui était son archiprêtre en 1995 ?
3	Quel Écossais a construit la cathédrale ?
4	Quelle impératrice russe l'a accrédité pour la construire ?

TAB. III.1 – Exemple d'un groupe de questions enchaînées tirées du corpus utilisé pour la campagne d'évaluation Clef 2007.

, les questions sont censées être liées uniquement à la première question ou à sa réponse.

Le tableau III.1 montre un des groupes de questions de ce corpus. La question 4 n'est pas directement liée à la question 1 car l'anaphore «l'(a)» n'a pas de référent dans la première question mais dans la réponse soit de la seconde soit de la troisième question. En effet, l'action d'accréditation se rapporte par définition à une personne. Or seules les réponses aux questions 2 et 3 sont des référents compatibles avec cette anaphore.

III.1.1.2 Étude des liens entre questions

Afin de tenter de comprendre et maitriser les questions enchaînées notre démarche commence par une analyse des catégories de groupes de questions enchaînées que nous pouvons rencontrer.

Comme il existe une distinction entre les mots d'une question et les liens qu'ils impliquent avec les autres questions nous introduisons la notion de «thème». À chaque terme est associé une sémantique. Nous appelons «thèmes» les sémantiques de chaque terme.

Nous pouvons diviser les groupes de questions enchaînées en trois grandes catégories.

1. Les groupes où toutes les questions de position 2+ ont un lien avec la première question qu'il faut rétablir pour les comprendre.
2. Les groupes où toutes les questions sont indépendantes, ou, plus exactement, peuvent s'interpreter indépendamment les unes des autres.
3. Les groupes où la nature des liens change d'une question à l'autre.

Voyons les plus en détail.

A) Liens forts

Il y a des groupes dont toutes les questions de position 2+ ont un lien avec la première question. Nous observons au moins deux variantes à ce modèle

simple (**Simple**). Sans que le lien aille vers une autre question, nous observons des cas où le lien va vers des thèmes qui se substituent (**Substitutif**) de question à question. De même, il est possible que les questions réutilisent de plus en plus de thèmes de la première question (**Incrémentale**).

Dans les exemples ci-dessous les thèmes réutilisés sont tous les mêmes (**Simple**).

Thème(s) à reprendre	n°	Question
	1	Quel est le métier de John Barbirolli ?
«John Barbirolli»	2	Qui était le soliste et violoncelle dans le Concerto pour violoncelle d'Elgar qu'il enregistra en 1965 ?
«John Barbirolli»	3	Citer le nom d'un orchestre de Manchester qu'il conduisit.
«John Barbirolli»	4	Citer le nom d'un orchestre américain qu'il conduisit.

Dans les exemples suivants, les thèmes réutilisés se substituent (**Substitutif**) les uns aux autres de question à question, bien qu'ils appartiennent tous à l'ensemble des thèmes de la première question. Ce type de groupe est fréquent dans le cas des questions complexes comme le montre les décompositions de questions évoquées à la section I.2.2.1 (page 34), cela donne autant de thèmes qui peuvent être approfondis.

Thème(s) à reprendre	n°	Question
	1	À combien de personnes a-t-on demandé de quitter leur domicile durant les inondations aux Pays-Bas, en hiver 1995 ?
«aux Pays-Bas» «en hiver 1995»	2	À quelle hauteur l'eau est-elle montée à Lobith durant les inondations ?
«les inondations» «aux Pays-Bas»	3	Combien tuent-elles de gens tous les ans ?

Enfin, dans ce dernier exemple, les questions réutilisent de plus en plus de thèmes de la première question (**Incrémentale**).

Thème(s) à reprendre	n°	Question
	1	Combien de places y a-t-il dans la voiture électrique Impact de General Motors ?
«voiture électrique Impact de General Motors»	2	Quelle est approximativement l'autonomie de la voiture en miles ?
«voiture électrique Impact de General Motors»	3	Combien de temps faut-il pour accélérer de 0 à 60 miles par heure ?
«réponse de 1» «voiture électrique Impact de General Motors»	4	Existe t-il un modèle avec plus de place ?

B) Liens absents

Il y a des groupes où aucune question n'a de lien vers une autre (**Aucune**). Si nous cherchions absolument à trouver des liens entre ces questions, nous devrions pouvoir en trouver (comme Royaume-Uni). Mais ces questions sont déjà intelligibles, bien formées et suffisamment précises pour que nous puissions nous accorder sur une réponse en les considérants isolément. Il n'y a donc pas de lien entre ces questions. Les multiples occurences du thème de «la ville de Perth» ne rendent pas nécessaire l'existence de liens puisque les informations sont explicites.

n°	Question
1	Dans quel pays du Royaume-Uni se trouve la ville de Perth ?
2	Quelle route est connue comme étant la route des "motor mile" de Perth ?
3	Nommer des employeurs à Perth.
4	Citer le nom d'un fleuve qui traverse Perth.

C) Lien de type variable

Il y a des groupes où le type des liens change de question à question. Nous essayons donc de proposer une catégorisation qui permet de représenter l'existence de chaque type de lien :

- Les liens Non Triviaux (**Non Trivial**), où une question a un lien avec une autre question que la première du groupe. Au moins, une question réutilise des thèmes d'une question de position 2+ :

Thème(s) à reprendre	n°	Question
	1	Où se trouve le musée de l'Ermitage ?
«l'Ermitage»	2	Qui était le directeur du musée en 1994 ?
«le directeur du musée en 1994» «l'Ermitage» «réponse de 2»	3	Nommé par qui ?

- Les liens d'une question avec réponse d'une question précédente. C'est la catégorie (**Avec réponse**). Au moins, une question a un lien avec la réponse d'une autre question :

Thème(s) à reprendre	n°	Question
	1	Où se trouve la cathédrale Sainte-Sophie en Russie ?
«réponse de 1» «en Russie»	2	Combien y a-t-il d'habitants dans cette ville ?

- Les groupes avec des questions ayant des liens vers la première question uniquement et d'autres n'ayant aucun lien. Ces groupes forment une catégorie spéciale ; (**Mélangé**). Il s'agit d'un mélange entre groupe de type **Simple** et groupe de type **Aucune** :

Thème(s) à reprendre	n°	Question
	1	Quel océan Steve Fosset a-t-il traversé en ballon en février 1995 ?
«Steve Fosset» «en février 1995»	2	Quel âge avait-il à ce moment-là ?
«Steve Fosset» «en février 1995»	3	Quel était son métier ?
	4	Quelle distance Richard Branson a-t-il parcouru en ballon en 1991 ?

Evidemment sans les précisions «en ballon» et «en 1991» la questions quatre n'aurait eu de sens qu'avec l'existence d'un lien vers la première question.

- La variante de ce mélange où non seulement une question n'a pas de lien, mais où créer un lien est incorrect. C'est une catégorie de groupe avec (**Remplacement**) Le thème a été complètement remplacé. Un des thèmes réutilisé dans une question de position 2+ n'est pas adapté à la résolution d'une question ultérieure et complètement remplacé par un autre thème :

Thème(s) à reprendre	n°	Question
	1	Quel était le nom de la barge qui a coulé à Porto Rico le 7 janvier 1994 ?
«Porto Rico» «le 7 janvier 1994»	2	Qu'a heurté la barge ?
«Porto Rico» «le 7 janvier 1994» «la barge»	3	Quelle quantité de mazout a été déversée dans l'eau ?
«Porto Rico»	4	Combien de miles de plage compte-t-elle en 1998 ?

Les différents types de groupes peuvent se combiner pour former un groupe de questions «Non Trivial» avec «Remplacement».

Nous pouvons observer que la nature des liens entre les questions n'est pas limitée aux co-références anaphoriques. Ainsi dans l'exemple précédent le lien entre les questions 1 et 3 est implicite ou situé sur un niveau logique assez élevé. Nous pouvons aussi observer que la deuxième question tend à suggérer un renforcement sur la thématique de «la barge», et que celle-ci serait reprise implicitement dans la question 3. Nous ne pouvons donc pas affirmer qu'il n'y a pas de lien (aussi) entre 2 et 3. Mais comme la question 1 contient toute l'information nécessaire à la résolution de 3, dans un premier temps nous ignorons ce détail.

Nous retrouvons une situation similaire dans l'exemple suivant :

Thème(s) n°	Question
1	Qui était le directeur du musée en 1994 ?
2	Nommé par qui ?

Plusieurs indices nous prouvent que la seconde question est liée à la première comme. Le fait que l'action de nommer (dans le sens «élire») s'applique préférentiellement aux personnes, que la juxtaposition d'une question «longue» et d'une question «courte» (en nombre de caractères), qui suggère une réutilisation très forte des mots dans un but d'économie.

Les limites des possibilités de liens sont probablement celles des inférences de l'esprit. Il est donc inutile de chercher à faire une liste exhaustive des types de liens possibles.

III.1.1.2.1 Analyse du corpus Nous avons annoté à la main le corpus ClefQA07-FR-EN . Nous pouvons alors observer les occurrences des différents

Groupe	de taille 4	de taille 3	de taille 2
Simple	10	3	5
Substitutif	0	0	0
Incrémental	5	0	0
Aucune	6	2	3
Non Trivial	2	0	0
Avec réponse	4	0	2
Mélangé	1	0	0
Remplacement	11	1	0

TAB. III.2 – Une classification des phénomènes de liens entre questions.

types de groupe, taille de groupe par taille de groupe dans le tableau III.2[3] (page 86).

Ce corpus propose plus de questions groupées que les questions des autres langues de la campagne Clef[4]. Les différentes complexités peuvent se combiner pour former un type complexe. Par exemple, les groupes de type «*Non trivial* et *Réponse de précédente*» sont les groupes avec des références vers des questions de position 2+ qui réutilisent une réponse. Les groupes de type *Aucune* sont intéressants dans la mesure où les questions les composant portent sur des thèmes communs que nous pourrions regrouper sous une étiquette «thème global».

III.1.1.2.2 Utilisation des catégories Dans le tableau III.2, si nous réalisons les totaux nous constatons qu'il y a plus de 53 groupes. Nous pouvons en compter 55. Certains groupes ont les attributs de plusieurs lignes du tableau, ils y figurent donc plusieurs fois. Grâce à ce tableau, nous comprenons mieux la variété des types de liens des questions. Le type 1.1 n'est pas représenté à ClefQA2007 , mais il existe dans d'autres corpus. Il est possible qu'une question avec un thème de **Substitutif** puisse servir d'introducteur à une question sur un thème connexe à celui déjà introduit, mais qui ne partage pas les mêmes thèmes que les questions précédentes (un **Remplacement**). Plus le groupe est grand plus cela a de chance de se produire. Cela peut avoir pour conséquence des différences de classification des groupes de questions.

[3]Un groupe de taille 2 n'est pas forcement de type *Simple* ou *Aucune*, car il est possible que les termes de la première (α) ne doivent pas être tous repris dans la seconde (β). Cela peut aussi être un groupe de type *Remplacement*.

[4]La campagne Clef est réalisée pour plusieurs couples de langues. Dans chaque couple les questions sont différentes.

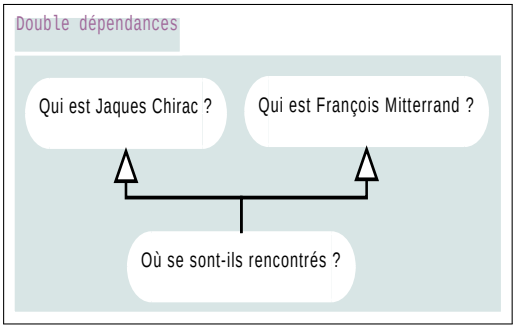

FIG. III.1 – Un groupe avec une question en position 3+ avec une dépendance vers deux questions n'ayant pas de lien entre elles.

III.1.1.3 Généralisation du problème

Nous pouvons aussi nous intéresser à des questions qui n'auraient pas été relevées dans les campagnes d'évaluation. Par exemple les doubles liens comme dans l'exemple suivant :

> Qui est Jacques Chirac ?
> Qui est François Mitterrand ?
> Où se sont-ils rencontrés ?

À notre connaissance il n'existe pas actuellement de système qui puisse gérer ce genre de construction de référent à double antécédent (figure III.1). Nous pouvons imaginer que d'autres problèmes plausibles, non révélés à cause des biais d'expérimentations doivent exister. Nous allons proposer un modèle qui permet de généraliser les cas identifiés dans notre corpus.

III.1.2 Formalisation en dépendances unitaires

Nous avons vu que nous pouvons ancrer la notion de lien entre questions dans un cadre de dialogue homme machine. Nous aimerions aussi disposer d'un système cohérent avec les divisions en modules traditionnellement réalisées dans le domaine des SQR ; car c'est la définition rigoureuse de modèles, qui (quand ils sont bien choisis) permet de gagner en modularité[5]. Nous choisirons alors de garantir la cohérence de la représentation des liens (entre questions) par une description probabiliste.

Si les données issues des liens entre questions peuvent améliorer la recherche de la réponse alors comment les organiser, ré-utiliser pour que cette recherche soit plus efficace ? Puis comment créer un modèle cohérent de ces liens ?

[5]Que ce soit dans un domaine de QR ou de DHM.

III.1.2.1 Réutilisation des éléments d'une questions passée

Comme nous l'avons déjà vu à la section II.1.2 (page 54), dans la littérature, nous trouvons essentiellement trois approches possibles pour réintroduire les éléments d'une question liée. L'ajout des mots de la question liée à la fin de la question, l'unification des termes dans une structure syntaxiquement correcte et la restriction de l'espace de recherche par réutilisation des documents trouvés à la première recherche. Voyons les limites de ces approches dans le cadre de la construction d'un système formel pour la représentation des liens entre les questions (pour les questions enchaînées).

Thèmes de l'exemple II.3 pour les questions liées à la quatrième du groupe :

1	Où se trouve [la cathédrale Sainte-Sophie en Russie] ? [Novgorod]
3	Quel [Écossais] a construit [la cathédrale] ? [Vladimir of Novgorod]
4	Quelle impératrice russe l'a accrédité pour la construire ?

III.1.2.1.1 Fusion syntaxique des questions

Outre les difficultés de formulation qui peuvent être rencontrées, cette approche est limitée à ce que nous pouvons restructurer dans une phrase sans trop perturber le système d'analyse de la question. La question «Quelle impératrice russe l'a accrédité pour la construire ?» peut se ré-écrire de nombreuses manières[6] :

- Ajout à la fin avec séparation des termes réutilisés par un ';' : *Quelle impératrice russe l'a accrédité pour la construire ; [un] écossais, Vladimir [of/de] Novgorod, cathédrale Sainte-Sophie [en] Russie ?*
- Unification syntaxique des termes : *Quelle impératrice russe a accrédité [[un] écossais/Vladimir [of/de] Novgorod] pour construire la cathédrale [Sainte-Sophie [en] Russie]] ?*

Pour les SQR(s) inter-lingues, la traduction de la réponse doit aussi être prévue. Ceci complexifie les problèmes précédents.

III.1.2.1.2 Fusionner des représentations des requêtes

Une autre stratégie consiste à ajouter les mots de la question liée directement dans la structure des termes destinés à la requête du moteur de recherche. Cet ajout ne permet pas de faire de distinction entre ce qui vient de la question courante ou de la précédente.

Dans le groupe du tableau II.3 (page 61), après la résolution d'anaphores, il n'est plus possible de savoir que le terme «cathédrale Sainte-Sophie Russie»

[6] «Vladimir of Novgorod» est la réponse de Musclef à la question 3 du tableau II.3 (page 61).

provient de la première question. Il est possible d'avoir simultanément deux documents dans lesquels nous trouvons qu'un Écossais a construit quelque chose et un autre dans lequel nous trouvons qu'une personne a construit une cathédrale en Russie. Mais il est possible que nous ne disposions pas de documents parlant simultanément d'Écossais, de cathédrale et de Russie. Quel poids accorder à ces différents termes quand nous ne savons pas d'où ils viennent ?

Par exemple : avec le thème de la cathédrale de Novgorod, dans l'ensemble des documents en anglais de ClefQA2007 , il était possible de trouver le nom d'une personne ayant construit la cathédrale[7]. Mais rien ne confirme ou n'infirme que cette personne est de nationalité écossaise[8].

Ce n'est donc pas satisfaisant, d'autant moins si les groupes de questions forment des séquences où les termes sont repris implicitement ou avec une variation/oubli partiel. Cette approche conduit à une perte de performance en fonction de la longueur des séquences de questions.

III.1.2.1.3 Mémorisation des documents des premières questions

Nous pouvons remarquer qu'avec cette stratégie, une question dépendant de n autres, dispose alors d'un ensemble de recherche réduit n fois. Il n'y a aucun moyen de re-ouvrir l'espace de recherche. Pour des séquences de questions avec des dépendances, cette stratégie n'est pas généralisable. Par exemple, imaginons que dans le groupe de questions sur la cathédrale Sainte-Sophie, nous ayons déjà analysé les dépendances entre questions et que nous puissions restreindre l'étude à la séquence de questions 1 3 4 (du tableau II.3 (page 61)). La résolution de la question 4 se fera sur les seuls documents qui parlent de «cathédrale Sainte-Sophie en Russie» et d'«Écossais ». Si la réponse est dans un document lié au nom de l'Écossais seul sans sa nationalité, alors la réponse ne sera pas trouvée.

L'impact de la taille des paragraphes est évidemment très important. Trop gros ils ne restreignent pas suffisamment la recherche, et la seconde question est traitée presque sans tenir compte des termes des précédentes. Trop petits ils la restreignent trop, et par conséquent il n'y a plus assez d'informations autour de la réponse. Pour trouver une taille de paragraphe adaptée aux questions enchaînées, il faudrait aussi que les questions réfèrent toujours à des informations proches dans les paragraphes des documents réponses des

[7]Vladimir of Novgorod, cf Article wikipédia anglais «Saint_Sophia_Cathedral_in_-Novgorod» de la cathédrale Sainte-Sophie de Novgorod

[8]Une analyse approfondie avec un moteur d'inférence aurait peut-être trouvé la nationalité de cette personne.

premières questions. Il faut aussi que cette proximité puisse être quantifiée afin de définir la taille adaptée des paragraphes.

III.1.2.1.4 Synthèse des 3 approches

Dans les 3 approches présentées ci-dessus, les termes ont la même importance, ce qui pose problème et ne permet pas une recherche efficace. Aucune de ces stratégies n'est satisfaisante. Plus généralement, la quantité de termes pour une recherche unique augmente linéairement avec la profondeur des dépendances entre questions. Le problème se pose pour le choix des termes : ne prendre que ceux de la question ? Ou prendre tous ceux de toutes les questions liées ? C'est un problème de contrôle du bruit par rapport au silence dans le nombre de documents retournés par le moteur de recherche.

III.1.2.1.5 Le but de la recherche des liens

Dans un groupe de questions enchaînées, les contraintes sur la recherche d'information sont les termes issus des liens entre questions ainsi que les termes de la question courante. Ainsi, nous voulons proposer une nouvelle structure permettant de lier des contraintes classiques de recherche d'information, à des contraintes concernant l'ordre dans lequel ces termes peuvent être relaxés en cas de silence. Nous voulons donc obtenir un ordre de relaxation des contraintes (les termes). L'utilisation de cette structure suppose de concevoir une stratégie de relaxation qui soit adaptée au domaine d'application du SQR.

Toute relaxation de contraintes sera alors réalisée en tenant compte des performances de la recherche. Une recherche plus performante pourra être mise en place et, éventuellement, une justification des termes choisis fournie à l'utilisateur. Dans le cas idéal, nous sommes aussi convaincus que les SQR capables de résoudre des questions enchaînées de cette manière sont une étape vers des systèmes de dialogue homme-machine en *domaine ouvert*.

Nous voulons donc créer une structure qui permet de représenter les dépendances d'un groupe afin d'améliorer la recherche dans les documents. Nous nous intéressons plus particulièrement à la partie structure et organisation du problème. C'est ce que nous allons présenter maintenant.

III.1.2.2 Présentation de la structure des dépendances

Nous avons montré les propriétés que devait avoir une structure pour représenter plus finement les dépendances. Nous allons maintenant la préciser avant d'expliquer comment la construire et d'évaluer sa capacité à trouver ces dépendances.

1	Où se trouve la cathédrale Sainte-Sophie en Russie ?
2	Qui était son archiprêtre en 1995 ?
3	Quel Écossais a construit la cathédrale ?
4	Quelle impératrice russe l'a accrédité pour la construire ?

TAB. III.3 – Exemple d'un groupe de questions enchaînées tirées du corpus utilisé pour la campagne d'évaluation Clef 2007.

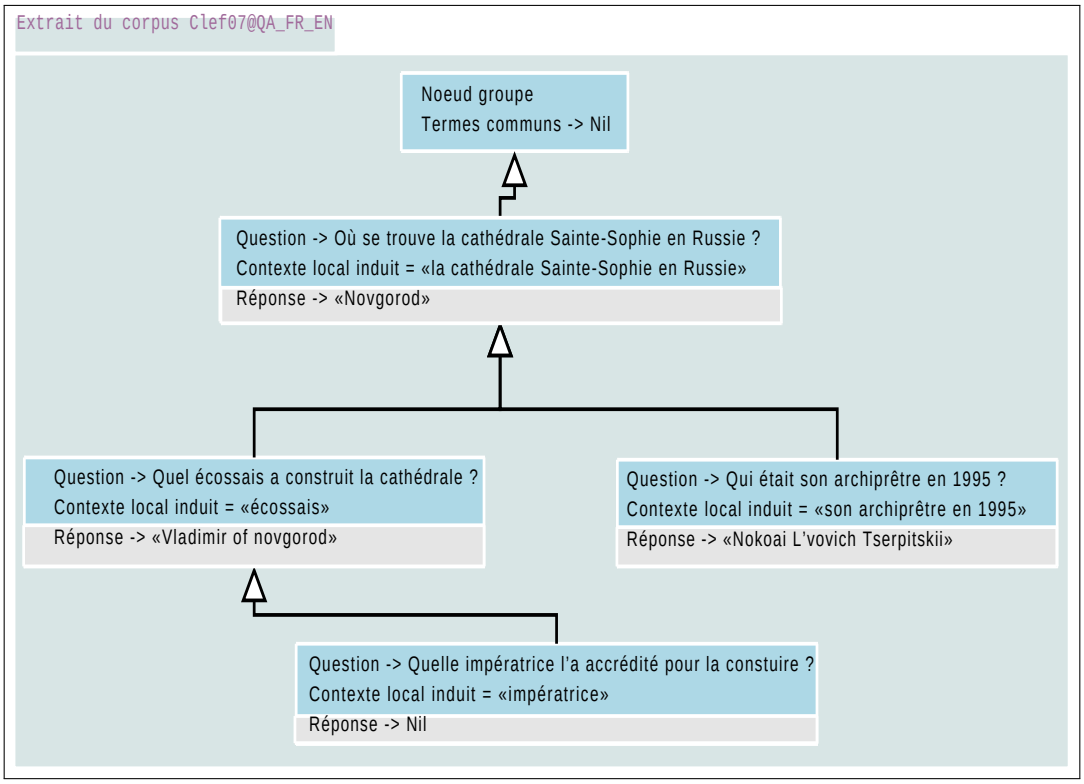

FIG. III.2 – L'arbre correspondant au groupe du tableau III.3

En s'inspirant des travaux sur les structures de dialogue [Vilnat, 2005] [van Schooten & op den Akker, 2005], de la nature séquentielle des groupes de questions et du partage des termes des questions déjà résolues du groupe, nous choisissons d'organiser la structure du contexte d'un groupe de questions en un arbre. La figure III.2 illustre l'arbre obtenu pour le groupe de questions du tableau III.3 (page 91).

contexte

À sa racine nous trouvons les termes communs à toutes les questions, les thématiques induites des termes forment le contexte commun induit. Ce sont

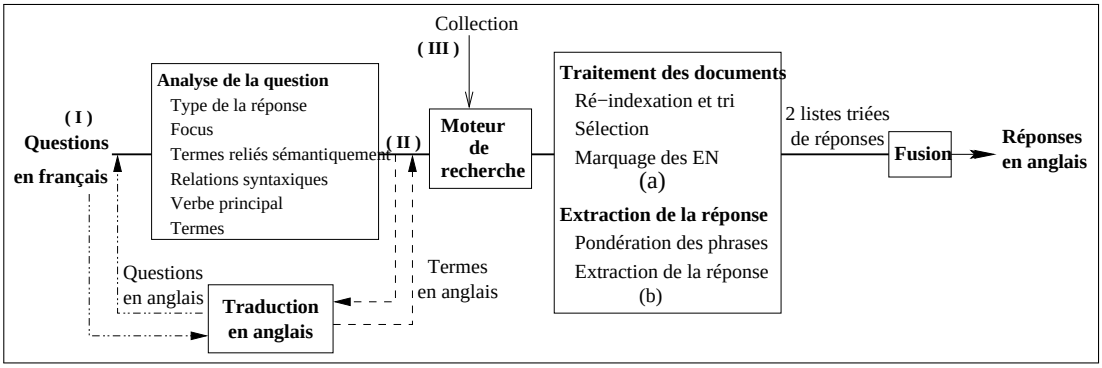

FIG. III.3 – Architecture du système Musclef en mode inter-lingue. Rappel de la figure II.2

les dépendances qui encodent la structure de l'arbre. Chaque branche est constituée d'un arbre de paires question/réponse. À chaque nœud sont indiqués la question et son contexte local induit. Un contexte local est composé d'une liste de termes faisant éventuellement référence à un nœud réponse.

La structure d'arbre nous permet de représenter les groupes où les questions ne reprennent que le contexte issu de la première question. L'ajout des éléments d'informations utiles à la recherche d'information à chaque nœud permet une représentation homogène des groupes où les questions réutilisent des contextes liés les uns aux autres. Les questions ne réutilisant pas le contexte des précédentes sont rattachées au nœud *groupe*. Ce nœud *groupe* peut également recevoir des éléments afin de contraindre l'espace de recherche à la manière des évaluations de Trec 2006 [9] [Hickl *et al.*, 2006].

III.1.2.3 Formalisation en probabilité du calcul des dépendances

Afin de résoudre le problème de la réutilisation de la méthode dans d'autres cadres logiques[10], il est intéressant d'obtenir un ancrage probabiliste du modèle ci-dessus. Soit α et β deux couples de questions et réponses d'un groupe donné, tel que α précède β.

Le calcul des dépendances unitaires peut être vu comme la probabilité d'association d'une dépendance de β vers α.

Nous pouvons formaliser la probabilité d'existence d'une dépendance en un calcul d'argMax sur une collection de traits.

Soit Γ l'ensemble des termes que l'utilisateur doit fournir dans ces 2 questions pour que la réponse à β puisse être trouvée. Γ dépend des stratégies

[9] Une thématique était donnée explicitement pour chaque groupe de questions.

[10] Par exemple dans le cadre d'un système à scénario, ou pour la construction d'un système plus général de résolution de lien...

du SQR utilisé ainsi que des corpus de documents dans lesquels la réponse est cherchée. La probabilité P à calculer est l'existence de l'évènement : β *est une sous-partie de* Γ *strictement plus petite que* Γ. Notons que même si Γ n'est pas optimum (l'utilisateur pourrait fournir plus d'informations), rien n'empêche d'avoir suffisamment d'informations pour que la probabilité d'association d'une dépendance soit maximalement correcte. Soit Ψ une collection de traits[11] munis d'une fonction d'évaluation[12] permettant de décrire l'apport et la capacité d'unification de β dans Γ.

Alors $P_{\beta\alpha}$ est la somme des plus grandes possibilités d'apport et capacité d'unification, soit :

$$\boxed{P_{\beta\alpha} = argMax(\Sigma_{Ti\in\Psi} eval(Ti, \alpha, \beta))} \qquad \text{(F0)}$$

Le calcul n'a de sens que si β est fixe et que nous recherchons le α donnant le meilleur argMax.

Il peut théoriquement exister des groupes de questions où il y a deux maximums équiprobables pour une question β. Comme il existe plusieurs ensembles de termes permettant d'obtenir la réponse, Γ, du point de vue de l'utilisateur, n'est pas un ensemble fermé. Si l'utilisateur pose deux questions alors, il divise aussi les termes entre les deux questions. [13] En conséquence, il est possible qu'il existe deux sous-ensembles de traits différents de Ψ permettant d'obtenir des P égales. C'est d'autant plus probable que Ψ est petit.

Ce modèle nous donne un cadre de calcul pour développer une expérience visant à établir la faisabilité du calcul des dépendances et du contexte.

III.1.3 Sélection des termes

Intéressons-nous maintenant au problème des choix des termes de α qui doivent être intégrés dans Γ. Les méthodes des SQR actuels utilisent des stratégies définies statiquement pour sélectionner les termes dans les questions. Le but de notre travail n'est pas de redéfinir les critères linguistiques qui permettent de juger de l'intérêt des termes. Donc toute stratégie de sélection de termes de α pour constituer l'ensemble Γ de β sera une stratégie dérivée de celles des SQR actuels. Par conséquent les termes de α sélectionnés pour compléter Γ constitueront un sous-ensemble (éventuellement égal) des termes qui auraient été sélectionnés lors de la résolution de α.

[11] Tous les traits sont détaillés à partir de la page 98

[12] type de la question, catégorie, ou des combinaisons plus complexes, traits issus de l'analyse de la question comme illustrée sur la figure III.3

[13] La division en deux questions est obligatoire dès qu'un terme d'une question β ne peut pas figurer dans une question α sans changer la réponse attendue.

III.1.3.1 Choix des termes

La difficulté de cette tâche est de ne pas sélectionner trop de termes. En effet si nous ne choisissons pas cette approche, alors autant prendre tous les mots de toutes les questions liées, mais alors sans pertinence de la requête. Ce problème n'a pas de solution générique, c'est aussi le cœur de la recherche de documents. Si les bons termes ne sont pas choisis, peu importe la complexité des stratégies déployées (synonymes, lemmatisations...) et la complexité des techniques de recherche (statistique, booléenne...), les résultats seront mauvais. Intéressons-nous d'abord à la sélection des termes pour la résolution d'une question simple. Si les documents contenant la réponse ne sont pas sélectionnés lors de la recherche des documents, c'est toute la chaîne de traitements qui sera en échec.

intérêt secondaire

Nous appelons *terme d'intérêt secondaire* un terme qui n'est pas indispensable à la sélection d'un document contenant la réponse. Un terme d'intérêt secondaire peut améliorer le classement de documents contenant la réponse, mais le gain est marginal. Nous appelons *terme non pertinent* un terme d'intérêt secondaire qui n'améliore que le classement des documents ne contenant pas la réponse. Nous appellons *terme indispensable* un terme sans lequel un document pertinent ne remonterait pas dans la *posting-list* dans la limite des n premiers documents imposée par la recherche.

non pertinent

indispensable

Il faut que tous les termes indispensables soient sélectionnés. Le système ne connait ni l'intérêt différentiel des termes, ni la probabilité qu'ils soient les seuls à remonter le score des documents (Cf section IV.1 page 114). Donc même les termes d'intérêt secondaire sont sélectionnés. Nécessairement des termes non pertinents vont être sélectionnés.

III.1.3.2 Importance des termes différents selon le lien

Nous avons vu que les liens entre questions et les partages de termes ne sont pas «*absolus*»[14]. Nous voyons alors que des termes de α peuvent avoir un intérêt différent au niveau de la résolution de β. Malgré un lien, des termes de α qui sont importants peuvent devenir d'un intérêt secondaire dans le cadre de β (le cas de la reprise d'un terme de α dans β est le cas le plus parlant), ils peuvent même devenir contre-productifs.

Par exemple dans le tableau III.4 les dépendances sont [1,3] et [1,4]. Dans la question 1 (α), le terme «domicile» [15] est indispensable pour spécifier le

[14]Il existe de nombreuses manières de liéer des questions, comme dans le tableau III.2 (page 86)

[15]Le terme «personne» peut être retiré grâce à des heuristiques s'appuyant sur le type de la réponse attendue et la position des verbes.

1	A combien de personnes a-t-on demandé de quitter leur domicile durant les inondations aux Pays-Bas en hiver 1995 ?
2	Quelle proportion des Pays-Bas est sous le niveau de la mer ?
3	A quelle hauteur l'eau est-elle montée à Lobith durant les inondations ?
4	Qui était le premier ministre des Pays-Bas à ce moment-là ?

TAB. III.4 – Premier groupe de questions enchaînées du corpus de la campagne ClefQA2007

type de la quantité de personnes. Mais ce même terme «domicile» semble non pertinent dans question 3. Pourtant, la dépendance est claire entre les questions 1 et 3.

III.1.3.3 Intérêt des termes pour la recherche

Il peut même y avoir plusieurs manières de comprendre le lien entre α et β, l'un sera peut-être préféré par la plupart des utilisateurs, mais le *statu quo* est possible. Un terme peut-être préféré à un autre, les deux étant mutuellement exclus, et chacun apportant un sens différent à la question.

Nous pouvons aussi évoquer le problème de la sélection des termes dans les cas d'apposition et de coordination. Nous n'avons pas conçu d'outils spécifiques pour résoudre le problème des coordinations entre les termes d'un couple $\alpha - \beta$. Tous les termes utiles sont sélectionnés, mais parfois nous sélectionnons plus de termes que ceux qui sont utiles. La difficulté d'interprétation du lien $\alpha - \beta$ et le problème de la définition de l'intérêt d'un terme en fonction du corpus de documents, rendent difficile la conception d'un test visant à évaluer la sélection des termes. Ce problème est résolu empiriquement. En effet la sélection de termes a essentiellement un impact sur la recherche des documents, or ces mêmes termes ont un impact qui diffère en fonction de la stratégie de recherche. Il y a une interaction entre la stratégie de recherche des documents et la stratégie de sélection des termes. C'est donc le couple «sélection de termes - stratégie de recherche» qui sera évalué globalement.

III.1.4 Compatibilité avec une structure de dialogue à venir

Cette formalisation en dépendance nous permet une compatibilité avec la classification de van Schooten et Rieks op den Akker [van Schooten & op den Akker, 2005].

Elle est aussi assez compatible avec des extensions de SQR-enchaînées vers des systèmes de dialogue homme-machine.

Au chapitre I.2 (page 26) nous avons vu rapidement les modèles structurels et dynamiques du dialogue. Si nous voyons les dépendances comme des relations d'une grammaire dialogique, alors comme dans un modèle structurel, il est évident que les dépendances peuvent être mises en forme de manière à être rapprochées d'une grammaire dialogique.

La formalisation en dépendances est aussi compatible avec un modèle de dialogue dynamique. Les dépendances sont toujours valides après la modification de l'interprétation d'une question. Si les dépendances avaient été définies par l'étude de l'ensemble des questions dont α dépend par rapport à β, alors la modification de l'interprétation d'une seule question dont α dépend pourrait remettre en cause la dépendance $\alpha - \beta$. C'est la construction des dépendances par des études de paire de couples $\alpha - \beta$ qui permet de garantir qu'une ré-interprétation de α (ou β) par le gestionnaire de dialogue ne romprait pas la dépendance.

En conséquence de ce résultat, nous pouvons dire que les dépendances peuvent être réorganisées dynamiquement comme dans les modèles de dialogues dynamiques (suite notamment a de nouvelles informations, compréhensions, interprétations...). Les dépendances peuvent être décorées avec toutes les informations de typages liés au dynamisme du système de dialogue. Au lieu d'être une charge supplémentaire, cela pourrait même être réutilisé en complément à l'ensemble des traits (Ψ).

III.2 Construction de la structure de dépendances

Nous avons décrit dans le paragraphe précédent la structure que nous voulions obtenir. Nous allons maintenant décrire le processus permettant de la construire. Pour celà, nous allons préciser les dépendances que nous recherchons puis les données permettant de les reconnaître. Enfin nous évaluons les résultats obtenus.

III.2.1 Trouver les dépendances

Nous présentons maintenant une technique pour trouver les dépendances entre les questions d'un même groupe. À notre connaissance il n'y a pas un phénomène linguistique unique qui permette de toutes les trouver et simultanément de n'en trouver aucune mauvaise.

Le calcul des dépendances via la méthode basée sur un ensemble de traits (Ψ , page 92) est axé sur les informations disponibles dans les SQR classiques.Ce calcul réutilise directement la sortie du module d'analyse de questions dy système existant.

Comme nous le verrons à la section III.2.2(page 98) les réponses sont aussi intégrées à la collection de méthodes Ψ des calculs de traits.

III.2.1.1 Vue en terme de probabilité

Comme vu à la section précédente, nous pouvons formaliser la probabilité d'existence d'une dépendance en un calcul d'argMax sur une collection de traits. Il est alors simple de définir une stratégie utilisant un seuil de probabilité en dessous duquel nous décidons que la dépendance n'existe pas. C'est une simplification de la méthode présentée dans [Séjourné, 2008].

L'algorithme de recherche de dépendance unitaire est générique par le type et le nombre de traits/critères linguistiques qu'il utilise. Le type de critère recherché est celui prenant en entrée un couple de questions(ou un résultat de l'analyse de ces questions) et produisant en sortie un score permettant d'estimer la probabilité d'existence d'une dépendance.

III.2.1.2 Définition du seuil de correction

Voyons comment nous pourrions définir le seuil de probabilité de l'existance d'une dépendance et comment nous pourrions le calculer. Ce sont les outils disponibles qui guident notre réflexion. Il existe de nombreux outils probabilistes pour déterminer des classes ou des schémas. Nous construisons

2 classes, l'une pour les probabilités suffisamment élevées, correspondant à des valeurs de confiance élevées pour l'existence des dépendances, et une classe pour les cas à faible probabilité d'existence. Par exemple, l'algorithme de maximisation d'espérance[16] permet d'apprendre cette limite si nous disposons d'un corpus d'apprentissage. Il faut chercher à utiliser l'algorithme pour calculer les moyennes et déviations standards pour chaque classe. Il est alors facile de déduire la limite entière la plus proche. Le seuil de sélection des dépendances est la limite de probabilité de confiance dans l'espérance de l'existence d'une dépendance. Nous revenons à ces calculs à la section suivante. Pour faire cet apprentissage il est indispensable d'avoir préalablement défini et calculé Ψ la collection de traits.

III.2.2 Calcul des dépendances

L'objectif de ces expérimentations n'est pas d'obtenir les meilleurs systèmes possibles pour calculer quelques traits. La nature exploratoire de ce travail dissuade de réaliser des systèmes très aboutis ou spécifiques pour les expérimentations. Par contre, il est opportun de les réaliser dans un même cadre logique et suivant des interfaces communes. Cela a pour premier effet de simplifier la recherche d'optimalité[17]. Cela a pour second effet de permettre une exploration plus rapide de l'espace des paramètres d'optimisation des traits les uns envers les autres.

III.2.2.1 Choix et calcul des traits

Suivent ici cinq traits mis en œuvre et utilisés pour les tests.

III.2.2.1.1 Anaphores co-référentes Un système de résolution d'anaphores inter-questions développé au laboratoire permet d'associer des éléments d'une question à l'autre. Les travaux de [Hernandez, 2004] ont inspiré ce système. Il comporte notamment :

- une résolution via des règles axées sur des éléments de morpho-syntaxe
- un système d'identification des termes pouvant servir de référent ou d'antécédent

[16] EM : Expectation-maximisation, Dempster *et al.* 1977.

[17] En TAL, il est difficile de prouver que les traitements sont optimaux, mais nous pouvons en avoir des indices. Il ne s'agit pas de prouver que notre algorithme est optimal, il s'agit d'avoir des cadres logiques permettant au concepteur des traits de savoir si leurs traits le sont.

– une logique d'association préférentielle de termes à base d'informations de genre, de nombre et de règles à base de redondance de noms propres et entités nommées.

Une anaphore entre 2 questions a une forte probabilité d'indiquer une dépendance entre ces 2 questions.

III.2.2.1.2 Type de question Dans Musclef chaque question a une catégorie [Ligozat, 2006]. Les catégories permettent de connaitre le type de la réponse attendue. Les catégories des questions détectées par Musclef peuvent être utilisées pour détecter les dépendances. Si deux questions ont des catégories identiques il y a beaucoup de chance qu'il n'y ait pas de dépendances de l'une vers l'autre :

Citer le nom d'un aliment contenu dans le régime alimentaire de base d'Asie du sud-est.
Citer le nom d'un aliment contenu dans le régime alimentaire de base d'Europe.

Les catégories du SQR Musclef possèdent deux niveaux. Par exemple, la catégorie «combien» possède deux formes étendues, la forme «combien de»[18] et la forme «combien autre». «combien» est la racine de la forme, «de» et «autre» sont des spécialisations. La probabilité d'existence d'une dépendance est maximale quand les racines sont différentes, cependant les spécialisations permettent de nuancer cet probabilité.

III.2.2.1.3 Entropie des caractères La longueur relative d'une question par rapport à une autre est significative pour distinguer l'existence d'un lien. Une question avec moins de *caractères* réintroduit probablement moins d'éléments et vraisemblablement réutilise plus les éléments déjà introduits. Comme c'est une mesure de l'entropie relative entre 2 questions, seules les différences importantes sont significatives, comme dans l'exemple suivant :

Quel était le nom de la barge qui a coulé à Porto Rico le 7 janvier 1994 ?
Qu'a heurté la barge ?

Nous avons privilégié le nombre de caractères plutôt que le nombre de mots (qui peuvent porter plusieurs sens à la fois) car ils permettent mieux de montrer les différences de longueur des questions. Arbitrairement, la différence choisie dans notre implémentation est le facteur deux. Il faut que α soit deux fois plus longue que β. Cette valeur pourrait être affinée à l'avenir.

III.2.2.1.4 Entités nommées Les entités nommées identiques d'une question sur l'autre sont des bons critères pragmatiques pour le calcul de probabilité d'existence d'un lien. Ce trait a tendance à montrer que les 2 questions ne sont pas liées, au contraire la répétition partielle d'une entité nommée dans

[18]La forme «combien de» est la forme par défaut, elle est notée juste «combien» dans les arcanes du système.

une question indique plutôt une dépendance vers la question dans laquelle elle est complète.

Qu'est-ce que Kia ?
Nommer un modèle de Kia.

Les deux questions peuvent probablement être traité indépendament. Dans les deux cas il s'agit de la même «Kia».

Quels étaient les noms complets de Flanders et Swann ?
Dans quelle ville Swann est-il mort ?

Ici, le second «Swann» est celui du couple «Flanders et Swann». Il faudra s'en souvenir pour éviter des confusions avec un autre «Swann».

III.2.2.1.5 Répétition de texte Les répétitions de segment de texte d'une question sur l'autre révèlent l'instanciation d'un thème global. Un apprentissage nous montre que si des segments communs de plus de 15 caractères sont répétés d'une question sur l'autre et si les segments ne sont en position préfixe ni dans l'une ni dans l'autre, alors il n'y a pas de dépendance unitaire entre les deux questions.

Quand l'homme politique irlandais Willie O'Dea est-il né ?
Où l'homme politique irlandais Willie O'Dea est-il né ?

Le système utilise une recherche du plus long segment commun entre les deux questions, puis il teste sa longueur et celles des préfixes.

III.2.2.2 De la théorie de la méthode à sa mise en œuvre

Les traits de Ψ utilisés dans notre expérimentation sont donc [*Anaphores co-référentes*, *Type de question*, *Entropie des caractères*, *Entités nommées*, *Répétition de texte*] présentés ci-dessus. Le calcul des dépendances se découpe en deux grandes étapes : d'abord, chercher le maximum de probabilité d'existence pour une question β donnée, puis regarder si cette probabilité est suffisamment significative. Nous allons faire la somme des évaluations des traits de Ψ afin de calculer tout l'espace des probabilités de dépendances de type $\beta - \alpha$.

Pour chaque couple de questions, chaque trait est évalué en un score (via leurs fonctions d'évaluation). Dans notre expérimentation les évaluations des traits ne sont pas directement des probabilités, mais des scores. Il faut donc normaliser ces scores pour les rendre comparables, puis les scores peuvent être convertis en probabilités.

III.2.2.2.1 Du trait au score normalisé pondéré À chaque combinaison $\beta - \alpha$ d'un groupe est associé un vecteur de scores. Pour le groupe en exemple du tableau II.3(page 61), les vecteurs sont ceux du tableau III.5. Le vecteur de la case de la première colonne et première ligne (1,2) est [1,1,0,0,0] ;

Num quest	1	2	3
2	[1,1,0,0,0]	X	X
3	[1,1,0,0,0]	[0,1,0,0,0]	X
4	[1,1,0,0,0]	[0,1,0,0,0]	[2,0,0,0,0]

TAB. III.5 – Les vecteurs de scores pour le groupe en exemple.

Num quest	1	2	3
2	[8,7,0,0,0]-> **15**	X	X
3	[8,7,0,0,0]-> **15**	[0,7,0,0,0]-> **7**	X
4	[8,7,0,0,0]-> **15**	[0,7,0,0,0]-> **7**	[16,0,0,0,0]-> **16**

TAB. III.6 – Les vecteurs de scores après harmonisation, pondération et projection.

il indique que seulement une anaphore et une différence de type de question ont été trouvées entre les questions 1 et 2. Le vecteur de la case (3,4) est [0,2,0,0,0] ; il indique qu'une anaphore double a été trouvée.

Il est alors évident qu'il faut que les scores utilisent un intervalle de valeurs commun, c'est la normalisation. Pour chaque trait nous connaissons le maximum, le minimum et le nombre de valeurs distinctes de la fonction d'évaluation(par construction). Ces connaissances servent à normaliser entre 0 et 1 chaque trait. La normalisation utilisée dans notre implémentation est [0.4,0.5,1,1,1].

Par exemple, la valeur maximum que donne le critère des types de questions[19] est 2. La normalisation pour ce critère est donc de «1/2».

Comme nous n'avons pas la même confiance *a priori* en chaque trait, nous ajoutons un facteur de pondération. La pondération utilisée dans notre implémentation est [20,14,5,-5,-10], elle est choisie par rapport à la confiance*a priori* en chaque critère. [20]

La fonction à optimiser est obtenue en réalisant la somme des composantes de chaque vecteur après normalisation et pondération. Nous obtenons alors les données du tableau III.6.

[19]Les types de questions sont comparées sur les deux niveaux de similarités : « définition-personne » «définition-autre»

[20]A la suite de notre expérimentation, nous avons mis en place un système permettant de tester un intervalle de pondération pour chaque trait. Si nous disposons de suffisamment de données d'apprentissage, il est possible de déterminer les meilleures pondérations. Ici c'est le corpus ClefQA08-FR-EN qui a été utilisé et qui est proche du ClefQA07-FR-EN .

Le vecteur de la case (1,2) devient donc [0.4x20x1, 0.5x14x1, 1x5x0, 1x-5x0, 1x-10x0] soit une somme de 15. Les valeurs normalisées-pondérées-extremums sont données par le vecteur [7,24,5,-5,-10]. La valeur maximum pour les calculs de probabilité d'existence est donc de :

$$\boxed{7x1 + 24x1 + 5x1 + -5x0 = 36}$$

III.2.2.2.2 Du score à la probabilité Le système utilise le score maximum pour la totalité des questions pour décider de la dépendance à retenir. Pour la simplicité des calculs, il n'est pas indispensable de ramener la probabilité d'existence d'une dépendance entre 0 et 1. En effet seules les valeurs relatives ont de l'importance ; nous cherchons le maximum et un seuil. Les projections dont la somme est négative sont supprimées, il n'y a donc pas de *probabilité négative*[21]. Ci-dessous, les valeurs de la fonction dont nous cherchons l'argMax :

Num quest	1	2	3
2	15	X	X
3	15	7	X
4	15	7	16

Nous pouvons en déduire la probabilité d'existence d'une dépendance unitaire pour chaque couple $\beta - \alpha$. Pour chaque question en position β nous connaissons le meilleur candidat α pour l'existence d'une dépendance entre les deux. Dans notre exemple les dépendances candidates sont donc [1,2] , [1,3] et [3,4].

III.2.2.2.3 Existence des dépendances Nous utilisons le seuil de probabilité de correction choisi à l'avance à l'aide d'un apprentissage supervisé. Cette valeur permet d'éliminer les combinaisons à probabilité trop faible.

Ce seuil est appris sur un corpus de dépendances annotées sous la forme '*1 : existe* ' et '*0 : existe pas*'[22]. Nous avons réalisé les calculs ci-dessus pour chaque groupe de questions. Nous avons alors créé une table qui indique pour chaque score l'existence ou non d'une dépendance. Le système weka [Frank *et al.*, 2005] nous a fourni un outil de classification statistique nous

[21] Pour peu que cela ait un sens.

[22] Corpus ClefQA07-FR-ES , qui est différent du corpus de test ClefQA07-FR-EN mais proche.

permettant alors d'évaluer ce seuil. Nous en déduisons un seuil[23] de 1. Pour tous les couples de type $\alpha - \beta$ où toutes les sommes pondérées normalisées de traits sont supérieures à 1, alors nous disons qu'il existe une dépendance.

Selon nos calculs les trois dépendances candidates [1,2] , [1,3] et [3,4] sont donc bien des dépendances puisque que chaque score est supérieur à 1.

III.2.2.2.4 Construction de l'arbre des dépendances Maintenant que nous connaissons les dépendances entre questions au sein du groupe de questions, nous pouvons construire la structure des dépendances. L'arbre est construit après un tri topologique des dépendances unitaires. Le tri topologique donne 2 ordres valides (et donc deux arbres, figure III.4).

FIG. III.4 – Exemple de deux constructions différentes, mais valides.

Le premier est toujours choisi, car il respecte l'ordre d'introduction des questions. Pour un groupe de questions donné, la structure arborée n'est pas forcément unique. Si une question '1' introduit les éléments A et B. Si une question '2' réutilise uniquement l'élément A et que la question '3' réutilise les éléments A et B. Alors, il est possible de rattacher la question '3' soit à la question '1' (stratégie de la première introduction d'un élément), soit à la question '2' (stratégie de la définition la plus récente d'un élément). Si nous ne disposons pas de critères sémantiques/pragmatiques, l'élément peut être décrit par les mêmes mots, mais correspondre à des sens différents. Étant donné le type des questions posées il peut être intéressant de privilégier la stratégie de la première introduction, alors que sur un système de dialogue nous préfèrerons peut-être la dernière afin d'avoir une représentation plus concise ne demandant pas de mémorisation des évènements anciens.

Nous pouvons alors décorer l'arbre avec les thèmes des questions, réponses et vecteur de scores. Les thèmes utiles sont repérés via un mécanisme basé

[23]Nous avons utilisé une collection de classifieurs. Tous indiquent une valeur tournant aux alentours de 1. Par exemple le K-moyenne, donne une moyenne de 0.77 et déviation standard de 0.53 pour la classe des «existe pas», une moyenne de 2.62 et déviation standard de 0.98 pour la classe «existe». Le point d'incertitude maximum est donc de : $(0.77 + 0.53 + 2.62 - 0.98)/2 = 1.48$ Comme le système travaille en nombres entiers la limite est arrondie à 1.

sur la morpho-syntaxe. Ils sont constitués des noms, adjectifs, nombres et déterminants les entourant. L'analyse des questions de Musclef permet de s'assurer que les constituants formant le *focus* [Ferret *et al.*, 2001] (cf I.1.3.1 page I.1.3.1) [24] sont présents dans les thèmes.De même, les entités nommées sont arbitrairement ajoutées dans la liste des éléments. Chaque élément produit est un sous ensemble contigu de la question. Tous les sous éléments contigus sont fusionnés pour former les thèmes.

Une fois l'arbre de dépendances construit, chaque thème d'une question est traité exactement de la même manière que dans la suite d'un SQR. Nous utilisons juste une numérotation des termes afin de les suivre dans les étapes de sélection des mots, traductions... Les termes ne sont instanciés que juste avant leur utilisation dans le moteur de recherche. Le vecteur de scores est ajouté afin de caractériser le type de dépendance entre les questions. Les réponses aux questions sont ajoutées ainsi que leur type (type de la réponse attendue de la question précédente). La figure III.5(page 105) montre cette structure pour le groupe de 4 questions de l'exemple en tableau II.3(page 61).

Les dépendances forment la structure de l'arbre. Les nœuds sont indifférenciés des feuilles, une question peut se raccrocher à n'importe quelle autre question. Les réponses aux questions déjà évaluées sont ajoutées dans l'arbre alors que celles qui ne sont pas encore évaluées (uniquement des feuilles) peuvent être absentes. La décoration de l'arbre n'est donc complétée qu'après la fin de l'exécution du SQR. Comme nous pouvons le voir dans la figure II.3(page 61,résultat de notre exemple), les thèmes sont associés au nœud qui porte la question dans laquelle la détection a été réalisée. D'autres informations endémiques de la relation de dépendance comme le vecteur de score peuvent être utiles dans la suite des calculs. Les autres informations concernant l'analyse de la question peuvent être obtenues par une correspondance avec le numéro de la question.

À l'aide de cette organisation, nous pouvons concevoir une nouvelle interrogation des documents tenant compte de l'origine des éléments. C'est l'objet du chapitre suivant. Mais nous allons d'abord évaluer l'apport des dépendances de l'arbre de dépendance.

III.2.3 Évaluation

Remarquons que les modifications que nous pouvons apporter à un SQR interactif portant sur des questions enchaînées ne sont évaluables normalement que face à une étude de satisfaction de l'utilisateur ou une amélioration

[24] Le focus d'une question : une phrase nominale qui a des chances d'être présente dans la réponse : Qui est le ministre ? « Le ministre » est...

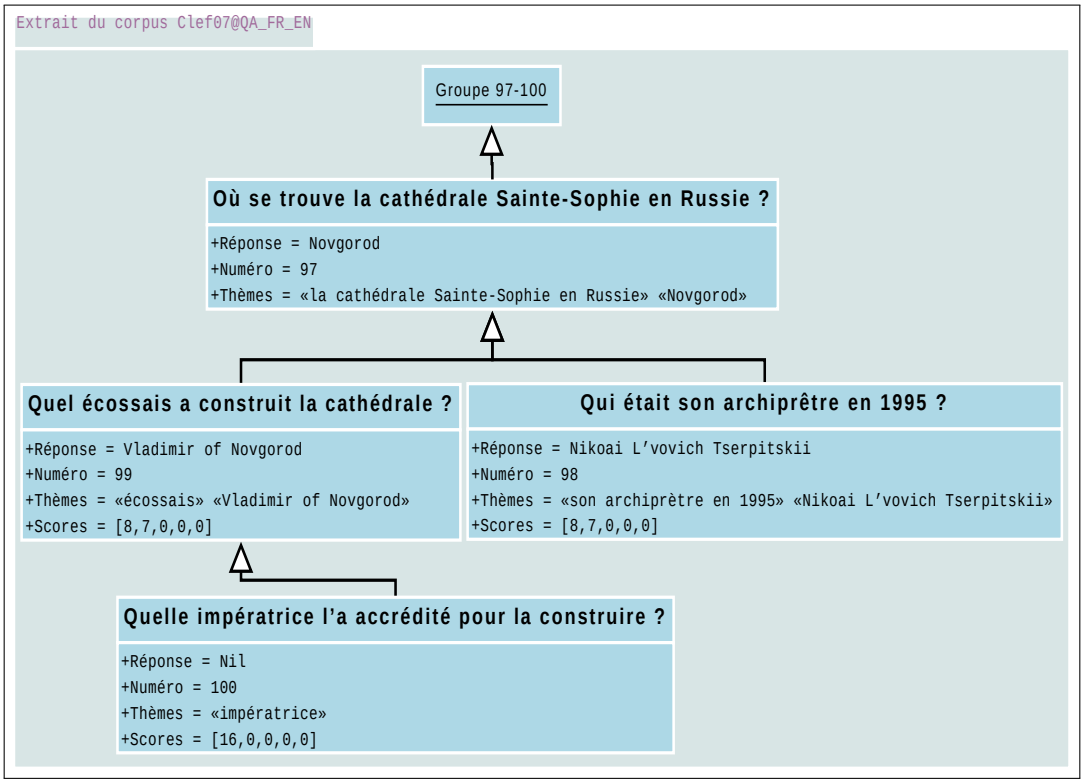

FIG. III.5 – La structure d'arbre pour un groupe de questions enchaînées.

des résultats sur la tâche SQR. Comme notre but n'est pas de gagner quelques % en F-mesure ou en MRR, mais d'explorer des possibilités d'interactions nouvelles et les contraintes liées, nous nous arrêtons un moment ici afin de tenter une évaluation des modifications que nous avons apportées à l'analyse des questions. Le but est de déterminer si avec des stratégies de *boosting* nous avons réussi à obtenir des résultats suffisamment corrects pour être déjà exploitables. Si c'est le cas alors nous aurons prouvé la faisabilité de notre approche. La formalisation de notre structure de contexte nous permet de mettre au point un prototocole pour une évaluation intermédiaire du système. Les dépendances entre questions sont sujettes à une évaluation. Le choix des termes qui composent les thèmes du contexte est lui lié au système et n'a pas lieu d'être évalué à cette étape.

III.2.3.1 Métrique, mesure et évaluation

Avec notre implémentation, les 122 questions du corpus ClefQA07-FR-EN situées en rang 2+ d'un groupe ont reçu un score qui a été comparé aux

Traits/Résultat	Rappel	Précision	F-mesure
type de question	0.695	0.633	0.663
type de question+entropie	0.717	0.634	0.673
anaphores co-référentes	0.717	0.857	0.781
anaphores+entropie	0.750	0.821	0.784
anaphores+type de question	0.880	0.704	0.782
entropie des caractères	0.152	0.736	0.252
Les 5 ensembles	0.739	0.883	**0.804**

TAB. III.7 – Les performances de découverte automatique de dépendances unitaires.

92 dépendances annotées manuellement. Nous avons annoté un total de 92 dépendances unitaires (III.1.1.1) [25]. Nous avons évalué notre approche sur ce corpus en comparant les dépendances unitaires annotées à celles trouvées automatiquement.

III.2.3.1.1 Les résultats chiffrés Soit «Commun» l'ensemble des dépendances communes entre l'ensemble des dépendances annotées «à la main» et l'ensemble de dépendances trouvées par le système. Le rappel est alors calculé en prenant le rapport de «Commun» sur le nombre total de dépendances annotées. La précision est calculée en prenant le rapport de «Commun» sur le nombre total de dépendances calculées. Nous avons utilisé la F-mesure calculée par la formule :

$$\boxed{(2 * \mathit{rappel} * \mathit{précision})/(\mathit{rappel} + \mathit{précision})}$$

Tous les calculs sont réalisés avec des nombres en précision double, mais les résultats sont arrondis avant le report dans le tableau III.7.

Tous les critères utilisés simultanément permettent une détection des dépendances unitaires dont les performances sont récapitulées dans le tableau III.7(page 106), pour la détection des 92 dépendances du corpus ClefQA07-FR-EN . Les traits «entités nommées» et «segments partagés» sont des traits construits pour restreindre la probabilité d'existence d'une association de dépendance. Il est donc normal que leur usage isolé soit sans intérêt.

[25] Nous pouvons remarquer ici une différence avec l'étude réalisée par Dominique Laurent *et al.* [Laurent *et al.*, 2007] qui compte 76 questions disposant d'un lien anaphorique car nous ne nous limitons pas aux anaphores coréférentes.

III.2.3.1.2 Échange des rôles des corpus Afin de rechercher de la significativité supplémentaire, nous avons calculé la valeur limite comme à la section précédente, mais sur le corpus d'évaluation. Les résultats suggèrent la même valeur limite. Nous avons utilisé le corpus d'évaluation pour rechercher la pondération qui maximise la F-mesure sur ce corpus. Cette pondération permet d'obtenir une F-mesure de 0.853 (meilleur sur-apprentissage), et aucun trait n'obtient de pondération nulle. Nous en déduisons qu'ils améliorent tous les résultats. Les deux plus efficaces sont la détection des anaphores et les différences de type de questions. Cependant, les autres traits peuvent être affinés et d'autres pourraient être envisagés, incluant des éléments syntaxiques et/ou sémantiques.

III.2.3.1.3 Dans la suite de Musclef En première approche pour l'utilisation des dépendances calculées précédemment, nous avons ajouté les termes du contexte à la liste des termes des questions ayant des dépendances. Nous avons pu observer que dans 48 cas nous obtenons des phrases candidates là où sans la méthode de sélection nous n'en obtenons aucune. Nous pouvons aussi observer que le déploiement de la méthode proposée est robuste à la traduction des termes dans les systèmes inter-lingues. Ceci nous incite à penser que nos résultats de sélection des dépendances sont corrects.

III.2.3.2 Augmentation de la robustesse des calculs

L'observation de ces résultats nous permet de faire quelques constatations qui permettent d'augmenter la robustesse des calculs.

III.2.3.2.1 Préférer l'ordre naturel

Remarquons d'abord que par construction une question en rang 3 d'un groupe de questions ne peut jamais avoir plus de dépendances(moins une) que son numéro de rang dans le groupe. Toutes les dépendances sont chronologiquement ordonnées comme dans une véritable interaction avec une machine. Il en résulte que toutes les questions de rang n de tous les groupes peuvent être traitées simultanément si toutes les questions de rang $n-1$ ont déjà été traitées. Il n'est donc pas nécessaire de connaitre les dépendances pour traiter les questions dans le « bon » ordre, mais si la dépendance n'a pas été trouvée préalablement, le moteur de recherche manquera quand même d'information.

III.2.3.2.2 La réponse inutile

Ceci nous amène à nous interroger sur les réponses aux questions. Notons que

si la réponse à la question précédente est théoriquement indispensable pour faire un bon calcul de dépendance, nous pouvons modérer cet «indispensable» de plusieurs manières.

Premièrement, la réutilisation de la réponse est un phénomène rare, 3 occurrences dans ClefQA07-FR-EN , aucune dans ClefQA08-FR-EN et aussi rare dans les autres couples de langues (moins de 3 ou absent).

Deuxièmement, la qualité globale d'extraction de la réponse (surtout en multilingue) n'est pas optimale, il vaut mieux ne pas accorder une trop grande confiance à la réponse. Cela est dû aux longues chaînes d'erreurs qui peuvent se produire dans un SQR. Même en mono-lingue les meilleurs systèmes ne permettent pas une confiance de plus de 80%.

Troisièmement, ce qui est vraiment important n'est pas le texte exact de la réponse, mais le type de la réponse[26]. Le type de la réponse est calculé à l'analyse de la question et n'est plus retravaillé par la suite du SQR. Nous en déduisons qu'utiliser les informations du type de la réponse à la place de la réponse elle-même est au moins aussi judicieux. En conséquence de ces trois observations, il est possible de réaliser un assez bon calcul de dépendance sans avoir réalisé la phase de recherche des documents/phrases/réponses.

III.2.3.2.3 Conséquence sur l'organisation du calcul

Nous avons exploité les remarques ci-dessus. Cela nous a permis de supprimer de l'interface d'entrée du module de calcul de dépendances, les liens logiques vers le calcul du texte de la réponse.

Le type de la réponse est connu dès l'analyse de la question, il est généralement bon.Nos traits n'utilisent pas les réponses et documents des questions précédentes. L'analyse de la question préalable à notre travail ne dépend pas des questions précédentes. Quelle que soit la question, il est alors possible de réaliser le calcul des dépendances sans recherche dans les documents.

Cette méthode n'est pas anodine sur le calcul : erreurs/succès. Si dans un groupe de questions, les questions disposent systématiquement d'une dépendance vers la réponse de la question précédente, alors une erreur portant sur la réponse à la première question devrait avoir un impact sur la réponse à la dernière question. Notre méthode de calcul qui utilise le type attendu de la réponse et non pas la réponse, donne des résultats indépendants de la correction de la réponse à la première question. Nous avons utilisé cette méthode sans l'évaluer, car comme expliqué ci-dessus, les cas de réutilisation de la réponse dans nos corpus sont rares, et la qualité globale de sélection

[26]L'important est d'établir l'existence de la dépendance. Sauf dans le cas particulier d'utilisation de traits nécessitant vraiment la réponse ; le type de la réponse est suffisant pour établir l'existence ou non de la dépendance

de la bonne réponse en rang 1 en multi-lingue ne nous incite pas à en tenir compte. [27]

III.2.3.3 Analyses des conséquences des erreurs

Nous pouvons nous demander, combien faudrait-il trouver de dépendances correctes, quelle F-mesure, pour que le système soit réutilisable ? Quelle F-mesure atteindre pour que la fiabilité soit suffisante pour construire d'autres travaux réutilisant ce calcul de dépendance, car certaines erreurs n'affectent pas les performances ?

La fiabilité dépend directement des conséquences des erreurs d'annotation en dépendance. Les conséquences dépendent de l'usage que nous voulons en faire. Nous voulons nous servir des dépendances notamment pour améliorer les performances de la recherche des documents. Il sera donc tout aussi intéressant de se demander (Chapitre IV page 113) quelles pourront être les modifications à apporter à la recherche des documents et les stratégies utilisées autour pour s'adapter au mieux aux dépendances disponibles.

Examinons les conséquences des erreurs possibles, à savoir l'ajout ou l'oubli d'une dépendance.

III.2.3.3.1 Ajouter une dépendance là où il n'y a pas de lien

C'est presque sans conséquence si les termes des deux questions sont les mêmes, mais que les types de questions sont différents. La sélection de la réponse n'utilise les informations de type de la question α que si des informations de type sont absentes de β. L'ajout de la dépendance peut-être assez grave si α contient une entité nommée rare. Cette erreur risque de trop forcer une sélection de documents non pertinents via l'ajout d'une entité nommée qui n'a que peu de chance d'être pertinente. Pour contrebalancer l'ajout erroné d'une dépendance, nous préférons des stratégies qui favorisent l'importance des données de β par rapport à celles de α en pondérant les analyses respectives.

III.2.3.3.2 Oublier une dépendance là où il y a un lien

Les erreurs d'annotations sont assez graves lorsque le système en oublie qui aurait permis de résoudre une question elliptique.[28] Il manque alors des infor-

[27] Rappelons que la réponse exacte n'est qu'une partie de la réponse d'un SQR. Il peut aussi fournir des documents candidats, des passages candidats et des phrases candidates.

[28] Le cas de la dépendance mal détectée sur une question elliptique est rare grâce aux traits concernant les types de question. Si nous disposons d'une définition précise de ce qu'est une question elliptique, il serait aussi possible de construire un trait favorisant

mations importantes pour l'extraction de la réponse si l'analyse de la question n'arrive pas à déterminer de type de réponse.

Un autre cas est l'oubli de la reprise de α entrainant l'absence d'une entité nommée capitale pour répondre à β. L'utilisation du système de résolutions d'anaphores contribue pour une bonne part à la solution de ce problème, mais il est complètement inefficace sur les dépendances purement implicites. Plus l'entité nommée est saillante dans le corpus plus l'erreur est grave. Il arrive aussi que la question β puisse être interprétée de deux manières différentes, dont l'une qui n'implique pas la présence de dépendance.

Une autre variation par rapport à la réponse optimale qui aurait été obtenue avec la détection de la dépendance $\alpha - \beta$ est liée à une perte d'entité(s) nommée(s) de α. C'est le cas où les entités nommées de α sont peu saillantes par rapport à celles de β. La perte des entités nommées de α va probablement changer l'ordre de sélection des documents. Si, comme c'est souvent le cas, le système utilise une stratégie avec limite du nombre des documents alors il est possible de perdre quelques documents contenant la réponse.

Nos erreurs sont réparties dans les diverses catégories ci-dessus. Ce qui est souvent en cause, c'est un problème sur un phénomène linguistique particulier, soit une erreur dans l'analyse des questions, soit d'un trait ad hoc manquant.

III.2.4 Améliorations possibles

Il existe de nombreuses perspectives pour poursuivre ce travail. Les améliorations peuvent suivre deux directions : ou bien une direction qualitative comme traiter des cas supplémentaires, par exemple la double référence vers deux couples distincts, ou une direction quantitative pour améliorer notamment la F-mesure, ou évaluer la sélection des thèmes.

III.2.4.1 Le cas de la double référence vers 2 couples distincts

Dans le cas des dépendances unitaires vers deux questions distinctes comme dans la figure III.1 (page 87), il existe un problème d'adaptation au modèle dialogue et au système de calcul. D'un point de vue purement probabiliste il est possible d'étudier l'ensemble des dépendances ayant une probabilité d'existence supérieure au seuil d'association. Cela permet non

la probabilité d'existence d'une dépendance dans le cas où une question elliptique est reconnue.

seulement de déterminer l'association d'une dépendance vers β, mais également de toutes les dépendances $\alpha(1)...\alpha(n)$ vers les n questions précédentes β.

Cela signifie que le système réalise un calcul d'argMax pour les n meilleures questions α.

Pour intégrer à un modèle de dialogue le type de schéma de dépendance de la forme $\alpha(1)...\alpha(n) - \beta$; nous définissons la *super-dépendance*[29] comme étant une dépendance vers un n-uplet de question-réponse. Chaque couple question-réponse appartient alors à un n-uplet et il n'existe plus que des super-dépendances entre n-uplets. Nous pourrions intégrer dans le modèle de dialogue les n-uplets de la même manière que nous l'avons réalisé pour des couples questions-réponses.

Il faudrait aussi une évaluer cette méthode et adapter les notions de précision/rappel. Nous pouvons nous demander si un ensemble de liens $[\beta - \alpha_1$, $\beta - \alpha_2]$ est équivalent à $[\beta - \alpha_2$, $\beta - \alpha_1]$. Nous n'avons pas procédé à des tests, car ceux-ci demandent probablement de nouveaux corpus et des modifications lourdes dans l'architecture d'évaluation[30].

III.2.4.2 Améliorations quantitatives

Tout d'abord, les outils réalisés pour le calcul des traits peuvent être améliorés. Par exemple, l'ajout d'une analyse grammaticale permettrait de mieux repérer les relations pour la recherche des anaphores. Il est aussi possible de gagner en précision sur la hiérarchie des catégories utilisée par Musclef pour gagner en précision sur la détection des dépendances. De même, le trait de partage de segment de texte ou d'entité nommée est grossier dans la mesure où il n'étudie que des égalités exactes. Des égalités basées sur des distances minimales de longueur d'édition pourraient être déployées facilement.

Deuxièment les coefficients de confiance interne des traits sont choisis par rapport à leur performance brute individuelle dans la tâche. Les interactions entre les traits ne sont pas prises en compte. Cela pourrait être optimisé sur un corpus d'apprentissage.

Troisièment, le système utilise 5 traits, mais la méthode déployée en tolère bien d'autres et des phénomènes linguistiques particuliers pourraient être pris en compte, car la méthode est générique.

Finalement, un système interactif pourrait prendre en compte les résultats d'une recherche de documents infructueuse, pour remettre en cause les dépendances calculées.

[29] super-dépendance : une dépendance utilisant un héritage de dépendance

[30] N'autoriser qu'une unique dépendance permet de nombreuses simplifications dans le code des programmes par rapport au cas à dépendances multiples.

III.3 Conclusion

Nous avons détaillé comment construire des relations représentant les liens entre les questions. Nous allons maintenant voir comment nous pourrions utiliser ces liens. Afin de mieux réussir la construction des dépendances, l'étude du moteur de recherche de documents est également une nécessité, en effet une méthode decorrélée de son usage est toujours plus difficilement généralisable à d'autres systèmes similaires. De plus comme nous l'avons montrer, connaître l'usage des requêtes par le moteur de recherche est indispensable pour comprendre les résultats de l'expérimentation des calculs des dépendances. Voyons donc maintenant comment l'analyse des questions, des dépendances et la recherche des documents interagissent et comment la recherche des documents peut être améliorée.

Chapitre IV

Recherche des documents avec un contexte

Pour traiter correctement les questions enchaînées des modifications doivent être introduites sur l'analyse des questions et le moteur de recherche.

Dans un premier temps, nous examinons différentes méthodes pour utiliser les dépendances et les intégrer au moteur de recherche des documents. Puis nous nous intéressons à une fonction de similarité des documents adaptée aux dépendances. Nous montrons ensuite comment organiser les documents et comment réaliser la sélection des termes en vue de la construction des requêtes. Nous abordons enfin la réorganisation des calculs pour la gestion des dépendances entre questions et le déploiement pratique de la stratégie de recherche de documents.

1	Où se trouve le musée de l'Ermitage ?
2	Qui était le directeur du musée en 1994 ?
3	Dans quel palais le musée est-il logé ?
4	Combien de chambres y a-t-il dans ce palais ?

TAB. IV.1 – Exemple d'un groupe de questions enchaînées tirées du corpus utilisé pour la campagne d'évaluation ClefQA2007 . Les dépendances qui correspondent aux liens entre questions de ce groupe sont visibles dans la figure IV.1 ci-dessous.

IV.1 Utilisations possibles du contexte

Il existe plusieurs méthodes pour prendre en compte les termes des questions liées. Prenons par exemple le groupe de questions du tableau IV.1 (page 114) tiré du corpus d'évaluation de SQR de la campagne ClefQA2007 .

Imaginons une stratégie simple à réaliser pour guider nos raisonnements. Elle sélectionne tous les termes des questions liées et les utilise comme un «sac de mots». Après les extensions classiques[1] d'un SQR, ce sac de mots forme la requête. Appelons cette stratégie «sac de mots».

Nous ne nous intéresserons qu'aux modifications dans le cadre des VSM avec une mesure de similarité basée sur le $tf.idf$ comme présenté à la section II.2 (page 69). La pondération d'un terme est une fonction qui permet de modifier l'importance d'un terme par rapport aux autres lors du calcul du score d'un document. La fonction de pondération peut être arbitrairement complexe, dépendre du corpus, des autres termes et évidemment de notre structure de dépendances. Nous prendrons le $tf.idf$ comme pondération de référence pour créer une sous-stratégie de «sac de mots». Nous évaluons nos résultats par rapport à cette référence. Il se trouve que le moteur de recherche Lucene implémente le $tf.idf$.

Notre structure de dépendances suggère partiellement son propre usage ; il est possible d'en tirer partie pour tenter d'améliorer la sélection des documents-réponses[2] par rapport à cette mesure.

documents-réponses

IV.1.1 Pondération en fonction du rang

rang

Pour chaque niveau de dépendance, nous appellerons son «rang» dans la structure des dépendances, la profondeur à laquelle se situe une question

[1] Synonymes, traductions ou autres extensions sémantiques.

[2] Les documents dont la réponse peut être extraite par des modules ultérieurs (pour une question donnée).

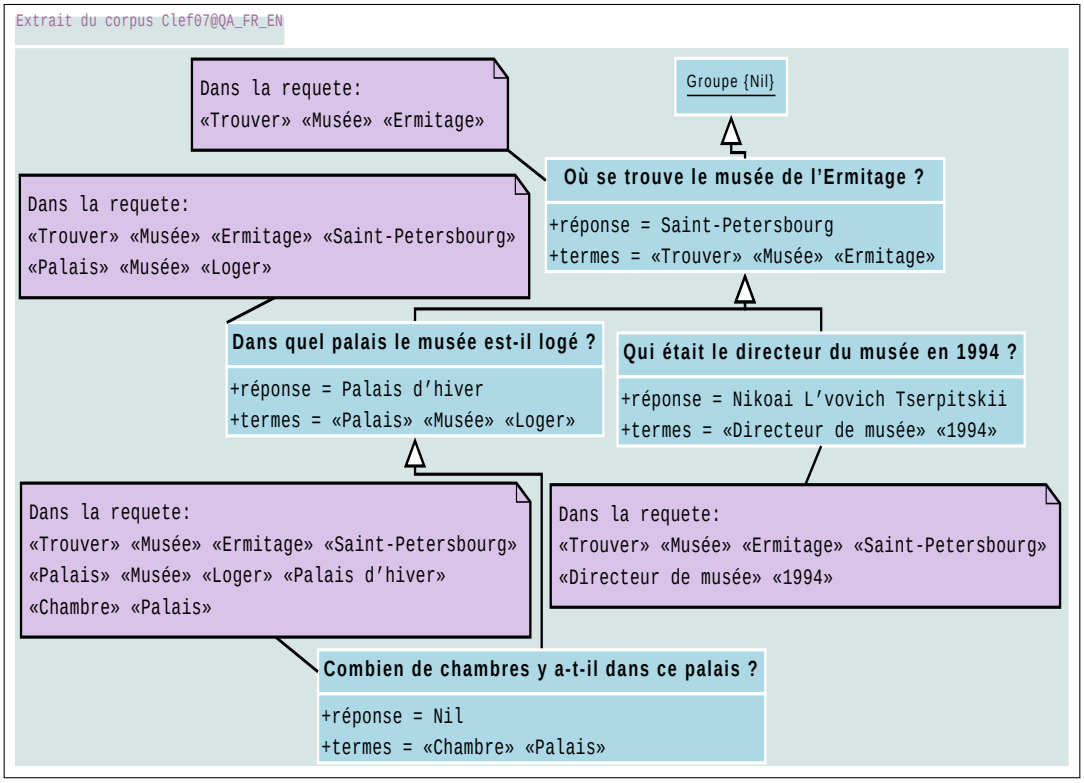

FIG. IV.1 – L'arbre correspondant au groupe du tableau IV.1

en partant de la dernière question posée. Nous pouvons alors formuler notre hypothèse :

> Plus le rang d'une question est élevé moins les mots de la question ont de chance d'être utiles pour sélectionner les documents-réponses. (Hyp A)

Nous nous intéressons alors plus particulièrement à la partie de la pondération qui permet de faire varier l'importance d'un terme par rapport à son $tf.idf$, Pour cela nous introduisons, une fonction multiplicatrice (fp) qui est ajoutée à la formule de pondération et qui ne dépend que du numéro du rang. Soit $fp(n)$ où n est le numéro du rang cette fonction. Diverses fonctions $fp(n)$ peuvent être appliquées (inverse, linéaire, binaire...). Soit $f(terme)$ la fonction qui associe un score à un terme pour chaque document.

Dans le groupe de questions du tableau IV.1 les termes avant traduction et expansion sont :

- 1) Trouver ; Musée ; Ermitage
- 2) Directeur de musée ; 1994
- 3) Palais ; Musée ; Loger

– 4) Chambre ; Palais

La sélection des mots (section I.1.1.3 page 19) est axée sur Musclef. Il faut aussi ajouter les réponses aux questions dans le cas de questions liées. Dans le cadre de la question 4 de l'exemple, le score d'un document via la pondération sera égal à :

$$\begin{aligned} score(document) = \; & fp(1) * (f(Chambre) + f(Palais)) \\ & + fp(2) * (f(Palais) + f(Musée) + f(Loger) + \\ & f(Palais\ d'hiver)) \\ & + fp(3) * (f(Trouver) + f(Musée) + f(Ermitage) + \\ & f(SaintPetersbourg)) \end{aligned}$$

Lors du calcul du score pour un document et un ensemble de termes, les termes peuvent être absents ou non pertinents. Notamment, l'ensemble des termes est perturbé par les erreurs de détection des dépendances. Ces aspects amplifient les problèmes de recherche des documents. Si le système attribue une fonction fp de la forme $fp(n) = 1/(n+1)$ alors il faut prendre en compte plusieurs phénomènes :

- Les niveaux des dépendances peuvent correspondre à des ensembles de termes de *tailles différentes*. Ces différences sont suffisantes pour qu'un rang ne comportant qu'un unique terme ait un impact nul face à un rang plus élevé, mais comportant plus de termes.
- Si le nombre de niveaux de dépendances devient important alors les termes de la question peuvent avoir un *poids total très faible*. Il peut devenir suffisamment faible pour être négligeable devant le poids des termes provenant des questions de la hiérarchie de dépendance.

Une bonne pondération doit gérer ces deux phénomènes, regardons les.

IV.1.1.1 Variabilité de taille des ensembles de termes

Une question comprenant beaucoup de termes peut être soit détaillée sur un thème particulier soit large et sur plusieurs thèmes. Si par exemple la question courante a une dépendance vers une des deux questions :

- *A* Qui est le président ?
- *B* Qui est le président de la SNCF en 2008 ?

Le terme *président* issu des termes de *A* doit-il avoir le même poids que celui issu de *B* dans la modification de la pondération ? Les mécanismes d'analyse vont collecter *président*, *SNCF* et *2008*, soit un terme dans un cas et 3 dans l'autre. L'impact des termes de la question courante est alors changé. Si la dernière question de l'utilisateur possède p termes alors l'impact de chacun des termes de cette question est $1/(p+q)$ ou q vaut soit 1 (dans A) soit 3

(dans B). L'impact des termes de la question de l'utilisateur varie en fonction du nombre de termes dans les questions, mais pas des dépendances.

Dans l'exemple le terme *Directeur de musée* est décomposé en de nombreux termes après la traduction et extension par synonyme. Voici une liste non exhaustive des ensembles des mots que le système Musclef peut en tirer : *administrateur, principal, proviseur, responsable, galerie, collection, conservatoire, cabinet, muséum, salon.*

Sans un contrôle arbitraire des termes en entrée du moteur de recherche, il est évident que même un coefficient 1/2 sur la contribution au score de chacun de ces termes va *écraser* la saillance des autres mots dans le résultat. Le problème ne provient pas tant du nombre exact de termes repérés dans une question en français que de la forte polysémie d'un mot, que ce soit au moment de sélectionner ses synonymes où au moment de le traduire.

Afin d'éviter la sur-représentation des rangs disposant de nombreux termes, il est possible d'intervenir à la sélection des termes pour limiter le nombre de mots par rang dans les dépendances. La limite doit être fixée de manière à être atteignable pour la majorité des questions afin d'obtenir une représentativité homogène.

IV.1.1.2 Poids total très faible

Étudions une hypothèse sur la classe des fonctions fp. Nous voulons que le poids de la dernière question posée ne soit jamais négligeable devant le poids total des termes présents dans toutes les questions qui y sont liées par des dépendances. Sachant que la série des $1/n$ diverge à l'infini, si une pondération des termes basés sur la profondeur de la dépendance est choisie alors il faut provoquer une décroissance de gain de probabilité liée au rang des dépendances, supérieur à $1/(n+1)$.

Si, dans l'arbre ci-dessus IV.1 (page 115), chaque terme donne le même $tf.idf$ alors le score via la fonction fp pour la question 4 peut s'écrire sous la forme :

$score(document)$	$= (1 * (1+1)) + (\frac{1}{2} * (1+1+1+1)) + (\frac{1}{3} * (1+1+1+1))$
	$= 2 + 2 + 1,3..$
	$= 5,3$
Ou avec éliminations des doublons :	
$score(document)$	$= (1 * (1+1)) + (\frac{1}{2} * (1+1+1)) + (\frac{1}{3} * (1+1+1))$
	$= 2 + 1.5 + 1$
	$= 4,5$

Au moment où la recherche des documents pour la questions 4 a lieu, sa réponse n'est pas encore connue, contrairement aux autres questions du contexte. Nous voyons que les rangs supérieurs à zéro ont un poids total supérieur au poids des termes de la question actuelle. Pourtant il y a des termes qui pour la question 4 sont secondaires (Trouver,Saint-Petersbourg)

Poids par terme dans le score total.	
Trouver ; Musée ; Ermitage ; Saint-Petersbourg	$\approx \frac{1}{12}$
Loger ; Musée ; Palais d'hiver ;	$\approx \frac{1}{8}$
Chambre ; Palais	$\approx \frac{1}{4}$

L'un des termes les plus importants «Ermitage» pour la sélection des documents contenant la réponse, n'a qu'un poids 1/12 dans un total normalisé entre zéro et un. Par contre, le terme «Chambre», important dans le cadre du «Palais d'hiver» mais secondaire sinon, a un poids fixe très important 1/4. Cet exemple nous montre qu'une fonction fp aussi «fortement» décroissante que $1/n$ est donc problématique par rapport à notre hypothèse qui veut que les termes ne soient jamais négligeables.

Si une fonction encore plus décroissante comme $1/(n+1)$ est utilisée, alors le poids relatif de la première question sera supérieur, mais les poids des questions des autres rangs encore plus faibles. Aucune de ces deux fonctions ne permet d'obtenir une pondération forte uniquement sur les termes indispensables.

Des fonctions linéaires décroissantes pourraient être envisagées, mais elles présentent les mêmes inconvénients que ceux que nous venons de constater. En outre avec une pente faible les fonctions fp linéaires donneront des résultats proches d'une fp constante et au-delà d'un certain rang le poids devient nul.

IV.1.2 Synthèse de la pondération par rang

L'objectif serait de trouver une pondération donnant de meilleurs résultats finaux que la pondération de référence.

Le premier obstacle que nous avons vu est le contrôle du nombre de termes dans les rangs. Ce contrôle peut-être réalisé en homogénéisant le nombre de termes par rangs lors de l'expansion par les synonymes.

Le second obstacle vient de la fonction de pondération. La fonction fp en $1/n$ n'est pas satisfaisante, car ses propriétés, notamment de divergences, posent des problèmes. Il n'est pas aisé de choisir une fonction fp qui *a priori* aurait des propriétés correctes pour chaque type de requêtes, documents, traductions, synonymes. Ceci nous incite déjà à chercher une autre solution.

De plus, nous savons que ces méthodes modifiant la pondération sont le contrepied des stratégies déployées par les systèmes utilisés à Trec (cf section II.1.5) où l'ensemble des documents était restreint mais où la fonction de score/probabilité n'était pas modifiée. Leur méthode revient à forcer la présence des termes contenus dans les premières questions dans les documents réponses. Ces termes ont donc un poids plus fort que ceux contenus dans la dernière question posée. C'est le contraire du comportement que nous cherchons où les rangs les plus anciens doivent avoir moins de poids. Pourtant, les systèmes utilisés pendant la campagne Trec obtiennent des résultats corrects. Nous en déduisons qu'il faut chercher une autre solution qu'une fonction fp.

Alors nous remettons en cause notre hypothèse initiale. Bien que simple à mettre en œuvre, nous renonçons à séparer la pondération en deux parties fp et $tf.idf$. Nous cherchons une méthode de pondération (toujours à base de $tf.idf$) plus dynamique qui tient mieux compte des liens entre questions via les termes qui les composent.

IV.2 Choix de la corrélation des termes

L'idée générale de l'approche est de renforcer la pondération des termes «liés» avec ceux des rangs précédents. Nous choisissons alors une forme de calcul de la pondération basée sur les co-occurrences de termes dans les documents.

Plutôt que de pondérer les scores des termes par une fonction dépendant uniquement du rang dans la structure des dépendances, nous généralisons son usage pour prendre en compte un autre aspect. Nous allons l'exploiter différement : nous allons tester la corrélation des termes d'un rang à l'autre, comme s'il s'agissait de paire de mots dont la distance dans un document est sans importance. Il s'agit de prendre deux termes de rangs consécutifs de la structure des dépendances et de regarder s'ils sont présents simultanément dans un document. Nous pouvons alors généraliser pour un nombre quelconque de rangs. Pour chaque terme d'un rang, il faut regarder s'il existe au moins un terme de chaque rang inférieur avec lequel il est présent dans le document dont il faut calculer le score. Des tests incrémentaux du rang, de la présence simultanée des termes servent alors de pondération dynamique. La pondération est dynamique, car elle dépend des liens entre les termes cherchés par l'utilisateur et trouvés dans le corpus.

Soit m le nombre maximum de termes qui peuvent être sélectionnés pour chaque rang. m est dans la mesure du possible une valeur à atteindre (mais à ne pas dépasser) pour le nombre de termes à sélectionner dans la stratégie de sélection de termes. Les variations sur le nombre de termes sélectionnés peuvent être réalisées sur le nombre de synonymes et de traductions retenus comme présentées à la section III.1.3 (page 93). Ceci fait écho aux méthodes de sélection de termes traditionnelles des SQR.

Nous avons étudié différentes correlations de termes, parmi lesquelles nous avons retenu celles qui permettent le mieux de tenir compte des rangs des termes. La généralisation dans une forme utilisable de cette pondération peut se faire de différentes façons. La corrélation qui permet que seule contribue au score du document ou bien la plus grande corrélation de termes ou bien chaque sous-corrélation de termes, n'est pas intéressante car elle revient simplement à chercher des n-uplets de termes indépendamment de l'existence d'une structure en dépendance. Celles qui posent le plus l'accent sur la corrélation d'un terme d'un rang avec un terme d'un autre rang sont les suivantes :

A) Lien unique pour tous les rangs
Un document est ajouté à la *posting-list* si et seulement il est composé d'au moins un terme de chaque rang de la structure. Ceci garantit qu'un terme de rang n ne peut avoir un poids relatif plus grand que la totalité des poids

des termes de rang $n-1$. Le nombre de termes pour chaque rang a ainsi une influence moins importante que dans les stratégies de pondération par rang de la structure. Cette stratégie renforce aussi l'impact des termes des questions liées (uniquement des termes corrélés avec ceux des rangs inférieurs).

De quel pays, Paris est-il la capitale ?
Quelles sont les spécialités locales ?
Est-ce qu'elles sont longues à préparer ?

Imaginons que les termes «Paris» et «capitale» soient très présents dans la collection. Alors, avec l'heuristique ci-dessus, les documents contenant «Paris» ou «capitale» ne pourront être sélectionnés que s'ils contiennent aussi les termes «spécialités»(ou un autre terme de rang 2) et «préparer» (ou un autre terme de rang 1) pour répondre à la dernière question.

Ainsi le nombre de documents contenant le terme «Paris» est toujours inférieur à la somme des nombres de documents contenant les termes des questions de rangs inférieurs.

B) Liens incrémentaux

Une variation de cette méthode de corrélation est de n'ajouter dans la *posting-list* que les documents qui respectent le critère de coorélation de termes au rang 1, puis d'ajouter ceux qui le respectent jusqu'au rang 2 et ainsi de suite jusqu'au rang le plus ancien. Ainsi, les contraintes d'existence des termes composant les questions les plus anciennes, sont moins fortes.

Il est possible de réaliser cette opération incrémentale sur des poids plutôt que sur des inclusions en *posting-list*. Avec une variante par poids les documents disposant d'une corrélation jusqu'au rang 3 auront en supplément le poids accordé pour une corrélation jusqu'au rang 2 et celle du rang 1. Au final, un terme aura un poids sur un document égal à la longueur de la plus longue chaine de corrélation qu'il permet de calculer.

C) Corrélation avec tout un rang

Une autre variation consiste à décider qu'un document n'est intégré dans la *posting-list* par sélection liée à un terme, que si tous les termes de tous les rangs précédents sont aussi présents dans le document. C'est une variation qui renforce encore plus le poids de la dernière question posée. Il est improbable de trouver dans un document (un passage de quelques centaines de caractères) d'un même auteur la totalité des synonymes pour un terme donné. C'est une variation qui impose une contrainte trop forte pour les termes des rangs supérieurs à 1.

Synthèse

Dans le but de trouver un juste milieu à la corrélation des termes, nous choisissons de nous intéresser à la variante à base de poids incrémentaux. Ce choix est une solution au problème de la section IV.1.1.2 (page 117). Dans le cas où il y a seulement deux termes importants l'un dans la dernière question l'autre dans la première, il n'y a plus de problème de choix de pondération. En effet, c'est ensemble que les termes ont le plus grand poids. Ils sont alors en «compétition» avec les corrélations de termes secondaires, mais c'est un problème classique qui n'est pas spécifique aux dépendances entre questions.

Intéressons-nous à la mise en œuvre de cette corrélation des termes.

Fonction TF	
base	$1/\#Term$
quantité d'information	$log(1 + 1/\#Term)$
racine carrée	$sqrt(1/\#Term)$

TAB. IV.2 – Plusieurs méthodes de calcul du Tf.

IV.3 Mise en œuvre de la corrélation des termes

IV.3.1 Rappel sur le $tf.idf$

Comme nous nous donnons pour cadre général le $tf.idf$ [Jones, 1972] nous devons choisir la formule de base que nous adapterons. La métrique du $tf.idf$ introduit par Salton [Salton & Yang, 1973] [Salton & Buckley, 1988], peut être déclinée en différentes formules, les tableaux IV.2 et IV.3 (page 124) présentent les plus utilisées [Manning *et al.*, 2008], avec les notations suivantes :

- $Term(t, D)$ = Nombre d'occurrences du terme t dans le document D. Notation abrégée «$\#Term$»
- $Docs(t)$ = Nombre de documents présentant au moins une occurrence du terme «t» dans une collection donnée. Notation abrégée «$\#Docs$»
- N = Nombre total de documents dans la collection.

Il est important de ne pas négliger la méthode de normalisation liée aux documents. Une partie de la difficulté réside dans la gestion de documents et de requêtes de tailles très variables. La méthode de normalisation est là pour pallier ces problèmes [Salton & Buckley, 1988]. De même, la normalisation existe sous de nombreuses formes comme celles présentées dans le tableau IV.4. Le score du document est multiplié par la méthode de normalisation. Par rapport au tableau IV.4, la méthode par défaut de Lucene[3] utilise une normalisation par «cosinus». Dans un souci de simplicité, nous choisissons pour notre nouveau modèle de ne pas normaliser les scores. Nous faisons l'hypothèse, que pour une grande majorité de nos documents, les problèmes ci-dessus sont rares car résolus en amont par la découpe des documents en paragraphes avant l'indexation.

Connaissant le cadre interactif que nous recherchons, nous pouvons proposer que les termes les plus récents soient plus importants que les plus anciens. Par ailleurs, connaissant le calcul de la probabilité d'association des dépendances et la confiance en la correction de l'association des dépendances nous suggérons le principe suivant :

[3]Nos expériences sont basées sur Lucene (Cf section II.2, page 69)

Fonction IDF	
base	$N/(1+\#Docs)$
quantité d'information	$log(1+N/(1+\#Docs)$
probabilité	$max(0, log((N-\#Docs)/(\#Docs)))$

TAB. IV.3 – Plusieurs méthodes de calcul de l'Idf.

Fonction de normalisation	
aucune	1
cosinus	$1/(\sqrt{(\Sigma_{i \in QueryTerms} t_i^2)})$
pivot	$1/u \mid u$ *est prédéfini pour le terme* $'t'$
longueur de D	$1/(charLength(D))^{\alpha} \mid \alpha < 1$

TAB. IV.4 – Plusieurs méthodes de calcul de la normalisation.

Un terme provenant d'une question d'un rang supérieur utilisé sans aucun terme de la dernière question, a moins de valeur qu'un terme de la dernière question utilisé sans aucun terme d'une question d'un rang supérieur. (Hyp B)

Le modèle à base de cooccurrences est construit sur l'hypothèse A globalement compatible avec celle ci-dessus. Le modèle que nous cherchons à créer peut donc s'appuyer l'hypothèse B dans le but de tenir compte de la présence simultanée des termes de la question et des termes des questions dépendantes dans les documents.

Les formalisations précédentes permettent de conserver une certaine compatibilité dans les méthodes de calcul pour de la modularité et des optimisations futures. Il nous suffit donc de dériver notre modèle à partir de variantes du tf et de l'idf qui conservent au maximum les informations. Les autres formes pourront être obtenues à partir de celles-ci ; nous choisissons donc les formes «quantité d'information» des tf et idf.

Passer de la forme «quantité d'information» à la forme «racine carrée» ne présente aucune difficulté (pour le tf), nous préfèrerons cette seconde forme par compatibilité avec Lucene. Ceci permettra d'obtenir des comparaisons plus fiables entre les méthodes de calcul lors des évaluations.

IV.3.2 Similarités avec le *phrase scoring*

La recherche de document qui se rapproche le plus du modèle ci-dessus est le *phrase scoring*(et *phrase queries*) [Manning *et al.*, 2008]. Avec le *phrase*

scoring, les documents sont indexés spécialement(par groupe de mots ou par position des mots) en vue de réaliser des interrogations où des mots doivent former une suite dans un document. Dans le *phrase scoring* il n'existe pas de notion de dépendance mais, comme dans le modèle à base de cooccurrences la présence simultanée d'un groupe de mots dans un document est obligatoire. Le modèle à base de cooccurrences n'impose pas que les mots se suivent et le modèle à base *phrase scoring* ne contrôle pas la fonction de pondération outre mesure[4].

IV.3.3 Variante du $tf.idf$

Le tf est construit sur la base de la fréquence des termes dans un document. L'*idf* est construit sur la base du nombre de documents contenant un terme par rapport au nombre total de documents. Comme nous ne cherchons pas seulement un terme, mais des corrélations de termes, nous allons étudier un score basé d'une part sur le nombre de documents contenant à la fois un terme de la question et des termes des questions dépendantes et d'autre part sur le nombre total de documents. Une solution est de réaliser une extension du $tf.idf$, qui tienne compte des rangs de la structure.

La « partie » tf est augmentée par les cooccurrences éventuelles des termes dans le document. La « partie » *idf* est réduite pour tenir compte de la quantité de documents qui présentent ces mêmes cooccurrences. Ces deux modifications sont réalisées pour refléter les stratégies proposées ci-dessus. Pour cela les fréquences liées aux termes des rangs les plus élevés sont multipliées aux fréquences des termes de premier rang à la manière d'une pondération.

Dans la suite nous utilisons les définitions suivantes. Soit t_{ij} le j-ème terme de rang i de la structure. Si $i = 1$ alors il s'agit d'un terme de la question. Soit $nombreDeRangs$ le nombre de rangs de la structure des dépendances.

IV.3.3.1 Fréquence des termes corrélés

Construisons un indicateur de la fréquence des termes de la question et de ceux de la structure dans un document, le Tf'. Nous n'accordons de l'importance à un terme de rang n que si un terme de rang $n-1$ de la structure est présent dans le document. Voici donc un système de fréquence des termes d'un rang pondéré par les fréquences des termes précédents :

[4]Les similitudes de cette méthode avec la nôtre fait qu'elles pourraient être implantées simultanément pour interroger sur des termes et non des mots. Mais elles introduisent des difficultés techniques supplémentaire.

$$
\begin{array}{|ll|}
\hline
Tf'(D) = & \sqrt{\Sigma_1^i \Pi_1^i \Pi_1^j (freq(t_{i,j}, D) + 1) - nombreDeRangs} \\
 & |i \in \text{rangs de la structure}, j \in \text{terme du } rang(i) \\
 & \\
freq(t, D) = \frac{1}{\#Term} \mid (\#Term > 0) & \\
freq(t, D) = 0 \mid (\#Term = 0) & \\
\hline
\end{array}
\quad \text{(F1)}
$$

C'est la somme des produits des fréquences d'un rang par le produit des fréquences des sous rangs. C'est une corrélation rang à rang.

Intuitivement, nous commençons par calculer l'impact pour les termes de rang 1, nous réalisons un produit des fréquences pour obtenir un impact global pour le rang. Par rapport au tf traditionnel, chaque rang est traité comme s'il s'agissait d'un terme unique(un super-terme), mais chaque rang(super-terme) est pondéré non pas par une valeur fixe, mais par le produit des fréquences de tous les sous-rangs(supers-termes) précédents. Il en résulte que moins les termes des premiers rangs sont présents, moins l'impact des termes des rangs les plus anciens est important. Notons que si un terme de rang n est absent, alors il représente un élément neutre pour l'opération de multiplication Π. Si tous les termes de rang n sont absents leur impact est exactement compensé par la soustraction finale du nombre de rangs. Si tous les termes de rang n sont présents le «+1» lié au rang est conservé[5]. Notons que seul l'ordre des documents nous intéresse, et que la conservation du «+1» ne change pas l'ordre des documents. Les différents rangs de la structure sont vus comme des super-termes. Sous cette forme nous retrouvons bien les formules classiques qui ne tiennent pas compte des dépendances. La fréquence corrélée de formule F1 est, pour une structure de profondeur 3 avec $m = 2$, défini comme suit :

Soit $t_{p,q}$ le q-ième terme du p-ième rang et $freq(x, D) + 1 = f(x)$ alors :

$$
\begin{array}{ll}
Tf'(D)^2 = & f(t_{1,1}) * f(t_{1,2}) \\
 & +f(t_{1,1}) * f(t_{1,2}) * f(t_{2,1}) * f(t_{2,2}) \\
 & +f(t_{1,1}) * f(t_{1,2}) * f(t_{2,1}) * f(t_{2,2}) * f(t_{3,1}) * f(t_{3,2}) \\
 & -3
\end{array}
$$

Nous voyons que les $i - 1 (i \in rangs\ de\ la\ structure)$ premiers termes du produit des rangs agissent comme une pondération définie dynamiquement.

Pour la dernière question de l'exemple du début du chapitre (tableau IV.1 page 114) nous avons $m = 3$. Il en résulte que la forme expansée de la formule est plus longue :

[5] $freq(t, D)$ pouvant être nul, or l'élément neutre de la multiplication est 1. Nous pouvons omettre le «+1» si la stratégie présentée en section V.3.1.1(page 143) est utilisée, ou qu'il n'y a pas de questions elliptiques.

Soit $t_{p,q}$ le q-ième terme du p-ième rang et $freq(x, D) + 1 = f(x)$ alors :	
$Tf'(D)^2 =$	$f(Chambre) * f(Palais)$ $+f(Chambre) * f(Palais) * f(Palais) * f(Musée) * f(Loger) * f(Palais\ d'hiver)$ $+f(Chambre) * f(Palais) * f(Palais) * f(Musée) * f(Loger) * f(Palais\ d'hiver) * f(Trouver) * f(Musée) * f(Ermitage) * f(Saint\text{-}Petersbourg)$ -3
Mais comme la stratégie de sélection des termes élimine les termes en double :	
$Tf'(D)^2 =$	$f(Chambre) * f(Palais)$ $+f(Chambre) * f(Palais) * f(Loger) * f(Musée) * f(Palais\ d'hiver)$ $+f(Chambre) * f(Palais) * f(Loger) * f(Musée) * f(Palais\ d'hiver) * f(Trouver) * f(Ermitage) * f(Saint\text{-}Petersbourg)$ -3

Le résultat exact pour l'exemple de l'arbre page 115 dépend du document. Remarquons que en rang 1, la réponse n'est pas encore calculée alors qu'elle l'est pour les autres rangs.

Le document D_1 contient la totalité des mots de la requête. En utilisant la notation $f(x) = x + 1$ où $x = freq(terme, D)$, le $Tf'(D_1)$ vaut :

$$\begin{aligned} Tf'(D_1)^2 &= f(1)^2 + f(1)^5 + f(1)^8 - 3 \\ &= 4 + 32 + 256 - 3 \\ &= 289 \end{aligned}$$

Supposons maintenant que nous calculons le Tf' du document D_2 qui est identique au document D_1 mais dont le mot *Loger* a été retiré. Nous obtenons les Tf' suivants :

$$\begin{aligned} Tf'(D_2)^2 &= f(1)^2 + f(0) * f(1)^4 + f(0) * f(1)^7 - 3 \\ &= 4 + 1 * 16 + 1 * 128 - 3 \\ &= 145 \end{aligned}$$

Nous voyons que le Tf' attribué aux documents respecte l'intuition que nous pouvions avoir sur les documents, car le $Tf'(D_1)$ est supérieur au $Tf'(D_2)$.

IV.3.3.2 Fréquence des documents avec termes corrélés

Construisons un indicateur de la fréquence des documents possédant des termes corrélés, l'Idf'. Soit $\odot$ l'opérateur binaire commutatif de corrélation de présence de deux termes dans un document. $docs(t_{i,j} \odot t_{x,y})$ désigne donc

le nombre de documents dans un corpus qui contiennent à la fois le y-ème terme du rang x de la structure et le j-ème terme du rang i de la structure. Un terme d'un rang donné de la structure n'est pris en compte que si au moins un terme de chaque rang inférieur (donc plus récent) est aussi pris en compte pour déterminer l'importance d'un nombre de documents. Dans le cas où tous les termes sont effectivement présents dans tous les documents contenant la bonne réponse, la quantité $docs(t_i)$ peut donc être substituée par $docs(t_i \odot t_{1,x} \odot t_{2,y} \odot ... \odot t_{n,z})$ où les valeurs x y ... z varient dans les limites possibles du rang concerné de la structure. Notons que les t_i de la requête sont intégrés aux calculs séparément les uns des autres. Obtenir tous les termes dans un même document est un cas idéal pour une requête idéale. Dans notre cas nous devons réduire nos contraintes sur la corrélation des termes, car nous ne sommes pas dans ce cas idéal. Nous pouvons relâcher des contraintes en autorisant certains termes des questions liées dans la structure à ne pas être corrélés aux autres. De cette manière, les corrélations de présence des termes sont moins fortes, et représentent des cas dégradés[6]. Plus la mesure est faible plus il existe un grand nombre de documents possédant ces termes corrélés.

Une solution est alors de prendre en compte toutes les corrélations impliquant au moins un terme d'un rang inférieur, et d'en faire la somme. Ainsi, s'il n'existe pas de corrélation, alors l'indicateur de fréquence des termes dans les documents sera faible. Par contre, s'il y a beaucoup de corrélations alors la valeur de l'indicateur sera renforcée par la fréquence de la corrélation la plus forte. Notons qu'une corrélation entre n termes implique l'existence d'une corrélation entre chaque sous groupe de $n-1$ termes... et récursivement.

Nous pouvons alors proposer la formule suivante comme indicateur de fréquence des documents :

$$
\begin{aligned}
Idf'(t_i) \;=\; & 1 + log(N) - log(1 + Docs(t_i) \\
& + \Sigma_1^x(Docs(t_i \;\odot\; t_{1,x}) \\
& \qquad \mid x \in t_1) \\
& + \Sigma_1^x \Sigma_1^y(Docs(t_i \;\odot\; t_{1,x} \;\odot\; t_{2,y}) \\
& \qquad \mid x \in t_1 \;,\; y \in t_2) \\
& + ... \\
& + \Sigma_1^x ... \; \Sigma_1^z(Docs(t_i \;\odot\; t_{1,x} \; ... \; t_{n,z}\;) \\
& \qquad \mid x \in t_1 \;,\; ... \;,\; z \in t_n,\; n \;=\; nombreDeRangs - 1) \\
& \qquad\qquad)
\end{aligned}
\tag{F2}
$$

Cette méthode de calcul se comprend en faisant une récursion. Pour un terme unique sans aucune dépendance nous retrouvons bien la formule de base. Imaginons maintenant que nous disposons d'un rang supplémentaire de dépendance. Le rang est ajouté à la partie précédente du calcul en faisant

[6] Ceci peut être utile comme précédemment, notamment pour les questions elliptiques

attention à la présence simultanée avec les termes de rangs inférieurs. Pour la présence simultanée, le système utilise l'opérateur de corrélation de présence. Chaque terme du rang est ajouté à son tour, en vérifiant la présence des termes de rangs inférieurs. La formule modélise bien cela sous la forme $\Sigma_1^x(\#docs(t_i \odot t_{1,x}) \mid x \in t_1$. L'addition ($\Sigma$) et la corrélation de présence($\odot$) étant commutatives, la généralisation pour des dépendances avec plus de rangs ne pose pas de problèmes particuliers.

IV.3.4 Variante du score

À notre variante du $tf.idf$ nous associons alors une variante de la méthode d'association de score au document. Traditionnellement le score est calculé par :

$$\begin{aligned} Score(Q,D) &= \Sigma_{t_i \in Q} tf(t_i,D) * idf(t_i) \\ &= \Sigma_{t_i \in Q} \sqrt{(1/(Term(t_i,D)))} * (log(N) - log(1+Docs(t_i))) \end{aligned} \quad \text{(F3)}$$

Par généralisation, la variante du score est définie via une relation entre la fréquence des termes et l'inverse de la fréquence des termes dans la collection. Bien que structurellement plus compliquée que les variantes des tableaux IV.2 et IV.3(page 123), c'est aussi une simple extension du $tf.idf$. Il n'est pas certain que la complexité combinatoire soit beaucoup plus élevée, car beaucoup de calculs peuvent être réalisés incrémentalement. De plus l'emploi de mémorisations en tables de hachage faible permet un très fort contrôle des principaux paramètres affectant la performance.

Le score d'un document est redéfini par :

$$\begin{aligned} score(Q,D) &= \Sigma_{t_i \in Q} Tf'(D) * Idf'(t_i) \\ &= \Sigma_{t_i \in Q} (F1) * (F2) \end{aligned} \quad \text{(F4)}$$

Seuls les poids des termes sont affectés, et non la normalisation des documents. Toutes les méthodes de normalisation traditionnelles restent donc utilisables[7]. Il n'y a pas de pondération supplémentaire, ce qui permet aux moteurs ne supportant pas la pondération «à la requête » de ne pas être pénalisés.

Nous avons proposé ici les versions, «*racine carrée*» et «*quantité d'information*» des Tf' et Idf'. La raison en est que nous voulons obtenir un modèle proche de celui de Lucene pour les tests. La version «*quantité d'information*» du Tf' peut s'obtenir simplement en remplaçant la fonction «*racine carrée*» par une fonction du «$log + 1$»[8]. À partir de la quantité d'informations, il est

[7]Notamment s'il ne peut pas être fait l'hypothèse de découpage en paragraphe.

[8]$Tf(D) = log(1+1/Term) = log(freq(D)+1)$ hors par construction nous avons choisi une étude à base de $\sqrt{(freq(D)+1)}$

facile de retrouver les autres variantes(section II.2 page 69) des deux tableaux et donc de s'adapter aux spécificités d'un corpus donné.

IV.3.5 Fondement du modèle à corrélation de termes

La mesure de similarité entre question et document n'est pas probabiliste, mais basée sur des études des discriminations des termes [Manning *et al.*, 2008] [Salton & Buckley, 1988] . Nous venons d'en proposer une extension avec pondération dynamique.

Comment procéder pour que notre modèle puisse être transféré à d'autres SQR que Musclef ? Ou à d'autres SQR utilisant des formalismes un peu différents ? Dans un but de compatibilité avec les autres SQR, il est plus profitable de chercher un modèle probabiliste comme cadre formel. Voyons comment nous pourrions le construire.

IV.3.5.1 Quelle probabilité est associée à cette pondération ?

Par rapport aux modèles probabilistes traditionnels, nous proposons un petit saut paradigmatique. Nous n'étudions pas la probabilité qu'a un document de contenir la réponse à une requête :

- puisque nous savons que nous prendrons plusieurs dizaines (centaines) de documents du corpus pour construire le résultat de la requête,
- puisque le reste du SQR est fait pour traiter ce genre de résultats constitués d'un certain (grand) nombre de documents.

Nous étudions la probabilité qu'un document contenant la réponse se situe dans les n premiers documents sélectionnés par le moteur de recherche.

Nous nous plaçons dans le cas où la réponse est présente dans le corpus. Supposons que l'un des documents contenant la réponse à une requête se situe dans la *posting-list*(c'est à dire qu'il existe un des termes de la requête présent aussi dans le document contenant la réponse). Si un moteur de recherche retourne l'ensemble des documents alors la probabilité de trouver la réponse à cette requête est de 1. En effet si nous disposons déjà d'un ensemble de termes permettant d'obtenir la réponse dans un document de la collection, alors la probabilité de sélectionner le document réponse en ne retournant que les documents de la *posting-list* est donc de 1. Puisque nous ne choisissons pas ici les termes de la requête, nous restreignons notre étude à ce cas.

Le score que nous associons aux documents est une confiance que le système a que les documents soient intéressants. Après normalisation entre zéro et un[9], nous représentons par une variable aléatoire X le fait que la réponse

[9]En utilisant la somme des scores de la *posting-list* et en divisant tous les scores par cette somme.

soit dans les n premiers documents[10]. Nous avons, si $n = 0$ alors $X = 0$ et aussi, si $n = \#documents$ alors $X = 1$.

Comme le score représente la confiance que le système a que les documents soient intéressants (c'est à dire que la réponse soit présente), nous étudions donc la probabilité que les n premiers documents soient intéressants. Par conséquent la confiance dans la présence de la réponse dans au moins 1 des n documents est la somme des scores normalisés entre zéro et un. Par hypothèse X dépend de n et de la *posting list*.

IV.3.5.2 Quel est l'impact sur n et sur la *posting-list* ?

Dans ce modèle qui tient compte de l'ajout de la structure des dépendances pour les questions, les termes de la requête ne sont pas changés, donc l'ensemble non ordonné des documents dans la *posting-list* n'est pas changé. Le n maximal, où la plus mauvaise stratégie d'attribution de score des documents est choisie systématiquement, est donc le même avec et sans la notion de dépendance. Seule la répartition des documents contenant la réponse est affectée. Seule l'espérance de trouver des documents contenant la réponse avant le n-ième document scoré pour des requêtes(tenant compte des dépendances) est modifiée. Il en résulte qu'il n'y a pas de nouveau document accessible par rapport à une requête sans pondération. Nous pouvons remarquer que l'Idf' et le Tf' proposés précédemment convergent vers l'idf et le tf traditionnel quand le nombre de rangs de la structure tend vers 1, c'est à dire quand il n'y a qu'une seule question. Nous en concluons, que dans le pire des cas il faut choisir un n identique à précédement (pondération classique).

[10]L'ensemble des termes du corpus de document est fini, de même pour les termes de l'utilisateur. En effet si un terme de la requête ne figure pas dans le corpus alors il ne figure pas non plus dans la *posting-list*. Il en résulte que la loi de la variable aléatoire est intégrable. Nous pouvons alors parler d'espérance mathématique.

Conclusion sur les corrélations de termes

L'étude des corrélations des termes des questions liées dans les documents d'une collection permet de construire un modèle à pondération dynamique pour la recherche des documents. Nous avons montré comment construire les classes de pondération à base de corrélation de termes pour tenir compte de différents cas de figure, selon l'utilisation envisagée du système. Nous avons instancié une nouvelle formule de calcul des scores des documents pour nos questions enchaînées. Nous avons alors vu qu'elle est compatible avec la normalisation (des scores) traditionnelle, avec la quantité de documents à sélectionner traditionnelle et avec la formalisation en probabilité proposée pour le calcul des dépendances.

Nous allons maintenant voir l'évaluation des modèles basés sur la corrélation des termes à base de dépendances entre questions pour les questions enchaînées.

Chapitre V

Évaluation de la recherche

Nous avons réalisé une évaluation de notre travail en l'intégrant au sein du SQR Musclef. Nous avons pour cela dû préparer les documents afin de les indexer par Lucene.

Nous avons ensuite inséré la construction de notre structure de dépendances et son utilisation lors de la recherche des documents.

Nous avons évalué les résultats sur un corpus annoté. La performance est estimée en fonction du gain obtenu en tenant compte des dépendances entre les questions.

V.1 Architecture de l'évaluation

L'architecture mise en place pour évaluer un SQR-enchaînées est sensiblement identique à celle d'un SQR classique. Il existe juste une étape supplémentaire de calcul des dépendances (ou juste une étape de résolution d'anaphores selon les systèmes). La figure V.1 (page 134) montre les différents aspects logiques de ce type d'évaluation mais, dans les limites de notre étude, nous nous arrêtons après la sélection des documents (des documents découpés préalablement en courts passages). En effet, dans le cadre de notre travail, nous n'avons pas modifié directement la méthode de sélection des phrases et réponses. Nous reviendrons sur cet aspect à la section V.3.5 (page 154).

Il faut préparer les corpus, que ce soit ceux de questions ou ceux de documents. Un aspect connexe de la préparation des documents et questions est la recherche (avec validation humaine) des réponses et des documents qui permettent de les obtenir. Ici les questions sont celles de ClefQA07-FR-EN

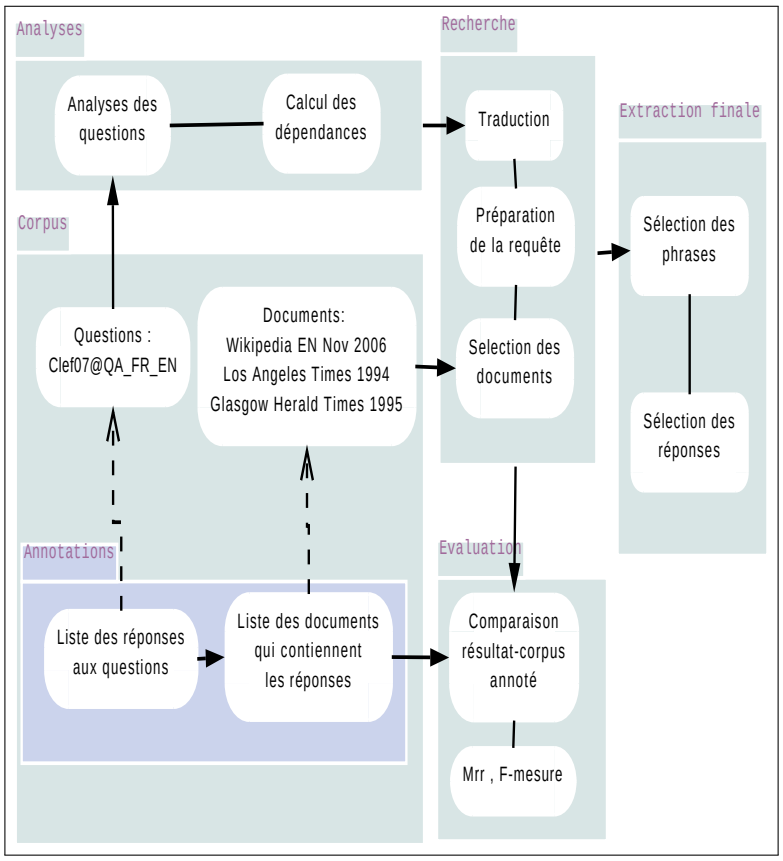

FIG. V.1 – Architecture d'évaluation pour les SQR adaptés aux questions enchaînées

et les documents sont une combinaison de la Wikipédia anglaise de novembre 2006 et d'une collection de journaux du Los Angeles Times (1994) et Glasgow Herald Times (1995).

Les traitements du SQR sont divisés en deux étapes : les analyses et la recherche. Les analyses se composent des analyses traditionnelles de questions et de celles spécifiques aux questions liées. La présence des questions liées est à prendre en compte lors de la recherche. La stratégie de préparation de la requête et la sélection des documents sont altérées. L'étape effective d'évaluation numérique peut alors avoir lieu par comparaison entre les documents sélectionnés et les versions validées «à la main». Des métriques comme le MRR et la F-mesure peuvent alors être mises en place pour donner des résultats supplémentaires.

V.2 Contraintes sur les documents

Nous allons d'abord nous intéresser à un aspect lié à l'ajout d'une structure de dépendance, la préparation des corpus jusqu'à l'indexation, et l'utilisation des termes dans la construction de requêtes pour une stratégie d'interrogation. Nous devons savoir dans le cas du domaine ouvert, quelles sont les hypothèses sur les documents, nous verrons alors comment utiliser nos outils.

Les documents nécessitent une préparation complexe avant leur utilisation dans un moteur de recherche. Même si la technologie des SQR permet de prendre en compte tout type de contenu, le format des documents lui-même doit respecter quelques règles élémentaires. Les documents doivent être :

- essentiellement constitués de texte brut en langue naturelle. Notamment les URL(s), formule(s) mathématique(s) ou autres ensembles spécifiques utilisant des caractères et mots mal codés sont exclus.
- munis d'une date de création.
- munis d'un titre.

Les documents doivent être suffisamment courts pour que les mesures de similarités soient significatives. Il est utile de savoir quelles sont les données qui sont éliminées et quelles sont les données qui sont indexées. Musclef ne dispose pas de stratégie permettant de gérer les tableaux et nous ne lui en avons pas ajoutée. Tous les tableaux sont donc supprimés des documents. Tous les documents sont transcodés en un encodage Latin1 (iso-8859-1). Si un mot comporte des caractères qui ne peuvent pas être encodés en Latin1 alors il existe deux cas : si une lettre de substitution d'un graphisme proche et canonique est connue, la conversion est alors transparente, sinon le mot est supprimé. A noter que les adresses des documents d'origine sont présentes dans un champ spécial des documents indexés. Il serait techniquement possible qu'une étape ultérieure des calculs puisse désigner un mot supprimé comme partie d'un terme réponse.

Il y a aussi quelques problèmes qui ne peuvent pas être traités à ce niveau d'analyse avec ce genre de techniques :

- les passages d'un document ayant des vocabulaires différents ne seront pas associés ; Ceci peut conduire à ne pas remonter des documents-réponses ;
- l'ordre d'apparition des termes dans le document est perdu[1].

[1]Les implémentations récentes proposent tout de même pour des coûts raisonnables (augmentation dans la limite de 2 fois la taille de l'index et 2 fois le temps de calcul), des méthodes pour réaliser des métriques de similarité tenant compte d'éléments d'ordonnancement des mots.

V.2.1 Le nettoyage des documents

Les documents ne sont jamais disponibles dans le format géré par l'index de document. Souvent les documents sont dans des formats originaux (xml ou autre) et dispersés sur plusieurs sources. Il y a 2 points critiques à observer :

- la sélection des données, visant à fixer les documents en opposition à une source dynamique (le web par exemple).
- l'unification des cadres logiques pour la sémantique des balises dans les documents structurés.

La sélection des données couvre les aspects techniques de bande passante, récursivité, recouvrement, gestion réseau, stockage et autres... Cela dépasse un petit peu le cadre de notre étude, mais il faut souligner qu'il ne s'agit pas d'une opération triviale qui peut être répétée à souhait ou d'une opération dont l'organisation peut être facilement modifiée pour des tests.

L'unification des cadres logiques est moins délicate sur le plan technique, mais elle est tout aussi critique pour réaliser une bonne interrogation. Il faut respecter deux contraintes pour unifier. Tous les formats de documents doivent pouvoir être injectés dans le format d'indexation. L'organisation des documents doit être similaire pour que des tâches comme le découpage des textes en documents de longueur fixe/connue ne brise pas (ou pas trop) de liens (anaphoriques ou autres).

Par exemple, une procédure cherchant à injecter les pages de la *Wikipédia* en respectant les contraintes énoncées plus haut, rencontrera successivement plusieurs problèmes pour lesquels une logique doit être définie :

- des problèmes d'ordre technique (gestion des formats, gestion mémoire, gestion des erreurs et malveillances ...)
- des problèmes de décisions statistiques. À partir de quel moment une section balisée (une référence par exemple) devient gênante pour la recherche d'information ? Si le seuil est trop bas nous risquons de perdre du texte contenant des réponses.
- des problèmes de résolutions de liens anaphoriques vers des données supprimées, car elles sont dans une langue spécialisée ou pointent vers des images.

Si les deux seules méta-informations choisies sont la *date de création* et le *titre* du document, cela signifie que la procédure de formatage devra spécifiquement sélectionner ces deux méta-informations. Mais cette même procédure devra éliminer les méta-informations qui ne sont pas correctes (auteurs, dernières dates de modifications, liens internes dans la collection ou lien externes, indicateurs de types de pages ...). Ceci implique de disposer d'un

cadre bien défini pour déterminer ce qui ne relève pas d'un certain type de méta-information en domaine ouvert[2].

V.2.2 Indexation des documents

La construction d'un index repose d'abord sur le hachage et la compression des termes. Le choix des structures de données exactes fait intervenir de nombreux paramètres et, parfois, des choix arbitraires. Les détails techniques peuvent subir des transformations radicales en fonction des possibilités de l'index (en quantité de données, en fonctionnalité secondaire, en export et intégration de type de données, portabilité, cryptage...).

V.2.2.1 Avec Lucene

Pour Lucene, un terme est une chaine de caractères (étiquettes), qui peut représenter un mot, ou bien une date, ou bien une URL, ou bien des méta-informations. Lucene n'a pas de support «natif» pour notre notion de terme.

V.2.2.1.1 Du terme au mot et inversement Afin d'obtenir un formalisme compatible entre Lucene et le reste du SQR les (multi-)termes sont transformés en suite de mots. Nous rappelons que la principale raison pour ne pas avoir fait cela plus tôt vient de la nécessité de leur traduction. Or comme l'index est déjà dans la langue cible, la traduction est un prérequis. La contrainte de traduction des (multi-)termes est moins nécessaire à cette étape.

Ce biais sur l'interrogation à base de mots plutôt que de termes est regrettable. La solution pour interroger sur des termes non découpés en mots aurait été de réaliser du *phrase scoring* comme vu à la fin de la section IV.3.2 (page 124). Il y a plusieurs raisons pour ne pas l'avoir fait :

- la réalisation technique est plus complexe (surtout en complément de nos modifications).
- ce n'est pas le fonctionnement d'origine de Musclef.
- comme cela affectera tous les tests, aucun ne devrait en tirer un avantage significatif.

Dans la suite, nous utilisons la notion de terme au sens de Lucene, comme une chaine de caractères.

[2]Même si les méta-informations jugées incorrectes pourraient être utiles si le système disposait d'une méthode pour les utiliser.

V.2.2.1.2 Les flux d'étiquettes Nous avons créé des modules destinés à transformer les documents en flux d'étiquettes transportant les termes. Ces étiquettes contiennent de nombreuses méta-informations, dont notamment les lettres de chaque mot. Chaque flux[3] est un pointeur vers un document. C'est sur ce flux que travaille le système d'indexation de Lucene, et toutes les informations qu'il contient sont réutilisables (sous réserve des bonnes options et réalisations de classes.). En manipulant correctement ces modules, il est possible de créer des flux ne représentant que les informations sur les différents aspects des documents que nous désirons. Par exemple, il est possible de créer un champ pour les titres, un pour les dates, un autre pour les entités nommées et un dernier pour les mots du corps du texte.

V.2.2.2 Conséquences des flux

Ce mécanisme de flux permet de laisser beaucoup de données dans les fichiers des documents et de sélectionner uniquement celles qui sont utiles. C'est un mécanisme transparent par rapport à d'autres moteurs qui ignorent automatiquement des balises.

Prenons l'exemple d'un document constitué d'une unique phrase (balisée, étiqueté par le treetagger, annotée en entité nommée) :

```
<document >
<date>6 juin 2009</date>
<title>Nom de chat.</title>
<text>
Mon DET mon
chat NN chat
s'PP se
appel VB appeler
<entiteNomme>
Mimiroux NP mimiroux
</entiteNomme>
. SENT .
</text>
<document>
```

Nous avons alors créé les flux de noms : *titre*, *date*, *lemmes*, *entiteNommes* et *source*. Notre répartition des données du document est alors :

- *titre* : Nom de chat
- *date* : 6 juin 2009

[3]Un flux est un «field» dans nomenclature de Lucene

- *lemmes* : mon chat se appeler mimiroux
- *entiteNommes* : mimiroux
- *source* : fichierDUnePhrase.txt

Dans la lignée du système Musclef, nous indexons les lemmes, mais pas les mots. Les mots peuvent être retrouvés au besoin (ainsi que la mise en page) grâce à la référence vers le fichier source. De ce fichier source, nous pouvons déduire les noms des différents fichiers résultats des analyses intermédiaires.

Nous pouvons donc orienter spécifiquement la sélection des documents sur les titres de documents (les titres des sections et sous-sections sont aussi ajoutés dans le cas de la Wikipédia), ou orienter la sélection vers les dates ou les entités nommées en fonction de l'importance que le développeur chargé du déploiement final choisit. Notre pondération initiale pour chaque flux donne tout le poids aux lemmes.

V.2.2.3 Modification des calculs

Si nous désirons utiliser les flux, il en résulte une modification de la sémantique des calculs et un coût en performance[4]. Si un terme $t1$ est relié à un flux $f1$ et un terme $t2$ à un flux $f2$ alors le Tf (et Tf') d'un document D donné peut être différent entre le flux $f1$ et le flux $f2$. Notre organisation des calculs doit donc prévoir un recalcul du Tf non pas pour chaque document seulement, mais pour tous les flux existants. La nouvelle organisation est très simple, mais largement suboptimale. Nous avons mis en place pour chaque document une boucle de calculs qui, pour chaque terme au sens Lucene, calcule l'Idf et le Tf, leur produit, puis leur somme cumulative dans le score du document pour chaque mot de la requête. Ce calcul pourrait largement être amélioré en séparant le calcul des Idf qui ne nécessite pas de référent à un document particulier. L'utilisation d'une petite mémoire résoud aisément les problèmes de complexité combinatoire qui pourraient survenir[5].

V.2.3 Stratégie d'interrogation

Le choix de notre stratégie de sélection de documents pour l'étude des questions enchaînées est celui de la requête unique. La stratégie ne relâche pas les termes tant qu'elle n'a pas obtenu suffisamment de documents. Avec notre méthode de recherche de documents en interrogation unique, nous ne

[4]L'existence des flux est une des raisons pour laquelle Lucene n'est pas le moteur de recherche le plus rapide possible.

[5]La complexité dans le pire des cas ne change pas, mais les cas moyens et médians sont très améliorés.

cherchons pas à obtenir des documents supplémentaires en essayant des requêtes alternatives. Il y a plusieurs raisons :

- Dans l'hypothèse où nous réaliserions plusieurs requêtes pour une même question, comment juger de l'efficacité d'un modèle si celui-ci est encapsulé dans une stratégie ayant un objectif redondant ? L'enjeu est de valider un modèle de recherche de documents. Si nous englobons ce modèle dans un autre, c'est l'ensemble qui est évalué ; ce n'est pas notre modèle. Il faut en une seule application du modèle mathématique tenter de trouver l'ensemble de documents intéressants pour pouvoir conclure. Si à la suite de cela, nous décidons qu'il vaut mieux faire plusieurs requêtes, rien ne nous empêche d'intégrer plus finement les multiples requêtes dans une pondération adéquate.
- Le fonctionnement à base de plusieurs requêtes n'est justifié que dans le cas d'interrogations booléennes. Dans le cas d'une interrogation à base de pondération si certains mots ne peuvent pas être trouvés dans la collection, ils n'empêchent pas les autres mots de rapporter des résultats. La stratégie d'interrogation multiple avec un fonctionnement en recherche booléenne bornée à n documents ne se justifie plus. Elle était déployée sur lorsque le système utilisait le moteur de recherche MG [de Kretser & Moffat, 2000].
- Certaines questions n'ont pas de réponse dans la collection de documents : nous notons « *nil* » ces réponses vides. Nous ne cherchons absolument pas à trouver des documents candidats pour les questions à réponses *Nil*. Dans une optique de campagne d'évaluation de SQR, étant donné la proportion de questions à réponses *Nil*, la stratégie qui consisterait à chercher des documents même pour ces questions, n'est pas la meilleure[6].

Il faut trouver une organisation des documents pour les indexer et une méthode de sélection des termes.

[6]D'un autre côté, les «Nil» au niveau de la sélection des documents ne sont pas considérés comme des bonnes réponses. Les raisons qui nous poussent par hasard à ne pas sélectionner de documents pour les questions à réponses «Nil» ne sont pas les bonnes. La question «Où est mort Jacques Chirac ?» admet la réponse «Nil», pourtant une requête avec ces mots doit rapporter des documents, car il existe des documents où il est par exemple question «d'hommage aux morts» et de «Jacques Chirac ». Ce n'est pas à la sélection des documents de déterminer si la réponse doit être «Nil» ou pas. Quand la sélection des documents ne donne aucun résultat, c'est plus probablement qu'un problème est survenu dans les étapes précédant l'interrogation du moteur (analyses des questions, sélection des dépendances, des termes, des traductions ...), mais s'il n'y a pas de documents c'est peut-être aussi que la question admet la réponse «Nil».

V.2.3.1 Organisation des documents

Les documents des collections d'évaluation V.3.3 (page 145) proviennent de la « Wikipédia » et d'un corpus d'articles du journal « LA-Times » et « GH-Times ». La mécanique des flux de Lucene nous aurait permis de séparer les documents en deux flux en fonction de leur origine. Pour des raisons historiques de développement de Musclef et de temps nécessaire aux développements (et analyse des conséquences), ce n'est pas la stratégie que nous avons retenue. Les documents sont fusionnés dans une représentation semblable et indexés sans distinction d'origine. L'avantage de cette méthode est de ne pas avoir à fusionner des listes de documents possédant un score, sans être certain qu'il n'existe pas des différences de sémantique dans les scores.

V.2.3.2 Sélection des termes ou choix de la requête

La construction de la requête est réalisée en deux étapes. Une première étape sélectionne les termes en fonction de leur catégorie (noms propres, nombres...) et ajoute un certain nombre de variantes de traductions et de synonymes. Cet ajout concerne les termes de la question elle-même et les termes des questions liées.

La seconde étape détermine quel « assemblage » de ces termes est utilisé pour construire la requête. Pour cela, le système sélectionne en priorité les élements les plus significatifs en privilégieant la catégorie « noms propres » ou la catégorie « nombres », sinon il sélectionne d'autres termes significatifs dans la question. Dans les deux cas, le système applique la même stratégie pour les termes des questions liées (mais les expressions figées entre guillemets sont conservées telles quelles) qui sont ajoutés avec les indications de provenance dans la requête[7]. Si l'analyse de la question ne parvient pas à fournir les informations nécessaires (et donc, altèrent la sélection des documents) alors elles sont remplacées quand cela est possible par celle de la question de rang supérieur.

[7] L'étude des corpus de questions ClefQA07-FR-ES et ClefQA07-FR-FR nous a montré que les expressions figées entre guillemets sont toujours des arrangements de mots traduits, tirés directement d'un document du corpus (dans la campagne ClefQA2007 .)

V.3 Déploiement d'un SQR enchaînées

Nous avons vu les modifications qu'il faut apporter aux composants d'analyse des questions pour réaliser un modèle d'interaction dans le cadre logique des SQR-enchaînées à base de dépendances. Nous avons deux points à évaluer :

- le gain des performances en sélection des documents
- l'impact sur les réponses courtes de l'apport des dépendances

V.3.1 Intégration du moteur de recherche

Comme souligné précédemment, les modifications à apporter au moteur de recherche sont profondément liées aux objectifs mêmes du SQR-enchaînées et aux premières étapes d'analyse des données (à savoir l'analyse de la question).

Les modules du SQR qui suivent l'extraction des documents (sélection des phrases, extraction des réponses...) sont conçus pour traiter le cas où aucun document n'aurait été retourné.

V.3.1.1 Décalage de rang

Si à un certain niveau de la structure de dépendances, la sélection n'a rapporté aucun terme, alors les calculs utilisant ces rangs peuvent être simplifiés de manière à ne pas avoir à rechercher des corrélations de présence de termes dans des documents en utilisant un ensemble de termes vide. En effet, si l'ensemble des termes est vide, alors il n'y a aucun document pouvant disposer d'au moins un terme ayant une corrélation de présence avec un terme d'un niveau antérieur. Toutes les fractions du score faisant appel à ces niveaux de la structure seront donc nulles. Il en résulte un score nul et donc un ensemble de documents systématiquement vide.

Par exemple, dans le cas de questions elliptiques ne donnant pas de termes sélectionnés, il peut arriver que l'ensemble des termes de la dernière question soit vide et que les questions liées contiennent des termes qui eux sont sélectionnés. Il en résulte que la totalité du score est nulle pour cette question elliptique.

Quelle pièce d'Alexeï Chipenko fut écrite en 1984 ? Qui est-il ?

Dans la seconde question, il n'y a pas de mots sélectionnés [8].

[8] La configuration de base exclut les recherches constituées d'un unique verbe de type être ou avoir.

Pour contrer ce problème, il suffit d'*oublier* les niveaux de la structure où aucun terme n'est sélectionné dans une question. Cela peut se faire en décalant d'un rang les niveaux de la structure pour éviter le niveau vide. L'ordre intrinsèque des documents sera conservé, car pour une requête tous les documents sont affectés par cette modification de la métrique[9]. Ceci fait perdre le score absolu des documents. Les scores de deux requêtes différentes sur le même corpus ne sont plus comparables. À cause de la nature même des niveaux de la structure des dépendances et des stratégies de sélection de documents, les scores n'étaient déjà plus comparables.

V.3.2 Modification du moteur de recherche

Le moteur de recherche doit être modifié de plusieurs façons pour admettre les nouvelles requêtes. Il doit d'abord accepter des requêtes dans un nouveau format. Ceci implique qu'il doit être en mesure de traiter une nouvelle structure de données pour représenter la requête. Le calcul de la mesure de similarité doit alors être adapté à ce format. Il faut intervenir directement sur la fonction qui calcule un score par rapport à une requête et un ensemble de documents.

C'est dans cette fonction de calcul des scores que nous réalisons l'implémentation des modifications du calcul du $tf.idf$, pour mettre en place le calcul utilisant une probabilité contextuelle. La première version de l'implantation du calcul des scores en fonction de la probabilité contextuelle conduit à un nombre des calculs bien plus grand que celui de la version non contextuelle (plus d'un ordre de grandeur). Un examen du détail des calculs réalisé par la machine à différents niveaux d'analyse révèle une très grande redondance des calculs d'une partie de la formule à l'autre.

V.3.2.1 Réduction du temps de calcul

En utilisant un système de table de hachage pour mémoriser les calculs, nous pouvons réaliser une réduction importante du nombre d'opérations.La table de hachage permettant le plus gros gain est celle qui mémorise la partie concernant l'inverse de la fréquence des termes dans les documents (Idf'). D'autres tables ont aussi été mises en place pour les opérations redondantes.

Afin de contourner le problème lié à la mémoire due à l'utilisation massive des tables de hachage, il existe une solution. Celle-ci consiste à choisir une table initialement grande devant le nombre de résultats à stocker, pour la factorisation de tous les calculs d'un unique groupe de questions. Puis au lieu

[9]La réponse sera cherchée sur les mêmes documents mais plus ou moins difficilement.

d'agrandir cette table de hachage quand son taux de remplissage devient fort, nous supprimons aléatoirement (pseudo) une portion (le tiers) du contenu de la table de hachage. Comme le programme est déjà obligé de tester la présence d'un résultat dans la table avant de décider s'il doit le calculer ou s'il peut le réutiliser directement, les utilisations de la table ne sont pas affectées. Statistiquement les résultats des calculs les plus réutilisés seront plus présents dans la table. Il en résulte un faible accroissement du temps de calcul, mais une charge mémoire beaucoup plus réduite. Cette heuristique des calculs permet de gérer complètement le ratio cpu/ram via le taux de remplissage [10]

Avec ces optimisations, le temps de calcul et l'utilisation de la mémoire de la version des scores utilisant les corrélations de termes n'est guère supérieur à celui du temps de calcul de la version ne les utilisant pas.

V.3.2.2 Toujours plus vite

Le moteur de recherche utilise une méthode de recherche qui pourrait associer un score à chaque document. Dans la pratique, des optimisations de la *posting-list* visent à réduire les calculs de scores aux documents ayant les meilleures chances d'avoir un haut score. Puis l'évaluation est paresseuse pour les autres documents. C'est l'ensemble de ces heuristiques qui permettent d'obtenir des temps de calcul raisonnables.

Gagner en vitesse de calcul, c'est s'autoriser plus de tests systématiques, soit pour les preuves de corrections (et débogages), soit pour l'analyse des performances sur de vastes ensembles de paramètres. Par exemple, gagner en vitesse permet de tester plus de métriques et de faire des corrélations entre elles.

V.3.3 Expérimentation, la méthodologie

Voyons maintenant la méthodologie retenue pour évaluer l'impact sur les performances de la recherche dans les documents.

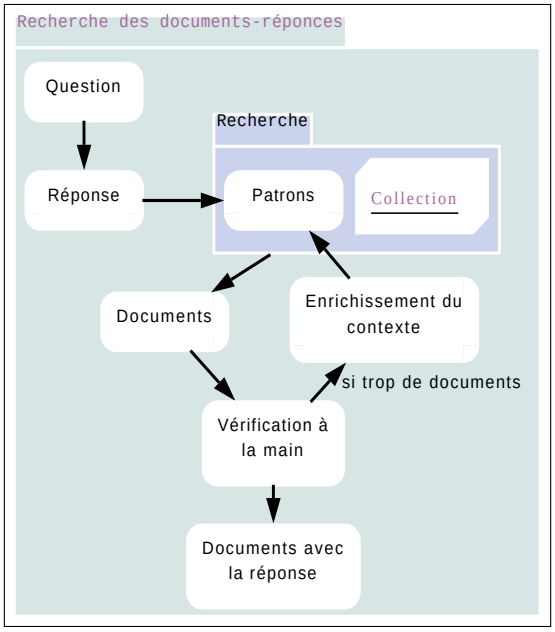

FIG. V.2 – Procédé de recherche des documents contenant les réponses pour constituer la référence d'évaluation.

V.3.3.1 Détermination de la présence de la réponse dans un document

Pour la préparation des corpus d'évaluation, nous avons développé une méthode adaptée à la tâche. La détermination de la présence de la réponse dans un document est réalisée partiellement à la main comme dans la figure V.2 (page 146). Dans un premier temps, une liste des réponses courtes[11] attendues est constituée pour chaque question. Ces réponses courtes sont trouvées en utilisant des méthodes traditionnelles semi-automatiques de recherche d'information. De ces réponses courtes, nous déduisons des ensembles de patrons figés qui permettent de les identifier dans des documents. Nous calculons alors l'ensemble des documents contenant ces patrons. Nous bou-

[10] Cette heuristique est dérivée du fonctionnement des tables de hachage faible (*weak Hash* ou *weak Pointers*). Dans les *weak Hash* la portion supprimée est décidée non pas aléatoirement, mais par le gestionnaire de mémoire du programme (disponible uniquement dans les langages qui en possèdent un). Dans certaines implémentations nous pouvons demander d'*enregistrer* la fonction dont le programme optimise le calcul, ceci afin de gérer de manière complètement transparente l'accès aux résultats.

[11] Les SQR ont pour objectif de trouver des réponses courtes. Donc il est inutile de trouver une stratégie pour trouver les réponses sophistiquées ou demandant des raisonnements.

clons alors sur deux opérations jusqu'à ce que la première hypothèse soit vérifiée :

- soit il y a suffisamment peu de documents ; nous vérifions «à la main» pour chaque document que le patron figé qui est trouvé correspond bien à la réponse.
- soit il y a trop de documents pour faire cette opération «à la main» pour chaque document, nous sélectionnons alors un petit cluster de documents que nous analysons à la main pour préciser les patrons.

Ces documents permettent de déterminer un ensemble de patrons secondaires « le contexte » qui doivent être présents dans le document pour que le patron réponse soit vraiment la réponse. Nous tenons évidemment compte des dépendances. Et nous recalculons l'ensemble des documents contenant les patrons avec « le contexte ». Nous obtenons alors 2 ensembles, un ensemble de documents contenant les réponses aux questions, un ensemble de patrons de réponses suivant une logique de type «et/ou» pour obtenir les documents contenant les réponses[12]

Cette méthode de recherche des « bons » documents n'est pas forcément optimale, mais issue de l'historique de la méthodologie Musclef d'évaluation des résultats. *In fine*, nous avons adapté le programme de sélection des documents dans la collection pour qu'il puisse évaluer les résultats retournés par les différentes versions des tests sur la recherche de documents.

V.3.3.2 Quelques caractéristiques du corpus d'évaluation

Notre évaluation a porté, comme au chapitre III sur les 200 questions du corpus ClefQA07-FR-EN en français avec réponse attendue à partir du corpus anglais de la Wikipédia de novembre 2006 et de l'année 1994 des journaux LA et GH. Nos patrons de bonnes réponses nous permettent de découvrir un maximum de 143 bonnes réponses et nous savons qu'il existe au moins 3 questions admettant une réponse « Nil »[13]

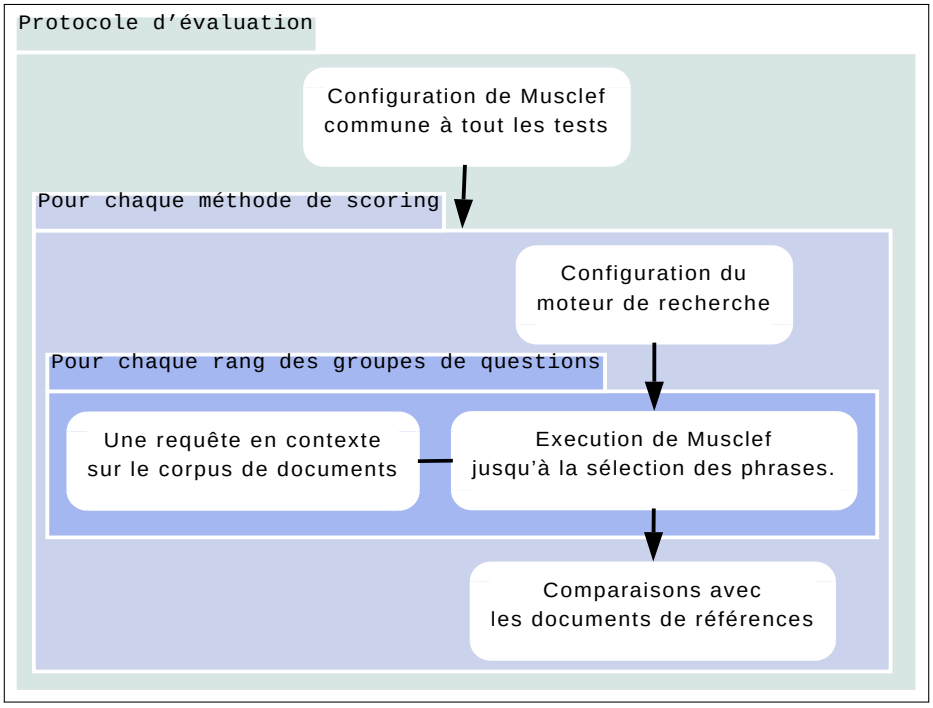

FIG. V.3 – Protocole d'évaluation.

V.3.3.3 Méthode d'évaluation

Comme le montre la figure V.3 (page 148), l'évaluation d'une stratégie se déroule en plusieurs étapes[14]. Des paramètres globaux à toutes les évaluations servent à configurer la plateforme Musclef. La configuration est réglée pour que les calculs s'arrêtent après la sélection des phrases (V.3.5.1 page 155). Pour chaque stratégie d'attribution des scores, un moteur de recherche est mis en place. La configuration Musclef est re-adaptée à son fonctionnement (notamment le format des requêtes). L'exécution a lieu et les résultats pour le corpus sont obtenus en un temps variant entre 20 minutes et 2 heures. Le programme de compte des bonnes réponses est alors lancé. D'autres programmes

[12]Pour une réponse, il peut y avoir plusieurs patrons testés séparément : cela fonctionne comme un opérateur « ou » ; pour une réponse, il peut y avoir plusieurs patrons secondaire « de contexte » : cela fonctionne comme un opérateur « et ».

[13]Soit 146 réponses identifiées, mais comme expliqué précédement nous ne comptons pas les «Nil».

[14]Les questions sont traitées en fonction de leur rang dans le groupe dont elles font partie. Les questions en rang 1 sont toutes traitées puis, celles de rang 2, etc... Ainsi, les questions dont les requêtes sont les plus courtes sont traitées les premières.

Stratégie	Nb réponses	MRR(Ok)	Moyenne(Ok)	MRR(All)	Moyenne(All)
A	88	0.19	15.86	0.083	62.98
B	100	0.17	18.99	0.088	59.49
C	102	0.19	15.86	0.083	62.98
D	99	0.19	18.44	0.094	59.63
E	123	0.12	84.13	0.079	90.24
F	123	0.09	205.3	0.060	164.8
G	104	0.18	22.43	0.094	59.66

TAB. V.1 – Statistiques sur les bons documents-réponses pour différentes stratégies d'attribution de scores avec Musclef.

d'analyses sont aussi appliqués à ce moment, notamment la sélection pour chaque question des scores ordonnés pour chaque document.

V.3.4 Évaluation des stratégies de calcul de score

Les résultats bruts de nos évaluations sont récapitulés dans le tableau V.1 (page 149) [Séjourné, 2009]. Antérieurement à ces travaux, des études sur Musclef ont montré qu'une sélection de 100 documents (passages) transmis à la sélection des phrases, est un bon choix. La sélection pour $n = 100$ ne se réalise vraiment que si la *posting-list* contient au moins n document.

V.3.4.1 Les métriques d'évaluation

Nous avons retenu deux métriques pour juger de la qualité de classement des documents. Le MRR qui concerne directement les performances utiles du système, car cette mesure favorise beaucoup les premiers bons documents. Et la Moyenne, qui donne un indice plus lisible des efforts à réaliser ultérieurement.

Le MRR(Ok) est calculé en ne tenant compte que des questions pour lesquelles au moins une réponse a été trouvée. C'est la moyenne des inverses des rangs des questions pour lesquelles un document-réponse a été trouvé dans les n premiers documents. De même pour la Moyenne(Ok) qui est une moyenne de rangs de document-réponse. Les MRR(All) et Moyenne(All) sont les approximations avec autant de décimales significatives que le MRR et la Moyenne traditionnels. Contrairement au MRR(Ok) si une réponse n'est pas dans les n premiers documents nous comptons simplement zéro. C'est ou bien la somme inverse des rangs des questions ou bien zéro, divisé par le

nombre total de questions. De manière similaire, la Moyenne(All) est calculée en comptant $n+1$ s'il n'y a pas de document-réponse dans les n premiers documents. Les calculs des All sont réalisés sur une base de 200 questions, mais ce qui est vraiment intéressant, c'est l'apport relatif des différentes méthodes. Il est facile de recalculer à partir des OK n'importe quel MRR ou Moyenne.

V.3.4.2 Les différentes stratégies de calcul des scores

Nous avons pratiqué des expérimentations et évaluations en calculant le score des documents de la *posting-list* avec différentes stratégies.

Dans la stratégie hors contexte (A du tableau V.1 page 149), les questions sont traitées de manière traditionnelle sans prise en compte des dépendances. Elles sont envoyées dans la méthode classique de Lucene. Analysons maintenant les résultats des stratégies utilisant les dépendances en rapport avec celui-ci.

V.3.4.2.1 Une ré-implémentation du tf.idf (B) C'est une ré-implémentation du $tf.idf$ où les informations de provenance des termes sont oubliées et où ils sont tous vus à égalité. Cette stratégie est donc vierge des modifications introduites dans le calcul par défaut de Lucene. Contrairement à cette méthode de Lucene, il n'est pas utilisé de coefficient de normalisation calculé à l'indexation pour représenter les variations sur les longueurs de documents. Ce travail est laissé à la partie traitant le découpage des documents en passage.

V.3.4.2.2 Stratégie C,E et D,F Les stratégies E et F ont été réalisées avec $n = 1000$ alors que les méthodes C et D ont été réalisées avec $n = 100$ Les stratégies C et E ont été réalisées avec la méthode par défaut de Lucene où l'origine des termes est oubliée. Les stratégies D et F ont été réalisées avec la méthode de *scoring*[15] présenté dans la section IV.3.3 (page 125). Les stratégies de pondération de référence sont celles réalisées avec les méthodes C et E.

V.3.4.3 Le boosting des classements

Si nous nous autorisons (en léger sur-apprentissage) à examiner nos résultats sur l'ensemble de nos corpus alors nous pouvons être tenté d'utiliser simultanément les méthodes aux performances similaires.

[15]Le *Scoring* est l'attribution des scores aux documents quand elle est réalisée d'une manière adaptée à un index.

V.3.4.3.1 La méthode de fusion (G) Nous avons remarqué que les méthodes C et D ont des moyennes d'OK très inférieures à 50, or nous sélectionnons plus de 100 documents[16]. Il est donc sans risque de : soit réduire le nombre de documents, soit prendre les 50 premiers des 2 méthodes. Ici nous avons pris les 50 premiers documents de C et D, et retiré les doublons. Il aurait été possible d'aller chercher plus de 50 documents une fois les doublons retirés. Il aurait aussi été possible de faire un mélange alternatif tenant compte des rangs des questions (en rang 1 et 2 nous prendrions les questions en rang 1 de chaque méthode, etc...), cela permettrait d'augmenter le MRR. En effet, les bonnes réponses sont classées en moyenne au rang 15-18 par les deux méthodes, cela remonterait leur rang moyen de 50+18 à 18+18. Les tests de fusion exacts n'auraient pas forcément été plus intéressants, car ce sont déjà nos meilleurs résultats. Nous observons le même MRR(All) que pour le système D car ce sont les réponses du système D qui ont été mises en première place.

Nous constatons que pour la stratégie A par rapport à la plus mauvaise méthode utilisant les dépendances, l'apport est de 12 nouvelles bonnes réponses, soit un gain de plus de 20% juste pour l'introduction basique des termes issus des dépendances entre questions.

V.3.4.3.2 Répartition des scores En optimisant encore les calculs, il est possible d'ajouter un enregistrement des scores rencontrés pour des analyses ultérieures. La figure V.4 (page 153) montre les scores bruts[17] pour chaque document pour chacune des 200 questions du corpus ClefQA07-FR-EN . Nous pouvons observer de grand écarts de répartition des scores. Comme ce graphique est difficilement lisible nous avons augmenté n de 100 à 1000 (pour un effet de recul) et trié les résultats par deux méthodes différentes.

Dans la figure V.5(page 154), nous observons les scores moyens des 1000 premiers documents triés par leurs scores pour les 200 questions du corpus. Ce document nous montre que contrairement à l'idée sur la régularité des résultats que pouvait nous donner la figure V.4, nous voyons que, d'un rang à l'autre, la différence entre les scores est faible (devant le score total) pour le classement des documents en moyenne. C'est ce qui donne cet aspect «lisse» au graphique car le graphe est suffisament grand pour qu'une vaste quantité de documents reçoive un score [18].

[16]Il y a 3057239 entrées (documents découpés en paragraphes) dans l'index.

[17]Scores calculés par la méthode de corrélation sans fusion (D). Avec la fusion les scores ne sont plus comparables d'une requête à l'autre.

[18]L'effet de «chute du score» à droite du graphique est lié à un effet de bordure de la moyenne sur l'ensemble d'échantillonnage des scores. L'effet d'augmentation au voisinage

Dans la troisième figure (V.6 page 155), nous avons fait la moyenne des scores des documents sélectionnés pour chaque question, et nous avons trié par ordre croissant le résultat. Nous pouvons observer des grandes différences de moyenne d'une question à une autre. Notamment à gauche du graphique nous observons un petit groupe de questions dont la moyenne est nulle (pas de document retourné). Nous observons aussi que pour les 80% des questions dont les moyennes des scores des documents sont médianes, les moyennes varient d'un facteur au moins 4. Nous observons aussi l'existence d'un ensemble de quelques questions qui ont des moyennes de score de document beaucoup plus élevées que l'écart-type. De ces deux observations, nous déduisons qu'il sera difficile de reconstruire une comparaison des scores des documents entre deux questions.

V.3.4.4 Analyse des résultats

Les gains en valeur absolue sont faibles. L'explication principale vient de la nature du corpus. Les questions du corpus ClefQA07-FR-EN (tableau III.2 page 86), est celui qui a notre connaissance comporte le plus de questions enchaînées. Cependant, toutes les questions ne rentrent pas dans ce cadre. Plus de la moitié des questions n'ont pas de dépendance. L'écart absolu entre les différentes stratégies en est donc réduit.

Introduire les termes pertinents en fonction des dépendances est un atout non négligeable. En revanche, il est plus difficile d'observer un gain modifiant les pondérations par défaut de ces termes. Si nous réduisons notre étude à l'ensemble des groupes de questions ayant une structure de dépendance non triviale, les résultats sont meilleurs, mais établis sur moins de données.

Nous avons procédé à une comparaison des résultats des stratégies C et G (ancienne et nouvelle avec fusion pour $n = 100$) sur l'unique base de la présence d'un document portant une réponse à une question. Observons uniquement l'ensemble des questions où une des deux stratégies a permis de trouver un document-réponse. Les différences ne sont pas nombreuses, mais nous constatons que la stratégie G réussit à trouver les documents-réponses des questions [178, 29, 12, 8] du corpus ClefQA07-FR-EN , là où la stratégie C réussit a trouver les documents réponses des questions [186, 115][19].

En comparant les numéros de questions des nouveaux documents-réponses obtenus, nous constatons qu'il ne s'agit, dans les deux cas, que de questions ayant des dépendances vers une question précédente.

du rang 1 est lié aux documents qui ne sont parfois qu'un petit nombre à être constitués d'une entité nommée dans un champ lexical particulier.

[19]Soit un différentiel de 2.

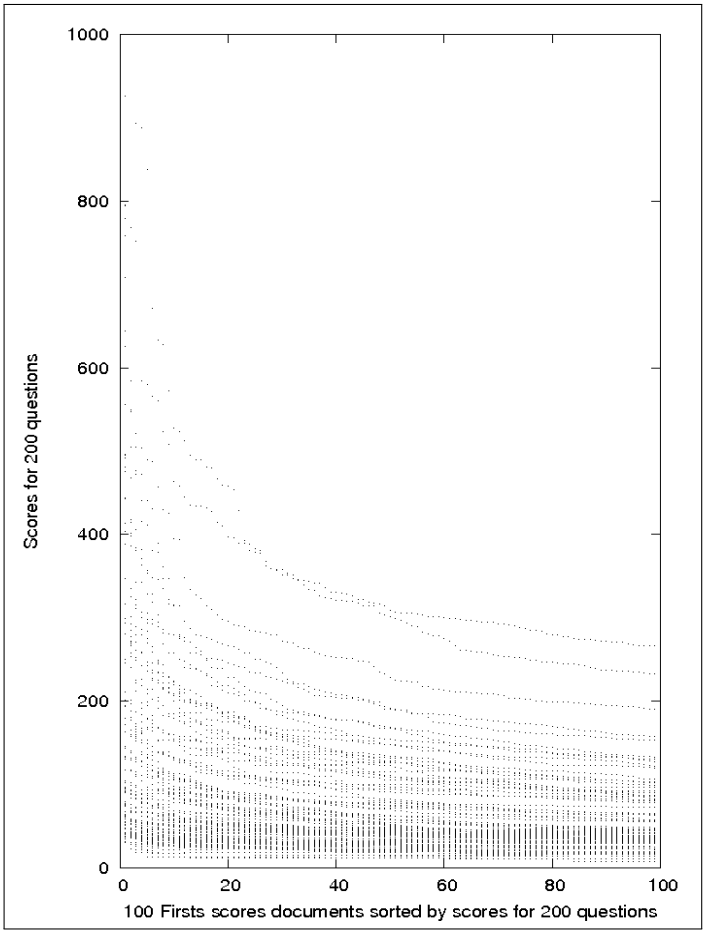

FIG. V.4 – Scores des 100 premiers documents sur 200 questions.

Nous réalisons la même comparaison entre les stratégies C et D. Nous observons que la stratégie C permet de trouver les documents-réponses des questions [196, 165, 154, 115] que la stratégie D ne trouve pas. De même la stratégie D permet de trouver les documents-réponses des questions [178, 177, 29, 12, 8, 123]. Or, dans ces questions, seules les questions numéros [178, 177, 29, 12, 8] et [196, 165, 154, 115] sont des questions ayant une dépendance, soit une question de plus dans le cas de la stratégie D.

Nous pouvons en déduire que par rapport à la pondération de référence (stratégie C), les nouvelles stratégies (D et G) permettent de gagner surtout des documents-réponses correspondant à des questions liées dépendantes d'une autre[20].

[20]Dans la limite où nous disposons de très peu de données.

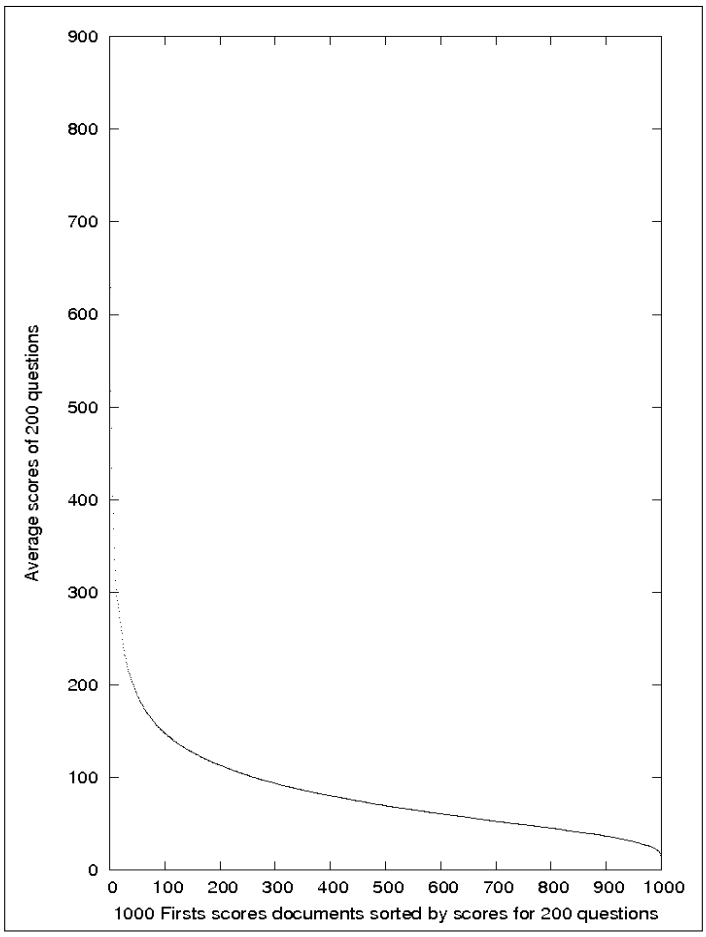

FIG. V.5 – Graphique donnant les scores moyens des 1000 premiers documents triés par leurs scores pour 200 questions.

Nous avions fait l'hypothèse que dans un cadre de véritable dialogue, les questions et les liens entre questions seraient moins biaisés que les groupes de questions enchaînées vus ci-dessus. Suite à cette expérimentation nous avons vérifié cette hypothèse dans l'annexe B (page 185) à l'aide de dialogues fabriqués.

V.3.5 Perspectives dans la suite des traitements

Nous pouvons maintenant nous interroger sur l'apport de ces traitements dans les performances du SQR et dans son intégration avec les outils l'entourant.

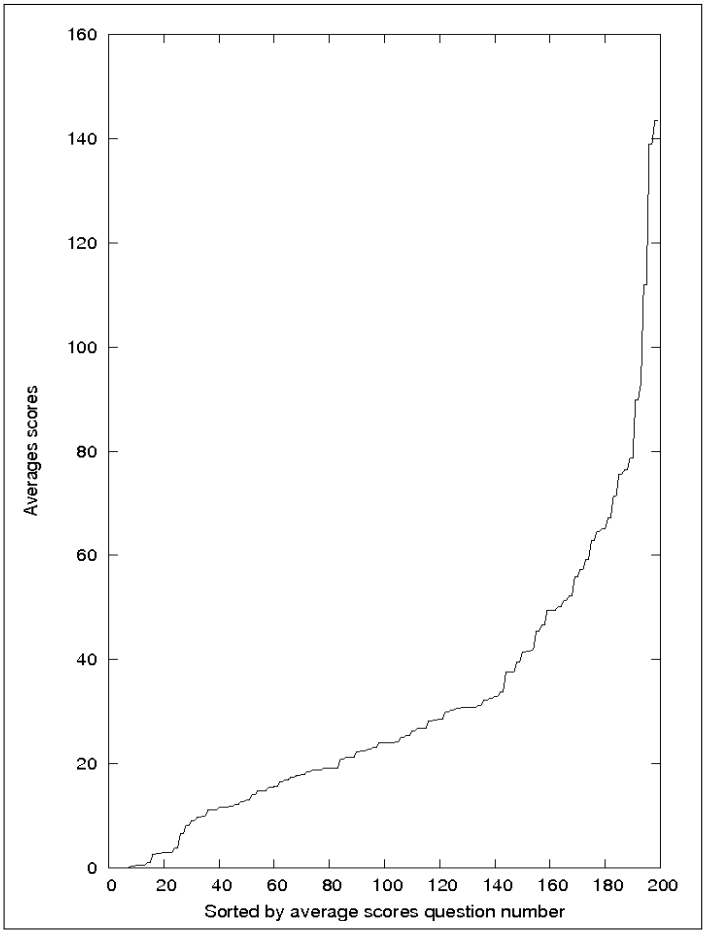

FIG. V.6 – Tri par score moyen de 1000 documents pour 200 questions.

V.3.5.1 La sélection des phrases

Les SQR recherchent la réponse exacte par des filtrages successifs. La sélection des documents prend 100 documents parmi plus de 3 millions. La sélection des phrases «réponses» est une étape qui réduit encore d'un facteur 5 la quantité de texte à analyser[21]

V.3.5.1.1 Modèle de recherche Après que les documents soient sélectionnés, ils sont découpés en phrases. Les phrases sont pondérée pour être à nouveau comparées en vue de sélectionner les meilleures pour la recherche de

[21]Dans le cas d'une évaluation avec la stratégie C, les documents sélectionnés représentaient 424Ko en moyenne (par question) et les phrases sélectionnées représentaient 115Ko en moyenne (par question).

la réponse. Musclef pose l'hypothèse que la réponse est contenue dans une unique phrase. De même que pour la recherche des documents, il faut décider de l'impact de la structure des dépendances pour la sélection des phrases. L'un des paramètres de sélection des phrases provient du score attribué par le moteur de recherche de document. Cette stratégie est similaire à celle déployée dans d'autres SQR. Comme nous ne disposons pas d'informations supplémentaires, l'impact de la structure des dépendances consistera uniquement en un impact indirect à travers le score attribué lors de la recherche de document.

De fait, la nouvelle stratégie est alors cohérente avec les stratégies de SQR avancées classiques, peu intrusive, ne modifie pas les interfaces et exploite, autant que possible, les nouvelles informations disponibles.

V.3.5.1.2 Évaluation au niveau des phrases Une fois l'étape de sélection des phrases réalisée par Musclef, nous avons appliqué sur un fichier qui contient toutes les phrases sélectionnées notre programme de décision de présence de la réponse à la question (section V.3.3.1 page 146). Pour chaque type de sélection de documents, nous avons réalisé ce test. Il en ressort que selon les tests 95% à 96% des bonnes réponses qui sont repérées juste après l'étape de sélection des documents sont encore présentes après l'étape de sélection des phrases. Cela est vrai que nous tenions compte ou non des dépendances, quelle que soit la méthode d'attribution de score. Nous en déduisons deux choses : premièrement la sélection des phrases n'a que peu d'impact sur les résultats finaux, secondement la méthode de sélection des phrases est complètement transparente à la méthode de sélection des documents puisque les écarts de résultats ne sont pas significatifs.

Comme les propriétés des mauvaises phrases sélectionnées ne sont pas évaluées, nous pouvons nous demander quel est l'impact sur la sélection des réponses courtes exactes ?

V.3.5.2 Sélection des « réponses exactes »

Nous n'avons pas réalisé d'expérience visant à améliorer la sélection des «réponses exactes» en tenant compte de notre nouvelle structure. La sélection des «réponses exactes» renvoie à des techniques très différentes de celles de sélection des documents/phrases. Même si le système réussit à sélectionner les bons documents puis les bonnes phrases, c'est un tout autre problème de choisir la bonne «réponse exacte» dans une phrase, parmi un ensemble de phrases aussi bien choisies soient-elles. Il ne nous a pas semblé judicieux de poursuivre l'expérimentation et l'évaluation sur la sélection des « réponses

Stratégie	dans les documents	5 premières réponses exactes	dont en première position
«brut (A)»	88	29	8
«référence (C)»	102	37	12
«fusion (G)»	104	37	12
«à la main»	100	39	11

TAB. V.2 – Résultat pour la sélection des « réponses exactes » à l'aide d'une métrique ancienne d'évaluation des réponses.

exactes » sans avoir une contribution originale bien formalisée utilisant mieux notre nouvelle structure.

Cependant dans le but d'obtenir une idée de l'impact du calcul des dépendances, nous avons réalisé quelques tests supplémentaires. Nous avons modifié le corpus de questions ClefQA07-FR-EN pour que toutes les questions puissent être comprises séparément les unes des autres (en utilisant les termes des questions liées)[22]. Nous avons utilisé Musclef sur ces questions modifiées et sur les questions non modifiées. Sur les questions non modifiées, nous avons utilisé les stratégies de calcul des scores C (stratégie de référence), la stratégie avec boosting (fusion G) et celle de base de Musclef qui ne prend pas en compte les dépendances (stratégie A). Tous les tests sont réalisés pour $n = 100$.

Les résultats peuvent alors s'observer dans le tableau V.2 (page 157). Les questions peuvent être ou bien «brutes» sans gestion des dépendances, ou bien «référence» avec la stratégie de pondération C, ou bien «fusion» avec la stratégie G, ou bien «à la main» avec des questions retravaillées pour inclure le contexte. Les quantités observées sont : soit le nombre de documents contenant la réponse après la sélection des dépendances, soit le nombre de réponses courtes exactes dans les 5 premiers résultats du SQR, soit la première réponse courte exacte proposée par le SQR.

Nous pouvons constater que les stratégies utilisant les dépendances sont très proches de la stratégie « à la main ». Pour autant qu'un si petit nombre de réponses soit significatif, nous pouvons conclure sur l'impact de la performance du calcul des dépendances. Avec un gain de 50% de bonnes réponses en première position pour un corpus comptant au moins de 50% de questions en rang 2+, notre méthode de calcul des dépendances semble adaptée au problème. D'un autre côté, le gain lié aux modifications sur le moteur de recherche est nul.

[22]disponibles en annexe

V.3.5.3 La boucle est bouclée

Il ne manque plus à notre modèle d'expérimentation que l'étape finale, ou la première étape selon le point de vue. La réponse choisie doit être associée à la question courante pour des résolutions futures. Les dépendances ne sont pas à mettre à jour, mais juste l'ensemble des termes liés à la question.

C'est la première « réponse courte exacte » qui est choisie pour être ajoutée à l'ensemble des termes de la réponse. Dans le domaine des SQR multilingues comme ici, il est notable que les performances finales de sélection des « réponses exactes » soient plus basses que dans les SQR mono-lingue. Dans la pratique nous constatons qu'il est contre-productif d'ajouter effectivement les termes de la « réponse exacte ». Si, pour le calcul des dépendances, il avait été possible de contourner le problème en utilisant le type de la réponse, au niveau du moteur de recherche la « réponse exacte » correcte à la dernière question est cruciale. Dans les 3 questions du corpus ClefQA07-FR-EN où il existe une dépendance vers une question précédente à cause du type de la réponse à la question précédente, il n'y a pas de document-réponse trouvé.

V.3.5.4 Conclusion de ces expérimentations

Notre étude des SQR à questions enchaînées dans un cadre d'interaction s'arrête ici. Nous avons montré un lien entre les systèmes d'analyse des dépendances entre questions et la sélection des documents. Nous avons vu les performances numériques respectives de ces deux apports et étudié leurs aspects théoriques. Ces études nous ont permis de voir les élements génériques de notre approche et nous pouvons en extrapoler de nombreux autres. Ceci nous assure des bonnes possibilitées d'intégration dans un cadre encore plus interactif, comme le dialogue homme-machine.

À court terme, la suite de ces expérimentations comporterait d'abord une étape de constitution de corpus ainsi qu'une réflexion sur cette constitution. En effet, un facteur limitant actuel est lié au sur-apprentissage sur nos corpus des campagnes Clef[23].

À plus long terme, nous pourrions alors envisager plusieurs axes, soit approfondir la sélection des termes avec une interaction entre la constitution de la *posting-list* et l'analyse de la question, soit approfondir l'analyse des possibilités de la corrélation des termes, ou bien étudier l'impact du contexte au niveau de la sélection des réponses courtes exactes.

[23] et Trec pour le développement

Chapitre VI

Conclusions et perspectives

Les travaux présentés dans ce mémoire se situent dans le contexte de la recherche d'information en domaine ouvert. Nous nous sommes placé plus particulièrement dans le cadre des systèmes de questions-réponses à questions enchaînées. Notre problématique principale a été la caractérisation des interactions entre questions et leurs conséquences dans un système de questions-réponses.

Nous ferons donc le point en deux étapes, d'abord nous verrons où nous nous sommes arrêté puis comment poursuivre.

Les contributions essentielles présentées dans ce mémoire de thèse sont les suivantes :

- **Formalisation en dépendance** : les groupes de questions liées par thématique partagent un nombre restreint de propriétés d'un couple de questions et réponses à l'autre. Ces propriétés peuvent être exprimées dans un formalisme probabiliste de dépendance d'informations d'une question pour la résolution d'une autre.
- **Analyse des groupes de questions** : nous avons réalisé un algorithme basé sur le modèle précédent afin d'analyser les groupes et d'en tirer toutes les dépendances. Nous avons alors défini une notion de contexte pour une question qui explicite son cadre d'interprétation. Nous avons mis en place une expérimentation visant à évaluer l'algorithme correspondant.
- **Recherche des documents** : nous avons proposé une analyse, une formalisation et un modèle pour la recherche des documents dans les questions enchaînées. Les modèles construits le sont en fonction des termes disponibles et de leurs stratégies de sélection. Nous avons pré-

senté quelques limites pour les stratégies existantes et la manière dont nous les contournons à l'aide des dépendances calculées.

Ce travail s'est appuyé dans un premier chapitre sur une étude des systèmes existants, leurs limites et les pièges à éviter vis-à-vis de notre objectif. Nous avons alors dans un second chapitre, étudié les questions enchaînées, leurs liens et comment les formaliser. Ceci nous a amené dans un troisième chapitre à analyser différentes options pour exploiter ce formalisme dans l'étape des systèmes de questions-réponses qui suit les analyses de questions. Nous avons alors proposé dans un quatrième chapitre le cadre d'évaluation qui semblait le plus judicieux. Nous allons approfondir ce bilan avant de dégager quelques perspectives.

VI.1 Bilan, où en sommes-nous ?

Il y a deux manières de voir nos résultats. Celle brute et chiffrée des méthodes de calcul des formalismes proposés, et celle qui ouvre le chemin aux améliorations. C'est de cette conjonction que nous pouvons analyser nos résultats et rappeler notre cadre de travail pour relativiser l'observable.

VI.1.1 Présentation des résultats

Nos résultats objectifs expérimentaux sont de deux sortes, ceux portant sur la *construction d'unités* pour représenter les liens entre les questions (dont nous appelons l'ensemble un contexte), et ceux portant sur l'*usage de ces unités* dans le cadre d'une recherche de documents dans un SQR interactif.

Construction de dépendances

Nous avons proposé un modèle de construction de liens entre questions basé sur la notion de dépendance.

Il existe une dépendance entre une question β et une question α quand certains termes de la question α ou de sa réponse sont indispensables pour obtenir la réponse à la question β. Sur la base de cette définition subjective d'une dépendance, nous avons constaté que cette formalisation est adaptée aux questions enchaînées et qu'elle offre une réponse aux problèmes des approches observés dans d'autres systèmes. Nous avons donné plusieurs indices solides montrant une compatibilité tant avec les modèles structuraux que dynamiques du dialogue.

Nous avons cherché à obtenir une mesure de la calculabilité de ce modèle. Sur le corpus ClefQA07-FR-EN avec un apprentissage sur les corpus FR_FR et FR_ES tous annotés à la main, nous obtenons un système multi-traits générique, optimisable à l'infini sans changement de structure et disposant d'une F-mesure de 0.8 avec une précision légèrement supérieure au rappel. Ayant discuté les erreurs possibles par rapport à leur fréquence et l'adaptabilité du modèle et des classes de méthodes de calcul, nous avons conclu que de meilleurs résultats passent par des analyses plus détaillées des questions et de meilleures connaissances sur les documents sur lesquels doivent être utilisées ces données, c'est à dire la manière dont seront formées les requêtes.

Pour montrer comment utiliser ces dépendances, nous nous sommes intéressé à la formulation des requêtes, et avons ouvert la problématique au moins aussi grande de l'utilisation des dépendances dans la recherche d'information.

Usage des dépendances

Quand nous désirons utiliser un modèle, de nombreuses possibilités sont ouvertes. Disposer des dépendances incite à les utiliser selon des méthodes jusque là inaccessibles pour obtenir de meilleurs résultats. Nous avons alors développé une série de systèmes utilisant (ou pas) les termes des questions des dépendances en tenant compte (ou pas) de leur hiérarchie induite.

Dans l'optique d'un usage plus général que celui de la campagne d'évaluation Clef, nous nous sommes intéressé au comportement des méthodes dans des cas limites et dans les cas classiques. La pondération des termes est apparue comme un élément central. La pondération est directement liée à la construction de la requête. N'ayant pas plus d'information à la construction de la requête qu'après le calcul des dépendances (utilisant l'analyse hors contexte des questions), la construction de la requête a été ramenée à sa plus simple expression dérivée des méthodes traditionnelles existantes dans le SQR utilisé pour l'expérimentation : Musclef. La pondération n'est pas un élément qui permet de faire apparaitre un terme inexistant, si un terme est oublié à l'analyse de la question il l'est pour toujours. La pondération sert uniquement à diminuer le bruit des documents non pertinents. C'est pour cela que les pondérations sont relatives d'un terme à l'autre.

Cherchant à exploiter les informations contenues dans les dépendances nous avons été en mesure de construire un modèle dynamique de la pondération des termes et des documents, basé sur la corrélation de présence de deux termes dans un document. Sur les mêmes corpus que ceux ayant servi à l'évaluation de détection des dépendances, nous avons évalué différentes méthodes de sélection des documents en utilisant les dépendances. L'évaluation de la méthode par corrélation par rapport à celles existantes a permis d'obtenir des résultats dignes de l'état de l'art. L'examen de la répartition des résultats a permis de déduire une heuristique permettant d'améliorer un petit peu les résultats en sélection des documents. Par manque de données pour l'expérimentation, par manque de temps pour les construire, nous n'avons pas pu nous intéresser à l'amélioration des techniques d'extraction de la réponse exacte via l'usage des dépendances.

Les résultats expérimentaux peuvent sembler décevants, mais le gain en performance brute n'était qu'un objectif secondaire.

VI.1.2 Analyse des résultats

Nous constatons que notre technique sans optimisation particulière permet d'obtenir des résultats ayant les mêmes performances que les systèmes de l'état de l'art. La couverture des résultats étant légèrement différente et

la répartition des bonnes réponses connues, il est alors possible de déployer une heuristique qui permet de dépasser les résultats initiaux. Notre gain est un petit gain en valeur absolue, mais supérieur à l'état de l'art.

Disponibilité des corpus

Les résultats sont essentiellement liés aux corpus de développements et d'évaluations disponibles. Les questions du corpus ClefQA07-FR-EN sont les plus complexes dont nous disposions. Or malgré cela, la structuration interne des groupes de questions ne concerne que quelques groupes. Les questions ayant été écrites pour et par des gens de la communauté TAL, elles sont pour la plupart bien formées et bien structurées. La résolution des anaphores, basée sur la grammaire et sur quelques propriétés de sémantique, est alors un trait favorisé. Nous pouvons supposer que des questions moins bien formées avec des groupes plus liés/complexes auraient très fortement baissé les résultats des stratégies simples vues au premier chapitre. Nous pourrions alors dire que nous avons réalisé un gain de robustesse du système.

Nous manquons de données dans nos corpus pour étudier plus d'optimisation. Nous n'avons pas pu établir la boucle : tests, analyse, modification. Nous ne disposons pas de suffisamment de corpus pour cela, la difficulté imposée par le domaine ouvert empêche notamment des analyses trop fines des questions et des corpus de documents. Nous ne voulions pas risquer la critique du surapprentissage, seule l'heuristique de la fusion des deux sources de résultats a été réalisée puisque les analyses montrent qu'elle est statiquement fondée.

Problème de définition et de technologie

Les résultats globaux ne sont pas très nombreux. C'est une thématique carrefour de plusieurs sujets, difficile à évaluer puisque la complexité a été réduite entre les campagnes ClefQA2007 et ClefQA2008 et n'a pas été renouvelée après TREC06 suite à l'uniformité des solutions déployées. Les difficultés sont donc aussi vastes que la définition de la tâche elle-même et la restriction aux domaines spécifiques de la jointure technologique.

Un tour d'horizon visant à constater expérimentalement et théoriquement ce qui peut faire l'objet de travaux de développement et d'optimisation ultérieurs nous semblait important. Où faire les gains ? De quelle nature peuvent-ils être ? Comment les évaluer par rapport à l'existant ? Ce sont des questions qui avant de trouver une réponse devaient être formalisées, une exploration mise en place et de premières expérimentation tentées. C'est le but du travail présenté dans ce mémoire qui ne peut pas être évalué numériquement.

Un cadre théorique

Les formalismes proposés ne sont directement liés à aucune théorie du dialogue ou des SQR et moteurs de recherche, mais nous les proposons comme socles pour la construction d'autres théories plus élaborées. L'analyse que nous avons menée montre bien qu'une théorie de structuration ne peut cependant pas être proposée indépendamment d'un cadre général d'utilisation de la structure et des méthodes de calcul associées (par exemple le dialogue pour la recherche d'information en domaine ouvert). Nos apports sont dans notre proposition de formalisation des calculs, via la formalisation en dépendances.

Tant au niveau du dialogue que de la recherche d'information, l'objectif n'est pas d'optimiser, mais d'ouvrir des voies. Les optimisations sont tellement nombreuses dans chaque cas qu'elles dépassent largement le cadre de ce mémoire. Ce sont pourtant des prérequis, mais les optimisations relèvent d'un travail d'ingénierie impliquant de nombreuses personnes, et ne faisant pas spécifiquement appel à des explorations supplémentaires.

VI.2 Perspectives : où allons-nous ?

Nous avons présenté de premiers travaux visant à formaliser un domaine et à explorer les difficultés qui s'y trouvent. Pour aller plus loin, il y a deux grands axes possibles. Celui de l'intégration plus fine des données dans les méthodes de recherche d'information. Et celui de l'intégration des questions enchaînées dans un cadre plus large d'interaction comme le dialogue ou une interface homme-machine plus élaborée.

Recherche d'information

L'objectif serait alors d'obtenir de meilleurs résultats à partir des données dont nous disposons, que ce soit par une meilleure organisation des calculs, ou une meilleure propagation des conséquences de l'existence d'un modèle d'enchaînement de questions. Comme nous l'avons vu, les dépendances ont surtout été réévaluées au moment de l'évolution de la requête (dans le flot du SQR). Nous n'avons pas tenu compte de l'impact de l'indexation des documents (pas spécifiquement pour les dépendances). Est-il possible de construire l'index différemment, de nettoyer les documents différemment afin de tenir compte dès l'indexation du type de calcul (un peu plus coûteux) que nous allons réaliser ? Est-il possible de réaliser les 2 stratégies de recherche qui composent notre heuristique finale en une seule construction de *posting-list* ? Est-il possible d'altérer une traduction d'un terme à l'aide des dépendances vers les autres questions ? Est-il possible de prendre en compte des méta-informations, décrites par l'utilisateur dans ces questions, pour altérer la recherche ou la requête ? Ce sont autant de pistes qui permettraient des gains chiffrés.

Expressivité accrue

L'expressivité des systèmes de dialogue suit *grosso modo* deux voies, celle de la complexification des énoncés qui peuvent être correctement analysés par un automate, et celle tout aussi empirique de la présentation d'interface homme-machine physique. Le dénominateur commun entre les deux approches est le paradigme du système de discussion via une ligne de commande utilisée notamment dans les *chatterbots*. Dans les deux cas, les développements se heurtent en particulier à la complexité d'obtention de corpus d'analyse des phénomènes à traiter.

Nos expérimentations ne sont pas tributaires du rendu des interfaces homme-machine car nous avons occulté la visualisation des résultats par l'utilisateur. C'est pourtant un domaine riche. Notamment l'interaction via

des widgets sur un écran, comme c'est le cas pour la plupart des SQR par écrit, n'est pas aussi simple à généraliser au modèle interactif. Une question importante est de savoir ce que le système fait des résultats déjà obtenus et auxquels l'utilisateur fait référence. Une autre question est de savoir si l'utilisateur à le droit de désigner via son pointeur les informations qu'il voit à l'écran afin de simplifier ses énoncés ? Les préciser ? Aller plus vite ? D'autres développements sont imaginables notamment à l'oral.

Vers d'autres intégrations

Si le modèle en dépendance peut sembler assez couvrant de l'ensemble des tâches liées à l'interaction autour des SQR qui peuvent se présenter, il faut garder à l'esprit que ce modèle ne ce fonde que sur les couches les plus simples de la sémantique. Les traits qui ont été déployés lors du calcul des dépendances utilisent comme données les plus complexes des invariants de la langue situés dans les «couches basses» d'analyse (genre, nombre, synonyme...). Imaginons qu'à l'avenir un utilisateur redéfinisse un terme pour l'employer dans une autre question, que se passe-t-il alors ? La réponse générale est que les dépendances déjà obtenues ainsi que les analyses sémantiques et pragmatiques qui peuvent être faites sont des données d'entrée pour le calcul des nouvelles dépendances.

Annexe A

Annexe des corpus de questions

Nous avons utilisé abondamment les questions des campagnes d'évaluation Clef de SQR multi-lingues. Voici les corpus des questions des années 2007 et 2008 pour des questions en français attendant des réponses en anglais à partir des corpus en anglais. Les corpus de documents sont constitués de journaux de 1994-1995 ainsi que de la wikipédia (anglaise) de novembre 2006.

Les questions Clef07-FR-EN

Les corpus sont dans le format, «numéro de question, numéro de groupe, texte de la question». Voici les données fournies :

```
001 00 À combien de personnes a-t-on demandé de quitter leur domicile durant les
inondations aux Pays-Bas, en hiver 1995 ?
002 00 Quelle proportion des Pays-Bas est sous le niveau de la mer ?
003 00 À quelle hauteur l'eau est-elle montée à Lobith durant les inondations ?
004 00 Qui était le premier ministre des Pays-Bas à ce moment-là ?
005 01 Quels étaient les noms complets de Flanders et Swann ?
006 01 En quelle année sont-ils devenus célèbres ?
007 01 Citer les langues dans lesquelles Flanders et Swann ont chanté.
008 01 En quelle année Flanders est-il mort ?
009 02 De quel instrument Swann jouait-il dans le duo "Flanders et Swann" ?
010 02 Dans quel pays Swann est-il né ?
011 02 Dans quelle ville Swann est-il mort ?
012 02 Comment s'appelait la femme de Swann ?
```

```
013 03 Pierre Beregovoy fut le premier ministre de quel pays ?
014 03 En quelle année Beregovoy a-t-il mis fin à ses jours ?
015 03 Quel était son parti politique ?
016 04 Combien de feux de brousse y a-t-il eu près de Sydney en janvier 1994 ?
017 04 Combien de maisons ont été détruites dans les banlieues de Sydney ?
018 04 Combien de prairies et de forêts ont été brûlées en Nouvelle-Galles du
Sud ?
019 04 Quelle était la vitesse du vent pendant les incendies ?
020 05 Quel était le nom de la barge qui a coulé à Porto Rico le 7 janvier 1994
?
021 05 Qu'a heurté la barge ?
022 05 Quelle quantité de mazout a été déversée dans l'eau ?
023 05 Combien de miles de plage compte Porto Rico ?
024 06 Qu'est-ce que le syndrome de la guerre du Golfe ?
025 06 Combien de personnes en ont été atteintes ?
026 07 À quel moment le toit du supermarché Casino s'est-il effondré à Nice ?
027 07 De quoi était fait le toit du supermarché ?
028 07 Combien de personnes ont été tuées dans l'accident ?
029 07 Dans quel pays se trouvait le supermarché ?
030 08 Quand l'église d'Angleterre a-t-elle ordonné la première femme prêtre ?
031 08 Combien de femmes ont été ordonnées ?
032 08 Qui était archevêque de Canterbury à ce moment-là ?
033 08 Dans quelle cathédrale les ordinations ont-elles eu lieu ?
034 09 Lister des marques d'huile d'olive.
035 09 Citer des pays dans lesquels l'huile d'olive est fabriquée.
036 09 Quelle quantité d'huile d'olive a été importée en Amérique en 1994 ?
037 09 Combien de variétés d'olive utilise-t-on pour fabriquer l'huile d'olive ?
038 10 A quel groupe appartient Skoda ?
039 10 Quel type de Skoda avait un moteur arrière ?
040 10 Dans quel pays fabrique-t-on les Skodas ?
041 10 Durant quel siècle fut fondé Skoda ?
042 11 Quel est le métier d'Alister McRae ?
043 11 D'où vient-il ?
044 11 Quel âge avait-il en 1995 ?
045 11 Dans quel type de voiture Alister McRae disputa-t-il le Rallye des 1000
lacs en Finlande en 1995 ?
046 12 Quel est le métier de Billy Connolly ?
047 12 Quelle est sa nationalité ?
048 12 Quel film de la BBC produit par Connolly a remporté le prix du meilleur
téléfilm à la cérémonie des BAFTA écossais en 1995 ?
049 13 Quel est le métier de Kiri Te Kanawa ?
050 13 Pour quelle maison de disques enregistra-t-elle la Bohème en 1994 ?
051 14 Quel est la profession de Michael Barrymore ?
052 14 Citer le nom d'un show télévisé qu'a animé Michael Barrymore.
053 14 Combien de temps son mariage a-t-il duré ?
054 14 Quel était le nom de sa femme ?
055 15 Quel est le métier de John Barbirolli ?
056 15 Qui était le soliste et violoncelle dans le Concerto pour violoncelle
d'Elgar qu'il enregistra en 1965 ?
```

057 15 Citer le nom d'un orchestre de Manchester que Barbirolli conduisit.
058 15 Citer le nom d'un orchestre américain que Barbirolli conduisit.
059 16 Dans quel pays du Royaume-Uni se trouve la ville de Perth ?
060 16 Quelle route est connue comme étant la route des "motor mile" de Perth ?
061 16 Nommer des employeurs à Perth.
062 16 Citer le nom d'un fleuve qui traverse Perth.
063 17 Combien de sites compte le système de parcs nationaux en Amérique ?
064 17 Combien y a-t-il d'hectares de parcs nationaux ?
065 17 Quel était le premier parc ?
066 17 Citer le nom des deux parcs établis le 31 octobre 1994.
067 18 Quel sorte d'animal Victor Bernal essaya-t-il d'acheter le 25 janvier 1993 ?
068 18 Sur quel type d'avion était l'animal ?
069 18 Qui prétendait être l'animal ?
070 18 A combien a été condamné Bernal ?
071 19 Quel océan Steve Fosset a-t-il traversé en ballon en février 1995 ?
072 19 Quel âge avait-il à ce moment-là ?
073 19 Quel était son métier ?
074 19 Quelle distance Richard Branson a-t-il parcouru en ballon en 1991 ?
075 20 Dans quel aéroport Carol Ann Timmel a-t-elle perdu son chat ?
076 20 Comment s'appelait le chat ?
077 20 Dans quel type d'avion le chat s'était-il perdu ?
078 20 Combien de temps l'avion a-t-il été retardé pendant que la recherche continuait ?
079 21 Qui a nationalisé le système ferroviaire de l'Argentine en 1947 ?
080 21 Quelle proportion de la voie fut considérée dangereuse en 1989 ?
081 21 Qui était le président de la commission nationale des chemins de fer en Argentine en 1994 ?
082 21 Combien de personnes ont travaillé pour les chemins de fer argentins en 1994 ?
083 22 Où se trouve le musée de l'Ermitage ?
084 22 Qui était le directeur du musée en 1994 ?
085 22 Dans quel palais le musée est-il logé ?
086 22 Combien de chambres y a-t-il dans ce palais ?
087 23 Qu'est-ce qu'un organophosphate ?
088 23 Citer le nom de pesticides contenant des organophosphates.
089 24 Combien de places y a-t-il dans la voiture électrique Impact de General Motors ?
090 24 Quelle est approximativement l'autonomie de la voiture en miles ?
091 24 Combien de temps faut-il pour accélérer de 0 à 60 miles par heure ?
092 24 Quelle est sa vitesse de pointe ?
093 25 Quels étaient les noms des deux bâteaux qui se sont heurtés dans le détroit de Bosporus en mars 1994 ?
094 25 Combien de personnes sont mortes dans la collision ?
095 25 Quel drapeau a été hissé par les deux bâteaux ?
096 25 Quelle est la longueur du détroit de Bosporus ?
097 26 Où se trouve la cathédrale Sainte-Sophie en Russie ?
098 26 Qui était son archiprêtre en 1995 ?
099 26 Quel Écossais a construit la cathédrale ?

```
100 26 Quelle impératrice russe l'a accrédité pour la construire ?
101 27 Que sont la salsa, le mambo et la samba ?
102 28 Qu'était "Brilliant Invader '95" ?
103 28 Quels pays y ont participé en plus du Royaume-Uni ?
104 28 Entre quels moments de la journée l'évènement a-t-il eu lieu ?
105 28 Nommer un type d'avion qui y participa.
106 29 Qui est Thom Rotella ?
107 29 D'où vient-il ?
108 29 Nommer un type de guitare joué par Rotella.
109 29 Nommer un club dans lequel il a joué.
110 30 Sur quelle station de radio Nicky Orellana est-il commentateur de
football ?
111 30 Où est-elle basée ?
112 31 Qu'est-ce que Kia ?
113 31 Quel Américain a été nommé à son conseil d'administration en septembre
1994 ?
114 31 Nommer un modèle de Kia.
115 31 Combien de voitures Kia construit-il par an ?
116 32 Qu'est-ce que Zanussi ?
117 32 Où est-il basé ?
118 33 Qui est Will Carling ?
119 34 Qui est Lester Piggott ?
120 34 Combien de vainqueurs classiques britanniques a-t-il monté ?
121 34 Combien de temps passa-t-il en prison pour infraction fiscale ?
122 35 Qui est Jonathan Edwards ?
123 36 Qui sont les "All Blacks" ?
124 36 Qui est leur capitaine ?
125 37 Qu'est-ce que l'arnica ?
126 38 Quelle entreprise fabrique du sirop de cassis ?
127 39 Qu'est-ce que Glaxo ?
128 40 Qu'est-ce que le Zantac ?
129 41 Qu'est-ce que Flexoset ?
130 42 Qu'est-ce que Lakeland Plastics ?
131 43 Donner des informations sur le yacht "Geronimo".
132 44 Qui est Keith Musto ?
133 45 Qu'est-ce qu'HSBC ?
134 46 Qui est Sir Denys Henderson ?
135 47 Qui est Nguyen Chi Thien ?
136 47 Quand a-t-il disparu ?
137 48 Qu'est-ce que l'Eurofighter ?
138 49 Qui a écrit la chanson "Dancing Queen" ?
139 49 Combien de personnes y avait-il dans le groupe ?
140 50 Qu'est-ce qu'un polygraphe ?
141 50 Quand fut-il inventé ?
142 50 Qui l'inventa ?
143 50 À quelle université l'inventeur étudiait-il à ce moment-là ?
144 51 Pour quelle université Aldrich H. Ames travaillait-il en Amérique ?
145 51 À quel pays Ames a-t-il vendu des informations ?
146 51 Quel était le nom de sa femme ?
```

147 51 Quand fut-il arrêté ?
148 52 Quand Yitzhak Rabin est-il né ?
149 52 Quand est-il mort ?
150 52 Quelle était sa profession quand il est mort ?
151 52 Combien de personnes se sont rendu à Jérusalem pour rendre hommage à Rabin quand il est mort ?
152 53 Qu'est ce que les "Frosties" ?
153 53 Quelle entreprise les fabrique ?
154 53 Quand ont-ils été lancés ?
155 53 Quelle est la mascotte des Frosties ?
156 54 Qu'est-ce que Sky Europe ?
157 54 Quand a-t-elle été établi ?
158 54 Combien de personnes emploie-t-elle ?
159 54 Combien d'itinéraires assure-t-elle ?
160 55 Quel président américain a dirigé les accords de Camp David ?
161 55 En quelle année les négociations ont-elles eu lieu ?
162 55 Combien d'accords ont été signés ?
163 55 A quelle date le "traité de paix israélo-égyptien" fut-il signé par la suite ?
164 56 Qu'est-ce que "l'effet de serre" ?
165 56 En quelle année l'effet de serre fut-il découvert ?
166 56 Qui l'a découvert ?
167 56 Citer le nom d'un gaz qui contribue au réchauffement climatique.
168 57 Qui était Stu Sutcliffe ?
169 57 Quand a-t-il rejoint les Beatles ?
170 57 De quel instrument jouait-il ?
171 57 Qu'a-t-il vendu au businessman John Moores pour acheter une guitare ?
172 58 Qu'est-ce qu'un "iPod" ?
173 58 Qui en fabrique ?
174 58 Quand ont-ils été lancés ?
175 58 Qui a fabriqué le disque dur pour la première génération iPod ?
176 59 Qu'est-ce que la "grippe bleue" ?
177 59 Citer le nom d'une profession pour laquelle les membres pourraient prétendre souffrir de la "grippe bleue".
178 60 Dans quel pays Edouard Balladur est-il né ?
179 60 Quand est-il devenu président de la France ?
180 60 Quel était son parti politique ?
181 60 De quelle entreprise était-il président entre 1968 et 1980 ?
182 61 Quel est le plus célèbre encas au Royaume-Uni ?
183 61 Combien de personnes au Royaume-Uni mangent des encas de la marque "Walkers" tous les jours ?
184 61 Quel est la plus célèbre saveur de chips au Royaume-Uni ?
185 61 Quelle entreprise au Royaume-Uni fabrique des chips faites maison ?
186 62 Citer le nom d'un bois utilisé pour construire des bâteaux.
187 62 Citer le nom d'un bois utilisé pour fabriquer le contreplaqué.
188 62 Citer le nom d'un bois utilisé pour fabriquer les violons.
189 62 Citer le nom d'un bois utilisé pour fabriquer du papier.
190 63 Citer le nom d'un aliment contenu dans le régime alimentaire de base d'Asie du sud-est.

```
191 63 Citer le nom d'un aliment contenu dans le régime alimentaire de base
d'Europe.
192 64 Nommer un type de plastique utilisé pour fabriquer le "plastique renforcé
de fibres de carbone" ?
193 64 Citer le nom d'un pont qui a été renforcé avec de la fibre de carbone.
194 64 Nommer un bien de consommation qui peut-être fait de fibres de carbone.
195 65 Quel régime de retraite affirme être "le deuxième régime de retraite le
plus grand en terme de fonds au Royaume-Uni" ?
196 65 À quelle compagnie acheta-t-il le centre commercial Forestside de Belfast
pour 50 millions de livres sterling en 1998 ?
197 65 Quelle organisation administra les "régimes de pension du NHS" ?
198 66 Combien de personnes ont remporté jusqu'ici la "médaille olympique Nobre
Guedes" ?
199 66 Qui l'a remporté en 1951 pour la voile ?
200 66 Qui l'a remporté en 2005 pour le judo ?
```

Les mêmes questions mais avec une transformation les rendant possiblement indépendantes les unes des autres.

```
À combien de personnes a-t-on demandé de quitter leur domicile durant les
inondations aux Pays Bas, en hiver 1995 ?
Quelle proportion des Pays-Bas est sous le niveau de la mer ?
À quelle hauteur l'eau est-elle montée à Lobith durant les inondations en hiver
1995 ?
Qui était le premier ministre des Pays-Bas en hiver 1995 ?
Quels étaient les noms complets de Flanders et Swann ?
En quelle année Flanders et Swann sont-ils devenus célèbres ?
Citer les langues dans lesquelles Flanders et Swann ont chanté.
En quelle année Flanders est-il mort ?
De quel instrument Swann jouait-il dans le duo "Flanders et Swann" ?
Dans quel pays Swann est-il né ?
Dans quelle ville Swann est-il mort ?
Comment s'appelait la femme de Swann ?
Pierre Beregovoy fut le premier ministre de quel pays ?
En quelle année Beregovoy a-t-il mis fin à ses jours ?
Quel était son parti politique ?
Combien de feux de brousse y a-t-il eu près de Sydney en janvier 1994 ?
Combien de maisons ont été détruites dans les banlieues de Sydney en janvier
1994?
Combien de prairies et de forêts ont été brûlées en Nouvelle-Galles du Sud à
Sydney en janvier 1994 ?
Quelle était la vitesse du vent pendant les incendies à Sydney en janvier 1994?
Quel était le nom de la barge qui a coulé à Porto Rico le 7 janvier 1994 ?
Qu'a heurté la barge à Porto Rico le 7 janvier 1994 ?
Quelle quantité de mazout a été déversée par la barge dans l'eau à Porto Rico le
7 janvier 1994 ?
Combien de miles de plage compte Porto Rico ?
Qu'est-ce que le syndrome de la guerre du Golfe ?
Combien de personnes ont été atteintes du syndrome de la guerre du Golfe ?
```

À quel moment le toit du supermarché Casino s'est-il effondré à Nice ?
De quoi était fait le toit du supermarché ?
Combien de personnes ont été tuées dans l'accident du supermarché Casino effondré à Nice ?
Dans quel pays se trouvait le supermarché Casino effondré à Nice?
Quand l'église d'Angleterre a-t-elle ordonné la première femme prêtre ?
Combien de femmes ont été ordonnées ?
Qui était archevêque de Canterbury à ce moment-là ?
Dans quelle cathédrale les ordinations ont-elles eu lieu ?
Lister des marques d'huile d'olive.
Citer des pays dans lesquels l'huile d'olive est fabriquée.
Quelle quantité d'huile d'olive a été importée en Amérique en 1994 ?
Combien de variétés d'olive utilise-t-on pour fabriquer l'huile d'olive ?
A quel groupe appartient Skoda ?
Quel type de Skoda avait un moteur arrière ?
Dans quel pays fabrique-t-on les Skodas ?
Durant quel siècle fut fondé Skoda ?
Quel est le métier d'Alister McRae ?
D'où vient le métier d'Alister McRae ?
Quel âge Alister McRae avait-il en 1995 ?
Dans quel type de voiture Alister McRae disputa-t-il le Rallye des 1000 lacs en Finlande en 1995 ?
Quel est le métier de Billy Connolly ?
Quelle est la nationalité de Billy Connolly ?
Quel film de la BBC produit par Connolly a remporté le prix du meilleur téléfilm à la cérémonie des BAFTA écossais en 1995 ?
Quel est le métier de Kiri Te Kanawa ?
Pour quelle maison de disques Kiri Te Kanawa enregistra-t-elle la Bohème en 1994 ?
Quel est la profession de Michael Barrymore ?
Citer le nom d'un show télévisé qu'a animé Michael Barrymore.
Combien de temps le mariage de Michael Barrymore a-t-il duré ?
Quel était le nom de la femme de Michael Barrymore ?
Quel est le métier de John Barbirolli ?
Qui était le soliste et violoncelle dans le Concerto pour violoncelle d'Elgar que John Barbirolli enregistra en 1965 ?
Citer le nom d'un orchestre de Manchester que Barbirolli conduisit.
Citer le nom d'un orchestre américain que Barbirolli conduisit.
Dans quel pays du Royaume-Uni se trouve la ville de Perth ?
Quelle route est connue comme étant la route des "motor mile" de Perth ?
Nommer des employeurs à Perth.
Citer le nom d'un fleuve qui traverse Perth.
Combien de sites compte le système de parcs nationaux en Amérique ?
Combien y a-t-il d'hectares de parcs nationaux en Amérique ?
Quel était le premier parc national en Amérique ?
Citer le nom des deux parcs nationaux en Amérique établis le 31 octobre 1994.
Quel sorte d'animal Victor Bernal essaya-t-il d'acheter le 25 janvier 1993 ?
Sur quel type d'avion était l'animal de Victor Bernal le 25 janvier 1993 ?
Qui prétendait être l'animal de Victor Bernal le 25 janvier 1993 ?

A combien a été condamné Victor Bernal ?
Quel océan Steve Fosset a-t-il traversé en ballon en février 1995 ?
Quel âge Steve Fosset avait-il en février 1995 ?
Quel était le métier de Steve Fosset ?
Quelle distance Richard Branson a-t-il parcouru en ballon en 1991 ?
Dans quel aéroport Carol Ann Timmel a-t-elle perdu son chat ?
Comment s'appelait le chat de Carol Ann Timmel ?
Dans quel type d'avion le chat de Carol Ann Timmel s'était-il perdu ?
Combien de temps l'avion a-t-il été retardé pendant que la recherche du chat de Carol Ann Timmel continuait ?
Qui a nationalisé le système ferroviaire de l'Argentine en 1947 ?
Quelle proportion de la voie du système ferroviaire fut considérée dangereuse en 1989 en Argentine?
Qui était le président de la commission nationale des chemins de fer en Argentine en 1994 ?
Combien de personnes ont travaillé pour les chemins de fer argentins en 1994 ?
Où se trouve le musée de l'Ermitage ?
Qui était le directeur du musée de l'Ermitage en 1994 ?
Dans quel palais le musée de l'Ermitage est-il logé ?
Combien de chambres y a-t-il dans ce palais ?
Qu'est-ce qu'un organophosphate ?
Citer le nom de pesticides contenant des organophosphates.
Combien de places y a-t-il dans la voiture électrique Impact de General Motors ?
Quelle est approximativement l'autonomie de la voiture Impact de General Motors en miles ?
Combien de temps faut-il à la voiture électrique Impact de General Motors pour accélérer de 0 à 60 miles par heure ?
Quelle est la vitesse de pointe de la voiture électrique Impact de General Motors?
Quels étaient les noms des deux bâteaux qui se sont heurtés dans le détroit de Bosporus en mars 1994 ?
Combien de personnes sont mortes dans la collision des deux bâteaux dans le détroit de Bosporus en mars 1994 ?
Quel drapeau a été hissé par les deux bâteaux dans le détroit de Bosporus en mars 1994 ?
Quelle est la longueur du détroit de Bosporus ?
Où se trouve la cathédrale Sainte-Sophie en Russie ?
Qui était son archiprêtre en 1995 de la cathédrale Sainte-Sophie en Russie?
Quel Écossais a construit la cathédrale Sainte-Sophie en Russie ?
Quelle impératrice russe l'a accrédité pour construire la cathédrale Sainte-Sophie en Russie ?
Que sont la salsa, le mambo et la samba ?
Qu'était "Brilliant Invader '95" ?
Quels pays ont participé à "Brilliant Invader '95"en plus du Royaume-Uni ?
Entre quels moments de la journée l'évènement "Brilliant Invader '95" a-t-il eu lieu ?
Nommer un type d'avion qui participa à "Brilliant Invader '95".
Qui est Thom Rotella ?
D'où Thom Rotella vient-il ?

Nommer un type de guitare joué par Thom Rotella.
Nommer un club dans lequel Thom Rotella a joué.
Sur quelle station de radio Nicky Orellana est-il commentateur de football ?
Où est-elle basée ?
Qu'est-ce que Kia ?
Quel Américain a été nommé au conseil d'administration de Kia en septembre 1994 ?
Nommer un modèle de Kia.
Combien de voitures Kia construit-il par an ?
Qu'est-ce que Zanussi ?
Où Zanussi est-il basé ?
Qui est Will Carling ?
Qui est Lester Piggott ?
Combien de vainqueurs classiques britanniques Lester Piggott a-t-il monté ?
Combien de temps Lester Piggott passa-t-il en prison pour infraction fiscale ?
Qui est Jonathan Edwards ?
Qui sont les "All Blacks" ?
Qui est le capitaine des "All Blacks" ?
Qu'est-ce que l'arnica ?
Quelle entreprise fabrique du sirop de cassis ?
Qu'est-ce que Glaxo ?
Qu'est-ce que le Zantac ?
Qu'est-ce que Flexoset ?
Qu'est-ce que Lakeland Plastics ?
Donner des informations sur le yacht "Geronimo".
Qui est Keith Musto ?
Qu'est-ce qu'HSBC ?
Qui est Sir Denys Henderson ?
Qui est Nguyen Chi Thien ?
Quand Nguyen Chi Thien a-t-il disparu ?
Qu'est-ce que l'Eurofighter ?
Qui a écrit la chanson "Dancing Queen" ?
Combien de personnes y avait-il dans le groupe "Dancing Queen" ?
Qu'est-ce qu'un polygraphe ?
Quand le polygraphe fut-il inventé ?
Qui l'inventa ?
À quelle université l'inventeur étudiait-il à ce moment-là ?
Pour quelle université Aldrich H. Ames travaillait-il en Amérique ?
À quel pays Ames a-t-il vendu des informations ?
Quel était le nom de la femme de Aldrich H. Ames?
Quand Aldrich H. Ames fut-il arrêté ?
Quand Yitzhak Rabin est-il né ?
Quand Yitzhak Rabin est-il mort ?
Quelle était la profession de Yitzhak Rabin quand il est mort ?
Combien de personnes se sont rendu à Jérusalem pour rendre hommage à Yitzhak Rabin quand il est mort ?
Qu'est ce que les "Frosties" ?
Quelle entreprise fabrique les "Frosties" ?
Quand les "Frosties "ont-ils été lancés ?

Quelle est la mascotte des Frosties ?
Qu'est-ce que Sky Europe ?
Quand Sky Europe a-t-elle été établi ?
Combien de personnes Sky Europe emploie-t-elle ?
Combien d'itinéraires Sky Europe assure-t-elle ?
Quel président américain a dirigé les accords de Camp David ?
En quelle année les négociations des accords de Camp David ont-elles eu lieu ?
Combien d'accords de Camp David ont été signés ?
A quelle date le "traité de paix israélo-égyptien" fut-il signé par la suite des accords de Camp David ?
Qu'est-ce que "l'effet de serre" ?
En quelle année l'effet de serre fut-il découvert ?
Qui a découvert "l'effet de serre" ?
Citer le nom d'un gaz qui contribue au réchauffement climatique.
Qui était Stu Sutcliffe ?
Quand Stu Sutcliffe a-t-il rejoint les Beatles ?
De quel instrument Stu Sutcliffe jouait-il ?
Qu'est ce que Stu Sutcliffe a vendu au businessman John Moores pour acheter une guitare ?
Qu'est-ce qu'un "iPod" ?
Qui fabrique les "iPod" ?
Quand les "iPod" ont-ils été lancés ?
Qui a fabriqué le disque dur pour la première génération iPod ?
Qu'est-ce que la "grippe bleue" ?
Citer le nom d'une profession pour laquelle les membres pourraient prétendre souffrir de la "grippe bleue".
Dans quel pays Edouard Balladur est-il né ?
Quand Edouard Balladur est-il devenu président de la France ?
Quel était le parti politique de Edouard Balladur ?
De quelle entreprise Edouard Balladur était-il président entre 1968 et 1980 ?
Quel est le plus célèbre encas au Royaume-Uni ?
Combien de personnes au Royaume-Uni mangent des encas de la marque "Walkers" tous les jours ?
Quel est la plus célèbre saveur de chips au Royaume-Uni ?
Quelle entreprise au Royaume-Uni fabrique des chips faites maison ?
Citer le nom d'un bois utilisé pour construire des bâteaux.
Citer le nom d'un bois utilisé pour fabriquer le contreplaqué.
Citer le nom d'un bois utilisé pour fabriquer les violons.
Citer le nom d'un bois utilisé pour fabriquer du papier.
Citer le nom d'un aliment contenu dans le régime alimentaire de base d'Asie du sud-est.
Citer le nom d'un aliment contenu dans le régime alimentaire de base d'Europe.
Nommer un type de plastique utilisé pour fabriquer le "plastique renforcé de fibres de carbone" ?
Citer le nom d'un pont qui a été renforcé avec de la fibre de carbone.
Nommer un bien de consommation qui peut-être fait de fibres de carbone.
Quel régime de retraite affirme être "le deuxième régime de retraite le plus grand en terme de fonds au Royaume-Uni" ?
À quelle compagnie acheta-t-il le centre commercial Forestside de Belfast pour

```
50 millions de livres sterling en 1998 ?
Quelle organisation administra les "régimes de pension du NHS" ?
Combien de personnes ont remporté jusqu'ici la "médaille olympique Nobre
Guedes" ?
Qui a remporté la "médaille olympique Nobre Guedes" en 1951 pour la voile ?
Qui a remporté la "médaille olympique Nobre Guedes" en 2005 pour le judo ?
```

Voici les dépendances que nous pouvons en déduire :

```
[1-4] -> [[1,3],[1,4]]
[5-8] -> [[5,6],[5,8]]
[9-12] -> []
[13-15] -> [[13,15],[13,14]]
[16-19] -> [[16,17],[16,18],[16,19]]
[20-23] -> [[20,21],[20,22]]
[24-25] -> [[24,25]]
[26-29] -> [[26,27],[26,29],[26,28]]
[30-33] -> [[30,31],[30,32],[30,33]]
[34-37] -> []
[38-41] -> []
[42-44] -> [[42,43],[42,44]]
[46-48] -> [46,47]
[49-50] -> [[49,50]]
[51-54] -> [[51,53],[51,54]
[55-58] -> [[55,56],[55,57],[55,58]]
[59-62] -> []
[63-66] -> [[63,64],[63,65],[63,66]]
[67-70] -> [[67,68],[67,69],[67,70]]
[71-74] -> [[71,72],[71,73]]
[75-78] -> [[75,76],[75,77],[75,78]]
[79-82] -> [[79,80]]
[83-86] -> [[83,84],[83,85],[83,85,86]]
[87-88] -> []
[89-92] -> [[89,90],[89,91],[89,92]]
[93-96] -> [[93,94],[93,95]]
[97-100] -> [[97,98],[97,99],[97,99,100]]
[102-105] -> [[102,103],[102,104],[102,105]]
[106-109] -> [[106,107],[106,108],[106,109]]
[110-111] -> [[110,111]]
[112-115] -> [[112,113]]
[116-117] -> [[116,117]]
[119-121] -> [[119,120],[119,121]]
[123-124] -> [[123,124]]
[135-136] -> [[135,136]]
[138-139] -> [[138,139]]
[140-143] -> [[140,141],[140,142],[140,143]]
[144-147] -> [[144,145],[144,146],[144,147]]
[148-151] -> [[148,149],[148,150],[148,151]]
[152-155] -> [[152,153],[152,154]]
```

```
[156-159] -> [[156,157],[156,158],[156,159]]
[160-163] -> [[160,161],[160,162],[160,163]]
[164-167] -> [[164,166]]
[168-171] -> [[168,169],[168,170],[168,171]]
[172-175] -> [[172,173],[172,174]]
[176-177] -> []
[178-181] -> [[178,179],[178,180],[178,181]]
[182-185] -> []
[186-189] -> []
[190-191] -> []
[192-194] -> []
[195-197] -> [[195,196]]
[198-200] -> [[198,199],[198,200]]
```

Les questions Clef08-FR-EN

```
001 00 Quelle était la nationalité de Jacques Offenbach?
002 00 Dans quelle ville est-il né?
003 00 De quel instrument jouait-il?
004 00 Citez les pianistes avec lesquels il a joué.
005 01 Quelle était la profession de Charles Wakefield Cadman?
006 01 Quel orchestre a-t-il fondé?
007 01 Combien de chansons a-t-il écrites?
008 01 Citez les films dont il a écrit la musique.
009 02 Quel compositeur a écrit "pacific 231"?
010 02 Qu'est-ce qui y est imité?
011 03 Quel est le nom de l'oeuvre la plus connue de Jeremiah Clarke?
012 03 Comment est-il mort?
013 04 Quelle est la profession de Richard Clayderman?
014 04 Combien de chansons a-t-il enregistrées?
015 04 Combien de disques de la "Ballade pour Adeline" ont été vendus?
016 05 En quelle année Emerson Lake &amp;amp; Palmer s'est-il formé?
017 05 Quelle est l'abréviation du groupe?
018 05 Citez deux instruments joués par Emerson.
019 06 Qui a sorti l'album "Songs in A Minor"?
020 06 Combien de disques ont été vendus le premier jour?
021 07 Quel est le vrai nom de "Common"?
022 07 A quelle date son premier album est-il sorti?
023 08 Qui était Charlotte Mew?
024 09 Où Alexandre Dumas Père était-il enterré jusqu'au 30 novembre 2002?
025 09 Où a-t-il été réenterré après le 30 novembre 2002?
026 10 Donnez une citation d'Alfred Tennyson.
027 11 Dans quelle ville W. H. Auden a-t-il grandi?
028 11 Quel était son poste à l'Université d'Oxford de 1956 à 1961?
029 12 Qu'est-ce que l'hégélianisme?
030 13 En quelle année le film "Les Vestiges du jour" est-il sorti?
031 13 Qui joue le rôle de Mr Stevens dans le film?
032 13 Qui a écrit la musique du film?
```

033 13 Qui a écrit le roman dont le film est inspiré?
034 14 Citez les deux studios qui ont sorti le film "Les Cendres d'Angela".
035 14 Dans quelle ville irlandaise le film se déroule-t-il?
036 15 Citez une société de production qui a co-produit le film "Evelyn".
037 16 Quel film a été décrit comme le 35ème plus grand film britannique de tous les temps?
038 16 Qu'est-ce que les rockers conduisent dans le film?
039 17 A qui appartient la voix entendue dans le film "Coeur de dragon"?
040 18 Citez trois architectes ou cabinets d'architecte irlandais.
041 19 Citez un édifice situé à Sydney en Australie qui a été conçu par Norman Foster.
042 20 Quel architecte a conçu les nouvelles Chambres parlementaires britanniques après leur destruction dans un incendie en 1834?
043 21 Qu'est-ce que Poundbury?
044 21 Quel architecte en a dessiné les plans?
045 21 Qui possède le terrain?
046 21 Quel prince anglais est associé au projet?
047 22 Combien d'églises de Londres ont été dessinées par Christopher Wren?
048 23 Quelle place de Londres a été conçue par Inigo Jones à la demande du Comte de Bedford?
049 24 Au Japon, qu'est-ce que le "bungo"?
050 25 Citez les quatre premières langues officielles de la République des Deux Nations.
051 26 Qu'est-ce que l'astronomie?
052 27 Qu'est-ce que la géomorphologie?
053 28 Qui a conçu le premier logiciel d'annotation de génomes?
054 29 A quelle date Mathieu Orfila a-t-il écrit son "Traité des poisons"?
055 30 En phylogénie, quand la loi de récapitulation d'Ernst Haeckel a-t-elle été communément admise?
056 31 Qu'est-ce que le télougou?
057 31 Par combien de personnes est-il parlé?
058 32 Donnez un autre nom pour "groupes phylogénétiques".
059 33 Citez trois personnes qui ont analysé les jeux de hasard.
060 34 Quelles sont les dates de naissance et de décès de Joseph Fourier?
061 35 Quand Joseph Fourier est-il né?
062 36 Qui a créé le terme "ethnographie de la communication" dans les années 60?
063 37 Où Martin Heidegger a-t-il obtenu un poste de professeur de philosophie?
064 38 Où Jacques Derrida est-il né?
065 38 Quand a-t-il donné sa conférence intitulée "Structure, signe, jeu dans les sciences humaines" à l'université John Hopkins?
066 39 A quel âge Nietzsche a-t-il obtenu une chaire à l'université de Bâle?
067 39 Quel âge avait-il quand il a décidé d'écrire "Ecce Homo"?
068 40 Quel est le parti de l'homme politique irlandais Bertie Ahern?
069 41 Quand l'homme politique irlandais Willie O'Dea est-il né?
070 42 Où l'homme politique irlandais Willie O'Dea est-il né?
071 43 Quel est le parti politique de Tony Blair?
072 44 A quelle école William Hague a-t-il été formé?
073 45 Sur quelle proportion du vingtième siècle le parti conservateur

britannique a-t-il été au pouvoir?
074 46 A quelle date le parti politique irlandais des Démocrates progressistes a-t-il été fondé?
075 46 Combien de sièges a-t-il gagnés lors des élections générales de 1987?
076 47 Quel est le titre du premier livre écrit en espéranto?
077 47 Quand a-t-il été publié?
078 47 Dans quelle université l'espéranto est-il utilisé?
079 47 Combien existe-t-il de terminaisons verbales en espéranto?
080 48 Quelle organisation en France s'occupe de la langue française?
081 48 Combien de membres compte-t-elle?
082 49 A quelle date eut lieu le premier tournoi de tennis à Wimbledon?
083 50 Combien existe-t-il de courts de jeu de paume?
084 51 Comment s'appelle la crosse utilisée dans le sport irlandais "hurling"?
085 52 Quel animal est monté par les joueurs de polo?
086 53 Quel véhicule est utilisé en motocross?
087 54 Combien de championnats du monde de "Superside" Steve Webster a-t-il gagnés entre 1987 et 2004?
088 55 En aviron, qu'est-ce qu'un "single scull"?
089 56 Quelle est la distance de la course de haies masculine courte?
090 57 Quelle est la distance d'une course de haies longue?
091 58 Dans quels pays la voile sur glace est-elle pratiquée?
092 59 Quelle est la largeur d'un voilier sur glace hollandais?
093 60 Quelle est la distance parcourue lors d'un marathon?
094 61 Quelle est la longueur d'une course de hamsters?
095 62 Qu'est-ce que le "Footspeed"?
096 63 Quel objet est lancé lors d'une partie de pétanque?
097 64 Qui est Barry McGuigan?
098 65 Qui est Steve Redgrave?
099 66 Quel club sportif privé Nigel Mansell possède-t-il?
100 67 Qui est Murray Walker?
101 68 En quelle année James Hunt a-t-il gagné le championnat du monde de Formule 1?
102 69 En quelle année John Curry est-il devenu champion du monde de patinage artistique?
103 70 Qu'est-ce qu'une "Chevrolet Sprint"?
104 71 Qu'est-ce qu'une "Bolwell Nagari"?
105 72 Où le constructeur de voitures Morgan Motor Company est-il basé?
106 73 Combien de roues possède la voiture F-4 construite par Morgan Motor Company?
107 74 Combien de voitures "Panther Rio" ont été construites?
108 75 Quand la Volkswagen Polo Playa a-t-elle été construite?
109 76 Quel constructeur automobile britannique a construit la Chevette?
110 77 Qu'est-ce qu'une quetsche?
111 78 Qu'est-ce qu'un néflier du Japon?
112 79 Quand a-t-on découvert que les racines de la chicorée contenaient de l'inuline?
113 80 Dans le Suffolk, où la rivière Deben prend-elle sa source?
114 81 Quelle est la longueur du pont Elizabeth II?
115 82 Combien d'acier fut utilisé pour construire le pont du Forth?

116 83 Quelle est la capitale de la République de Namibie?
117 84 Quelle est la superficie de la France métropolitaine?
118 85 Quelle est la capitale de la Lettonie?
119 86 Où se trouve le siège du gouvernement fédéral de Malaisie?
120 87 Quel est le magasin le plus grand du Royaume-Uni?
121 88 Qui fit construire une horloge sur la devanture du magasin "Fortnum & Mason" en 1964?
122 88 Qu'est-ce qui fait la célébrité de ce magasin?
123 89 Quelle hauteur peut atteindre l'avocatier?
124 90 Citez une plante que l'on cultive pour son sucre.
125 91 Citez une maladie que l'ail peut aider à soigner.
126 92 Quelle entreprise fabrique la console de jeu appelée "Wii"?
127 92 Qui a conçu cette console?
128 92 Citez un type de média accepté par cette console.
129 93 Qu'est-ce que la "Xbox"?
130 93 A quelle manette le magazine "Game Informer" a-t-il décerné le titre de "Blunder of the year"?
131 94 Qu'est-ce qu'un "lecteur MP3"?
132 95 Qu'est-ce qu'une "Livery Company"?
133 96 Que sont les Salésiens de Don Bosco?
134 97 Quelle est la longueur de l'autoroute M56 en Angleterre?
135 98 Qu'est-ce que la CAMRA?
136 99 Quel type d'organisation est le RYA?
137 100 Qui était Bruce Lee?
138 101 Quel est le surnom de Ian Thorpe?
139 101 Combien pesait-il à la naissance?
140 102 Où est le siège de l'Association Internationale des Fédérations d'Athlétisme?
141 103 Que signifie CORGI?
142 103 De quel organisme de surveillance gouvernemental le CORGI dépend-il?
143 104 Qel groupe allemand a sorti un album disponible uniquement sur clé USB?
144 105 Quelle autoroute circule autour de Manchester en Angleterre?
145 106 Quelle est la population de Huddersfield?
146 107 A quelle date la ville romaine d'Exeter a-t-elle été fondée?
147 108 Qui ordonna la construction du Countess Wear sur le fleuve Exe au 13ème siècle?
148 109 Quel était le nom de code de la "Nintendo DS"?
149 109 Quelle est la version améliorée, parue en 2006?
150 110 Combien de personnes faut-il dans une voiture pour emprunter la voie réservée au covoiturage sur l'autoroute de Santa Monica?
151 111 Que portent les femmes au bal masqué de Damas de Caridad?
152 112 Combien mesure Matt Steffe?
153 113 Quel studio a produit le film "The Toxic Avenger"?
154 114 De quel instrument Betty Bryant joue-t-elle?
155 115 Quelle institution a organisé l'opération Pothole?
156 116 Combien coûte une entrée aux "Area Code Games" de Blair Field?
157 117 Qui a créé la série télévisée "Brady Bunch"?
158 118 De quoi Tom Gullikson a-t-il été le capitaine en septembre 1994?
159 119 Quelle est la largeur du magasin "House of Fabrics" à Simi Valley?

160 120 Pour quelle entreprise Neil Sweig travaille-t-il?
161 121 Combien mesure Jerod Ward?
162 122 Citez un poisson qui peut être pêché près des Îles Revillagigedo.
163 123 Où le "Brand XXIV" a-t-il eu lieu?
164 124 De combien la taxe sur la cigarette prélévée par l'état californien a-t-elle augmentée le 1er janvier 1994?
165 125 Qu'est-ce que l'aquavit?
166 126 Quelle est la profession d'Art Shell?
167 127 Qu'est-ce que la disneyisation?
168 128 Qui est Richard D. Farman?
169 129 Qui est James L. Armstrong?
170 130 Citez les films qui ont été présentés à la sixième édition du Festival International du Film de Palm Springs?
171 131 Citez un morceau de l'album "MoJazz Christmas" composé par le pianiste Eric Reed.
172 132 Pour quelle entreprise Dick Mason travaille-t-il?
173 133 Quelle est la profession de Jerry Hickman?
174 134 Quelle est la population de Starbuck, Minn.?
175 135 Combien de chansons sont jouées dans le spectacle musical "Closer than Ever"?
176 136 Qui a composé la musique des chansons du spectacle musical "Closer than Ever"?
177 137 Combien de parties sont disputées chaque année sur le terrain de golf "Collingtree Park" près de Northampton?
178 138 Combien y avait-il de joueurs lors de l'Open d'Andalousie disputé au club d'Islantilla en 1995?
179 139 Quand Bob Steward a-t-il été diplômé de l'Ecole des Beaux-Arts de Glasgow?
180 140 En quelle année le Brunswick City Hotel ouvrira-t-il à Glasgow?
181 141 En quelle année la société "In Video" de David McWhinnie a-t-elle été créée?
182 142 Quand William McIntosh Millar a-t-il été décoré de l'OBE?
183 143 Où David S. Forsyth a-t-il tenu une bibliothèque de lecture publique?
184 144 Où Muriel Herkes vit-elle?
185 145 Où le congrès du Parti travailliste écossais s'est-il déroulé en 1995?
186 146 Qui préside les courses de chevaux d'Uttoxeter et de Newcastle?
187 147 Quel était le poste occupé par Gerry Purnell sur l'HMS Indefatigable en 1945?
188 148 Qui est Time Fawcett?
189 149 Qu'est-ce que le NFUS?
190 150 Quel type d'organisation est Camelot?
191 151 Citez les livres écrits par Nicholas Ind.
192 152 De quelle boisson le nom signifie-t-il "eau de vie"?
193 153 Qui est le directeur des services généraux du groupe HOWDEN?
194 154 Quel genre d'animal est le "Letsbeonestaboutit"?
195 155 Qui était le secrétaire d'état pour l'Ecosse en 1995?
196 156 Qu'a conçu Jill Vandebrand?
197 157 Quand Bath Press Group a-t-il été fondé?
198 157 Combien de livres publie-t-il chaque année?

```
199 158 Quel opéra de Puccini fait intervenir la viole d'amour?
200 159 Combien de joueurs y a-t-il dans une rencontre de simple au tennis?
```

Ce corpus est moins intéressant car les groupes sont plus petits et mettent en jeu moins de mécanismes linguistiques que le Clef07.

```
[1-4] -> [[1,2],[1,3],[1,4]]
[5-8] -> [[5,6],[5,7],[5,8]]
[9-10] -> [[9,10]]
[11-12] -> [[11,12]]
[13-15] -> [[13,14]]
[16-18] -> [[16,17],[16,18]]
[19-20] -> [[19,20]]
[21-22] -> [[21,22]]
[24-25] -> [[24-25]]
[27-28] -> [[27,28]]
[30-33] -> [[30,31],[30,32],[30,33]]
[34-35] -> [[34,35]]
[37-38] -> [[37,38]]
[43-46] -> [[43,44],[43,45],[43,46]]
[56-57] -> [[56,57]]
[64-65] -> [[64,65]]
[66-67] -> [[66,67]]
[74-75] -> [[74,75]]
[76-79] -> [[76,77],[76,78],[76,79]]
[80-81] -> [[80,81]]
[121-122] -> []
[126-128] -> [[126,127],[126,128]]
[129-130] -> [[129,130]]
[138-139] -> [[138,139]]
[141-142] -> [[141,142]]
[148-149] -> [[148,149]]
[197-198] -> [[197,198]]
```

Annexe B

C'est sur des dialogues construits

Nous avons imaginé des cas d'utilisations en situations réelles de questions enchaînées. Une mise en forme de la réponse est ajoutée. Les thématiques sont choisis parmis celles des questions de la campagnes ClefQA07-FR-EN , mais elles sont modifiées pour être plus réalistes. Par réalistes nous voulons dire qu'elles semblent mieux correspondre à l'interaction dont diposerait un utilisateur avec ce genre de système. Une des caractéristiques des ces dialogues qui nous interresse particulièrement est que la totalité des questions en rang 2+ sont liées d'une manière ou d'une autre aux questions précédentes.

User	Qu'est-ce qu'un polygraphe ?
System	C'est un «lie detector».
User	Qui l'a inventé ?
System	C'est «William Moulton Marston» qui a inventé le polygraphe.
User	En quelle année ?
System	Le polygraphe a été inventé en «1913».
User	Quel etait son métier ?
System	«William Moulton Marson» était «Psychology student».
User	A quelle université étudiait-il ?
System	«William Moulton Marson» était à l'université «Harvard University».

User	Qui est Alister McRae ?
System	Alister McRae est «Scottish rally driver».
User	Avec quel voiture a-t-il déjà gagné ?
System	«Alister McRae» a gagné avec une «Nissan Sunny».
User	Quand est-il née ?
System	«Alister McRae» est née le «20 December 1970».
User	Pour quelles marques a-t-il travaillé ?
System	«Alister McRae» a travaillé pour «Nissan», «Volkswagen», «Subaru», «Hyundai», «Mitsubishi».
User	Quel course a-t-il gagné ?
System	«Alister McRae» a gagné «British Rally Championship» avec «Nissan Sunny».
User	En quelle année a-t-elle eu lieu ?
System	Le «British Rally Championship» gagné par «Alister McRae» avec «Nissan Sunny» était en l'année «1995».

Les réponses affichés par le système dans les exemples ci-dessus sont celles qui servent à l'évaluation des documents-réponses. Les questions de l'utilisateur sont évaluées via les mêmes procédures que celles vus au chapitre V (page 133). Comme nous avons vérifié que pour chaque question la réponse se trouve effectivement dans le corpus de recherche et que les thématiques des questions sont choisis parmis celles des questions de premier groupes pour lesquelles des résultats étaient obtenus, nous avons de bonne chances d'obtenir les documents-réponses pour toutes les questions. Avec la stratégie d'interrogation n'utilisant pas les corrélations de présence des mots pour les questions liées, le Mrr est de 0.152 et la moyenne de 23. La stratégie d'interrogation utilisant une fusion de celle traditionnele et celle avec corrélation obtient un Mrr de 0.23 et une moyenne de 18.5. Il n'y a pas de différence entre le Mrr(Ok) et le Mrr(All) puisque dans les 2 cas, pour l'intégralité des questions un document-réponse était présent dans les n-premiers résultats.

Bibliographie

[Abella & Gorin, 2008] Abella A. & Gorin L. A. (2008). Method for dialog management. US Patent 7403899 07/22/2008.

[Abney, 1996] Abney S. (1996). Partial parsing via finite-state cascades.

[Abney, 1997] Abney S. (1997). The scol manual - version 0.1b.

[Ait-Mokhtar *et al.*, 2002] Ait-Mokhtar S., Chanod J.-P. & roux C. (2002). Robustness beyond shallowness : incremental dependency parsing. *special issue of the NLE Journal.*

[Allen *et al.*, 1996] Allen J. F., Miller B. W., Ringger E. K. & Sikorski T. (1996). Robust understanding in a dialogue system. *ACL-96*, p.5̃.

[Allwood *et al.*, 2000] Allwood J., Hilland D., Selcon S., Taylorand R., Oviattand S., Murrayand A., de Ramand S. C., Buntand H., Sadekand D., Bilangeand E., Luzzatiand D., Vilnat A., Beunand R., Edwards J., Sinclairand D., Tatham M., Morton K., Lewisand E., Maybury M., Leeand J., Junquaand J., Cuxacand C., Edmondson W., Teil D., Bellikand Y., Gavignet F., Guyomard M., Siroux J., Boudreau G., McCannand C., Leeand J., Gaiffe B., Pierrel J.-M., Romaryand L., Taylor M., Waughand D., Dattaand A., Brooke M., Tomlinsonand M. & Benot. C. (2000). *The structure of multimodal dialogue II.* M. Martin Taylor and F. Neel and Don G. Bouwhuis.

[Amaral *et al.*, 2007] Amaral C., Cassan A., Figueira H., Martins A., Mendes A., Mendes P., Pinto C. & Vidal D. (2007). Priberam'a question answering system in qa@clef2007. *Priberam Informatica.*

[Benamara, 2004] Benamara F. (2004). *WEBCOOP : Un système de question réponse coopératif sur le web.* PhD thesis, Université de Toulouse Paul Sabatier.

[Bennacef *et al.*, 1994] Bennacef, S.K./Bonneau-Maynard, H./Gauvain, J.-L./Lamel & W L. (1994). A spoken language system for information retrieval. *ICSLP International Conference on Spoken Language Processing . Yokohama,.*

[Bobrow *et al.*, 1977] Bobrow D. G., Kaplan R. M., Kay M., Norman D. A., Thompson H. S. & Winograd T. (1977). Gus, a frame-driven dialog system. *Artif. Intell.*, **8**(2), 155–173.

[Boni & Manandhars, 2005] Boni M. D. & Manandhars S. (2005). Implementing clarification dialogues in open domain question answering. *Journal for Natural Language Engineering*, p.3̃1.

[Bourdil *et al.*, 2004] Bourdil G., Elkateb F., Ferret O., Grau B., Illouz G., benoît Mathieu, Monceaux L., Robba I. & Vilnat A. (2004). Answering in english questions asked in french by exploiting results from several sources of information. *CLEF*, p.1̃2.

[Bunt, 1996] Bunt H. C. (1996). Dynamic interpretation and dialogue theory. In M. M. Taylor, F. Néel, & D. G. Bouwhuis, Eds., *The Structure of Multimodal Dialogue, Volume 2*. Amsterdam : John Benjamins.

[Buscaldi *et al.*, 2007] Buscaldi D., annd Paolo Rosso Y. B. & Sanchis E. (2007). The upv at qa@clef 2007. *Universidad Politcnica de Valencia.*

[Caelen, 2003] Caelen J. (2003). *Dialogue homme-machine et recherche d'information, Chapitre 7, dans Assistance intelligente à la recherche d'information.* Hermès.

[Chiori *et al.*, 2003 4] Chiori H., Takaaki H., Hideaki I., Eisaku M. & andFurui Sadaoki K. S. (2003-4). Study on spoken interactive open domain question answering. *Spontaneous Speech Processing and Recognition (SSPR)*, p. 111–114.

[Cowart, 2001] Cowart W. (2001). Talkbot. 2006 :http :// www.frontiernet.net/ wcowart/.

[Cowart *et al.*, 2006] Cowart W., Copple K., Cowart W. & Taylor C. (2006). Verbot's chatterbox challenge. 2009 :http :// www.chatterboxchallenge.com/.

[Cunningham *et al.*, 2002] Cunningham H., Maynard D., Bontcheva K. & Tablan V. (2002). Gate : A framework and graphical development environment for robust nlp tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics.*

[Cutting, 2000] Cutting D. (2000). lucene. 2009 :http :// lucene.apache.org/java/docs/. 2000-03-30 First open source release.

[Dang *et al.*, 2006] Dang H. T., Lin J. & Kelly D. (2006). Overview of the trec 2006 question answering track. *The Fifteenth Text REtrieval Conference.*

[de Kretser & Moffat, 2000] de Kretser O. & Moffat A. (2000). Needles and haystacks : A search engine for personal information collections. Proc. 23nd Australasian Computer Science Conference, Canberra, Australia, 58-65.

[Ferret *et al.*, 2001] Ferret O., Grau B., Hurault-Plantet M., Monceaux L., Robba I. & Vilnat A. (2001). Finding an answer based on the recognition of the question focus. *Text retrieval conference, TREC 10.*

[Frank *et al.*, 2005] Frank E., Hall M. A., Holmes G., Kirkby R., Pfahringer B., Witten I. H. & Trigg L. (2005). Weka - a machine learning workbench for data mining. In O. Maimon & L. Rokach, Eds., *The Data Mining and Knowledge Discovery Handbook*, p. 1305–1314. Springer.

[Galibert *et al.*, 2005] Galibert O., Illouz G. & Rosset S. (2005). Ritel : dialogue homme-machine à domaine ouvert. *TALN-RECITAL 2005*, **1**, 439.

[Gomez *et al.*, 2005] Gomez J. M., Montes M., Sanchis E. & Rosso P. (2005). Jirs, a passage retrieval system for multilingual question answering. *In 8th International Conference of Text, Speech and Dialogue 2005(TSD'05)*, p. 443–450.

[Grau *et al.*, 2005] Grau B., Ligozat A. L., Robba I., Vilnat A., Kateb F. E., illouz G., Monceaux L., Paroubek P. & Pons O. (2005). De l'importance des synonymes pour la sélection de passages en question-réponse. *CORIA*, **1**, 71–84.

[Grau *et al.*, 1994] Grau B., Sabah G. & Vilnat A. (1994). Control in man-machine dialogue. *Think, Tilburg University, The Netherlands*, **1**(5), 32–55.

[Greenwood *et al.*, 2006] Greenwood M., Stevenson M. & Gaizauskas R. (2006). The university of sheffield's trec 2006 q&a experiments. *The Fifteenth Text REtrieval Conference.*

[Grice, 1975] Grice H. P. (1975). Logic and conversation. *Syntax and semantics 3rd speech acts New York, academic press*, p. 41–58.

[Grosz & Kraus, 1993] Grosz B. & Kraus S. (1993). Collaborative plans for group activities. *IJCA*, **1**.

[Harabagiu *et al.*, 2005] Harabagiu S., Hickl A., Lehmann J. & Moldovan D. (2005). Experiments with interactive question-answering. *43 rd annual meeting of the ACL*, p. 205–214.

[Hartrumpf *et al.*, 2007] Hartrumpf S., Glökner I. & Leveling J. (2007). Coreference resolution for questions and answer merging. *QA@CLEF2007*.

[Hernandez, 2004] Hernandez N. (2004). *Description et détection automatique de structures de TEXTe*. PhD thesis, University de Paris-Sud XI LIMSI/CNRS. 2009 :http :// www.limsi.fr /Individu/hernandz/research/Hernandez-these.tar.gz.

[Hickl *et al.*, 2004] Hickl A., Lehmann J., Williams J. & Harabagiu S. (2004). Experiments with interactive question answering in complex scenarios. In S. Harabagiu & F. Lacatusu, Eds., *HLT-NAACL 2004 : Workshop on Pragmatics of Question Answering*, p. 60–69, Boston, Massachusetts, USA : Association for Computational Linguistics.

[Hickl *et al.*, 2006] Hickl A., Williams J., Bensley J., Roberts K., Shi Y. & Rink B. (2006). Question answering with lcc's chaucer at trec 2006. *15th Text REtrieval Conference, Gaithersburg*, p.1̃.

[Hori *et al.*, 2003] Hori C., Hori T., Tsukada H., Isozaki H., Sasaki Y. & Maeda E. (2003). Spoken interactive odqa system : Spiqa. *ACL-2003 Interactive Poster and Demonstration Session*.

[Humphreys *et al.*, 1998] Humphreys K., Gaizauskas R., Azzam S., Huyck C., Mitchell B., Cunningham H. & Wilks Y. (1998). Description of the lasie-ii system as used for muc-7. In *In Proceedings of the Seventh Message Understanding Conferences (MUC-7* : Morgan.

[Hutchens, 1994] Hutchens J. (1994). Megahal. 2009 :http :// megahal.alioth.debian.org/.

[Informatique, 2006] Informatique D. (2006). Antidote rx. 2009 :http :// www.druide.com/.

[Jacquemin, 1999] Jacquemin C. (1999). Fastr. 2009 :http :// www.limsi.fr/Individu/jacquemi/FASTR/.

[Jones, 1972] Jones K. S. (1972). A atatistical interpretation of term specificity and its application inretrieval. *Journal of Documentation*, **28**(1), 11–21.

[Kristjansson *et al.*, 2004] Kristjansson T., Culotta A., Viola P. & McCallum A. (2004). Interactive information extraction with constrained conditional random fields. *Proceedings of AAAI-2004*.

[Laurent & Al., 2006] Laurent D. & Al. (2006). Cordial 2006. 2009 :http :// www.synapse-fr.com/.

[Laurent & Séguéla, 2005] Laurent D. & Séguéla P. (2005). Qristal, système de questions-réponses. *TALN 2005*. 2009 :http :// www.qristal.fr/.

[Laurent *et al.*, 2007] Laurent D., Séguéla P. & Nègre S. (2007). Cross lingual question answering using qristal for clef 2007. *Synapse Développement*.

[Lehuen, 1997] Lehuen J. (1997). *Un modèle de dialogue dynamique et générique intégrant l'acquisistion de sa compétene linguistique*. PhD thesis, Université e Caen. Le système COALA : 2009 :http :// www-ic2.univ-lemans.fr / lehuen/recherche/these.

[Lemeunier, 2000] Lemeunier T. (2000). *L'intentionnalité communicative dans le dialogue homme-machine en langue naturelle*. PhD thesis, LIUM-CNRS FRE 2730 Université du Maine. 2009 :http :// www-lium.univ-lemans.fr / lemeunie/biblio.php.

[Ligozat, 2006] Ligozat A.-L. (2006). *Exploitation et fusion de connaissances locales pour la recherche d'informations précises*. PhD thesis, Université de Paris-Sud XI LIMSI/CNRS.

[Lin *et al.*, 2003] Lin J., Quan D., Sinha V., Bakshi K., Huynh D., Katz B. & Karger D. R. (2003). What makes a good answer ? the role of context in question answering. *Ninth IFIP TC13 International Conference,INTERACT 2003*.

[Loebner, 2006] Loebner D. (2006). The loebner prize in artificial intelligence. 2009 :http :// www.loebner.net/Prizef/loebner-prize.html.

[Manning *et al.*, 2008] Manning C. D., Raghavan P. & Schütze H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. 2009 :http :// www-csli.stanford.edu / hinrich/information-retrieval-book.html.

[McCracken *et al.*, 2006] McCracken N. J., Diekema A. R., Ingersoll G., C. S., Harwell, Allen E. E. & Liddy O. Y. E. D. (2006). Modeling reference interviews as a basis for improving automatic qa systems. *HLT-NAACL*, p. 17–24.

[Moldovan *et al.*, 2004] Moldovan D., Harabagiu S., Clark C., Bowden M., Lehmann J. & Williams J. (2004). Experiments and analysis of lcc's two qa systems over trec 2004. *Text REtrieval Conference*.

[Monceau, 2002] Monceau L. (2002). *Adaptation du niveau d'analyse des interventions dans un dialogue. Application à un système de question - réponse*. PhD thesis, Université de Paris XI LIMSI/CNRS. 2009 :http :// www.sciences.univ-nantes.fr /info/perso/permanents/monceaux.

[Page & Brin, 1998] Page L. & Brin S. (1998). google. 2009 :http :// www.google.com.

[Parmentier, 2005] Parmentier F. (2005). Ector. 2005 :http :// ector.sourceforge.net/.

[Penas *et al.*, 2007] Penas A., Forner P. & Giampiccolo D. (2007). Guidelines for participants in qa at clef 2007. *CELCT, Trento(IT) and UNED, Madrid*, p.1̃.

[Rosset *et al.*, 2007] Rosset S., Galibert O., Adda G. & Bilinski E. (2007). The limsi qast systems : Comparison between human and automatic rules generation for question-answering an speech transcriptions. *ASRU, Kyoto, Japan.*

[Rosset *et al.*, 2006] Rosset S., Galibert O., Illouz G. & Max A. (2006). Integrating spoken dialog and question answering : the ritel project. *INTERSPEECH ICSLP.* 2009 :http :// www.limsi.fr /Individu/rosset/ritel/imix.pdf.

[Rosset & Petel, 2006] Rosset S. & Petel S. (2006). The ritel corpus - an annotated human-machine open-domain question answering spoken dialog corpus. *International Conference on Language Resources and Evaluation.*

[Rosset & Tribout, 2005] Rosset S. & Tribout D. (2005). Détection automatique d'actes de dialogue par l'utilisation d'indices multiniveaux. *TALN-05*, **1**.

[Roulet *et al.*, 1985] Roulet E., Auchlin A., Moeschler J., Rubattel C. & Schelling M. (1985). *L'articulation du discours en français contemporain.* Peter Lang Verlagsgruppe. 3e édition : 1991.

[Sabah *et al.*, 1997] Sabah G., Vivier J., Vilnat A., Pierrel J.-M., Romary L. & Nicolle A. (1997). *Machine, Langage et Dialogue.* l'Harmattan.

[Salton & Buckley, 1988] Salton G. & Buckley C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**(5), 513–523.

[Salton & Yang, 1973] Salton G. & Yang C. S. (1973). On the specification of term values in automatic indexing. *Departement of Computer Science Cornell University Ithaca New York 14850.* Technical Report 73-173.

[Scherpbier & Hutchison, 1995] Scherpbier A. & Hutchison G. (1995). htdig. 2009 :http :// www.htdig.org/.

[Schmid, 1994] SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. *Internatonal Conference on New Methods in Language Processing, Manchester*.

[Small *et al.*, 2004] SMALL S., STRZALKOWSKI T., JANACK T., LUI T., RYAN S., SALKIN R., SHIMIZU N., KANTOR P., KELLY D., ROBERT RITTMAN, WACHOLDER N. & YAMROM B. (2004). Hitiqa : Scenario based question answering. *HLT-NAACL*, p. 52–59.

[Sven, 2006] SVEN H. (2006). Extending knowledge and deepening linguistic processing for the question answering system insicht. *6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005, Vienna, Austria*, p. 361–369.

[Séjourné, 2005] SÉJOURNÉ K. (2005). *Du moteur de question réponse au dialogue à but*. PhD thesis, University de Paris-Sud XI LIMSI/CNRS. mémoire de stage de DEA/master 2eme année.

[Séjourné, 2008] SÉJOURNÉ K. (2008). Une structure pour les questions enchainées. *RECITAL, Avignon, 9-13 juin*.

[Séjourné, 2009] SÉJOURNÉ K. (2009). Exploitation d'une structure pour les questions enchaînées. *TALN 2009 - Session poster, Senlis, 24-26*.

[Turing, 1950] TURING A. (1950). Computing machinery and intelligence. *Mind*, **59**, 433–460. Le test de Turing.

[van Schooten & op den Akker, 2005] VAN SCHOOTEN B. & OP DEN AKKER R. (2005). Follow-up utterances in qa dialogue. *TALN-05*, **1**(46(3)).

[Vilnat, 2005] VILNAT A. (2005). *Habilitation à diriger les recherches : Dialogue et analyse de phrases*. PhD thesis, University de Paris-Sud XI LIMSI/CNRS. 2009 :http :// www.limsi.fr /Individu/anne/HDR/MemoireHDR.pdf.

[Webb, 2006] WEBB N. (2006). Interactive question answering proceedings of the workshop. *HTL-NAACL*.

[Weizenbaum, 1966] WEIZENBAUM J. (1966). Eliza. 2009 :http :// www-ai.ijs.si /elizacgi-bin/eliza_script.

[Whittaker *et al.*, 2006] WHITTAKER E., NOVAK J., CHATAIN P. & FURUI S. (2006). Trec2006 question answering experiments at tokyo institute of technology. *The Fifteenth Text REtrieval Conference*.

[Zhou *et al.*, 2006] ZHOU Y., YUAN X., CAO J., HUANG X. & WU L. (2006). Fduqa on trec2006 qa track. *15th Text REtrieval Conference, Gaithersburg*, p. 1026–1033.