



Approches supervisées et faiblement supervisées pour l'extraction d'événements et le peuplement de bases de connaissances

Ludovic Jean-Louis

► To cite this version:

Ludovic Jean-Louis. Approches supervisées et faiblement supervisées pour l'extraction d'événements et le peuplement de bases de connaissances. Autre [cs.OH]. Université Paris Sud - Paris XI, 2011. Français. NNT : 2011PA112288 . tel-00686811

HAL Id: tel-00686811

<https://theses.hal.science/tel-00686811>

Submitted on 11 Apr 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS 11 - PARIS SUD
ÉCOLE DOCTORALE EDIPS

T H È S E

pour obtenir le titre de

docteur en Sciences

de l'Université de Paris 11 - Paris Sud

Mention : Informatique

Présentée et soutenue par

LUDOVIC JEAN-LOUIS

**Approches supervisées et faiblement supervisées pour
l'extraction d'événements complexes et le peuplement de
bases de connaissances**

Thèse dirigée par Olivier FERRET

préparée au Laboratoire Vision et Ingénierie des Contenus

soutenue le 15 Décembre 2011

Jury :

<i>Rapporteurs :</i>	Patrice BELLOT	-	Université Aix-Marseille
	Adeline NAZARENKO	-	Université Paris Nord
<i>Examineurs :</i>	Pierre ZWEIGENBAUM	-	LIMSI CNRS
	Claude DE LOUPY	-	Syllabs
<i>Directeur :</i>	Olivier FERRET	-	CEA LIST
<i>Encadrant :</i>	Romarc BESANÇON	-	CEA LIST

Remerciements

Ce travail n'aurait pas été possible sans l'aide de mes encadrants Romaric Besançon (encadrant scientifique) et Olivier Ferret (directeur de thèse). Je voudrais les remercier chaleureusement pour leurs conseils avisés, leurs critiques constructives ainsi que pour les nombreuses discussions très enrichissantes que nous avons eues au cours de ces trois années. Ces échanges ont très souvent été fondamentaux pour l'avancement de mes travaux.

Un grand merci à tous les membres permanents du laboratoire Vision et Ingénierie des Contenus (LVIC) pour leurs conseils, encouragements et leur disponibilité. Merci à Gaël de Chalendar, Olivier Mesnard, Hervé Le Borgne, Adrian Popescu, Bertrand Delezoïde, Nasredine Semmar, Faïza Gara, Géraud Canet et Meriama Laïb-Boukhari.

Je voudrais aussi remercier tous les membres non permanents, stagiaires, CDDs et thésards que j'ai rencontrés durant ces trois années de thèse et qui ont contribué d'une manière ou d'une autre au succès de ce travail. Aussi, un grand merci à mes colocataires de bureau, Christophe Servan, Soumya Hamlaoui, Aymen Shabou, Dmitri Voitsekhovitch, Claire Mouton, Bassem Makni, Kris Jack, qui ont dû me supporter et qui m'ont soutenu dans les périodes difficiles. Je dois aussi saluer les membres du bureau voisin Nhung Pham, Mohamed Ouazara, Zied Boulila, Albert Savary pour leur soutien, nos discussions et les sorties que nous avons partagées. J'adresse des remerciements particuliers à Wei Wang, Dhafer Lhabib et Adrien Durand qui ont contribué à l'implémentation d'une partie de ce travail.

Merci également à l'ensemble du jury de thèse pour avoir bien voulu examiner mes travaux. En particulier, merci à Pierre Zweigenbaum pour sa relecture détaillée ainsi que ses commentaires qui ont permis d'améliorer le manuscrit.

Je voudrais enfin remercier chaleureusement ma famille et mes amis pour leur soutien indéfectible.

Résumé

La plus grande partie des informations disponibles librement sur le Web se présentent sous une forme textuelle, c'est-à-dire non-structurée. Dans un contexte comme celui de la veille, il est très utile de pouvoir présenter les informations présentes dans les textes sous une forme structurée en se focalisant sur celles jugées pertinentes vis-à-vis du domaine d'intérêt considéré. Néanmoins, lorsque l'on souhaite traiter ces informations de façon systématique, les méthodes manuelles ne sont pas envisageables du fait du volume important des données à considérer.

L'extraction d'information s'inscrit dans la perspective de l'automatisation de ce type de tâches en identifiant dans des textes les informations concernant des faits (ou événements) afin de les stocker dans des structures de données préalablement définies. Ces structures, appelées *templates* (ou formulaires), agrègent les informations caractéristiques d'un événement ou d'un domaine d'intérêt représentées sous la forme d'entités nommées (nom de lieux, etc.).

Dans ce contexte, le travail de thèse que nous avons mené s'attache à deux grandes problématiques :

- l'identification des informations liées à un événement lorsque ces informations sont dispersées à une échelle textuelle en présence de plusieurs occurrences d'événements de même type ;
- la réduction de la dépendance vis-à-vis de corpus annotés pour la mise en oeuvre d'un système d'extraction d'information.

Concernant la première problématique, nous avons proposé une démarche originale reposant sur deux étapes. La première consiste en une segmentation événementielle identifiant dans un document les zones de texte faisant référence à un même type d'événements, en s'appuyant sur des informations de nature temporelle. Cette segmentation détermine ainsi les zones sur lesquelles le processus d'extraction doit se focaliser. La seconde étape sélectionne à l'intérieur des segments identifiés comme pertinents les entités associées aux événements. Elle

conjugue pour ce faire une extraction de relations entre entités à un niveau local et un processus de fusion global aboutissant à un graphe d'entités. Un processus de désambiguïsation est finalement appliqué à ce graphe pour identifier l'entité occupant un rôle donné vis-à-vis d'un événement lorsque plusieurs sont possibles.

La seconde problématique est abordée dans un contexte de peuplement de bases de connaissances à partir de larges ensembles de documents (plusieurs millions de documents) en considérant un grand nombre (une quarantaine) de types de relations binaires entre entités nommées. Compte tenu de l'effort représenté par l'annotation d'un corpus pour un type de relations donné et du nombre de types de relations considérés, l'objectif est ici de s'affranchir le plus possible du recours à une telle annotation tout en conservant une approche par apprentissage. Cet objectif est réalisé par le biais d'une approche dite de supervision distante prenant comme point de départ des exemples de relations issus d'une base de connaissances et opérant une annotation non supervisée de corpus en fonction de ces relations afin de constituer un ensemble de relations annotées destinées à la construction d'un modèle par apprentissage. Cette approche a été évaluée à large échelle sur les données de la campagne TAC-KBP 2010.

Abstract

The major part of the information available on the web is provided in textual form, *i.e.* in unstructured form. In a context such as technology watch, it is useful to present the information extracted from a text in a structured form, reporting only the pieces of information that are relevant to the considered field of interest. Such processing cannot be performed manually at large scale, given the large amount of data available. The automated processing of this task falls within the Information extraction (IE) domain.

The purpose of IE is to identify, within documents, pieces of information related to facts (or events) in order to store this information in predefined data structures. These structures, called *templates*, aggregate fact properties – often represented by named entities – concerning an event or an area of interest.

In this context, the research performed in this thesis addresses two problems :

- identifying information related to a specific event, when the information is scattered across a text and several events of the same type are mentioned in the text ;
- reducing the dependency to annotated corpus for the implementation of an Information Extraction system.

Concerning the first problem, we propose an original approach that relies on two steps. The first step operates an event-based text segmentation, which identifies within a document the text segments on which the IE process shall focus to look for the entities associated with a given event. The second step focuses on template filling and aims at selecting, within the segments identified as relevant by the event-based segmentation, the entities that should be used as fillers, using a graph-based method. This method is based on a local extraction of relations between entities, that are merged in a relation graph. A disambiguation step is then performed on the graph to identify the best candidates to fill the information template.

The second problem is treated in the context of knowledge base (KB) population, using a large collection of texts (several millions) from which the information is extracted. This extraction also concerns a large number of relation types (more than 40), which makes the manual annotation of the collection too expensive. We propose, in this context, a distant supervision approach in order to use learning techniques for this extraction, without the need of a fully annotated corpus. This distant supervision approach uses a set of relations from an existing KB to perform an unsupervised annotation of a collection, from which we learn a model for relation extraction. This approach has been evaluated at a large scale on the data from the TAC-KBP 2010 evaluation campaign.

Table des matières

Table des matières	vii
Table des figures	xi
Nomenclature	xi
1 Introduction	1
1.1 Motivations et objectifs	1
1.2 Organisation du document	6
2 Extraction d'information : vue d'ensemble du domaine	7
2.1 Extraction d'information	7
2.2 Le contexte de l'extraction d'information	9
2.2.1 Contextes d'utilisation	9
2.2.2 Les types de documents utilisés	10
2.3 Les tâches en extraction d'information	11
2.3.1 Présentation des tâches	11
2.3.2 Architecture type	14
2.4 Approches pour l'extraction d'information	16
2.4.1 Approches symboliques	16
2.4.2 Approches statistiques	17
2.4.2.1 Méthodes supervisées	18
2.4.2.2 Méthodes semi-supervisées	19
2.4.2.3 Méthodes non supervisées	21
2.5 Reconnaissance des entités nommées	22

TABLE DES MATIÈRES

2.5.1	Présentation du problème	22
2.5.2	Les approches pour la détection des entités nommées . . .	23
2.5.2.1	Utilisation de règles	24
2.5.2.2	Utilisation d'apprentissage statistique	26
2.6	Coréférence entre entités nommées	28
2.6.1	Présentation du problème	28
2.6.2	Les approches pour la résolution de coréférence	30
2.6.2.1	Utilisation d'approches linguistiques	31
2.6.2.2	Utilisation d'approches statistiques	32
2.7	Extraction de relations entre entités nommées	33
2.7.1	Présentation du problème	33
2.7.2	Les approches pour l'extraction de relations entre entités nommées	36
2.7.2.1	Approches à base de patrons	36
2.7.2.2	Approches à base de classifieurs statistiques . . .	39
2.8	Construction des <i>templates</i> sur les événements	41
2.8.1	Présentation du problème	41
2.8.2	Les approches pour la construction de <i>templates</i>	42
2.9	Notre problématique d'extraction d'information	43
3	La segmentation des textes en événements	47
3.1	Introduction	47
3.1.1	Qu'est-ce qu'un événement ?	48
3.1.2	Les informations discursives et les événements	50
3.2	Segmentation des textes et extraction d'information	53
3.3	La segmentation en événements à partir d'indices temporels . . .	54
3.4	Modèle discursif sous-jacent à la segmentation événementielle . . .	56
3.5	Modèles de segmentation événementielle	57
3.5.1	Une segmentation fondée sur les temps verbaux: le modèle HMM	58
3.5.2	Élargissement des indices temporels: le modèle CRF	59
3.5.3	Modèle MaxEnt	61
3.5.4	Approches heuristiques	62

TABLE DES MATIÈRES

3.6	Pré-traitement des documents	63
3.7	Évaluation des méthodes de segmentation	64
3.7.1	Les corpus d'évaluation	65
3.7.2	Évaluation intrinsèque de la segmentation en événements .	66
3.7.3	Évaluation de la segmentation pour l'extraction d'information	70
3.8	Conclusions	72
4	Le rattachement des entités aux événements	75
4.1	Introduction	75
4.1.1	Bases de données et <i>templates</i>	77
4.1.2	Les relations complexes	78
4.2	Graphes d'entités nommées	81
4.3	Application du rattachement à l'extraction des événements	83
4.3.1	Construction du graphe d'entités	84
4.3.2	Sélection des entités et remplissage des <i>templates</i>	87
4.4	Application et évaluation de l'approche de rattachement	90
4.4.1	Construction du graphe d'entités	91
4.4.2	Sélection des entités et remplissage des <i>templates</i>	94
4.4.3	Impact de la segmentation sur le rattachement	95
4.4.4	Analyse d'erreurs	96
4.5	Conclusions	98
5	Peuplement de bases de connaissances	101
5.1	Introduction	101
5.2	Le peuplement de bases de connaissances	104
5.3	Lien entre peuplement de KB et question-réponse	104
5.4	Vue d'ensemble de l'approche pour l'extraction de relations	105
5.4.1	Apprentissage des patrons de relations	107
5.4.2	Filtrage pour l'apprentissage des patrons de relations . . .	109
5.4.3	Extraction des relations	112
5.4.4	Amélioration par l'utilisation d'un filtrage générique de re- lations	113
5.5	La campagne d'évaluation TAC-KBP	115

TABLE DES MATIÈRES

5.5.1	TAC-KBP 2009 – 2010	115
5.5.2	TAC-KBP 2011	119
5.6	Évaluation de l’approche dans le cadre de TAC-KBP	120
5.6.1	Les données	120
5.6.2	Métriques d’évaluation TAC-KBP 2010	121
5.6.3	Évaluation de l’apprentissage des patrons	125
5.6.4	Évaluation de l’extraction des relations	126
5.6.4.1	Recherche des phrases candidates	127
5.6.4.2	Extraction de relations	128
5.6.5	Vue d’ensemble des résultats pour TAC-KBP 2011	131
5.7	Aperçu des systèmes utilisés pour TAC-KBP	133
5.8	Discussion sur les résultats de TAC-KBP	136
5.9	Conclusions	138
6	Conclusion	143
6.1	Bilan des résultats	143
6.2	Analyse de notre contribution	145
6.3	Perspectives	151
	Liste des publications	155
	References	157

Table des figures

1.1	Exemple de <i>template</i> à compléter pour l'acquisition d'entreprise	2
1.2	Exemple de <i>template</i> instancié pour l'acquisition de société	2
2.1	Exemple de reconnaissance d'entités nommées	24
2.2	Exemple de patron pour la détection des montants des transactions	25
2.3	Exemple de chaînes de coreférences	31
3.1	Organisation des événements dans les articles de presse	51
3.2	Exemple d'annotation d'événements à partir d'une dépêche de presse	57
3.3	Illustration de la segmentation de textes en événements avec le modèle HMM	59
3.4	Exemple d'analyse linguistique produite par LIMA. L'analyse est présentée en format tabulaire, avec le séparateur « », pour chaque mot du document : la première colonne désigne la position du premier caractère du mot dans le texte; la deuxième colonne contient le mot issu du texte, la troisième colonne contient le lemme ainsi que la catégorie morphosyntaxique; la dernière colonne désigne le temps grammatical lorsque le mot est un verbe.	64
4.1	Approche en deux étapes pour l'extraction de relations complexes	78
4.2	Exemple de graphes d'entités nommées au niveau des phrases . . .	82
4.3	Exemple de graphe d'entités nommées au niveau du document . . .	85
4.4	Répartition des erreurs, par type	97
5.1	Architecture générale du système	107

TABLE DES FIGURES

Chapitre 1

Introduction

1.1 Motivations et objectifs

Les premiers travaux s'intéressant au traitement automatique des textes en langue naturelle datent du milieu des années 60 et se focalisaient en partie sur les aspects syntaxiques de la langue, sans véritablement intégrer l'aspect informationnel des textes. À la suite des travaux menés à la fin des années 70 et au début des années 80 sur la compréhension de texte, les chercheurs se sont intéressés à des formes plus limitées de compréhension dans la perspective de répondre à la question : *comment extraire des informations utiles à partir des textes écrits en langue naturelle ?* Ainsi, les travaux se sont orientés dès la fin des années 80 vers l'*extraction d'information*, notamment avec les campagnes d'évaluation MUC *Message Understanding Conference*, suivies par les campagnes ACE *Automatic Content Extraction*, organisées en partie par le NIST.

L'extraction d'information a pour objectif d'identifier les informations structurées contenues dans les textes, typiquement en vue de les stocker dans des structures de données préalablement définies. En pratique, il s'agit de compléter à partir de documents des *templates*, ou formulaires, caractérisant les informations que les utilisateurs souhaitent voir extraire d'un texte en relation le plus souvent avec un événement. Les informations ainsi extraites par le biais de ces *templates* viennent généralement alimenter des bases de données ou des bases de connaissances.

1. INTRODUCTION

Pour illustration, prenons l'extrait de dépêche ci-dessous, à partir duquel nous cherchons à instancier un *template* dans le domaine de l'acquisition de sociétés (cf figure 1.1).

Rachat de BEA Systems par Oracle : c'est fait. Au prix fort.
Le géant de la base de données annonce en effet ce mercredi 16 janvier avoir finalisé un accord en vue du rachat de BEA Systems. Oracle a accepté les conditions de BEA Systems, fixant le montant du rachat à 8,5 milliards de dollars.
Il s'agira d'un des plus grands coups de Larry Ellison, le p-dg d'Oracle. Le précédent record était de 10 milliards pour l'acquisition de PeopleSoft en 2004.

Template pour l'acquisition d'entreprises

Événement	Acquéreur	Société acquise	Date	Montant
<EVENT>	<ORG>	<ORG>	<DATE>	<MONEY>

FIG. 1.1 – Exemple de *template* à compléter pour l'acquisition d'entreprise

Le résultat du processus d'extraction d'information est illustré par le tableau suivant (cf figure 1.2) :

Template pour l'acquisition d'entreprises

Événement	Acquéreur	Société acquise	Date	Montant
rachat	Oracle	BEA Systems	mercredi 16 janvier	8,5 milliards de dollars
acquisition	Oracle	PeopleSoft	2004	10 milliards

FIG. 1.2 – Exemple de *template* instancié pour l'acquisition de société

Du point de vue applicatif, l'argument mis en avant en faveur de l'extraction d'information est le caractère limité des capacités de traitement d'un être humain :

[Cowie and Lehnert, 1996] mentionnent ainsi qu’un être humain est incapable de lire, comprendre et synthétiser de grands volumes de données sans erreurs, et surtout sans faire d’omissions.

Un système d’extraction d’information est traditionnellement associé à une chaîne de traitement composée de différents modules, dont chacun est responsable d’une tâche accomplissant un traitement linguistique. Ces tâches sont généralement appliquées à un document, même si certaines peuvent s’appliquer à plus : on fait alors l’hypothèse que les valeurs dans le *template* proviennent de plusieurs documents. Dans [Cunningham, 2005] l’auteur propose un découpage du processus d’extraction en cinq tâches parmi lesquelles la reconnaissance des entités nommées et l’identification des relations entre ces entités nommées. L’ensemble des tâches est décrit dans le chapitre 2.

Sur l’extrait d’article ci-dessus, la tâche de reconnaissance des entités nommées revient à associer les valeurs [*Oracle* ; *PeopleSoft*] à des organisations, les valeurs [*8,5 milliards de dollars* ; *10 milliards*] à des montants et les valeurs [*mercredi 16 janvier* ; *2004*] à des dates. La tâche d’identification de relations consiste à identifier la présence/absence de relations sémantiques entre les entités nommées, ce qui permet de les associer correctement dans le *template*. Toujours dans l’extrait de dépêche, deux types de relations entre les entités sont illustrés : les relations *locales* et les relations *globales*. Les relations *locales* s’expriment à l’intérieur d’une même phrase, par exemple : «*Rachat de BEA systems par Oracle*». Les relations *globales* s’expriment au niveau du document, par exemple :

«... annonce ce mercredi 16 janvier avoir finalisé un accord en vue du rachat de *BEA Systems*. *Oracle* a accepté les conditions de *BEA Systems*, fixant le montant du rachat à 8,5 milliards de dollars.».

La finalité du processus d’extraction est de représenter sous une forme structurée les propriétés d’un fait ou d’un événement donné, ces propriétés étant exprimées par des entités spécifiques. Dans l’extrait ci-dessus, il s’agit de repérer deux acquisitions d’entreprises ayant des dates et des montants distincts. Si faire un tel repérage semble intuitif pour un humain, plusieurs questions se posent lorsqu’il doit être fait de façon automatique, en particulier pour des événements de même nature ; *quels sont les critères à prendre en compte pour structurer les informations ? quelles stratégies utiliser pour regrouper ces informations ?*

1. INTRODUCTION

Dans cette thèse, la première problématique que nous abordons concerne l'extraction d'événements complexes, c'est-à-dire des événements étant décrits par plusieurs entités nommées ayant chacune un rôle déterminé dans l'événement. Plus précisément, l'extraction d'événement vise à remplir un *template* pour un événement donné lorsque plusieurs événements comparables sont mentionnés dans un même document.

Cette problématique n'est pas totalement nouvelle et des solutions ont été proposées, celles-ci peuvent être catégorisées selon deux paradigmes, les approches à base d'ingénierie des connaissances et les approches à base d'apprentissage statistique. Les approches relevant de l'ingénierie des connaissances se caractérisent par l'utilisation d'un ensemble de règles créées par des experts d'un domaine. Les approches à base d'apprentissage statistique cherchent à diminuer l'intervention des experts en utilisant des méthodes statistiques.

Les systèmes à base de règles servent généralement à constituer un ensemble de règles ou patrons d'extraction totalement attaché à un domaine, soit de façon manuelle [Hobbs, 1993] ou de façon automatique [Aone and Ramos-Santacruz, 2000]. Les systèmes à base d'apprentissage statistique quant à eux, se servent des traits caractéristiques contenus dans des exemples, dits d'entraînement, pour apprendre à reconnaître les informations à extraire. Un des premiers systèmes utilisant ce paradigme pour l'extraction d'information est *AutoSlug* qui est présenté dans [Riloff, 1993].

Dans les deux cas, les paradigmes sont appliqués au niveau des phrases une à une et sans tenir compte des liens au niveau textuel (au niveau du document) entre les différentes phrases. La conséquence est que les *templates* conçus à partir de cette méthode sont incomplets puisqu'ils ne contiennent pas les relations distantes entre les entités. Il apparaît donc nécessaire d'introduire des stratégies d'appariement (ou de rattachement) des entités au niveau textuel afin de compléter le *template*.

Les stratégies jusqu'ici utilisées reposent le plus souvent sur des ensembles de règles d'agrégation fortement liés à un domaine, ce qui rend les règles performantes même si elles sont faiblement portables. En revanche, ce n'est que récemment que quelques travaux ont tenté de traiter ce problème en utilisant des approches plus globales à base d'apprentissage statistique [Gu and Cercone, 2006;

Patwardhan and Riloff, 2007].

De même que les approches au niveau du document, nous proposons dans cette thèse une solution pour l'extraction d'information se caractérisant par deux étapes : la *segmentation des textes en événements* et le *rattachement des entités* aux événements. La segmentation en événements vise à découper les textes en segments homogènes sur le plan de leur statut événementiel, l'objectif étant à la fois d'identifier les zones de texte sur lesquelles le processus d'extraction doit se focaliser et de limiter l'espace de recherche des entités liées à un événement. Le rattachement des entités aux événements permet à partir du résultat de la segmentation d'associer à chaque événement principal d'un texte toutes les entités qui lui sont associées.

De façon complémentaire, nous abordons la problématique du peuplement de bases de connaissances. Notre motivation est de permettre de situer un événement dans un contexte plus général que celui du document ayant servi pour son extraction. Plus précisément, il s'agit de présenter des connaissances de type encyclopédique sur chacune des entités ayant un rôle dans la description d'un événement. Par exemple dans le *template* de la figure 1.2, l'une des entreprises concernée par le rachat est PeopleSoft, qui est une société américaine fondée en 1985. Ce type d'information se trouve typiquement dans des bases de connaissance, telles que les encyclopédies en ligne. Malheureusement, ce type de bases peuvent être incomplètes, voire contenir des informations obsolètes. Par conséquent, il est utile de pouvoir les enrichir en se servant des connaissances contenues dans des textes. Récemment, des travaux se sont intéressés à la construction de systèmes permettant de remplir une bases de connaissances à partir de corpus, en particulier au travers de la tâche *Knowledge Base Population (KBP)* de la campagne d'évaluation *Text Analysis Conference (TAC)*.

Le problème du peuplement de base de connaissance est présenté comme une évolution de l'extraction d'information et de la tâche de question-réponse. Il ne s'agit plus seulement d'inclure les informations extraites à une base de connaissance existante : il faut aussi prendre en compte les notions de nouveauté (ajoute-t-on une information nouvelle par rapport à l'existant ?) et de redondance (existe-t-il déjà dans la base de connaissances une forme équivalente de l'information que l'on souhaite ajouter ?) ; il faut extraire plusieurs dizaines de relations à partir

1. INTRODUCTION

d'un corpus dépassant le million de documents. Dans cette thèse, la solution proposée pour cette problématique s'appuie sur une approche de type « supervision distante » pour réaliser un apprentissage de patrons lexico-syntaxiques de relations à large échelle.

1.2 Organisation du document

Le chapitre 2 revient de façon plus approfondie sur l'extraction d'information : en plus des principales approches et principaux systèmes utilisés dans ce domaine, les approches plus directement liées à la notre y sont décrites.

L'approche d'extraction proposée pour tenter de résoudre la problématique d'extraction d'événements complexes repose sur les principes présentés dans le chapitre 2 ainsi que nos expérimentations. Cette approche se compose de deux étapes successives qui font chacune l'objet d'un chapitre. La segmentation des textes en événements est décrite dans le chapitre 3, puis le rattachement des entités aux événements dans le chapitre 4. Pour chacun de ces chapitres, les objectifs de notre démarche, les travaux associés ainsi que les expérimentations sont explicités.

Concernant la problématique de peuplement de bases de connaissances, le chapitre 5 présente un prototype développé en vue d'enrichir une base de connaissance existante. Ce prototype repose sur une approche faiblement supervisée pour l'extraction de relations qui est présentée dans ce même chapitre. Cette tâche étant au cœur de la campagne d'évaluation *KBP*, ce chapitre décrit de façon détaillée le contexte de cette campagne ainsi que les approches développées par les participants. Enfin, le chapitre 6 présente les conclusions ainsi que les perspectives de ce travail de recherche.

Chapitre 2

Extraction d'information : vue d'ensemble du domaine

2.1 Extraction d'information

L'extraction d'information peut être définie de façon générale comme l'extraction d'informations structurées à partir de textes en langue naturelle, donc non structurés. Cette tâche recouvre le plus souvent l'identification automatique de certaines entités, relations ou événements définis dans les textes, et est liée aux domaines du traitement automatique des langues et de l'intelligence artificielle. Son but est plus globalement de faciliter l'accès à l'information à un lecteur humain parmi une masse importante de documents textuels disponibles au format électronique et de trouver les informations spécifiques dont il a besoin. Le domaine de l'extraction d'information diffère du domaine de la recherche d'information, qui vise à retrouver *un ensemble de documents pertinents* en rapport avec une requête donnée, et laisser le lecteur chercher l'information voulue dans les documents retournés. Dans le cas de l'extraction d'information, on cherche à fournir directement au lecteur cette information, extraite automatiquement du texte, ce qui est plus ambitieux et plus difficile. L'étape de recherche d'information est par contre complémentaire et peut venir en amont du processus d'extraction pour fournir à ce processus un ensemble de documents pertinents par rapport à une thématique donnée.

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

L'information extraite est en général structurée pour être fournie à un système de visualisation afin d'être présentée à un utilisateur ou stockée dans une base de données qui peut être interrogée par des requêtes formelles (par exemple une base de données relationnelle).

Le domaine de l'extraction d'information s'est développé à la fin des années quatre-vingt et au début des années quatre-vingt-dix avec les conférences MUC (*Message Understanding Conferences*), un ensemble de campagnes d'évaluation dédiées à l'extraction d'information, qui ont défini les différentes tâches du domaine ainsi que les protocoles et les métriques pour l'évaluation de ces tâches. La dernière conférence MUC a été organisée en 1998, avant que d'autres campagnes sur l'extraction d'information suivent, comme les campagnes ACE (*Automatic Content Extraction*), puis les campagnes TAC (*Text Analysis Conference*). Une tâche archétypique de l'extraction d'information est le remplissage automatique d'un formulaire (ou *template*) qui résume les informations clés contenues dans un texte en fonction des centres d'intérêt fixés par un utilisateur. Dans cette section, nous présentons un panorama général de l'extraction d'information qui s'appuie en partie sur des états de l'art existant dans ce domaine [Chang et al., 2006; Cowie and Lehnert, 1996; Cunningham, 2005; Grishman, 1997; McCallum, 2005; Mooney and Bunescu, 2005; Sarawagi, 2008; Simões et al., 2009; Turmo et al., 2006; Uren et al., 2006].

Plus précisément, nous présentons à la section 2.2 le contexte de l'extraction d'information, en indiquant quelles sont ses applications ainsi que les types de documents sur lesquels elle porte. Dans la section 2.3, nous présentons les différentes tâches de l'extraction d'information. La section 2.4 présente les principes génériques des différentes approches utilisées pour ces différentes tâches et nous détaillons l'utilisation de ces méthodes appliquées aux différentes tâches de l'extraction d'information dans les sections 2.5, 2.6, 2.7 et 2.8. Enfin, nous détaillons les spécificités de la problématique d'extraction d'information que nous traitons dans cette thèse et les grandes lignes de la méthode utilisée dans la section 2.9.

2.2 Le contexte de l'extraction d'information

2.2.1 Contextes d'utilisation

Il existe, dans différents domaines, des systèmes intégrant des composantes utilisant l'extraction d'information. Si une part importante de ces systèmes a été développée dans un contexte de veille, par exemple pour des organisations gouvernementales, une part significative a été consacrée à d'autres applications. [Sarawagi, 2008] recense quatre types d'usage pour les systèmes d'extraction d'information :

Les applications pour les entreprises Ces applications sont le plus souvent utilisées dans un contexte de veille : une entreprise cherche par exemple à s'informer de la façon dont elle est perçue ou comment elle se situe vis-à-vis de ses concurrents. Les domaines d'utilisation de ces applications sont le suivi d'événements d'actualité, le suivi de clients, la normalisation de données (suppression d'éventuels doublons dans une base de données) et enfin le suivi de petites annonces. Les documents utilisés sont principalement des dépêches de presse.

La gestion des données personnelles Il s'agit de systèmes permettant d'organiser les documents d'un utilisateur en fonction des informations qu'ils contiennent. Ils permettent d'établir des liens entre les contenus de différents documents. Contrairement aux moteurs de recherche internes qui permettent d'indexer du contenu provenant de différents types de documents (courriels, fichiers textes, etc.), ce type de systèmes se concentrent plus sur l'organisation et la synchronisation de contenus venant de plusieurs sources.

Les applications scientifiques Il s'agit de systèmes utilisés pour aider les chercheurs en leur présentant des informations synthétiques extraites d'articles scientifiques. Par exemple, dans le domaine de la bioinformatique, ces systèmes peuvent utiliser les articles présents dans des bases de données bibliographiques afin de détecter des noms de gènes/protéines et relever les interactions mentionnées entre ces éléments.

Les applications orientées Web Il s'agit de sites Web utilisant comme source d'information le contenu d'autres pages Web. Les informations issues de ces

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

contenus sont centralisées et structurées en fonction d'un besoin donné. Les domaines d'utilisation sont la création de bases de données de citations (ou d'opinions d'utilisateurs), la centralisation d'événements concernant une communauté (par exemple le regroupement des informations concernant des conférences scientifiques), la création de comparateurs de prix, la création de publicités ciblées.

D'autres exemples d'applications sont présentés dans [Cunningham, 2005; McCallum, 2005].

2.2.2 Les types de documents utilisés

Les documents utilisés pour l'extraction d'information sont très variés en fonction du domaine d'application du processus d'extraction : il peut s'agir d'écrits journalistiques (articles, dépêches de presse), d'articles scientifiques ou rapports spécialisés (par exemple pour l'extraction d'information dans le domaine médical), d'écrits narratifs (romans, textes anciens) ou de correspondances entre personnes (courriel, sms, forums), etc. Plus généralement, il s'agit d'utiliser les textes bruts pour détecter les informations pertinentes pour l'application visée.

La plupart des systèmes utilisent des documents de même nature (par exemple seulement des textes journalistiques, articles et dépêches de presse), mais dans une perspective générique, il conviendrait de pouvoir traiter des documents de natures différentes. Les organisateurs de la campagne d'évaluation sur le peuplement de base de connaissances TAC-KBP ont d'ailleurs proposé dans cette optique d'utiliser un corpus composé de documents de différentes natures (articles de presse, pages Web, etc.).

Le contenu des textes bruts auquel nous nous intéressons pour extraire des informations est exprimé en langage naturel à travers de phrases, propositions, etc. Néanmoins, ce contenu n'est pas toujours exclusivement composé de phrases et peut dans certains cas contenir des métadonnées sur la mise en forme ou la structure du document (ou des informations). On parle alors de documents semi-structurés.

Les métadonnées apportent au lecteur des connaissances supplémentaires sur le contenu du texte brut. Une illustration de document semi-structuré est présentée

dans l'extrait ci-dessous, il s'agit d'une adresse décrite selon les métadonnées *nom*, *rue*, *ville* et *code postal* :

Nom : Laboratoire Vision et Ingénierie des Contenus

Rue : 18 route du Panorama

Code postal : 92265

Ville : Fontenay-aux-Roses

Afin d'extraire des informations à partir de documents semi-structurés de ce type, les systèmes doivent tirer partie des métadonnées présentes mais aussi des indices de mise en forme ou de positionnement des informations. Par exemple pour extraire des adresses, il est important de noter que le code postal est le plus souvent mentionné avant le nom de la ville dans une adresse. De plus, les retours à la ligne indiquent que l'on change de type d'information (le destinataire et le code postal ne sont pas sur la même ligne).

Sous un autre angle, lorsque le contenu du texte brut ne contient pas de métadonnées, les documents sont dits non-structurés. Les documents sont alors exclusivement composés de phrases en langage naturel. Notons cependant que certains documents non-structurés ont une structure qui peut être exploitée pour extraire des informations. Par exemple, si l'on considère les dépêches de presse, les phrases sont généralement regroupées en paragraphes ; elles contiennent le plus souvent un titre, etc. Afin d'extraire les informations à partir de textes non-structurés, les approches utilisées sont différentes de celles appliquées aux documents semi-structurés : les indices de mise en forme des informations ne jouent pas en effet de rôle prépondérant puisque les informations sont disséminées dans des phrases.

2.3 Les tâches en extraction d'information

2.3.1 Présentation des tâches

En fonction de la nature de l'information extraite, plusieurs tâches différentes ont été définies dans le domaine de l'extraction d'information. Nous présentons ici une vue générale des cinq tâches généralement retenues [Cunningham, 1997,

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

2005; Turmo et al., 2006]. Ces tâches seront présentées plus en détail dans la suite de ce chapitre.

- La reconnaissance des entités nommées (NER *named entity recognition*) : cette tâche concerne l'identification de certaines entités spécifiques dans les textes, en leur associant un type défini. Ces entités peuvent être relativement générales, comme les dates, les noms d'organisations, de lieux, ou dépendantes du domaine, par exemple des montants monétaires dans le domaine financier, des noms de protéines ou de médicaments dans le domaine médical. Dans le domaine d'application de cette thèse, la surveillance des événements sismiques, ces entités seront en particulier la magnitude, les coordonnées géographiques de l'événement etc. ;
- La résolution de la coréférence (*Coreference Resolution*) : cette tâche correspond à l'association de plusieurs mentions ou occurrences d'entités référant à la même entité. Cette tâche inclut la mise en correspondance de noms d'entités en prenant en compte leurs variations possibles et la mise en correspondance des pronoms reprenant des entités dans les textes (résolution d'anaphores pronominales). La résolution de coréférence d'entités est aussi appelé suivi d'entité (*Entity Tracking*), par exemple dans le cadre des campagnes ACE ;
- L'identification des attributs associés aux entités (*template element construction*) : cette tâche a pour objet d'associer des informations complémentaires aux entités nommées, en essayant d'extraire les mentions explicites de propriétés associées aux entités. Par exemple, les personnes peuvent se voir associer des alias, des titres, etc ;
- L'identification des relations entre les entités nommées (*template relation construction*) : cette tâche consiste à extraire des relations existant entre deux entités dans un texte ;
- l'identification d'événements (*scenario template construction*) : cette tâche consiste à remplir automatiquement une structure d'information représentée sous la forme d'un formulaire (*template*), associant différents éléments d'information à un événement donné. Par exemple, pour un événement d'acquisition entre deux entreprises, la structure d'information contiendra les noms des deux entreprises (l'entreprise qui achète et celle qui est achetée),

la date, le montant financier etc.

Dans cette séparation entre les différentes tâches, chaque tâche de la liste utilise les résultats des tâches précédentes : par exemple, on a besoin de reconnaître les entités spécifiques pour faire la coréférence entre entités. De même, on a besoin d'avoir des relations entre entités pour construire des structures d'information plus complexes. Néanmoins, les frontières entre les différentes tâches sont parfois ténues. Par exemple, les tâches d'identification d'attributs pour des entités et d'extraction de relations entre entités sont très similaires. En effet, entre un nom de personne et un nom d'entreprise, on peut imaginer avoir une relation *travaille_pour*, qu'on chercherait à extraire des textes dans un cadre d'extraction de relations. Or, cette relation étant statique, on pourrait, en considérant une orientation particulière sur la relation, la voir comme la définition d'un attribut *est_employé_de* associé aux personnes. Ces deux tâches concernent donc l'identification de relations entre les entités. La différence que l'on peut faire est que pour l'identification d'attributs, ces relations sont de nature attributive (la date de naissance est une relation attributive entre une entité de type *Personne* et une entité de type *Date*) alors que dans le cadre de l'extraction de relations pour la construction de *templates*, ces relations peuvent être de nature événementielle (une acquisition entre entreprises est une relation événementielle entre deux entités de type *Organisation*). Une autre façon de voir la différence entre ces relations est de considérer que les relations attributives sont en général des relations valides indépendamment du domaine considéré, alors que les relations événementielles sont relatives au domaine. La forme que peut prendre l'expression de ces relations peut être différente selon leur nature, mais les approches pour faire cette extraction restent similaires et dans le reste de ce travail, nous présenterons les deux tâches comme une seule tâche d'extraction de relations.

De la même façon, la construction de *templates* peut être vue comme une généralisation de l'extraction de relations entre entités, en considérant que le formulaire décrivant un événement est une relation n-aire entre plusieurs entités. On parle parfois d'extraction de relations complexes [McDonald et al., 2005].

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

2.3.2 Architecture type

Nous avons décrit précédemment les différentes tâches en matière d'extraction d'information. Afin de détecter les informations pertinentes contenues dans un document les systèmes d'extraction d'information s'appuient sur des connaissances linguistiques obtenues en appliquant différents traitements : par exemple déterminer la catégorie morpho-syntaxique associée à un mot. L'ensemble de ces traitements linguistiques s'inscrivent dans une architecture type des systèmes d'extraction d'information que nous décrivons ici.

Plus globalement, les systèmes appliquent différents traitements linguistiques au niveau des mots ou des phrases (niveau local) d'un document. Par la suite, les résultats de ces premiers traitements sont agrégés (ou combinés) pour repérer des informations exprimées au-delà d'une seule phrase ou au niveau du document dans son ensemble : par exemple une relation sémantique entre deux entités nommées peut être exprimée sur plusieurs phrases.

[Grishman, 1997] résume le processus d'extraction d'information en trois temps : premièrement, la détection des faits à travers les traitements linguistiques au niveau local ; deuxièmement, le regroupement des faits identifiés avec des faits existants (éventuellement la création de nouveaux faits) ; enfin, les informations liées aux faits pertinents sont transformées pour correspondre au format des *templates*.

Les systèmes d'extraction d'information sont le plus souvent des systèmes modulaires dont les architectures sont très hétérogènes. L'intérêt des systèmes modulaires est qu'ils permettent d'effectuer certaines tâches de façon indépendante. D'autre part, les modules peuvent, de façon naturelle, être appliqués en cascade, afin qu'un module utilise comme entrée la sortie du module précédent [Hobbs, 1993]. Une part importante de ces systèmes s'appuie sur un ensemble de composants qui sont responsables de traitements plus ou moins élaborés et sur lesquels nous nous appuyons pour mettre en avant une architecture commune.

Le premier composant, que nous appelons *initialisation*, est chargé de la détection dans le texte brut des frontières des mots, des fins de phrases (éventuellement des fins de paragraphes) ainsi que de certains traitements linguistiques (lemmatisation, tokenisation, détection des catégories morpho-syntaxiques). Le second module, que nous appelons *identification des liaisons*, est responsable de

la détection puis de la caractérisation des relations entre les mots : il peut s’agir de relations syntaxiques, relations de coréférence ou de l’identification des rôles sémantiques. Ce module permet d’établir des liens entre tous les mots d’un document sans se focaliser uniquement sur les entités nommées. Le troisième module, que nous appelons *traitement des entités nommées*, est dédié au repérage et à la caractérisation des entités nommées, au repérage des relations sémantiques entre les entités (relations inter et intra phrastiques), au remplissage du formulaire à partir des entités pertinentes identifiées.

D’autres travaux [Grishman, 1997; Hobbs and Riloff, 2010; Hobbs, 1993; McCallum, 2005; Turmo et al., 2006] ont proposé des architectures pour ce type de systèmes qui peuvent être assez proches de ce que nous proposons ici. [Hobbs, 1993] suggère une architecture composé de 10 modules. Parmi eux, les modules *découpage en zone de texte* et *prétraitement* effectuent des traitements équivalents à notre *initialisation*. À la différence de notre phase *initialisation*, [Hobbs, 1993; Turmo et al., 2006] suggèrent un module supplémentaire destiné au filtrage des phrases pertinentes d’un document. De plus, la détection des entités nommées est réalisée lors du prétraitement.

[Turmo et al., 2006] propose un module *analyse syntaxique* regroupant deux modules issus de [Hobbs, 1993] (les modules *analyse syntaxique superficielle* et *analyse syntaxique*). Ces deux modules étaient initialement séparés car peu de systèmes (à cette époque) étaient capables de produire des arbres syntaxiques complets : ils retournaient plutôt des fragments d’arbres syntaxiques. Dans notre cadre, le second module regroupe à la fois l’analyse syntaxique et la résolution de la coréférence.

[Grishman, 1997] propose de distinguer les composants d’un système type d’extraction selon deux niveaux : le niveau *local* et le niveau *discursif*. Le niveau *local* concerne les traitements applicables sur chaque phrase indépendamment des autres. Ce niveau effectue des traitements que nous proposons dans le module *initialisation* et certains traitements d’autres modules : la détection des entités nommées dans le module *traitement des entités nommées* et l’analyse syntaxique dans le module *identification des liaisons*.

Le niveau *discursif* concerne des traitements qui s’effectuent au-delà de l’espace de la phrase. Il regroupe la résolution d’anaphores pronominales (et/ou de

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

groupes nominaux) et l'inférence d'éventuels liens ou connaissances sur les entités, que nous incluons dans les modules *identification des liaisons* et *traitement des entités nommées* respectivement. Enfin les *templates* sont générés grâce à la sortie de l'ensemble des traitements.

2.4 Approches pour l'extraction d'information

Les approches utilisées pour résoudre les problématiques en matière d'extraction d'information peuvent être présentées en fonction de plusieurs aspects. Par exemple, dans l'état de l'art proposé par [Sarawagi, 2008], les méthodes peuvent se différencier selon plusieurs axes : on relève par exemple l'opposition entre les approches manuelles et les approches à base d'apprentissage et l'opposition entre les approches à base de règles ou les approches statistiques. Notons qu'en général, les méthodes dites manuelles et les méthodes à base d'apprentissage requièrent toutes les deux une forme d'intervention humaine : soit pour écrire les règles, soit pour annoter des données d'entraînement.

Dans [Hobbs and Riloff, 2010], les approches sont réparties en approches manuelles, c'est-à-dire, reposant sur des patrons et des règles implémentés via des automates à états finis (*Cascaded Finite-State Transducers*) et en approches à base d'apprentissage, reposant sur des algorithmes statistiques. Ici, le choix fait consiste à distinguer les approches *symboliques* (à base de règles ou de patrons) des approches *statistiques*. Dans cette section nous présentons les objectifs et le fonctionnement général de chacune de ces approches. L'application plus précise de ces approches pour les tâches de l'extraction d'information sera présentée plus en détail dans les sections suivantes.

2.4.1 Approches symboliques

Les systèmes symboliques, ou systèmes à base de règles, reposent sur un ensemble de règles (parfois des expressions rationnelles) ou de patrons : une règle est en général, dans ce contexte, assimilable à un patron contextuel, le plus souvent composé de mots et d'autres attributs issus des traitements linguistiques. Les règles peuvent être écrites manuellement par des experts ou alors acquises au-

tomatiquement [Riloff, 1993; Soderland, 1999]. D'un point de vue historique, les premiers systèmes développés pour l'extraction d'information étaient des systèmes symboliques parce que ce genre de système est relativement simple à développer, à interpréter et à modifier pour des humains et qu'il permet d'obtenir un niveau de précision relativement élevé (supérieur à 90 % pour la détection des entités nommées).

Ces systèmes sont généralement composés d'un ensemble conséquent de règles, avec un recouvrement possible entre certaines d'entre elles. Par conséquent, il est nécessaire d'imposer un ensemble de contraintes (ou de politiques) sur leur déclenchement. Il existe plusieurs formalismes de représentation des règles, dont Common Pattern Specification Language (CSPL) [Appelt and Onyshkevych, 1998], WHISK [Soderland, 1999] ou JAPE [Cunningham et al., 2000]. [Muslea, 1999] propose une revue de différents types de patrons d'extraction générés par des systèmes utilisant de l'apprentissage automatique.

Notons que les systèmes symboliques n'excluent pas l'apprentissage. En effet, de tels systèmes peuvent apprendre des règles à partir d'un corpus d'entraînement. Concernant les méthodes d'acquisition de ces règles, [Sarawagi, 2008] distingue l'approche de construction ascendante des règles *Bottom-up rule formation* de l'approche descendante *Top-down rule formation*. La première se sert de règles très spécifiques (donc avec une faible couverture, mais une bonne précision), qui sont généralisées entre elles, les règles spécifiques étant supprimées au fur et à mesure du processus de généralisation. Au final on obtient un ensemble réduit de règles avec une couverture et un niveau de précision équivalent à l'ensemble de départ. L'approche descendante vise à considérer une règle très générique, donc avec une large couverture mais une faible précision, puis à la spécialiser de plus en plus jusqu'à l'obtention d'un niveau de précision acceptable tout en conservant une bonne couverture.

2.4.2 Approches statistiques

Les systèmes symboliques sont caractérisés par une certaine flexibilité (ajout et/ou modification rapide de nouvelles règles) qui leur permet de pouvoir prendre en considération à la fois les cas très fréquents et les cas très rares dans le corpus

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

utilisé pour générer les règles. En revanche, ils sont très liés au domaine ayant servi au développement des règles, ce qui rend leur adaptation à de nouveaux domaines coûteuse : il faut souvent réécrire un nouvel ensemble de règles adaptées. Les méthodes à base d'apprentissage statistique permettent de pallier en partie ce problème. L'idée générale de ces méthodes est de définir un modèle de classification statistique, ou classifieur (*classifier*), capable d'apprendre à associer automatiquement des classes à des éléments.

Les méthodes à base d'apprentissage statistique se distinguent par le degré de supervision qu'elles requièrent, c'est-à-dire par la proportion d'exemples annotés dont elles ont besoin pour fixer les paramètres du modèle. Ainsi, on distingue trois principales approches d'apprentissage par ordre décroissant de leur degré de supervision :

- les approches supervisées correspondent à un fort degré de supervision ;
- les approches semi-supervisées correspondent à un degré de supervision moyen à faible ;
- les approches non supervisées correspondent à un degré de supervision très faible à nul.

Ces différentes approches sont présentées plus en détails dans les sections suivantes.

2.4.2.1 Méthodes supervisées

Dans l'approche supervisée, les paramètres du modèle sont exclusivement établis en fonction des caractéristiques des classes fixées par les exemples d'apprentissage (source de supervision). En pratique, il s'agit d'utiliser un corpus dit d'apprentissage contenant des exemples annotés dans des contextes variés afin d'apprendre les paramètres d'un modèle. Par la suite, lorsque l'on présente un nouvel exemple au modèle, il utilise les valeurs des paramètres établies lors de la phase d'apprentissage pour décider de la catégorie à attribuer.

Les méthodes supervisées se distinguent principalement par la manière de prendre la décision de classification et la manière de modéliser les traits caractéristiques associés aux différentes classes. La sélection d'une classe¹ plutôt

¹Éventuellement plusieurs classes dans le cadre de la classification multi-classes

qu’une autre repose sur l’un des quatre éléments suivants : une probabilité dans le cadre des modèles probabilistes ; une frontière et une fonction noyau dans le cadre des modèles utilisant des séparateurs à vaste marge (SVM) ; une règle dans le cadre des modèles à base de règles ou d’arbres de décisions ; une distance et un nombre k de voisins dans les modèles de k -plus proches voisins.

Nous présentons brièvement dans le tableau 1 quelques algorithmes de classification supervisée, en distinguant les modèles *génératifs* des modèles *discriminants*. Un modèle génératif est un modèle qui repose sur l’estimation de deux probabilités à partir des données d’apprentissage : $P(Y)$, la probabilité d’obtenir une classe donnée ; $P(X|Y)$ probabilité de générer une valeur lorsque l’on observe une catégorie donnée [Mitchell, 1997]. À l’opposé, un modèle discriminant dans le cadre probabiliste repose sur l’estimation d’une seule probabilité : $P(Y|X)$, la probabilité d’obtenir une classe sachant la valeur d’une observation de l’élément que l’on cherche à catégoriser.

Notons que l’ensemble des modèles probabilistes du tableau 1 sont décrits et comparés d’un point de vue plus théorique dans [Klinger and Tomanek, 2007]. Aussi [Bird et al., 2009; Kotsiantis et al., 2007; Mitchell, 1997; Wainwright and Jordan, 2008] présentent des inventaires plus complets en matière d’apprentissage supervisé.

2.4.2.2 Méthodes semi-supervisées

Les algorithmes d’apprentissage supervisé utilisent exclusivement des corpus d’entraînement annotés. Cependant, ce type de corpus est coûteux à produire à cause du temps nécessaire pour réaliser l’annotation manuelle, qui peut être complexe pour les tâches avancées de l’extraction d’information comme le remplissage de *templates*. Les approches semi-supervisées permettent de réduire cette dépendance en utilisant à la fois des exemples annotés et non annotés [Chapelle et al., 2006; Zhu, 2005].

L’idée générale de ces approches est d’apprendre les paramètres du modèle de façon itérative : (i) on commence par entraîner un classifieur (de façon supervisée) sur une partie annotée du corpus d’apprentissage ; (ii) ce classifieur est ensuite utilisé pour étiqueter les exemples non annotés ; (iii) les exemples ayant un

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

Algorithmes	Famille de décision	Description
Bayésien Naïf (<i>Naive Bayes</i>) [Rish, 2001]	Probabiliste	Modèle génératif utilisé pour la classification mono-classe
Maximum d'Entropie (<i>MaxEnt</i>) [Berger et al., 1996]	Probabiliste	Modèle discriminant utilisé pour la classification mono-classe
Chaînes de Markov Cachées (<i>Hidden Markov Models – HMM</i>) [Rabiner, 1989]	Probabiliste	Modèle génératif utilisé pour la classification de séquences, c'est le modèle de séquence associé au <i>Naive Bayes</i>
Champs Conditionnels Aléatoires (<i>Conditional Random Fields – CRF</i>) [Lafferty et al., 2001; Sutton and McCallum, 2007]	Probabiliste	Modèle discriminant utilisé pour la classification de séquences, c'est le modèle de séquence associé au <i>MaxEnt</i>
Arbres de décision C4.5 [Quinlan, 1996]	Règles	Modèle utilisé pour la classification mono-classe
Séparateurs à vaste marge (<i>Support Vector Machine – SVM</i>) [Burges, 1998; Cortes and Vapnik, 1995]	Plans séparateurs	Modèle utilisé pour la classification mono et multi classes

TAB. 1 – Exemples d'algorithmes d'apprentissage supervisé

score de confiance au-dessus d'un seuil prédéfini sont ensuite réinjectés parmi les exemples annotés ; (iv) le processus est répété tant qu'il y a des exemples non annotés dans le corpus d'apprentissage. Ce procédé d'apprentissage semi-supervisé est appelé auto-apprentissage (*self-teaching* ou *bootstrapping*) [Zhu, 2005]. Dans la même contribution, l'auteur présente le co-apprentissage (*co-training*) qui est un cas particulier d'apprentissage semi-supervisé, qui repose sur l'entraînement itératif de deux ou plusieurs classifieurs, chacun utilisant un ensemble d'attributs différent. À la fin du processus, les modèles sont combinés.

2.4.2.3 Méthodes non supervisées

De même que les approches semi-supervisées, les approches non supervisées visent à réduire la quantité d'exemples annotés nécessaire pour l'apprentissage. Dans ce cas précis, l'objectif est de ne pas en utiliser du tout. Cet objectif passe par le regroupement automatique des exemples non annotés pour permettre l'émergence de classes homogènes qui traduisent l'organisation naturelle des données [Xu and Wunsch, 2005].

Ainsi, l'idée générale de ces approches de regroupement (ou *clustering*) est de regrouper les exemples non annotés en fonction des caractéristiques qu'ils partagent. Les groupes ainsi constitués sont appelés *clusters*. Un algorithme de *clustering* s'appuie sur une mesure de similarité entre les exemples et cherche à optimiser deux facteurs :

la cohésion à l'intérieur d'un *cluster* il s'agit de maximiser les similarités entre tous les éléments d'un même *cluster* ;

la disjonction entre les *clusters* il s'agit de maximiser les dissimilarités entre les éléments d'un *cluster* et ceux des autres *clusters*.

Les méthodes de *clustering* se distinguent par : (i) la manière de calculer la similarité (ou distance) entre les éléments à regrouper, (ii) la manière de construire les *clusters*, (iii) la manière de définir le nombre de *clusters* à construire. Concernant le deuxième point, les méthodes peuvent globalement se diviser entre celles qui cherchent à construire une arborescence entre les exemples (*clustering* hiérarchique, *hierarchical clustering*) et celles qui cherchent à répartir les exemples

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

en un nombre déterminé de *clusters* sans faire l'hypothèse de l'existence d'une arborescence (partitionnement, *partitionnal clustering*).

Concernant le premier point, le calcul de la similarité entre les éléments se fait en utilisant des mesures de distance ou de similarité entre tous les éléments deux à deux, qui s'appuient en général sur une représentation ensembliste ou vectorielle des exemples. Des exemples de mesures généralement appliquées sont : le cosinus de l'angle entre les vecteurs représentant les exemples, la distance euclidienne, la distance de Mahalanobis, la distance de Manhattan, etc. [Xu and Wunsch, 2005]. Enfin les résultats fournis servent à construire une matrice de similarité à partir de laquelle le *clustering* est effectué.

Il existe de nombreux algorithmes de *clustering*. Nous présentons dans le tableau 2 quelques exemples d'algorithmes, repris de [Xu and Wunsch, 2005]. Des inventaires plus complets sur les méthodes non supervisées sont proposés dans [Berkhin, 2002; Xu and Wunsch, 2005].

Algorithmes	Type de <i>clustering</i>
Plus proche voisin (<i>Single linkage</i> ou <i>Nearest neighbour</i>)	Hiérarchique agglomératif
K-Means	Diminution de l'erreur quadratique
Triangulation de Delaunay	Théorie des graphes
Sous graphes connectés (<i>highly connected subgraphs</i>) [Khuller, 1997]	Théorie des graphes
Markov Clustering (MCL) [Dongen, 2000]	Théorie des graphes

TAB. 2 – Méthodes non supervisées (*clustering*)

2.5 Reconnaissance des entités nommées

2.5.1 Présentation du problème

Dans [Ehrmann, 2008] la définition proposée pour une entité nommée est : «*toute expression linguistique qui réfère à une entité unique du modèle de manière autonome*». Dans la définition, le «*modèle*» fait référence à une description préalable des informations pertinentes pour un contexte applicatif donné. Cela revient

le plus souvent, en pratique, à déterminer la liste des types d'informations, en rapport avec un domaine particulier, que l'on souhaite extraire.

Ainsi ce modèle permet d'inclure (ou d'exclure) des types d'informations que l'on cherche à détecter automatiquement : par exemple, il peut s'agir de retrouver des dates, des noms de personnes/sociétés ou des quantités. En d'autres termes, la tâche de reconnaissance d'entités nommées consiste à repérer puis à catégoriser des expressions lexicales dans un texte. De façon intuitive, on pourrait penser à construire un dictionnaire gigantesque dans lequel on associerait toutes les expressions lexicales possibles à un type d'entité donné. Il suffirait alors de parcourir les documents à la recherche de ces expressions pour identifier les entités.

Néanmoins, cette approche n'est pas envisageable car une même valeur d'entité peut appartenir à plusieurs types d'entités distincts. Dans ce cas, seul le contexte autour de l'entité permet de choisir le type d'entités le plus pertinent. Par exemple «*Renault*» peut faire référence à soit à la société, soit à un nom de famille. Il est donc nécessaire de prendre en compte les informations présentes dans le contexte antérieur et le contexte postérieur de l'entité afin de supprimer l'ambiguïté.

En pratique, le repérage des entités peut souvent être divisé en deux sous-problèmes : (i) définir les frontières de l'entité (déterminer quels mots composent l'entité), (ii) définir le type qui doit lui être associé. L'exemple de la figure 2.1 illustre la reconnaissance d'entités nommées dans le contexte financier. Dans ce contexte applicatif, les informations à distinguer sont de type *date* ($< D >$), *organisation* ($< O >$) et *montant de transaction* ($< M >$).

2.5.2 Les approches pour la détection des entités nommées

L'identification des entités nommées (EN) constitue une tâche fondamentale en matière d'extraction d'information. Cette tâche a été l'objet de plusieurs campagnes d'évaluations, parmi elles MUC-7¹, CoNLL-2002², CoNLL-2003³, ESTER⁴, etc. Ces travaux ont permis de traiter cette problématique dans divers

¹http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html

²<http://www.clips.ua.ac.be/conll2002/ner/>

³<http://www.clips.ua.ac.be/conll2003/ner/>

⁴http://www.afcp-parole.org/camp_eval_systemes_transcription/

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

Rachat de <O>BEA Systems</O> par <O>Oracle</O> : c'est fait. Au prix fort.		
Le géant de la base de données annonce en effet ce <D>mercredi 16 janvier</D> avoir finalisé un accord en vue du rachat de <O>BEA Systems</O>. <O>Oracle</O> a accepté les conditions de </O>BEA Systems</O>, fixant le montant du rachat à <M>8,5 milliards de dollars</M>.		
Il s'agira d'un des plus grands coups de Larry Ellison, le p-dg d'<O>Oracle</O>. Le précédent record était de <M>10 milliards </M> pour l'acquisition de <O>PeopleSoft</O> en <D>2004</D>.		
Date : <D>	Organisation : <O>	Montant : <M>

FIG. 2.1 – Exemple de reconnaissance d'entités nommées

domaines (terrorisme, gestion de successions), langues (anglais, espagnol, japonais, français) et type de documents (articles de presse, courriers électroniques) [Nadeau and Sekine, 2007].

Ainsi, de nombreux systèmes faisant de la reconnaissance d'entités nommées (*Named Entity Recognition* – *NER*) ont été développés. On peut noter qu'à partir du milieu des années 90, les systèmes à base de règles conçues manuellement ont fait place aux systèmes statistiques. Ce changement s'est effectué en partie grâce à un point fort des systèmes statistiques : ils sont plus rapidement adaptables à de nouveaux domaines que leurs équivalents symboliques.

2.5.2.1 Utilisation de règles

Nous présentons de façon succincte quelques propriétés d'un système symbolique de reconnaissances d'entités nommées. Le tableau 3 reprend les principaux attributs utilisés par les systèmes à base de règles pour détecter les entités nommées [Sarawagi, 2008].

L'exemple de la figure 2.2 illustre la construction de règles pour la détection des entités de type MONTANT à partir de deux phrases. Notons que dans cette même figure, les exemples sont utilisés pour créer une expression rationnelle qui recouvre les deux formes d'apparition de l'entité. La démarche suivie dans ce cas est une démarche généralisante : on combine les règles obtenues pour chaque

Types d'attributs	Description
Forme de surface	Le/les mots qui composent l'entité
Capitalisation	Divers indices sur la forme de capitalisation des mots (première lettre en majuscule, tout le mot en majuscule), présence de caractères spéciaux, présence de ponctuation, etc.
Catégorie morpho-syntaxique	La catégorie morpho-syntaxique des mots
Présence dans une liste fermée de valeurs (gazee-ter)	Il s'agit de vérifier la présence de l'entité dans une liste connue d'entités (exemple liste de noms de lieux)
Autres pré-traitements	Il s'agit d'utiliser les attributs provenant d'autres traitements linguistiques précédents (par exemple l'analyse syntaxique)

TAB. 3 – Attributs des règles utilisés par les systèmes symboliques

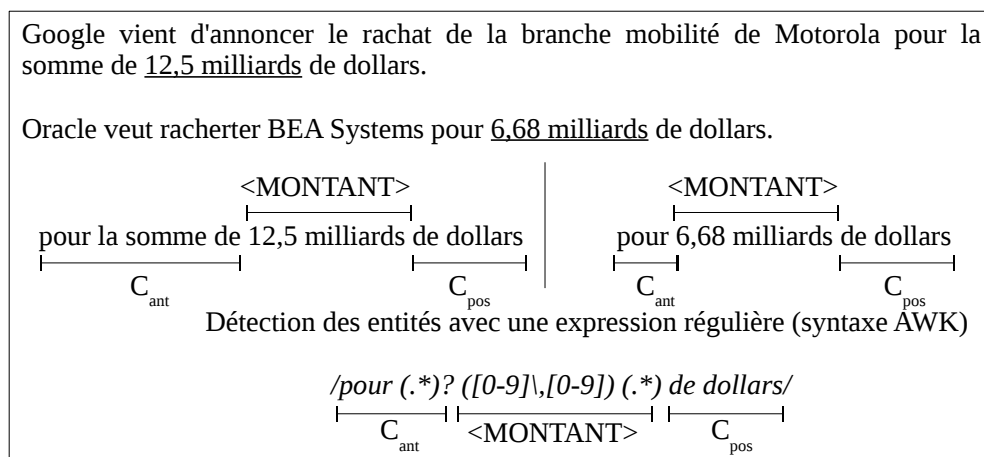


FIG. 2.2 – Exemple de patron pour la détection des montants des transactions

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

passage contenant des mentions d'entités afin d'obtenir des règles recouvrant le plus de formes possible. La règle générée dans l'exemple précédent est assez basique puisqu'elle utilise très peu d'informations linguistiques. En pratique les systèmes symboliques de référence utilisent des règles plus élaborées qui peuvent soit être créées manuellement comme dans GATE [Cunningham et al., 2002], soit construites automatiquement, par exemple dans les systèmes AutoSlug [Riloff, 1993], Rapiér [Califf and Mooney, 1997], LP² [Ciravegna, 2001] ou plus récemment TEXTMARKER [Atzmüller et al., 2008], [Reiss et al., 2008].

Les systèmes à base de règles permettent également d'introduire de façon simple des connaissances encyclopédiques existantes sous la forme de listes ou de dictionnaires qui intègrent des entités ou des éléments d'entités déjà connus. On peut, par exemple, introduire des listes de noms de villes connues pour la reconnaissance des noms de lieux, ou utiliser des listes de prénoms connus pour la reconnaissance des noms de personnes. Notons que ce type de connaissances peuvent aussi être intégrées dans le cadre des systèmes par apprentissage statistique en tant qu'attributs d'une observation.

2.5.2.2 Utilisation d'apprentissage statistique

Les systèmes à base d'apprentissage statistique abordent la reconnaissance des entités comme un problème de classification. Il s'agit alors d'apprendre les paramètres d'un modèle capable de déterminer pour chacun des mots d'un document s'il appartient à une entité, et si oui, quel type d'entité doit lui être associé.

En ce qui concerne la délimitation de la frontière des entités, la représentation adoptée vise à spécifier si un mot fait ou non partie d'une entité et, dans le cas positif, préciser s'il amorce l'entité ou s'il fait partie du corps de l'entité. Les classes à identifier pour les mots sont alors les classes *Begin* pour un mot débutant une entité, *Inside* pour un mot étant à l'intérieur d'une entité et *Outside* pour un mot n'appartenant pas à une entité. Ce modèle est connu sous le nom de modèle *BIO* (*Begin* – *Inside* – *Outside*) [Sarawagi, 2008]. Néanmoins, cette méthode ne distingue pas les mots qui terminent une entité des mots qui sont à l'intérieur d'une entité. La méthode *BCEO* (*Begin* - *Continue* - *End* - *Other*) permet de faire cette distinction [Sarawagi, 2008].

Le tableau 4 montre un exemple d’application de ces deux méthodes. Notons que, dans l’exemple, l’annotation prend en compte à la fois la détermination de la frontière d’une entité et la détermination du type d’entité.

Mots	Notation BIO	Notation BCEO
Le	O	O
précédent	O	O
record	O	O
était	O	O
de	O	O
10	B-MONTANT	B-MONTANT
milliards	I-MONTANT	E-MONTANT
pour	O	O
l’	O	O
acquisition	O	O
de	O	O
PeopleSoft	B-ORGANISATION	B-ORGANISATION
le	O	O
31	B-DATE	B-DATE
décembre	I-DATE	C-DATE
2004	I-DATE	E-DATE
.	O	O

TAB. 4 – Exemple de délimitation des frontières des entités nommées avec les modèles BIO et BCEO

Pour ce qui est des types d’entités à associer aux mots, le problème est traité par des approches génératives ou discriminantes. Le plus souvent, les modèles sont adaptés à la classification de séquences, afin de prendre en compte l’aspect séquentiel des phrases et les dépendances éventuelles entre les types d’entités (par exemple il y a une probabilité non nulle que les chiffres suivant une entité de type organisation soient une date). Il s’agit alors de trouver, pour une phrase x contenant n mots notés $x = \{x_1, \dots, x_n\}$, la séquence cachée des types d’entités nommées $y = \{y_1, \dots, y_n\}$ qui lui correspond. L’intérêt des modèles de séquence est qu’ils permettent de prendre en compte les dépendances entre les catégories d’entités successives y_i et y_{i-1} . Selon les modèles, le choix d’une catégorie d’entité nommée se fait à partir des observables x_i (ex. HMM), ou à partir d’un ensemble de traits caractéristiques sur les x_i (aussi appelés *features*) (ex. CRF). Les

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

features permettent de modéliser les propriétés intrinsèques des mots de l'entité et des mots dans le voisinage de l'entité.

Les *features* généralement utilisés pour la détection des entités sont de trois types [Sarawagi, 2008] :

- les *features* sur les mots, ou les formes de surface (ex. la catégorie morpho-syntaxique associées aux mots) ;
- les *features* orthographiques (ex. le mot commence-t-il par une lettre majuscule ?) ;
- les *features* utilisant des dictionnaires de valeurs (ex. le mot appartient-il à une liste de noms de pays ?).

Plusieurs systèmes de reconnaissance des entités nommées ont été construits à partir de tels *features*, en utilisant des modèles différents : algorithme MaxEnt [Borthwick, 1999], algorithme HMM [Bikel et al., 1999; Seymore et al., 1999], algorithme CMM (*Conditional Markov Model*) [Malouf, 2002], algorithme MEMM (*Maximum Entropy Markov Model*) [McCallum et al., 2000], algorithme CRF [Lafferty et al., 2001]. [Nadeau and Sekine, 2007] présente un inventaire des approches utilisées pour la détection des entités nommées, en particulier les approches statistiques, entre 1991 et 2006.

2.6 Coréférence entre entités nommées

2.6.1 Présentation du problème

Dans [Ng, 2010], la résolution de la coréférence est définie comme une tâche dont l'objectif est de déterminer dans un texte les groupes nominaux faisant référence au même objet. L'objet qui est repris par plusieurs expressions est appelé *antécédent* et les expressions lexicales qui pointent vers lui sont appelées *anaphores*. Le terme de *coréférence* est utilisé lorsque l'anaphore et son antécédent font référence au même objet [Mitkov, 1999]. Par exemple dans : «*L'éditeur de bases de données et de progiciels Oracle retire son offre de rachat de BEA Systems. Il avait fixé au 28 octobre la date limite pour que BEA accepte sa proposition de 6,7 milliards de dollars.*», le pronom *il* et l'antécédent *Oracle* sont en relation de coréférence.

En résumé, la résolution de l'anaphore revient à trouver l'antécédent associé à une anaphore. Le langage naturel étant très riche, les phénomènes de référence peuvent s'exprimer de différentes manières en fonction du type d'antécédent (pronom ou groupe nominal défini) ou en fonction de la position de l'antécédent (intraphrastique ou interphrastique). Le nombre de relations de coréférence entre deux entités dépend de la langue et du degré de granularité considéré. Les premiers travaux issu de MUC-6¹ se sont concentrés sur 7 types de coréférence (apposition, métonymie, coréférence pronominale, groupes nominaux renvoyant à une même catégorie d'objet, reprise d'un groupe nominal par une forme définie, utilisation de pluriel pour faire référence à une catégorie d'objet, reprise d'une grandeur par sa valeur). Par la suite, [Bagga, 1998] montre que sur 11 types de coréférence identifiés comme les plus fréquents sont (*cf.* exemple 1, page 29) les variations de noms, (*cf.* exemples 2 et 4) les coréférences pronominales, et (*cf.* exemple 3) la répétition du groupe nominal. Dans le cadre de l'extraction d'information et du traitement des entités nommées, le cas de la variation des noms propres (exemple : SAP, System Analysis and Program Development, SAP AG) est particulièrement traité pour la coréférence entre entités nommées.

De nombreux travaux se sont focalisés sur la résolution des anaphores pronominales, ce type d'anaphore concerne tous les pronoms à la troisième personne (singuliers/pluriels) et vise plus particulièrement les pronoms personnels (*cf.* exemple 2), démonstratifs (*cf.* exemple 4), réfléchis (*cf.* exemple 5) ou possessifs (*cf.* exemple 6). Sous un autre angle les anaphores dites définies sont portées par des groupes nominaux définis au lieu de pronoms. En particulier, elles peuvent utiliser les liens sémantiques entre les antécédents et les anaphores : lien de méronymie/holonymie (*cf.* exemple 7) ou lien d'hyperonymie/hyponymie (*cf.* exemple 8).

1. *éditeur, éditeur de bases de données, firme, etc.*

Formes alternatives du même groupe nominal

2. *[Google]* a annoncé qu'[il] va racheter Motorola Mobility.

Le pronom personnel *il* reprend *Google*

3. Oracle le *[géant américain]* des bases de données a proposé de racheter BEA

¹http://cs.nyu.edu/cs/faculty/grishman/C0task21.book_1.html

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

Systems pour 6,68 milliards de dollars. L'offre du *[géant américain]* à 17 dollars par action a été rejetée.

Reprise à l'identique de *géant américain*

4. Ce rachat apparaît comme stratégique. *[Celui-ci]* intervient après l'acquisition de Business Objects par SAP en Octobre dernier.

Celui-ci reprend *le rachat*

5. *[Le rachat]* *[s]*'est déroulé après l'accord des actionnaires.

se reprend *le rachat*

6. *[Oracle]* a dépensé plus de 20 milliards de dollars en acquisition pour étoffer *[son]* catalogue.

Son reprend *Oracle*

7. Oracle l'éditeur de *[logiciel]* propose de racheter BEA Systems. L'éditeur de *[bases de données]* a fait une offre de plusieurs milliards de dollars.

logiciel est en relation méronymique avec *bases de données*.

8. Il s'agit d'une des plus grandes *[acquisitions]* du PDG d'Oracle Larry Ellison.

acquisitions est en relation hyponymique avec *le rachat* : une acquisition est une catégorie de rachat.

Dans le cadre de la construction de *templates*, l'intérêt du traitement de la coréférence est de permettre d'identifier les chaînes de coréférence. Ces chaînes sont équivalentes à des regroupements d'expressions ayant un lien de coréférence entre elles. Par exemple, l'extrait de la figure 2.3 contient trois chaînes de coréférence C1-C3-C4-C5-C6, C2-C9 et C7-C8 entre des anaphores nominales. Nous détaillons dans la section suivante les méthodes utilisées pour la résolution de la coréférence.

2.6.2 Les approches pour la résolution de coréférence

De même que pour la reconnaissance des entités nommées, la résolution de coréférence a été l'objet de plusieurs campagnes d'évaluations, parmi elles MUC-

Pendant la crise, les affaires continuent. *[Google]*_{C1} le prouve en rachetant pour 8,6 milliards d'euros *[Motorola Mobility]*_{C2}, afin de *[se]*_{C3} renforcer dans un secteur où *[il]*_{C4} a fait irruption fin 2007 grâce au système d'exploitation Android. *[Google]*_{C5} peut se permettre un tel investissement en période de crise financière : *[son]*_{C6} trésor de guerre est estimé à quelque *[27,6 milliards d'euros]*_{C7}. *[Cela]*_{C8} lui permet de surpayer sans sourciller ce *[fabricant de téléphones mobiles]*_{C9}, avec un prix d'achat des actions 63% plus élevé que leur dernier cours de Bourse.

FIG. 2.3 – Exemple de chaînes de coreférences

6¹, MUC-7², ARE³ et plus récemment SemEval-2010⁴.

Les approches pour la résolution de coréférence sont généralement classées en deux catégories, d'une part les approches ayant un fondement linguistique [Mitkov, 1999; Poesio et al., 2010] et d'autre part les approches à base d'apprentissage statistique [Ng, 2010]. Nous décrivons brièvement ces deux approches dans cette section.

2.6.2.1 Utilisation d'approches linguistiques

La résolution de coréférence à partir d'approches linguistiques est un problème qui a été étudié de façon exploratoire entre la fin des années 60 et le milieu des années 90. Les approches développées cherchaient alors à résoudre le problème à partir de trois types d'information distincts : les informations syntaxiques, les connaissances du monde réel, et les informations discursives.

La première approche linguistique performante a été définie à la fin des années 70 [Hobbs, 1978]. Elle se concentrait sur la résolution de coréférence pronominale en utilisant les informations syntaxiques produites à partir de phrases manuellement annotées. L'idée principale était de sélectionner le meilleur antécédent possible en appliquant certaines contraintes sur les groupes nominaux, par exemple

¹http://cs.nyu.edu/cs/faculty/grishman/C0task21.book_1.html

²http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html

³Anaphora Resolution Exercise : <http://clg.wlv.ac.uk/events/ARE/>

⁴Semantic Evaluation : <http://stel.ub.edu/semeval2010-coref/>

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

les accords en genre ou en nombre, la proximité (en nombre de phrases) par rapport à l'anaphore, etc.

Par la suite, d'autres approches reposant aussi sur les informations syntaxiques [Kennedy and Boguraev, 1996; Lappin and Leass, 1994] ont été appliquées à plus large échelle, et cette fois à partir d'un résultat d'analyse syntaxique produit par un analyseur automatique. À l'opposé, d'autres approches telles que [Baldwin, 1997; Mitkov, 1998] ont cherché à minimiser l'apport de connaissances extérieures ou des informations syntaxiques utilisées. L'intérêt de ce type de méthodes est d'obtenir un bon niveau de résultat avec des méthodes simples qui reposent sur un nombre de traitements linguistiques limité.

2.6.2.2 Utilisation d'approches statistiques

L'usage de méthodes statistiques pour la résolution de coréférence s'est répandu au milieu des années 90 alors que les campagnes telles que MUC-6 et MUC-7 permettaient d'accéder à une quantité importante de données annotées. Dès lors, de nombreuses approches, dans un premier temps supervisées [Connolly et al., 1994; Soon et al., 2001; Yang et al., 2003], puis non-supervisées [Cardie and Wagstaff, 1999; Ng, 2008] ont été développées.

Les approches supervisées pour la résolution de coréférence se répartissent en trois grandes catégories : la première s'appuie sur un classifieur binaire pour déterminer si un couple (anaphore, antécédent potentiel) est en situation de coréférence. Le principe est de comparer une anaphore avec chacun des antécédents potentiels se trouvant dans les phrases précédentes. Les couples annotés positivement par le classifieur sont alors considérés comme coréférents et constituent une chaîne de coréférence. Du fait des comparaisons deux à deux entre les anaphores et les antécédents, ce type de modèle est appelé *mention-pair* [Ng and Cardie, 2002; Soon et al., 2001].

Dans ce type de modèle, la décision de classification est prise pour deux entités données sans tenir compte des précédentes décisions, ce qui peut entraîner des contradictions dans les chaînes de coréférence. Les approches de type *entity-mention* ont été proposées afin de diminuer ces incohérences. Dans ces approches, la méthode consiste à intégrer une anaphore dans une chaîne de coréférence déjà

constituée en s’assurant qu’elle est compatible avec les autres antécédents de la chaîne.

Un autre inconvénient de la classification binaire sur laquelle repose l’approche *mention-pair* est que le résultat ne permet pas d’ordonner les antécédents potentiels les uns par rapport aux autres. En effet, le classifieur évalue la probabilité qu’un antécédent soit associé à une anaphore, et non pas le meilleur antécédent parmi tous antécédents potentiels. Pour pallier ce défaut, les approches de type *mention-ranking* proposent de trier tous les antécédents potentiels en fonction de leurs probabilités d’être en relation de coréférence avec l’anaphore. Suite à ce classement, l’antécédent potentiel ayant le meilleur rang est conservé. Le lecteur pourra se référer aux travaux de Vincent Ng pour des inventaires détaillés sur la résolution de la coréférence à partir d’approches supervisées [Ng, 2010] et non supervisées [Ng, 2008].

2.7 Extraction de relations entre entités nommées

2.7.1 Présentation du problème

L’objet de la tâche d’extraction de relations entre entités nommées (ou *template relation construction – template relation production*) est d’établir des liens sémantiques entre les entités nommées. Historiquement, cette tâche a été définie lors de la campagne MUC 7 avec pour objet d’identifier trois relations prédéfinies (*LOCATION_OF*, *EMPLOYEE_OF*, *PRODUCT_OF*) entre les formulaires d’attributs extraits sur les entités. Par la suite, les travaux se sont développés en augmentant le nombre de relations à extraire (24 relations dans la tâche *Relation Detection and Characterization* de la campagne ACE ¹), et en changeant le domaine d’application des relations, par exemple le biomédical : campagnes LLL², ou plus récemment i2b2³. Le tableau 5 présente quelques relations sémantiques ainsi que leurs types pour l’extrait ci-dessous, pour l’extraction d’information

¹<http://projects.ldc.upenn.edu/ace/docs/EnglishRDCV4-3-2.PDF>

²<http://genome.jouy.inra.fr/texte/LLLchallenge/>

³<https://www.i2b2.org/NLP/Relations/>

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

dans le domaine sismique :

Un [*tremblement de terre*] de magnitude [*5*] sur l'échelle ouverte de Richter s'est produit [*dimanche*] dans le [*nord-ouest de l'Iran*], a rapporté la télévision d'état sans fournir de bilan à ce stade. Le [*séisme*] a touché la ville de [*Khalkhal*], dans la province d'Ardebil, à [*16 H 41*] locales (12 H 11 GMT), selon la même source.

Relation (R)	Entité (e_1)	Entité (e_2)
lieu d'un événement	tremblement de terre (Événement)	nord-ouest de l'Iran (Lieu)
date d'un événement	tremblement de terre (Événement)	dimanche (Date)
lieu d'un événement	séisme (Événement)	Khalkhal (Lieu)
date d'un événement	séisme (Événement)	16 H 41 (Heure)

TAB. 5 – Exemple de relations entre entités nommées

Les exemples de relations du tableau 5 montrent des relations typées binaires, qui impliquent donc seulement deux entités nommées. Néanmoins, les relations entre les entités ne sont pas toujours binaires et peuvent impliquer plus de deux entités : par exemple dans la dernière phrase, il y a une relation ternaire entre les entités *séisme* – *Khalkhal* – *16h41*. De nombreux travaux abordent le problème des relations binaires entre les entités, mais peu se sont intéressés aux relations d'ordre supérieur, parmi eux [Afzal, 2009; Liu et al., 2007; McDonald et al., 2005]. Les relations complexes entre un sous-ensembles d'entités peuvent par ailleurs être vues comme des remplissages de *templates* partiels, qui rendent donc l'étape de construction de *template* plus immédiate par fusion des *templates* partiels. De façon théorique, un *template* complet peut même être vu comme une seule relation complexe regroupant toutes les entités concernées.

Un autre aspect dont il faut tenir compte pour extraire des relations est la distance entre les entités. Dans les exemples du tableau 5, toutes les relations sont exprimées dans la même phrase, cependant dans l'exemple précédent on peut noter qu'il existe une relation indirecte entre la magnitude 5 du séisme et la ville de *Khalkhal* qui a été touchée. Cette relation indirecte est mise en évidence en tenant compte du lien de coréférence entre *tremblement de terre* et *séisme* et aussi de la relation sémantique entre la magnitude 5 et le *tremblement de terre*. Plus

généralement l'exemple illustre l'intérêt d'utiliser les liens de coréférence entre les entités nommées pour l'extraction de relations sémantiques. Dans cette optique [Chan and Roth, 2010] suggèrent d'exploiter les liens de coréférence entre entités pour l'extraction de relations, ils proposent aussi d'intégrer des connaissances extérieures venant d'une source encyclopédique (Wikipedia).

La problématique d'extraction de relations entre entités nommées recouvre en fait plusieurs problématiques différentes, selon le degré d'information disponible en entrée du système :

- dans un premier cas, on cherche à déterminer le type de relation R à attribuer au couple d'entités (e_1, e_2) ;
- dans un deuxième cas, on cherche à déterminer l'entité e_2 en connaissant le type de relation R et l'entité e_1 ;
- un troisième cas d'application de l'extraction de relations concerne le domaine dit ouvert : les systèmes d'extraction de relations entre entités sont souvent, dans la pratique, limités à un nombre prédéterminé de relations et sont le plus souvent adaptés à des domaines spécifiques. Mais en domaine ouvert, on cherche à extraire autant de relations que possible sans *a priori* ni sur le type de la relation R , ni sur le type des entités e_1 et e_2 ;
- dans un quatrième cas, on cherche à déterminer si une relation existe ou non entre deux entités sans pour cela s'intéresser au type de la relation.

Cette section aborde l'extraction de relations dans le cas le plus fréquent, c'est-à-dire où l'on cherche à extraire une relation typée entre deux entités de types prédéfinis. Le chapitre 5 présente plus en détail le deuxième cas d'utilisation de l'extraction de relation, qui en fait une évolution du problème du question-réponse. La problématique du question-réponse a pour objet de retrouver dans une collection de documents une réponse à une question formulée en langage naturel. À la différence de la recherche d'information, la question ne se limite pas à un ensemble de mots-clés. De plus, il ne s'agit pas uniquement de retrouver un ensemble de documents pertinents mais la ou les réponses pertinentes contenues dans ces documents.

2.7.2 Les approches pour l'extraction de relations entre entités nommées

Le point commun des approches pour l'extraction de relations est qu'elles cherchent à capturer les indices contenus dans le voisinage des entités : en avant, entre, ou après les deux entités. Ce contexte sert à faire le lien entre les deux entités et le type de relations que l'on veut caractériser.

Les stratégies d'extraction de relations commencent généralement par identifier les formes alternatives du contexte entre/avant/après les deux entités (pour un type de relations donné). Ce contexte est ensuite modélisé, en général sous la forme de patrons lexico-syntaxiques ou à l'aide de classifieurs statistiques. Les patrons ou classifieurs sont ensuite appliqués à de nouveaux exemples afin d'extraire les relations pour le type de relation concerné. D'autres approches existent pour l'extraction de relations, comme, par exemple, l'utilisation d'heuristiques spécifiques, comme dans le système *RelEx* [Fundel et al., 2007]. Dans ce système, la sélection des relations repose sur quatre règles qui sont appliquées à des arbres de dépendances afin d'extraire des interactions entre des protéines à partir de résumés d'articles scientifiques. Dans les approches à base de patrons ou de classifieurs, les indices les plus fréquemment utilisés sont pour la plupart issus de trois types d'informations linguistiques : les formes de surfaces, les catégories morpho-syntaxiques, les informations syntaxiques (arbres syntaxiques ou graphes de dépendances).

Dans cette section, nous nous concentrons sur les approches d'extraction de relations dans un contexte supervisé mais cette extraction peut également être faite de façon non supervisée comme dans les travaux de [González and Turmo, 2009; Hasegawa et al., 2004; Hassan et al., 2006; Shinyama and Sekine, 2006]. Nous distinguons les approches supervisées pour l'extraction de relations en fonction du mécanisme qu'elles utilisent : soit des patrons de relations soit des classifieurs statistiques.

2.7.2.1 Approches à base de patrons

Les approches à base de patrons de relations datent du début des années 90. Ces approches cherchaient à utiliser un minimum de connaissances extérieures

tout en étant le moins dépendantes possible du genre des textes. Dans cette perspective, ce type d'approche cherche à capturer, par des patrons, les redondances dans les phrases qui contiennent les relations.

Ainsi un des premiers travaux sur l'extraction de relations avait pour but d'extraire des relations d'hyponymie entre des termes [Hearst, 1992], en utilisant des patrons lexico-syntaxiques pour capturer des relations entre des groupes nominaux au sein de la même phrase. Son approche utilise comme éléments de départ des patrons lexico-syntaxiques qui sont écrits manuellement. La phase suivante consiste à collecter des phrases contenant des occurrences de la relation d'hyponymie à partir d'instances de la relation (ex. : *England, country*).

Les occurrences collectées servent ensuite de base pour écrire de nouveaux patrons à partir des points communs observés dans les occurrences. Enfin, les nouveaux patrons sont appliqués pour retrouver de nouvelles instances, et le processus est réappliqué de façon itérative depuis la phase de collecte des occurrences ci-dessus. En résumé, la méthode proposée par Hearst utilise une forme d'auto-apprentissage qui part d'instances connues de relations pour découvrir de nouvelles instances. L'exemple ci-dessous reprend un exemple de patron lexical issu de [Hearst, 1992].

Texte : ... *most European countries, especially France, England and Spain.*

Patron : *NP {,} especially {NP,* {or|and} NP*

Relations d'hyponymie : (France, European country), (England, European country), (Spain, European country)

Dans la même perspective, le système DIPRE (*Dual Iterative Pattern Relation Extraction*) proposé par [Brin, 1999] permet d'extraire des relations entre les titres de livres et leurs auteurs. DIPRE utilise comme documents des pages web renvoyées par un moteur de recherche et des patrons sous forme d'expressions rationnelles. La démarche suivie par DIPRE est aussi fondée sur une forme d'auto-apprentissage : il s'agit d'utiliser des instances de relations sous la forme de couples d'entités pour retrouver des occurrences de relations contenues dans des pages web, ces occurrences servent à générer des patrons représentatifs des relations (les patrons sont évalués pour s'assurer qu'ils aient une couverture suffisante). Par la suite les patrons sont appliqués afin de trouver de nouvelles instances de relations. Ces nouvelles instances sont réintroduites dans le système

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

pour une nouvelle itération. Le processus se termine lorsque le nombre d'instances renvoyé par le système est supérieur à une limite prédéfinie.

À la différence de Hearst, [Brin, 1999] utilise des patrons qui sont générés automatiquement : il s'agit d'exploiter les régularités contenues dans les occurrences de relations pour créer des patrons. Pour cela, les occurrences de relations sont découpées automatiquement en dissociant les mots apparaissant entre les deux entités qui servent d'exemples de ceux apparaissant avant la première entité (préfixe) et de ceux apparaissant après la deuxième entité (suffixe). Par la suite, les occurrences de relations dont les parties préfixe et suffixe sont similaires sont regroupées et le contexte entre les deux entités est utilisé pour construire une expression régulière. Pour illustration considérons les phrase suivantes pour la construction d'un patron (on fait l'hypothèse qu'elles sont renvoyées par un moteur de recherche) :

- (i) *[The Lord of the Rings]_{e1} is a high fantasy epic written by [J. R. R. Tolkien]_{e2}.*
- (ii) *[The Silmarillion]_{e1} is a collection of ^{context} [J. R. R. Tolkien]_{e2} 's mythopoeic works, ^{suffix} edited and published posthumously ...*
- (iii) *... better known by its abbreviated title _{prefix} [The Hobbit]_{e1} , is a fantasy novel written by _{context} [J. R. R. Tolkien]_{e2}.*

Suite au processus de génération des patrons, on obtient une expression rationnelle de la forme «*is a(. * ?) (written by) ?*».

À la manière de Brin, [Agichtein and Gravano, 2000] propose un système d'extraction de relations nommé *SnowBall* qui s'appuie sur un ensemble d'exemples de relations sous forme de couples d'entités. Si le principe général de l'auto-induction de relations est conservé, la distinction entre DIPRE et SnowBall se fait déjà au niveau de la génération des patrons de relations.

Dans SnowBall les patrons sont aussi représentés par des tuples contenant les parties *prefix*, *middle*, *suffix* introduites dans DIPRE, mais cette fois les patrons sont générés de façon itérative : les mots contenus dans chacune de ces parties *prefix*, *middle*, *suffix* sont pondérés à chaque itération par le nombre d'occurrences (normalisé) du mot dans la partie concernée.

Dans DIPRE les occurrences de relations sont regroupées en fonction du contexte entre les entités et l'ordre relatif des entités. Dans SnowBall, le regrou-

pement se fait en prenant en compte les parties préfixes et suffixes : un score de similarité est attribué à chaque paires d'occurrences de relations. En pratique, le calcul de similarité se base sur le produit scalaire des vecteurs de composantes (préfixe, contexte, suffixe). Chaque groupe d'occurrences sert ensuite à la génération d'un patron et un score de confiance est associé à chaque patron généré.

Pour ce qui concerne l'extraction de nouvelles relations, SnowBall commence par rechercher des occurrences potentielles de relations, ces occurrences sont transformées sous forme de tuples et ensuite comparées aux patrons préalablement générés. Lors de la comparaison des occurrences avec les patrons, les occurrences ayant un score de similarité inférieur à un seuil donné sont éliminées. Enfin, parmi les occurrences conservées, celles qui sont (1) issues de patrons ayant un score de confiance élevé et (2) issues d'un grand nombre de patrons sont conservées.

2.7.2.2 Approches à base de classifieurs statistiques

Les approches à base de classifieurs statistiques pour la détection de relations sont souvent séparées en deux catégories : d'une part, les méthodes reposant seulement sur un ensemble de traits caractéristiques de la relation (*Feature-based*) [Kambhatla, 2004] et d'autre part, les méthodes utilisant un noyau en plus des traits caractéristiques, à savoir les méthodes à base de noyaux (*Kernel-based*) [Zelenko et al., 2003].

Cette section présente ces deux catégories et s'appuie en partie sur l'état de l'art présenté dans [Bach and Badaskar, 2007].

Les approches fondées sur *features* visent à traiter l'extraction de relations comme un problème de classification. Ainsi, il s'agit d'apprendre un modèle pour chaque type de relations que l'on cherche à extraire (classification binaire). Afin d'être compréhensible par le modèle, les relations sont représentées sous forme de *features* et le modèle renvoie une réponse positive ou négative selon que la relation appartient ou non au type visé.

Plusieurs modèles ont été utilisés pour cette tâche avec des niveaux de résultats différents : [Kambhatla, 2004] utilise un modèle Maximum d'Entropie et obtient un score de 52,8 % (F1-Mesure) sur le corpus ACE 2004, [GuoDong et al., 2005]

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

utilisent un modèle SVM et obtiennent un score de 55.5 % sur ce même corpus. Il faut souligner que les *features* utilisés diffèrent dans les deux modèles. En particulier, [Kambhatla, 2004] n'utilise pas de *features* concernant des ressources sémantiques telles que des listes de noms de lieux. Ce résultat souligne l'importance du choix des *features* dans les approches *Feature-based*. On pourrait penser de façon intuitive que l'ajout d'un grand nombre de *features* complexes améliore les résultats. Malheureusement, ce n'est pas toujours le cas. Par exemple, [Jiang and Zhai, 2007] démontrent que ce type d'ajout n'améliore pas beaucoup les performances et que dans certains cas, cela pourrait les faire diminuer. Les auteurs soulignent aussi qu'il est possible d'obtenir un score élevé (68 % de F1-Mesure) à partir d'un ensemble de *features* composé ni d'informations syntaxiques, ni d'informations de dépendances. Les *features* qu'ils utilisent reposent sur les n-grammes de mots (unigramme, bigramme, trigramme). L'inconvénient de ces *features* lexicalisés est leur dépendance vis-à-vis du corpus d'apprentissage, et donc au domaine considéré, par un phénomène de sur-spécialisation du modèle d'apprentissage.

Dans une autre perspective, les approches (*Kernel-based*) proposent de traiter l'extraction de relations comme un problème de calcul de similarité entre un ensemble annoté d'occurrences de relations (l'ensemble est composé d'exemples positifs et négatifs de la relation visée) et une nouvelle occurrence de la relation. L'idée est de comparer la nouvelle occurrence de relation à celles qui sont annotées comme positives ou négatives et de déterminer ainsi de quelle catégorie la nouvelle occurrence est plus proche.

Le calcul de la similarité est effectué en s'appuyant sur une fonction noyau, qui permet d'appliquer des transformations (linéaires ou non) au vecteur de *features* qui représente l'occurrence de relation afin de le projeter dans un espace de grande dimension. Les transformations appliquées par le noyau reviennent à appliquer un produit scalaire dans un espace de grande dimension. Lorsque l'apprentissage du modèle est terminé, l'étiquetage d'une nouvelle occurrence revient à déterminer si sa distance vis-à-vis des exemples positifs est supérieure à celle vis-à-vis des exemples négatifs. Le problème, dans ce cas, ne se situe plus seulement au niveau de la sélection des *features* (comme dans les approches *Feature-based*) mais aussi dans le choix de la fonction noyau.

L'intérêt des méthodes à base de noyaux est qu'elles peuvent être appliquées directement sur des structures complexes (des graphes ou des arbres) contrairement aux méthodes *Feature-based*. Pour cela, ces méthodes font appel à des noyaux particuliers : le *String Kernel* permet de calculer des similarités entre des chaînes de caractères, le *Tree Kernel* est quant à lui appliqué à des arbres.[Bunescu and Mooney, 2006, 2005; Moncecchi et al., 2010; Moschitti, 2006; Zelenko et al., 2003] présentent d'autres systèmes utilisant des méthodes à base de noyaux.

Concernant les performances, [Tikk et al., 2010] présentent une étude comparative des approches à base de noyaux pour une tâche d'extraction de relations entre des protéines à partir de documents scientifiques. Les auteurs utilisent pour leurs comparaisons neuf noyaux différents et les testent sur cinq corpus différents. La première conclusion de cette étude est que les résultats des méthodes à base de noyaux ne sont supérieurs que de quelques points¹ aux résultats obtenus en utilisant une approche à base de règles qui s'appuie seulement sur les arbres de dépendances syntaxiques [Fundel et al., 2007]. Il faut préciser que les méthodes à base de noyaux sont dépendantes du corpus utilisé pour l'apprentissage, ce qui entraîne des baisses de performance lorsque l'on change de corpus. robuste à ce type de phénomène. général réécrire des règles différentes Une autre conclusion de cette étude est que, pour les relations entre protéines, les noyaux utilisant les arbres de dépendances obtiennent de meilleurs résultats que ceux utilisant les arbres syntaxiques.

2.8 Construction des *templates* sur les événements

2.8.1 Présentation du problème

La construction des *templates* ou *scenario template construction*, vise à compiler les informations en sortie de tous les processus précédents (extraction d'entités nommées, résolution de coréférence, extraction de relations) pour compléter des scénarios d'événements. Plus précisément, les informations sont regroupées dans

¹Sur les cinq corpus utilisés, la meilleure approche à base de noyaux obtient une F1-mesure supérieur à 10 points sur 2 corpus, autrement la différence est en moyenne de 3,6 points.

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

des formulaires (ou *templates*) concernant des faits ou événements auxquels on s'intéresse. Enfin, le remplissage des *templates* sert aussi d'étape de normalisation des données, il s'agit alors d'uniformiser les valeurs (ou *fillers*) dans les champs (ou *slots*) (dates, chiffres, codes produits, etc.) ou de vérifier que les valeurs extraites sont valides (par exemple, qu'elles appartiennent à des listes fermées de valeurs).

En pratique, les champs à remplir dans les *templates* sont définis en amont du processus d'extraction et sont totalement liés au domaine d'application du processus. Par exemple, pour la campagne MUC 6, les scénarios à extraire concernaient les mutations de personnes à responsabilité dans les sociétés. Notons que dans le cas où un document évoque plusieurs événements d'intérêt, un *template* différent doit être complété pour chacun d'eux. En conséquence, un système doit être capable d'identifier (en amont) si un document traite ou non du type d'événement que l'on cherche à repérer et, si oui, le nombre d'événement distincts qui y sont mentionnés.

2.8.2 Les approches pour la construction de *templates*

Afin de traiter cette problématique, une démarche courante, proposée par exemple dans les systèmes FASTUS [Appelt et al., 1995], IE² [Aone et al., 1998] ou REES [Aone and Ramos-Santacruz, 2000], est de s'appuyer essentiellement sur les informations qui ont été repérées au niveau de chaque phrase (les entités et les relations entre entités), de façon indépendante, et de se servir de ces informations pour construire des *templates* partiels pour chaque phrase. Les liens de coréférence sont ensuite intégrés par des règles afin de regrouper les *templates* partiels et reconstituer des scénarios d'événements. Nous rappelons que les règles dans ce cas sont assez dépendantes du domaine, et servent souvent à vérifier qu'il n'y a pas d'incohérence lorsque les deux *templates* sont fusionnés.

Dans [Chieu and Ng, 2002], une approche utilisant des classifieurs à maximum d'entropie est proposée pour extraire des *templates* mais toujours en faisant la même hypothèse que les approches précédentes, à savoir : un *template* est associé à une phrase, ce qui revient à chercher à compléter tous les champs à partir du contenu d'une seule phrase.

Plus récemment quelques travaux se sont intéressés à traiter le remplissage de *templates* de façon plus globale avec pour objet d'intégrer des informations venant de plusieurs phrases. Dans [Chieu et al., 2003], les auteurs présentent le système ALICE (*Automated Learning-based Information Content Extraction*) qui est utilisé sur le corpus MUC-4 pour extraire des *templates* concernant des attaques terroristes en Amérique latine. ALICE repose sur deux composantes principales : (1) un classifieur statistique¹ qui sert à trouver dans une liste d'entités d'un type donné, celle ayant la meilleure valeur possible ; (2) un gestionnaire de *template* qui est responsable d'attribuer une valeur à un champ d'un *template* existant ou alors à un champ d'un nouveau *template*. Les décisions du gestionnaire de *template* reposent sur trois types de contraintes génériques concernant les entités de type date et lieu et des termes spécifiques (*seed words*), par exemple pour vérifier qu'il n'y a pas d'incompatibilité entre la date associée au *template* courant et la date associée à la valeur d'un nouveau champ. Enfin les résultats reportés sont comparables, bien qu'ils soient inférieurs, à ceux obtenus par des systèmes fondés sur des heuristiques utilisés lors de la campagne MUC-4.

En suivant la même idée, dans [Patwardhan, 2008; Patwardhan and Riloff, 2007, 2009], la méthode proposée essaie de compléter les *templates* en utilisant des indices extra-phrastiques. Par exemple, [Patwardhan and Riloff, 2007] propose d'utiliser un classifieur (SVM) pour identifier les phrases pertinentes pour compléter un *template*, puis des patrons sont appliqués pour extraire les valeurs des champs du *template*. L'approche est évaluée sur le corpus MUC-4 et comparée à un système existant présenté dans [Riloff, 1993], avec des résultats supérieurs et comparables à ceux de [Chieu et al., 2003] sur ce même corpus.

2.9 Notre problématique d'extraction d'information

Le but de l'extraction d'information est d'extraire automatiquement des informations structurées à partir de textes. Dans notre étude, nous nous intéressons

¹[Chieu et al., 2003] expérimentent plusieurs modèles (SVM, MaxEnt, Naive Bayes, Decision Tree) dans l'article, le meilleur résultat est obtenu par le modèle MaxEnt

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

plus précisément à l'extraction d'événements complexes dans des dépêches d'actualité pour une application de veille événementielle sous la forme du remplissage automatique de formulaires ayant une structure prédéfinie, aussi appelée *templates*, rendant compte des caractéristiques des événements. Les caractéristiques des événements auxquelles nous nous intéressons sont préalablement définies par des utilisateurs et peuvent varier en fonction du domaine de veille considéré. Ici, un événement complexe fait référence à la mise en correspondance de plusieurs entités nommées pour décrire un fait d'actualité. section précédente, à savoir : la distinction de différents événements dans une même dépêche, et le fait que les informations relatives à un même événement peuvent être réparties à plusieurs endroits du texte.

Le fait de travailler essentiellement à partir de dépêches d'actualité, qui se justifie dans le cadre d'une application de veille événementielle, a aussi une influence sur les traitements proposés, à cause de la nature stylistique de ces documents. En effet, ce type de textes présente les informations en les structurant dans le but de répondre à des questions informatives (*quand ? où ? combien ? qui ?*) sur un événement [Lucas, 2004]. Les informations que nous recherchons sont donc en général présentes de façon explicite dans les textes (les informations sont organisées de sorte qu'il n'y ait pas d'ambiguïtés entre les différentes informations mentionnées dans un même document) et notre objectif est de compiler ces informations pour les associer à l'événement principal faisant l'objet de la dépêche.

Par ailleurs, une des caractéristiques des dépêches de presse est de faire fréquemment référence à plusieurs événements de même nature, en général pour donner des points de comparaison par rapport à l'événement faisant l'objet de l'actualité. Dans le cadre applicatif de veille que nous considérons, les utilisateurs s'intéressent essentiellement à l'événement le plus récent, celui qui fait l'actualité. La première difficulté est donc de pouvoir sélectionner dans les dépêches les passages (phrases ou groupes de phrases) les plus pertinents en rapport avec l'événement que l'on veut traiter : lorsque plusieurs faits d'actualité sont mentionnés dans un même document, ils sont le plus souvent présentés dans des passages différents qu'il faut différencier. Lorsque les passages faisant référence au fait principal ont été ciblés, le problème est alors de sélectionner les informations les plus pertinentes à l'intérieur de ces passages.

Une deuxième difficulté, évoquée dans la section précédente, vient du fait que les informations qui caractérisent l'événement principal peuvent être mentionnées au travers de plusieurs phrases, alors que de nombreux systèmes utilisent uniquement des informations locales (au niveau de la phrase) pour compléter les *templates*. [Stevenson, 2006] estime qu'environ 60 % des faits contenus dans les corpus MUC(4-6-7) peuvent être extraits en se restreignant aux informations locales, ce qui nécessite d'envisager une approche permettant d'aller au-delà de la phrase pour permettre l'extraction des informations pour les 40 % des faits restants.

La démarche que nous proposons pour traiter ces problématiques se compose de deux étapes qui caractérisent notre processus de remplissage des *templates* : la première étape, appelée *segmentation événementielle* des textes, a pour objet de cibler les phrases les plus pertinentes en lien avec l'événement principal d'un document. La motivation est de centrer la phase de remplissage de *templates* autour des seuls passages liés à l'événement principal au lieu de l'ensemble du document. Il faut noter que les segments de textes se référant à un même événement ne sont pas forcément contigus. Dans cette perspective, la segmentation en événements s'appuie sur les indices temporels (temps grammaticaux, expressions temporelles, présence/absence de dates) évoqués dans les documents et les exploite afin de définir des segments de document faisant référence à un même type d'événement.

La deuxième étape, dite de *remplissage des templates*, exploite les segments en lien avec l'événement principal (identifiés lors de la segmentation événementielle) afin de détecter les relations sémantiques entre les entités nommées dans la perspective de sélectionner les entités pertinentes pour compléter le *template*. Précisément, il s'agit d'exploiter les relations sémantiques entre entités exprimées aussi bien au niveau de chaque phrase, aussi appelé niveau phrastique (niveau local), qu'au niveau de plusieurs phrases à savoir le niveau discursif (niveau global). Il faut noter que les relations au niveau global sont établies en se servant des relations déjà détectées au niveau local. Les relations entre entités sont représentées de façon globale par un graphe d'entités, à partir duquel la sélection des entités pertinentes pour le remplissage des *templates* est effectuée, en utilisant les propriétés du graphe.

Dans le cadre de l'extraction d'événements dans les dépêches d'actualité, les *templates* que nous cherchons à compléter ont une forme prédéfinie qui implique

2. EXTRACTION D'INFORMATION : VUE D'ENSEMBLE DU DOMAINE

un nombre limité de relations entre les entités : le nombre de relations entre les entités dépend du type de fait d'actualité à décrire et de la structure du *template* associé à ce fait. Plus la structure du *template* contient de champs à renseigner et plus il y a de relations impliquées. Ici, nous avons considéré une structure de *template* impliquant moins d'une dizaine d'entités. Dans une autre perspective, la problématique, complémentaire par rapport à celle-ci, du peuplement de bases de connaissances se situe dans un cadre où un grand nombre de relations entre les entités sont possibles : par exemple si l'on considère les types d'entités personnes et organisations plusieurs dizaines de relations sont possibles. Les problèmes qui se posent sont donc différents et doivent être traités d'une autre façon. Étant donné qu'il faut traiter un grand nombre de relations entre des entités nommées, le premier problème est d'obtenir des exemples manuellement annotés de ces relations (contrairement à la problématique de remplissage des *templates*). Le second problème concerne le passage à l'échelle de l'approche à utiliser pour traiter efficacement plusieurs dizaines de milliers de documents. Dans cette perspective, nous proposons une *approche faiblement supervisée* pour l'extraction de plusieurs dizaines de types de relations à partir d'un corpus de grande taille composé de près de deux millions de documents. L'aspect faiblement supervisé de cette approche lui permet de ne pas dépendre de la disponibilité d'un grand nombre d'occurrences de relations préalablement annotées, à la différence de la problématique de remplissage des *templates*.

Chapitre 3

La segmentation des textes en événements

Dans ce chapitre, nous présentons une étude sur la segmentation des textes en événements. Cette segmentation vise plus précisément à découper les textes en segments (zone composée d'une ou plusieurs phrases), faisant référence à un même fait d'actualité. Plus globalement, son but est d'identifier les zones de texte les plus pertinentes pour le processus d'extraction.

3.1 Introduction

Nous avons vu à la section 2.8 que l'étape finale du processus d'extraction d'information était la construction de *templates*. Ces *templates* regroupent les informations concernant des événements mentionnés dans les textes. La difficulté rencontrée pour construire un tel *template* est que les informations que l'on cherche à repérer et à rassembler sont souvent dispersées dans tout le document : le plus souvent une seule phrase ne contient pas toutes les informations utiles pour le remplissage d'un *template*, il faut alors explorer les informations de plusieurs phrases.

La démarche précédemment adoptée [Aone et al., 1998; Aone and Ramos-Santacruz, 2000; Appelt et al., 1995] pour cette étape repose sur deux phases : (1) la construction d'un *template* intermédiaire pour chaque phrase d'un document ;

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

(2) la fusion (*merging*) des *template* intermédiaires pour constituer un *template* final. Un point faible de cette démarche, identifié dans [Humphreys et al., 1997], est qu'elle conduit à une surgénération d'instances d'événements. Autrement dit, elle conduit à construire plus de *templates* intermédiaires que nécessaire avant l'étape de fusion. [Humphreys et al., 1997] insiste aussi sur l'importance de pouvoir identifier les phrases faisant référence au même événement afin de fusionner correctement les *templates* intermédiaires : «*it is crucial that multiple references to the same event be correctly identified and merged*». Un autre aspect dont il faut tenir compte pour compléter les *templates* est la présence d'informations discursives. L'idée est de pouvoir tirer parti de ce type d'informations afin de déterminer la présence de plusieurs événements.

Dans notre approche d'extraction d'information, l'étape de segmentation des textes que nous proposons permet de tenir compte de la présence de plusieurs événements comparables (plusieurs faits d'actualités) dans un même document lors de la fusion des *templates* intermédiaires. L'avantage par rapport à l'approche classique est d'intégrer un critère relatif au type d'événements auquel fait référence le *template* intermédiaire. Ce critère s'appuie sur les indices temporels trouvés dans le document pour associer une phrase à un type d'événement.

3.1.1 Qu'est-ce qu'un événement ?

L'objet de la segmentation est de repérer les passages pertinents dans les documents en vue de la phase de remplissage de *template* concernant un événement. La notion d'événement peut être ambiguë à ce stade puisqu'elle n'a pas été définie : quelles sont les types d'événements concernés ? Comment les événements sont-ils représentés ? Cette section vise à apporter des précisions sur l'utilisation des événements dans un cadre général et dans le domaine de l'extraction d'information. Enfin elle aborde les limitations des événements pour notre démarche de segmentation.

Événement est un mot polysémique qui vient du latin *evenire* (advenir). Événement est associé à plusieurs sens : par exemple en physique, il désigne un «*phénomène considéré comme localisé et instantané, survenant en un point et*

un instant bien déterminé»¹. En terme journalistique, un événement désigne un fait marquant d'actualité, qui peut être prévisible ou imprévisible : catastrophe naturelle, résultat sportif, résultat scientifique, résultat politique, etc.

Notons qu'en règle générale les notions d'événement et de catégorie d'événements sont confondues. En fait, un événement désigne une instance d'une catégorie d'événements donnée : par exemple l'événement *séisme du 12 janvier 2010 à Haïti* est une instance de la catégorie tremblement de terre.

En matière d'extraction d'information, par exemple dans le cadre des campagnes MUC, un événement est lié à une catégorie d'événement dont il s'agit de retrouver les occurrences dans un document. Plus précisément, la catégorie d'événement est associée à une structure de donnée appelée *template*, qui regroupe les informations pertinentes concernant le déroulement de l'événement. Selon les éditions de MUC, les événements considérés ont concerné des faits d'actualité différents : mutation de managers, attaques terroristes, etc. Dans un contexte journalistique, les documents relatent généralement plusieurs faits d'actualité et par conséquent, plusieurs événements, soit de même catégorie, soit de catégories différentes. Lors des campagnes MUC, les participants devaient créer un *template* à chaque fois qu'ils trouvaient une instance d'une catégorie d'événements donnée dans un document.

Les informations associées aux événements dans les *templates* prennent en général la forme d'entités nommées et selon les cas, la notion d'événement se matérialise par une relation qui peut être soit entre entités nommées, soit portée par un verbe ou un déverbal ou encore s'étendre au-delà d'un contexte phrastique.

Dans notre contexte de travail, un événement est considéré comme une occurrence d'un fait, de façon similaire aux conférences MUC. Dans notre étude, nous considérons des événements ponctuels datés et les structures de *templates* que nous considérons pour décrire les événements doivent donc contenir une date. Les événements sont liés à des mentions qui servent de marqueurs pour la catégorie d'événements. Par exemple, *séisme*, *secousse*, *tremblement de terre*, *réplique* sont des mentions d'événements pour la catégorie tremblement de terre. Dans ce qui suit, nous considérerons ces mentions comme des entités nommées du point de vue de leur identification.

¹<http://www.larousse.fr/encyclopedie/nom-commun-nom/événement/50167>

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

Les deux phrases suivantes illustrent la catégorie d'événements à laquelle nous nous attachons dans notre contexte applicatif et la façon dont ces événements peuvent être relatés dans des articles de journaux :

- «*Le séisme de 2010 à Haïti est un tremblement de terre crustal d'une magnitude de 7,0 à 7,3 survenu le 12 janvier 2010 à 16 heures 53 minutes, heure locale*»
- «*Le séisme du 11 mars 2011 de la côte Pacifique du Tohoku au Japon est un tremblement de terre d'une magnitude 9,0, survenu au large des côtes nord-est de l'île de Honshu*»

3.1.2 Les informations discursives et les événements

Selon [Kitani et al., 1994], les informations discursives sont nécessaires pour identifier les passages se référant aux mêmes événements : «*discourse segmentation is necessary to identify portions of text containing related pieces of information*».

Néanmoins, dans les approches existantes [Aone et al., 1998; Aone and Ramos-Santacruz, 2000], hormis les informations de coréférence, peu d'informations discursives sont utilisées lors de la fusion des *templates* intermédiaires. De plus, l'exploitation des informations de coréférence intervient seulement à la fin du processus d'extraction. [Crowe, 1995] fait de ce point de vue exception en faisant intervenir des informations discursives en amont de la construction des *templates* : son approche dont la perspective générale est de segmenter les textes en fonction des événements qui les composent, utilise les marqueurs discursifs.

En particulier, ne sont pas pris en compte pour la fusion des *templates* intermédiaires l'organisation des événements ou la structure événementielle des documents. Les stratégies existantes cherchent plutôt à déterminer des critères de compatibilité entre les *templates* en faisant intervenir des indices (ex. : dates) mentionnés dans les phrases supports des *templates* intermédiaires [Chieu et al., 2003].

Après avoir considéré 150 articles de presse provenant du corpus MUC-5, [Kitani et al., 1994] suggèrent qu'il existe deux structures typiques pour l'organisation des événements dans les documents. La figure 3.1 représente de façon

schématique les deux structures, *Str-1* et *Str-2*, proposées par [Kitani et al., 1994].

Dans *Str-1*, les événements sont décrits de façon séquentielle : une fois qu'un événement a été entièrement décrit, aucune référence ne lui est faite dans le reste du document. Dans *Str-2*, deux types d'événements sont considérés : un événement principal et plusieurs événements secondaires. Les événements secondaires sont décrits entièrement et de façon séquentielle alors que les éléments associés à l'événement principal peuvent être répétés ou dispersés dans le texte.

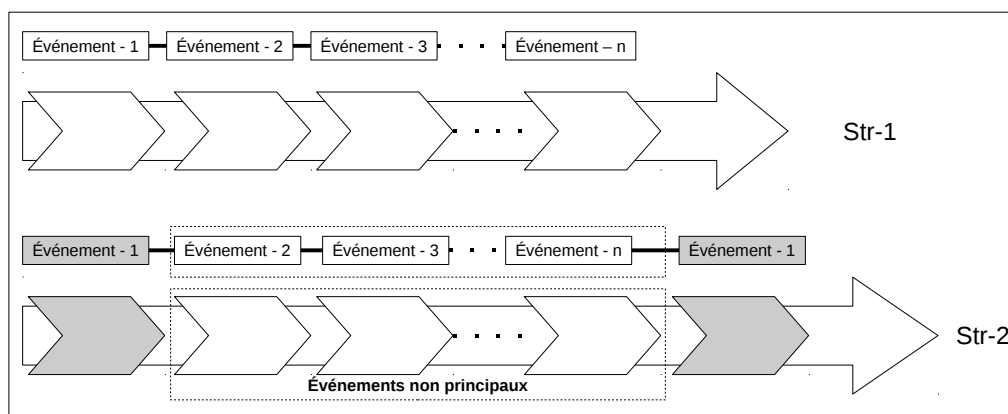


FIG. 3.1 – Organisation des événements dans les articles de presse

Dans la même perspective, [Lucas, 2004] propose de différencier les textes à structure simple et les textes à structure complexe : les premiers sont centrés sur un seul événement décrit selon un seul point de vue ; les seconds font référence à plusieurs événements parmi lesquels se distingue un événement principal auquel se subordonnent les autres événements.

De façon plus générale, [Cicurel, 1993] propose une étude sur les événements-catastrophes dans des dépêches de presse. En considérant particulièrement le cas des tremblements de terre, [Cicurel, 1993] suggère que le discours est organisé de façon à permettre au lecteur de reconnaître des scénarios cognitifs et par conséquent, faciliter la mémorisation du contenu des articles. [Cicurel, 1993] suppose aussi l'existence d'une «*structure discursive de l'événement identifiable dans les articles de presse*» et propose de représenter les événements-catastrophes à partir d'une liste composée de 7 éléments (appelés superstructures) :

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

Événement-noyau : l'événement central qui fait l'objet de l'article et qui est rapporté au lecteur.

Les conséquences : un état des lieux des conséquences humaines ou matérielles engendrées par l'événement-noyau.

Les événements antérieurs : événements précédant l'événement-noyau.

La périodicité de l'événement-noyau : il s'agit de donner des éléments sur la probabilité qu'un événement comparable à l'événement-noyau se reproduise.

L'arrière plan : explications permettant la compréhension du phénomène lié à l'événement-noyau.

La réaction verbale : citations recueillies auprès de personnes liées de façon directe ou non (témoins, experts, responsables politiques) à l'événement-noyau.

Les histoires parallèles : il s'agit de fournir un contexte historique concernant le thème de l'événement sans qu'il existe de rapport nécessaire avec l'événement-noyau. C'est le cas par exemple d'un événement ayant marqué la mémoire collective.

La segmentation en événements que nous opérons prend place dans un contexte de veille au sein duquel les utilisateurs sont essentiellement intéressés par les faits les plus récents : le but est alors de fournir à ces utilisateurs une information synthétique relative à ces événements récents à partir de leur évocation dans des dépêches d'actualité. En d'autres termes les utilisateurs ne sont intéressés que par les informations liées à la structure discursive *événement-noyau* de la dépêche. Nous nous différencions en cela des approches générales d'extraction d'événements qui cherchent à associer des informations à tous les événements évoqués dans un texte.

Une des caractéristiques des dépêches de presse est de faire fréquemment référence à plusieurs événements de même nature, en général pour donner des points de comparaison par rapport à l'événement faisant l'objet de l'actualité : il s'agit des événements associés à la superstructure *événements antérieurs*. Dans notre cas, il s'agit d'extraire uniquement des informations concernant les structures *événements-noyaux*, les autres sont vues comme des sources de perturbation

pour cette extraction. Dans ce chapitre, nous nous focaliserons sur l'utilisation d'indices temporels pour la segmentation événementielle en nous intéressant particulièrement au cas de textes à structure complexe, c'est-à-dire contenant un événement principal ainsi qu'un ou plusieurs événements secondaires.

3.2 Segmentation des textes et extraction d'information

Un des problèmes mis en avant dans l'introduction de ce chapitre est que le processus de construction des *templates* repose largement sur les éléments repérés au niveau phrastique et peu sur les éléments discursifs. Afin d'intégrer cette dimension discursive, quelques travaux [Crowe, 1995; Gu and Cercone, 2006; Naughton, 2007; Patwardhan and Riloff, 2007] ont proposé de segmenter le texte en amont de la construction des *templates*. L'hypothèse sous-jacente à cette segmentation est que toutes les régions d'un document ne sont pas pertinentes pour trouver les informations liées aux *templates*. Il convient donc de se focaliser sur les plus importantes. Ainsi, parmi les 7 structures discursives énumérées à la section 3.1.2, les informations contenues dans les structures *arrière plan* et *histoires parallèles* ne devraient pas être présentes dans les *templates*. En revanche, celles associées à l'événement *noyau* sont indispensables. L'intérêt, dans notre cas, est de pouvoir remplir les *templates* en écartant les informations issues des structures discursives différentes de l'événement *noyau* que nous savons non pertinentes pour l'utilisateur.

En suivant cette idée, [Crowe, 1995] propose une approche à base d'heuristiques pour le rattachement d'une proposition à un événement. Cette approche repose sur trois modules d'analyse traitant respectivement les informations temporelles, spatiales et les marqueurs discursifs ainsi que sur un gestionnaire d'événements intégrant les résultats de ces trois modules pour assumer la tâche de segmentation. Plus précisément l'intégration des résultats des modules précédents est faite en appliquant des heuristiques dans l'idée d'exploiter la structure du document. Ces heuristiques sont appliquées au niveau des paragraphes et des phrases afin d'attribuer un événement à une proposition. Par exemple,

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

toutes les propositions, autres que celles contenues dans la première phrase d'un paragraphe, sont associées au même événement que la première proposition de la phrase précédente. L'approche a été testée sur une centaine de documents annotés manuellement et provenant du corpus MUC-4 mais les performances détaillées du système ne sont pas rapportées.

[Naughton, 2007] représente une tendance un peu différente dans la mesure où l'objectif principal de la segmentation des textes y est de différencier des segments faisant référence à des événements et des segments à caractère non événementiel (passages qui ne font pas référence à un événement). Cette segmentation est réalisée par un modèle statistique d'étiquetage des phrases en quatre types (nouvel événement, continuation d'un événement, référence à un événement déjà mentionné, sans événement) implémenté par un automate probabiliste. Les états de l'automate correspondent aux catégories d'événements et les transitions sont associées à la succession des phrases. L'approche a été testée sur un corpus d'articles de presse concernant la guerre d'Irak et le meilleur score en terme de F1-mesure rapporté par l'auteur est de 62 % pour l'ensemble des catégories.

Toujours concernant les approches statistiques, [Patwardhan and Riloff, 2007] proposent d'appliquer des patrons d'extraction de *templates* seulement sur les passages les plus pertinents. Ces passages sont extraits des corpus MUC-4 et ProMed en utilisant un classifieur de type SVM. Plus précisément, le classifieur attribue la classe *pertinente* ou *non pertinente* à chaque phrase d'un document. Le classifieur utilise comme ensemble de *features* tous les unigrammes des mots issus des phrases de l'ensemble d'apprentissage. Concernant les performances du classifieur, les auteurs rapportent respectivement 89 % et 53 % de F1-Mesure pour la détection des phrases non pertinentes et pertinentes sur le corpus MUC-4.

3.3 La segmentation en événements à partir d'indices temporels

Compte tenu de l'étroite dépendance existant entre les dimensions temporelle et événementielle des textes [Pustejovsky et al., 2005], l'utilisation d'indices temporels pour mettre en évidence la segmentation des textes en événements ap-

paraît comme une voie intéressante. Les relations entre temps et segmentation du discours ont été principalement abordées dans le domaine de la linguistique textuelle et de la psycho-linguistique au travers de l'étude du rôle discursif des adverbes de temps placés à l'initiale des propositions. Dans le domaine de la psycho-linguistique, [Bestgen and Vonk, 2000] montrent ainsi l'existence d'une corrélation entre la présence de tels adverbes temporels et les changements de thème. Du point de vue linguistique, [Ho-Dac and Péry-Woodley, 2008] décrivent pour leur part une situation plus complexe dans laquelle le rôle discursif de ces adverbes est dépendant du type des textes. Toujours concernant l'aspect linguistique, [Charolles, 1995] suggère que ce type d'indices participe à l'organisation de la cohésion du discours. En particulier, ces indices servent à introduire des cadres de discours dans le but de délimiter des cadres temporels, spatiaux, ou modaux.

Pour aller dans ce dernier sens, notre intérêt pour les indices temporels est lié au style particulier des dépêches de presse. Ces textes contiennent des événements de même nature qu'il faut pouvoir distinguer (*cf.* section 3.1.2). En règle générale, les cadres spatio-temporels sont particulièrement intéressants pour cette tâche [Xu, 2011]. Dans la pratique, nous avons observé que pour les textes journalistiques dans notre domaine d'application, à savoir les événements sismiques, les cadres temporels étaient suffisants pour opérer cette distinction : les événements évoqués comme points de comparaison avec l'événement noyau sont généralement des événements s'étant déroulés dans la même zone géographique, même ville, région, canton (ou une zone géographique proche) mais à une date antérieure (il peut, par exemple, s'agir du dernier événement marquant ou d'un événement ayant eu lieu quelques années, mois, semaines avant l'événement principal). Par conséquent, l'aspect temporel est plus discriminant que l'aspect spatial. Dans la figure 3.2, on peut ainsi noter que les phrases relatant un séisme de magnitude 6,8 sont associées au cadre temporel daté du 16 juillet 2007 alors que l'événement noyau est mentionné dans un autre cadre.

Notre segmentation en événements repose plus généralement sur l'hypothèse que dans les dépêches de presse, les changements d'événements sont liés à la présence de marques linguistiques indicatrices de ruptures temporelles. Il est donc possible d'utiliser ces marques pour détecter les changements intervenant sur le plan événementiel. La figure 3.2 illustre ce point de vue : le temps d'énonciation

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

utilisé pour l'événement noyau (passé composé) est ainsi différent de celui utilisé pour l'événement antérieur (plus-que-parfait). Nous reviendrons plus en détail dans la section suivante sur les indices temporels utilisés dans notre approche.

3.4 Modèle discursif sous-jacent à la segmentation événementielle

Une segmentation des textes présuppose un certain modèle discursif. Dans notre cas, celui-ci est déterminé à la fois par la dimension discursive considérée, ici la dimension événementielle, le type des documents cibles, en l'occurrence des articles de journaux et des dépêches de presse, et le contexte applicatif, celui de l'extraction d'information. À la base de ce modèle, un texte est vu comme une séquence de phrases où chaque phrase est associée à un événement ou à une absence d'événement. Nous faisons l'hypothèse comme dans les travaux existants qu'une phrase est associée à un seul événement. Cette hypothèse est simplificatrice mais ne s'avère en pratique pas trop réductrice dans les domaines d'application que nous avons testés. Comme nous nous intéressons au repérage des entités associées au seul événement principal de la dépêche, nous ne distinguons pas les autres événements antérieurs. Ainsi, nous proposons d'adopter un modèle non hiérarchique, à mi-chemin en termes de complexité entre le modèle de [Kitani et al., 1994] et celui de [Cicurel, 1993]. Plus précisément, ce modèle distingue trois types de segments discursifs de même niveau :

- **Événement principal** : segment regroupant un ensemble contigu de phrases faisant référence à l'événement-noyau du texte ;
- **Événement secondaire** : segment regroupant un ensemble contigu de phrases en rapport avec une information associée à un événement différent de l'événement principal (dans le cas présent, ce sont des événements antérieurs) ;
- **Contexte** : segment regroupant un ensemble contigu de phrases n'appartenant ni à l'*Événement principal* ni à un *Événement secondaire*.

Le tableau 1 montre la correspondance entre les types de segments que nous distinguons et les structures discursives proposées dans [Cicurel, 1993] et reprises

dans la section 3.1.2. La figure 3.2 montre un exemple de l'application de notre modèle discursif à une dépêche de presse.

Très violent séisme dans le nord du Japon: 3 morts, 65 blessés, 12 disparus Un violent séisme a frappé samedi le nord du Japon, avec un premier bilan de trois morts, 65 blessés, et au moins 12 disparus, provoquant des glissements de terrain, défonçant des routes et faisant tanguer les immeubles. La secousse, d'abord annoncée de magnitude 7 sur l'échelle ouverte de Richter, a été révisée à la hausse à 7,2. Elle s'est produite à 08 H 43 (23 H 43 GMT), à 10 km seulement de profondeur, aux confins des préfectures d'Iwate et de Miyagi (nord-est), à environ 500 km de Tokyo, où elle a également été ressentie. Le premier bilan officiel fait état de trois morts, un pêcheur de 55 ans emporté dans un glissement de terrain, un sexagénaire écrasé par un camion et un ouvrier de 48 ans tué par des chutes de pierres sur un chantier. Par ailleurs, douze personnes sont portées disparues à Kurihara (préfecture de Miyagi): trois ouvriers ensevelis sous des décombres sur un chantier et neuf personnes dans un hôtel de sources thermales balayé par un glissement de terrain. L'agence nationale des désastres a fait état pour sa part de 65 blessés, alors que les médias japonais en décomptent plus d'une centaine. Environ 29.000 foyers autour de l'épicentre sont privés d'électricité. "Pendant quelques secondes, la terre a rugi. Je me demandais ce qu'il allait se passer", a raconté Mikiko Sugawara, 54 ans, dont le mari gère un hôtel à Kurihara.	Événement principal
Situé à la jonction de quatre plaques tectoniques, le Japon subit des milliers de tremblements de terre chaque année.	Contexte
Le dernier tremblement de terre meurtrier s'était produit le 16 juillet 2007 à Niigata (centre).	
Une secousse de magnitude 6,8 avait fait onze morts et plus de 1.000 blessés.	Événement secondaire
La centrale nucléaire de Kashiwazaki-Kariwa, la plus puissante du monde, avait été endommagée et est depuis arrêtée.	Contexte

FIG. 3.2 – Exemple d'annotation d'événements à partir d'une dépêche de presse

3.5 Modèles de segmentation événementielle

Dans la section précédente, nous avons présenté le modèle discursif sous-tendant notre segmentation des textes. Dans cette section, nous présentons les différentes méthodes que nous avons conçues, implémentées et évaluées pour opérationnaliser ce modèle. Le point commun à toutes ces méthodes est d'envisager la segmentation des textes en événements comme un problème de classi-

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

Catégories [Ce travail]	Superstructure discursive [Cicurel, 1993]
Événement principal	Événement-noyau, conséquences, réactions verbales
Événement secondaire	Événements antérieurs
Contexte	Périodicité de l'événement-noyau, arrière plan, histoires parallèles

TAB. 1 – Correspondance entre les types de segments distingués et les structures discursives de [Cicurel, 1993]

fication de séquences dans lequel l'objectif est d'attribuer un type d'événement à chaque phrase en se fondant sur des indices temporels. La segmentation est donc réalisée non pas en modélisant explicitement les frontières des segments mais indirectement, au travers du changement du type des unités de base constituant les segments. Un modèle graphique d'annotation de séquences apparaît comme particulièrement adapté à ce contexte : il permet en effet de typer les unités discursives de base (*i.e.* les phrases) tout en prenant en compte des contraintes de contiguïté entre ces unités assurant la délimitation de segments cohérents. Dans cette perspective, nous décrivons ici deux approches de segmentation en événements fondées respectivement sur les modèles de Markov Cachés (*Hidden Markov Model*, ou HMM) et les Champs Aléatoires Conditionnels (*Conditional Random Fields*, ou CRF). La première s'appuie sur la seule succession des temps grammaticaux des verbes tandis que la seconde permet d'intégrer un ensemble plus large d'indices temporels. En complément, nous proposons de comparer ces modèles avec un modèle discriminant de type maximum d'entropie (*MaxEnt*). L'idée est dans ce dernier cas est de pouvoir valider l'intérêt d'utiliser un modèle de classification de séquences par rapport à un modèle «local» ou non séquentiel.

3.5.1 Une segmentation fondée sur les temps verbaux : le modèle HMM

Les modèles de Markov Cachés constituent un modèle de classification de séquences [Rabiner, 1989] très largement utilisé en TAL (reconnaissance d'entités nommées, désambiguïsation morpho-syntaxique, etc.) et déjà appliqué à la segmentation de textes, notamment à la segmentation thématique [Yamron et al., 1998]. Les HMM sont des automates stochastiques à états finis permettant de déduire des séquences d'états non observables (ou états cachés) à partir de

séquences de données observées (observables). Ici, l'objectif est de déterminer à partir d'un texte donné, considéré comme une séquence de phrases, la séquence de catégories d'événements associée.

Nous faisons l'hypothèse que notre segmentation est un processus markovien, c'est-à-dire que l'état associé à l'observable courant ne dépend que des observables précédents et de l'état précédent : nous proposons d'utiliser les marqueurs temporels (temps grammaticaux) comme observables, les catégories d'événements constituant les états cachés. Les matrices de transitions (une pour les états, une pour les observables) sont obtenues à partir d'un corpus de textes annotés manuellement. Une illustration du modèle HMM que nous utilisons est proposée à la figure 3.3.

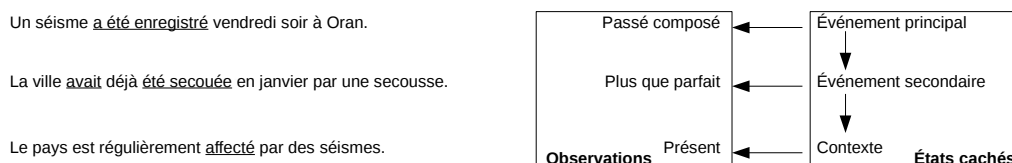


FIG. 3.3 – Illustration de la segmentation de textes en événements avec le modèle HMM

Une des contraintes induites par l'utilisation des HMM est que pour une séquence d'observations donnée, le calcul de la séquence d'états correspondante ne considère que l'état précédent et ne prend pas en compte les dépendances existant entre l'état précédent et la séquence d'observations. Dans notre contexte, un HMM ne permet donc pas d'intégrer d'autres critères que les temps grammaticaux qui servent ici d'observables. À l'inverse les CRF permettent de prendre en considération davantage d'informations sur les textes par l'entremise de *features*.

3.5.2 Élargissement des indices temporels : le modèle CRF

Depuis leur introduction en 2001, les CRF [Lafferty et al., 2001] ont été appliqués à tout un ensemble de problèmes dans le domaine du TAL : segmentation de textes [Hirohata et al., 2008], extraction de relations [Banko and Etzioni, 2008],

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

etc.

Un point commun entre les modèles CRF et HMM est qu'ils appartiennent à la classe des modèles graphiques dont l'objet est de représenter des dépendances entre des variables aléatoires. Dans le cas des modèles HMM, ces dépendances sont dirigées et elles permettent de lier l'observation courante et l'observation précédente afin d'instaurer une séquence. Dans le cas des modèles CRF, les dépendances ne sont pas dirigées. En théorie, il est donc possible d'établir une dépendance entre l'état courant et n'importe quel état. Par conséquent on peut appliquer les modèles CRF à des structures plus complexes telles que des graphes. Dans le cas de notre segmentation les modèles CRF linéaires de niveau 1 semblent le plus adapté pour modéliser le problème : le niveau 1 fait référence à la dépendance qui est instaurée entre l'état courant et l'état précédent.

La principale différence entre les modèles HMM et CRF est que le premier est un modèle génératif alors que le second est un modèle discriminant (cf. section 2.4.2.1). Cette différence a un impact direct en termes de capacité de modélisation : les CRF permettent en effet de représenter chacune des séquences d'observations, sous la forme d'un vecteur de *features* dont les composantes sont issues de tests atomiques effectués conjointement sur les observations et les états. Ces *features* offrent la possibilité d'intégrer des connaissances et des informations variées dans les modèles. Nous avons ainsi choisi d'enrichir notre précédent modèle de segmentation en événements en y intégrant les *features* suivants :

- **le temps des verbes** : comme avec notre modèle HMM, nous faisons l'hypothèse que les changements de temps grammaticaux, en particulier lorsqu'ils concernent des temps du passé, sont corrélés avec les changements d'événements dans le type de textes que nous considérons. Nous prenons en compte cette dimension dans notre modèle CRF en utilisant un *feature* binaire pour chaque temps grammatical possible. En pratique, chaque temps grammatical est associé à un *feature*, ce *feature* vaut 1 lorsque la phrase contient au moins un verbe conjugué au temps considéré, sinon le *feature* vaut 0 ;
- **la présence d'une date** : si une phrase contient une date antérieure à la date de l'événement principal, il est probable qu'elle fasse référence à un événement secondaire. Nous exploitons cette caractéristique de façon

limitée en utilisant un *feature* pour indiquer la présence ou l’absence d’une entité nommée de type date dans la phrase (dans le modèle actuel, la valeur de la date n’est pas utilisée) ;

- **les expressions temporelles** : ce *feature* est utilisé pour prendre en compte la présence d’une expression de localisation temporelle dans une phrase. Pour cela, nous utilisons un dictionnaire d’expressions que nous avons constitué manuellement à partir du corpus présenté dans [Laporte et al., 2008]. Le dictionnaire contient des expressions telles que : *au début de l’année, ces dernières années*. Plus précisément, ce dictionnaire contient environ 2100 entrées.

Le CRF étant un modèle de séquences, il permet de prendre en compte les liens de dépendance entre les états successifs : ici, cette capacité permet de rendre compte des dépendances entre les types d’événements successifs.

3.5.3 Modèle MaxEnt

De même que les deux modèles précédents, le modèle du Maximum d’Entropie (*MaxEnt*) [Ratnaparkhi, 1998] est largement utilisé dans le contexte du TAL. Ce modèle peut être vu comme le pendant discriminant du classifieur naïf de Bayes [Klinger and Tomanek, 2007], qui est pour sa part un modèle génératif. Parallèlement, les CRF peuvent être vus comme le modèle de séquences associé au MaxEnt.

Dans le cas présent, nous avons choisi d’utiliser le modèle MaxEnt comme point de comparaison. Offrant comme les CRF la capacité à prendre en compte des *features*, il devrait permettre de mesurer spécifiquement l’intérêt de considérer dans un modèle de segmentation événementielle un ensemble d’informations temporelles plus large que la simple succession des temps grammaticaux du modèle HMM. Le fait de ne pas être un modèle de séquences devrait quant à lui permettre de juger l’intérêt de cette dimension par comparaison avec le modèle CRF. Afin que cette comparaison soit pertinente, les modèles MaxEnt et CRF s’appuient sur les mêmes *features* (c’est-à-dire ceux décrits dans la section précédente : temps des verbes, présence d’une date et expressions temporelles). Seule l’absence de prise en compte des dépendances entre états successifs différencient les deux modèles.

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

3.5.4 Approches heuristiques

En complément des approches statistiques précédentes, nous avons également considéré deux approches alternatives de segmentation fondées sur des heuristiques liées au type de textes (dépêches de presses) et au domaine considérés (domaine sismique) : *HeurSeg*, issue d’une application existante d’extraction d’information dans le domaine des événements sismiques¹ développée spécifiquement pour ce domaine ; *ParaSeg*, qui repose sur la seule structuration en paragraphes des documents.

L’heuristique *HeurSeg* utilise comme critère principal la présence et la valeur des dates selon les principes suivants :

- des dates ayant des valeurs différentes (il s’agit de vérifier que la valeur normalisée en jour/mois/année de chaque entité de type date ne soient pas égales) correspondent à des segments différents (le segment principal étant celui de date la plus récente)
- les frontières de segments (fin d’un segment et début d’un autre) servent à symboliser la portée (ou étendue de validité) d’une date par rapport au document. La délimitation de ces frontières s’appuie sur la structure du texte en phrases et en paragraphes ainsi que sur la présence d’autres entités caractéristiques du domaine entre les dates.

L’heuristique *ParaSeg* exploite quant à elle le fait que dans les articles de journaux ou les dépêches de presse, l’information principale apparaît en premier lieu. Elle détermine donc la catégorie d’événements d’une phrase en fonction de la position de son paragraphe d’appartenance : les phrases sont classées comme appartenant à l’événement principal seulement si elles appartiennent aux deux premiers paragraphes du document ; autrement elles sont classées comme événement secondaire.

¹Cette application est actuellement utilisée par les analystes du Laboratoire de Détection et de Géophysique du CEA.

3.6 Pré-traitement des documents

Notre processus d'extraction d'information a vocation à être intégré à une application de veille concernant le domaine sismique. Dans ce contexte, les événements cibles sont principalement recherchés dans des articles ou des dépêches issus de flux d'informations du Web (par exemple flux en ligne de l'Agence France Presse). Avant de pouvoir utiliser ces documents, quelques traitements préalables sont ainsi nécessaires, traitements qui ont en pratique une influence non négligeable sur le résultat final d'une telle application. Dans le cadre de notre étude, nous nous sommes affranchis de certains de ces traitements : les documents issus d'un flux d'informations doivent en pratique être filtrés pour s'assurer qu'ils font référence à la catégorie d'événements ciblée¹.

Nous ferons ici l'hypothèse d'un filtrage parfait dans la mesure où nous nous focalisons sur les capacités d'extraction d'information. Pour les autres traitements, l'approche adoptée a simplement consisté à s'appuyer sur l'application de veille déjà existante. Les documents exploités pour la segmentation événementielle sont donc le produit de modules existants (collecte de documents, etc.) et gardent la trace des imperfections de ces modules. Ces imperfections touchent au plus bas niveau le nettoyage des documents et leur normalisation, c'est-à-dire la suppression de tous les éléments textuels indésirables résultant du contexte de collecte de ces documents (publicités, commentaires, balises HTML, etc.) mais également le repérage d'éléments d'information spécifiques pouvant être exploités au niveau de l'extraction d'information (titre, date, etc.). À un plus haut niveau, la reconnaissance des entités nommées du domaine visé est un de ces traitements amont sources d'erreurs.

Dans ce manuscrit, nous ne nous focalisons pas sur les phases de collecte et filtrage des documents, ni sur la détection des entités spécifiques. Il faut cependant souligner qu'elles ont toutes une influence certaine sur la qualité et la pertinence du processus d'extraction. Ces différents aspects sont traités de façon plus explicite dans [Besançon et al., 2011].

La segmentation des textes en événements repose sur le repérage et l'utili-

¹Typiquement, un document relatant un «séisme politique» n'est pas pertinent pour la veille sismique.

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

sation des informations temporelles présentes dans les textes. Ces informations peuvent être rapidement obtenues en appliquant un analyseur linguistique. Dans notre cas, tous les traitements linguistiques sont effectués en utilisant l'analyseur LIMA présenté dans [Besançon et al., 2010]. Plus spécifiquement les traitements se résument à appliquer les étapes suivantes : tokenisation, détection des fins de phrases, désambiguïsation morphosyntaxique, détection des verbes et de leurs temps reconnaissance des entités nommées. La figure 3.4 montre un exemple d'analyse linguistique d'un fichier produite par LIMA.

Très violent séisme dans le nord du Japon: 3 morts, 65 blessés, 12 disparus			
Un violent séisme a frappé samedi le nord du Japon, avec un premier bilan de trois morts, 65 blessés, et au moins 12 disparus, provoquant des glissements de terrain, défonçant des routes et faisant tanguer les immeubles.			
La secousse, d'abord annoncée de magnitude 7 sur l'échelle ouverte de Richter, a été révisée à la hausse à 7,2.			
1	Très	très#L_ADV_MODIF_ADV_OU_ADJ	NONE
6	violent	violent#L_ADJ_QUALIFICATIF_EPITHETE_PRENN	NONE
14	séisme	séisme#L_NC_GEN	NONE
21	dans	dans#L_PREP_GENERAL	NONE
26	le	le#L_DET_ARTICLE_DEF	NONE
29	nord du Japon	nord du Japon#L_NP_GEN	NONE
42	:	:#L_PONCTU_FORTE	NONE
44	3	3#L_DET_NUMERAL_CARD	NONE
.....			
89	séisme	séisme#L_NC_GEN	NONE
96	a	avoir#L_VERBE_AUXILIAIRE_INDICATIF	L_PRES
96	a frappé	frapper#L_VERBE_PRINCIPAL_INDICATIF	L_PC
98	frappé	frapper#L_VERBE_PRINCIPAL_PARTICIPE_PASSE	L_PASS
105	samedi	samedi#L_NC_MESURE	NONE
...			

FIG. 3.4 – Exemple d'analyse linguistique produite par LIMA. L'analyse est présentée en format tabulaire, avec le séparateur «|», pour chaque mot du document : la première colonne désigne la position du premier caractère du mot dans le texte ; la deuxième colonne contient le mot issu du texte, la troisième colonne contient le lemme ainsi que la catégorie morphosyntaxique ; la dernière colonne désigne le temps grammatical lorsque le mot est un verbe.

3.7 Évaluation des méthodes de segmentation

Cette section présente les résultats que nous avons obtenus en appliquant les approches statistiques et heuristiques pour la segmentation des textes en

événements. Pour la mise en œuvre des modèles statistiques, nous avons utilisé deux implémentations de référence : NLTK¹ pour le modèle HMM et CRF++² pour le modèle CRF. Le modèle MaxEnt a quant à lui été implémenté grâce à l'outil proposé par Dekang Lin³.

L'évaluation réalisée comporte deux volets : d'une part, une évaluation intrinsèque de la segmentation (les segments trouvés par la méthode sont-ils corrects ?) ; d'autre part, une évaluation au niveau de l'application visée, c'est-à-dire relative à l'impact de la qualité de la segmentation sur l'extraction des informations de l'événement principal dans les textes.

3.7.1 Les corpus d'évaluation

Pour l'évaluation des modèles, nous avons utilisé un corpus de 501 dépêches de presse en langue française concernant les événements sismiques. Ces dépêches ont été recueillies entre fin février 2008 et début septembre 2008 en provenance pour partie d'un flux de dépêches AFP (1/3 du corpus) et pour l'autre partie de dépêches collectées sur Google Actualités (2/3 du corpus). Ces dépêches évoquent 142 événements sismiques principaux différents. On y retrouve à la fois des dépêches ayant une structure simple (1 seul événement) et une structure complexe (plusieurs événements) : 252 dépêches (50 %) mentionnent au moins un événement secondaire.

Le corpus a été annoté manuellement en entités nommées et en mentions d'événements par des analystes du domaine mais uniquement pour les mentions et les entités liées à l'événement principal de chaque document. En revanche, les annotateurs pouvaient annoter plusieurs entités du même type s'ils les jugeaient équivalentes en tant qu'information apportée sur l'événement principal. De telles alternatives sont présentes en particulier lorsque plusieurs niveaux de granularité de l'information sont donnés : un séisme peut ainsi être localisé par une ville, une région ou un pays. Les informations associées à un événement sismique sont présentées dans le tableau 2, avec leur distribution dans le corpus. On remarque que la distribution des entités nommées n'est pas homogène : il y a beaucoup de

¹<http://www.nltk.org/>

²<http://crfpp.sourceforge.net/>

³<http://webdocs.cs.ualberta.ca/~lindek/downloads.htm>

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

noms de lieux (28,6 %) et très peu de coordonnées géographiques (0,9 %).

Type d'entité	Nombre	Nature
EVENT_TYPE (mention d'événement)	499	type d'événement (séisme, tsunami ...)
LOCATION	947	lieu de l'événement
DATE	470	date de l'événement
TIME	345	heure de l'événement
MAGNITUDE	484	magnitude
DAMAGES	531	dégâts causés par l'événement
GEO.COORDINATES	30	coordonnées géographiques

TAB. 2 – Distribution des entités nommées dans le corpus de référence : 3 306 entités dans 501 dépêches

Pour évaluer notre approche de segmentation en événements, nous avons annoté manuellement en segments événementiels une sous-partie de notre corpus composée de 140 dépêches principalement sélectionnées parmi les dépêches évoquant au moins un événement secondaire : la motivation principale était d'annoter aussi bien des documents évoquant un seul événement que des documents évoquant plusieurs événements. Notons que lorsqu'un document contient plusieurs événements secondaires, ces derniers ne sont pas différenciés entre eux. Le tableau 3 montre la distribution des événements sur la sous-partie annotée. La catégorie d'événements la plus représentée est *Événement principal* (70 %), de plus, 64 % des documents contiennent à la fois un événement principal et un événement secondaire, ce qui est cohérent avec l'aspect très factuel des dépêches de presse. La catégorie *Événement secondaire* regroupe sans distinction tous les événements différents de l'événement principal. Il est à noter que parmi les dépêches sélectionnées, le nombre réel d'événements secondaires distincts évoqués peut s'élever jusqu'à 4, avec un nombre moyen de 1,7 événements secondaires évoqués par article.

3.7.2 Évaluation intrinsèque de la segmentation en événements

L'objectif de cette première évaluation est de déterminer la capacité des différents modèles présentés à retrouver une segmentation événementielle de réf-

1659 événements sismiques dans 140 dépêches		
Type d'événement	Nombre de phrases	Proportion
Événement principal	1168	70 %
Événement secondaire	287	17 %
Contexte	213	13 %

TAB. 3 – Distribution des types d'événements dans le corpus de référence

rence. Étant donné que nous avons abordé cette segmentation comme un problème de classification, nous avons choisi des mesures d'évaluation en cohérence avec cette approche plutôt que de faire appel à des mesures très liées à la délimitation de segments comme celles utilisées en segmentation thématique par exemple. Nous avons ainsi adopté les mesures classiques de précision/rappel, avec la F1-mesure, la moyenne harmonique entre précision et rappel, comme synthèse. Plus précisément, ces mesures se définissent ici pour chaque type d'événement E_1 par :

$$\text{Précision} = \frac{\text{Correctes}}{\text{Annot}_{seg}} \quad (3.1)$$

$$\text{Rappel} = \frac{\text{Correctes}}{\text{Ref}_{seg}} \quad (3.2)$$

$$\text{F1-mesure} = 2 * \frac{\text{Précision} * \text{Rappel}}{(\text{Précision} + \text{Rappel})} \quad (3.3)$$

$$\text{Correctes} = \text{nombre de phrases correctement étiquetées } E_1 \quad (3.4)$$

$$\text{Annot}_{seg} = \text{nombre de phrases étiquetées } E_1 \text{ par le segmenteur} \quad (3.5)$$

$$\text{Ref}_{seg} = \text{nombre de phrases de type } E_1 \text{ dans le corpus} \quad (3.6)$$

Le tableau 4 donne les résultats en termes de précision/rappel (notés P., R.) sur notre corpus d'évaluation des modèles de segmentation HMM, MaxEnt et CRF que nous avons présentés précédemment. Le tableau 5 reprend ces résultats en termes de F1-mesure. Compte tenu de la taille limitée du corpus annoté et de la nécessité d'en consacrer une part significative pour entraîner les modèles, nous avons opté pour une procédure de validation croisée, classique dans un tel cas de figure. Plus précisément, les résultats donnés ont été obtenus en exploitant 4/5 du corpus pour la phase d'apprentissage et 1/5 pour la phase de test. Ils correspondent donc à des moyennes sur les 5 configurations possibles. Ces résultats sont

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

complétés par ceux des deux approches heuristiques de segmentation : *HeurSeg* et *ParaSeg*.

Le premier fait notable que laisse apparaître le tableau 4 est que tous les modèles considérés (HMM, MaxEnt, CRF) obtiennent globalement des performances supérieures à nos deux approches heuristiques (ParaSeg, HeurSeg). Ce constat doit être néanmoins nuancé en fonction des types d'événements et des modèles. Ainsi, dans le cas du modèle HMM, il semble que le seul critère utilisé, en l'occurrence la succession des temps des verbes, ne soit pas suffisant pour discriminer les types d'événements de façon fiable : si l'événement principal est correctement reconnu (88 % de F1-mesure), les autres types d'événements, et plus particulièrement les événements secondaires, le sont nettement moins. Les approches discriminantes (MaxEnt et CRF) obtiennent de ce point de vue de meilleurs résultats globaux que le modèle HMM. Cette supériorité s'affirme particulièrement au niveau de la précision, le rapport précision/rappel étant inversé par rapport au modèle HMM. Parallèlement, les résultats obtenus confirment notre hypothèse concernant la prise en compte des dépendances entre les états successifs : dans le tableau 5, le modèle MaxEnt obtient des résultats moyens inférieurs à ceux du CRF. De façon globale, on peut noter que le modèle CRF permet d'obtenir une meilleure segmentation en événements et la comparaison avec les modèles HMM et MaxEnt suggère que cette supériorité puise ses racines dans sa capacité à intégrer un large ensemble d'informations tout en les intégrant dans un modèle de séquences.

Enfin, le tableau 5 montre également que l'identification des segments de type *Événement principal* est nettement meilleure que celle des autres types de segments. Leur forte prévalence dans le corpus d'entraînement conjuguée à l'utilisation de méthodes statistiques n'est certainement pas étranger à ce fait.

	ParaSeg (%)		HeurSeg (%)		HMM (%)		MaxEnt (%)		CRF (%)	
Type d'événement	R.	P.	R.	P.	R.	P.	R.	P.	R.	P.
Principal	6,1	63,9	82,8	64,7	83,0	93,6	94,8	78,7	98,7	87,4
Secondaire	86,9	12,4	23,5	43,4	37,8	9,6	33,6	54,7	52,6	95,8
Contexte	0,0	0,0	16,9	21,7	49,1	40,0	22,0	84,2	69,3	93
Moyenne	31,0	25,4	41,1	43,3	56,6	47,7	50,1	72,5	73,5	92,1

TAB. 4 – Résultats de la segmentation en événements (Rappel-Précision)

	ParaSeg	HeurSeg	HMM	MaxEnt	CRF
Type d'événement	F. (%)	F. (%)	F. (%)	F. (%)	F. (%)
Événement principal	11,1	72,6	88,0	86,0	92,7
Événement secondaire	21,7	30,5	15,3	41,6	67,9
Contexte	0,0	19,0	44,1	34,9	79,4
Moyenne	10,9	40,7	49,1	54,2	80

TAB. 5 – Résultats de la segmentation en événements (F1-Mesure)

La comparaison de ces résultats avec ceux d'autres travaux est difficile car, ainsi que nous l'avons vu précédemment, notre segmentation événementielle ne recoupe pas directement de travaux existants. Nous donnons néanmoins quelques éléments de résultats concernant le travail le plus proche afin de mettre en perspective nos résultats. Dans [Patwardhan and Riloff, 2007], les auteurs entraînent ainsi un modèle SVM de classification de phrases dans une perspective globale proche de la nôtre. Leur objectif est comparable au notre, il s'agit de focaliser le processus d'extraction d'information mais ils cherchent plus directement à déterminer si une phrase est pertinente ou non afin de lui appliquer des patrons d'extraction. Cette classification est en outre faite sans passer par une segmentation en types d'événements. À la différence de nos modèles, leur modèle SVM est de ce fait entraîné à partir de *features* fortement lexicalisés à savoir les unigrammes des mots : la modélisation directe de la pertinence d'une phrase oblige en effet à tenir compte du domaine considéré, celui-ci étant caractérisé par les mots pleins des documents. L'évaluation de leur processus de sélection de phrases a été réalisée pour la langue anglaise sur les corpus MUC-4 et ProMed et donne des scores de 63 %|46 %|53 % (R.|P.|F.) sur des documents concernant des actes terroristes (MUC-4) et 72 %|41 %|52 % pour des documents médicaux (ProMed). Ces résultats ont un profil inverse par rapport à nos meilleurs classifieurs, c'est-à-dire avec un rappel supérieur à la précision, mais compte tenu des différences entre les deux travaux, il est difficile de déterminer si l'usage de *features* lexicalisés est à l'origine de cette différence.

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

3.7.3 Évaluation de la segmentation pour l'extraction d'information

L'objectif de la segmentation en événements est de constituer des segments de texte faisant référence à un seul type d'événements. Les segments délimités sont ensuite utilisés pour rattacher les entités aux événements (le rattachement se fait à l'intérieur d'un segment). Pour évaluer l'impact de cette segmentation sur le rattachement, nous faisons appel dans un premier temps à une heuristique simple, fondée sur l'hypothèse que les informations contenues dans les dépêches sont organisées en fonction de leur importance dans l'actualité : les informations les plus importantes (généralement associées à l'événement principal) sont citées avant les informations subordonnées (associées à un événement secondaire ou au contexte). Nous utilisons donc l'heuristique suivante : *pour chaque type d'entité, est choisie la première entité trouvée dans le segment.*

Type d'entité	NoSeg		HeurSeg		HMM		CRF	
	R. (%)	P. (%)	R. (%)	P. (%)	R. (%)	P. (%)	R. (%)	P. (%)
DAMAGES	83,5	77,9	76,3	74,4	69,9	65,1	80,2	75,3
DATE	38,4	35,9	69,3	65,0	48,9	45,6	64,4	60,1
EVENT.TYPE	82,1	81,6	79,3	78,8	59,2	58,8	76,7	76,2
GEO.COORDINATES	86,7	96,3	66,7	74,1	86,7	96,3	83,3	92,6
LOCATION	41,0	40,9	56,0	55,9	61,2	61,1	57,4	57,3
MAGNITUDE	93,5	93,0	86,3	85,9	66,7	66,3	86,7	86,1
TIME	61,1	51,2	56,4	49,2	78,8	71,5	63,4	55,5
Tous	66,6	63,5	71,0	68,6	63,4	61,2	71,7	68,8

TAB. 6 – Résultats du rattachement des entités à l'événement principal

Le tableau 6 illustre l'intérêt de notre segmentation événementielle en donnant à la fois les résultats du rattachement des entités à l'événement principal dans le cas où il n'y a pas de segmentation en événements (*NoSeg* : dans ce cas, on considère le document comme un seul segment, toutes les phrases étant donc associées à l'événement principal) et lorsque la segmentation est réalisée par l'heuristique de segmentation *HeurSeg* de la section 3.5.4 ou par les modèles HMM et CRF. Il faut d'abord souligner que l'approche sans segmentation permet d'obtenir un niveau de rattachement déjà élevé (et même supérieur au HMM : +2,7 % en F1-mesure) que la segmentation fondée sur l'heuristique *HeurSeg*

	NoSeg	HeurSeg	HMM	CRF
Type d'entité	F. (%)	F. (%)	F. (%)	F. (%)
DAMAGES	80,6	75,3	67,4	77,7
DATE	37,1	67,1	47,2	62,2
EVENT_TYPE	81,8	79,0	59,0	76,4
GEO_COORDINATES	91,2	70,2	91,2	87,7
LOCATION	41,0	55,9	61,1	57,3
MAGNITUDE	93,3	86,1	66,5	86,4
TIME	55,7	52,6	75,0	59,2
Tous	65,0	69,8	62,3	70,2

TAB. 7 – Résultats du rattachement des entités à l'événement principal (F1-Mesure)

améliore de façon conséquente (+6,2 % en F1-mesure par rapport à l'approche basique). Les résultats du tableau montrent que l'approche sans segmentation est particulièrement inadaptée pour traiter les entités de types date et lieux. Ces types d'entités sont *a priori* plus difficiles à traiter puisque les documents contiennent généralement plusieurs dates et plusieurs lieux. Cela montre une des limites de cette stratégie de rattachement qui n'est pas visible si on considère d'autres types d'entités comme par exemple les magnitudes. La raison est que pour les entités de type magnitude, il y a beaucoup moins d'ambiguïtés de rattachement : généralement la magnitude du séisme principal est mentionnée très tôt dans les documents. Ainsi, le risque d'erreur est faible si on sélectionne toujours la première. Toujours concernant ces deux types d'entités, on peut observer que le modèle à base de CRF est beaucoup plus performant que l'approche sans segmentation : +25 % en F1-mesure par rapport à l'approche sans segmentation pour les dates.

Les résultats montrent un gain important du modèle à base de CRF par rapport à l'heuristique *HeurSeg* sur les catégories *GEO_COORDINATES* et *TIME*, respectivement +17,6 % et +6,5 % en F1-mesure, ce qui permet de compenser les écarts sur les autres catégories. Abstraction faite des variations selon les types d'entités, le modèle à base de CRF donne pour sa part des résultats aussi bons

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

(et même un peu meilleurs) que ceux obtenus avec la segmentation heuristique *HeurSeg*. Son avantage est néanmoins de constituer une approche générique ne dépendant pas du domaine considéré.

3.8 Conclusions

Le constat à l'origine de ce premier chapitre est que les entités nommées servant pour la construction des *templates* sont dispersées dans différentes parties du texte. En conséquence, lorsque l'on cherche à compléter un *template* concernant un événement donné, il faut pouvoir déterminer les phrases faisant référence à cet événement afin de ne pas considérer les entités associées à d'autres événements.

Dans ce chapitre, nous avons proposé plusieurs approches pour la segmentation des textes en événements dans le but de faciliter le rattachement des entités pertinentes à l'événement principal du texte. Nous avons vu que concernant le style journalistique, les textes utilisent des structures discursives particulières pour rendre compte des événements (en particulier pour les catastrophes). L'idée de la segmentation que nous proposons est de s'appuyer sur des indices temporels afin d'identifier ces structures discursives dans les textes. Dans cet esprit, nous avons vu que les indices temporels étaient pertinents pour distinguer les types d'événements.

En termes de mise en œuvre, nous avons traité la problématique de la segmentation des textes en événements comme un problème de classification où l'objectif est de déterminer un type d'événement associé à chaque phrase. Nous avons proposé et évalué plusieurs modèles : principalement un modèle HMM, qui utilise comme seul critère de décision la succession des temps des verbes dans un texte et un modèle CRF, qui intègre pour sa décision un ensemble plus large d'indices temporels (expressions temporelles, dates).

En évaluant les différents modèles sur un corpus de dépêches concernant les événements sismiques, nous avons montré que le modèle CRF obtient de meilleurs résultats pour la segmentation événementielle des textes. Par ailleurs, nous avons évalué l'impact de la segmentation des textes en événements sur l'identification des entités pertinentes rattachées à l'événement principal de la dépêche et nous avons montré que le modèle à base d'apprentissage par CRF permet d'obtenir

des résultats équivalents (et même un peu meilleurs) à ceux obtenus avec un système utilisant une heuristique ad hoc propre au domaine tout en adoptant une approche beaucoup plus générique.

Suite aux évaluations, nous avons constaté que la principale source d'erreurs se situait au niveau du rattachement des entités, pour laquelle nous utilisons par défaut une heuristique assez simple. Dans le chapitre suivant, nous présenterons les méthodes que nous avons développées afin de remplacer cette heuristique et d'améliorer ainsi l'étape de remplissage de formulaire.

3. LA SEGMENTATION DES TEXTES EN ÉVÉNEMENTS

Chapitre 4

Le rattachement des entités aux événements

La segmentation en événements que nous avons proposée au chapitre précédent repose sur des indices temporels et a pour but d'identifier les segments de texte pertinents pour retrouver les entités associées à l'événement principal d'un texte. L'étape de rattachement des entités que nous présentons dans ce chapitre vise à compléter le processus de remplissage du *template* associé à cet événement. Plus spécifiquement, cette étape concerne la sélection des entités venant compléter les champs associés à ce *template*.

4.1 Introduction

La finalité du processus d'extraction d'information tel que nous l'envisageons ici est, pour chaque texte ayant un événement principal d'un type donné, d'instancier le *template* représentant ce type d'événements avec les informations spécifiques de l'événement évoqué par ce texte. Cette instanciation peut être vue également comme le fait d'associer à un événement les entités les plus pertinentes pour sa description. Les événements tels que nous les avons décrits à la section [3.1.1](#) peuvent en effet être considérés comme des configurations de relations entre les caractéristiques des événements (représentés par des entités nommées) et les événements eux-mêmes. Les *templates* ne sont qu'une représentation différente

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

de ces relations.

Nous avons vu à la section 2.8 que les *templates* sont le plus souvent construits à partir de *templates* intermédiaires eux-mêmes issus d’informations exprimées au niveau phrastique. Ce processus est donc ascendant et constructif mais il est globalement guidé par la structure *a priori* du template représentant le type d’événement considéré. Il se démarque de ce point de vue d’approches beaucoup moins supervisées proposant de générer automatiquement des *templates* pour un domaine donné [Filatova et al., 2006; Li et al., 2010] et donc d’extraire des événements et leurs informations associées sans structure fixée *a priori* [Chambers and Jurafsky, 2011].

Dans les travaux existants, la construction de ces *templates* intermédiaires n’utilise que très peu les informations de nature discursive. Dans le cas présent, nous suggérons de faire reposer cette construction pour partie sur les segments issus de la segmentation en événements au lieu de considérer les phrases hors de tout contexte. Cette segmentation permet en effet de sélectionner les parties de texte en rapport avec un type d’événements spécifique, ce qui évite la construction de *templates* intermédiaires pour les autres types d’événements.

Beaucoup des approches existantes de construction de *templates* [Aone and Ramos-Santacruz, 2000; Chieu et al., 2003] sont également limitées par une hypothèse simplificatrice considérant les relations entre entités dans la même phrase, donc les *templates* intermédiaires, comme des relations simples, donc binaires. Cette hypothèse est plus ou moins le fruit de la nature du processus d’extraction de relations à ce niveau, reposant sur des mécanismes (patrons lexicaux ou classifieurs) adaptés à l’extraction de relations binaires. Néanmoins, même à ce niveau, les événements peuvent être exprimés par des relations élaborées faisant intervenir plus de deux entités. L’extrait de texte suivant en est une illustration : «*Un séisme d’une magnitude de 5,5 degrés sur l’échelle ouverte de Richter a été enregistré vendredi soir dans la région d’Oran (430 km à l’ouest d’Alger), a annoncé la radio publique*». Dans [McDonald et al., 2005], ce type de relations est appelé *relation complexe*. Il s’agit de relations *n*-aires impliquant *n* entités nommées¹. Dans l’extrait précédent, on peut noter que chaque entité joue un rôle particulier vis-à-vis de la description du séisme, par exemple l’entité de type *date* indique

¹Pour la relation complexe de l’exemple, $n = 5$

la date d'occurrence de l'événement. De la même façon, les entités de type *nom de lieu* servent à indiquer la localisation de l'événement. Dans cette étude, nous considérons une relation complexe comme une relation n -aire entre des entités ayant des rôles différents. De plus, nous considérons que chaque type d'entité est associé à un seul rôle. Notons que dans un cadre plus général, un même type d'entité pourrait être associé à plusieurs rôles : par exemple les entités de type *nom de lieu* pourraient jouer les rôles de lieu d'occurrence de l'événement et de lieu où l'événement a été ressenti.

Cette notion de relation complexe n'est, à vrai dire, pas très différente de la notion de *template*, elle-même proche des structures manipulées dans le monde des bases de données (*i.e.* les tables des bases de données). Plus qu'une différence de nature formelle, ces différences de terminologie dénotent des travaux intervenant dans des contextes spécifiques. Dans la section suivante, nous évoquerons ainsi quelques travaux issus des bases de données et liés à la problématique du remplissage de *template* avant de revenir à la sphère de certains travaux en TAL s'appuyant sur la notion de relation complexe.

4.1.1 Bases de données et *templates*

Comme nous l'avons vu précédemment, les *templates* ont le plus souvent une structure «statique», généralement définie en amont du processus d'extraction. Du point de vue des bases de données relationnelles, le contenu d'un *template* est assez comparable à celui d'un tuple ou d'un enregistrement. De même, la structure d'un *template* (liste de ses champs) peut être convertie assez directement en une table relationnelle.

Dans le contexte de ce parallèle, quelques travaux [Feng et al., 2007; Mansuri and Sarawagi, 2006; Wick et al., 2006] se sont intéressés à l'extraction de tuples d'entités nommées directement à partir des documents, leur finalité étant bien évidemment de stocker ces tuples dans une base de données. [Wick et al., 2006] abordent ainsi le problème de l'extraction d'enregistrements à partir d'un corpus de pages personnelles appartenant à des membres de la communauté universitaire (professeurs, administratifs, étudiants, etc.). Leur objectif est de repérer les informations (noms, adresses, téléphones, etc.) contenues dans ces pages permettant

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

d'alimenter une base de contacts. Les auteurs utilisent pour ce faire une mesure de compatibilité des informations couplée à un algorithme de partitionnement de graphe : la mesure permet de construire un graphe de compatibilité des informations extraites des textes au sein duquel les tuples sont délimités grâce à l'algorithme de partitionnement de type hiérarchique agglomératif). L'approche, représentative de ce type de travaux, présente l'intérêt et l'originalité de considérer la tâche de remplissage de *template* (ou plus précisément de tuple) selon un point de vue global, ce qui permet d'intégrer plus facilement à ce niveau des contraintes entre informations extraites.

4.1.2 Les relations complexes

Comme nous l'avons mentionné ci-dessus, la notion de relation complexe, introduite en extraction d'information par [McDonald et al., 2005], recouvre des relations faisant intervenir n entités (pas nécessairement du même type). Elle a été reprise par la suite par différents travaux, dont [Afzal, 2009; Liu et al., 2007; Wick et al., 2006].

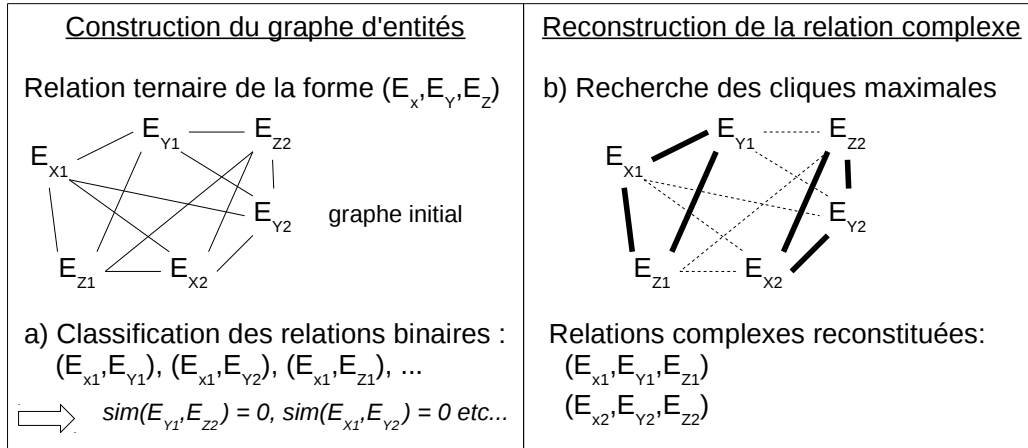


FIG. 4.1 – Approche en deux étapes pour l'extraction de relations complexes

Pour traiter ce type de relations, la stratégie proposée dans [McDonald et al., 2005] consiste à décomposer les relations d'ordre supérieur en plusieurs relations binaires. Par exemple, une relation ternaire peut être représentée par 3 relations

binaires : $(a, b, c) \Leftrightarrow (a, b), (a, c), (b, c)$. De là, [McDonald et al., 2005] proposent une méthode en deux étapes pour extraire les relations complexes : la première vise à construire un graphe à partir des relations binaires identifiées entre les paires d’entités. Plus spécifiquement, les auteurs utilisent un classifieur (de type maximum d’entropie) pour déterminer s’il existe ou non une relation entre deux entités. Le score de confiance donné par le classifieur est en outre utilisé pour pondérer les arcs du graphe. La seconde étape vise à retrouver toutes les cliques maximales (sous-graphes complets ayant le maximum de nœuds possible) pour reconstruire la relation complexe. Un score, égal à la moyenne géométrique des poids sur les arcs de la clique, est ensuite attribué à chaque clique maximale. Les cliques ayant un score inférieur à une valeur limite sont éliminées. La figure 4.1 illustre ces deux étapes.

Une approche très similaire à celle de [McDonald et al., 2005] a été proposée par [Afzal, 2009]. Les principales différences entre les deux approches concernent trois points. Le premier concerne les modèles utilisés pour la classification des relations binaires. [Afzal, 2009] obtient ses meilleurs résultats avec un modèle de type arbre de décision ¹ alors que [McDonald et al., 2005] s’appuient sur un classifieur de type maximum d’entropie. En deuxième lieu, [Afzal, 2009] ne réalise aucun filtrage des cliques maximales pour la reconstruction des relations complexes : la clique retenue est celle dont le produit des pondérations des arcs est maximal. Enfin, [Afzal, 2009] utilise le corpus MUC-6 concernant les mouvements de dirigeants alors que [McDonald et al., 2005] se servent d’un corpus dans le domaine biomédical.

Dans [Liu et al., 2007], les relations complexes sont appliquées comme dans le cas de [McDonald et al., 2005] au domaine biomédical. Plus précisément, il s’agit d’identifier des relations ternaires entre une protéine, un organisme et la localisation de la protéine dans l’organisme. Toujours dans la même perspective que [McDonald et al., 2005], la relation ternaire est décomposée en relations binaires. En revanche, l’identification de ces relations binaires présente la particularité de s’appuyer sur des *features* syntaxiques inspirés de la tâche d’attribution de rôles

¹Les *features* considérés dans [Afzal, 2009] sont quasi identiques à ceux de [McDonald et al., 2005], à l’exception d’un seul, qui n’a pas été repris. Outre les arbres de décision, [Afzal, 2009] a expérimenté avec de moins bonnes performances un classifieur de type maximum d’entropie et un classifieur bayésien naïf.

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

sémantiques (*semantic role labelling*). Par ailleurs, le classifieur utilisé est ici de type SVM. Une autre différence importante avec [McDonald et al., 2005] réside dans les contraintes posées pour la reconstruction des relations complexes. L’objectif est en effet de regrouper les relations binaires de type protéine-organisme (PO) et protéine-localisation (PL)¹ non seulement à condition que les relations soient dans la même phrase mais également que la protéine identifiée soit commune aux relations PO et PL. Enfin, [Liu et al., 2007] montrent que l’utilisation des informations syntaxiques améliore de façon conséquente les performances par comparaison avec de simples *features* lexicaux.

Il faut souligner que [Afzal, 2009; Liu et al., 2007; McDonald et al., 2005] exploitent globalement le même type d’approche mais que celle-ci s’applique à l’identification de relations complexes faisant intervenir des entités se trouvant à l’intérieur d’une même phrase. La phase de reconstruction des relations complexes n’est donc pas directement applicable à notre problème de remplissage de *templates* puisque nous ne nous limitons pas au repérage des relations entre entités à l’intérieur des phrases mais visons surtout celles exprimées à l’échelle textuelle. Néanmoins, nous proposons de nous inspirer de cette approche et de considérer la construction de *template* comme un problème de construction de relation complexe. L’idée est d’assimiler les événements à des relations complexes pour lesquelles le degré de la relation (arité) est égal au nombre de rôles à compléter dans le *template* (nombre de champs dans le *template*). L’extraction de «ces relations complexes» est également abordée en utilisant une méthode s’appuyant sur des graphes : un premier graphe d’entités est construit à partir du résultat de la segmentation en événements, puis plusieurs stratégies de rattachement indépendantes du domaine sont appliquées pour reconstruire la relation complexe. Avant de détailler la construction de ce graphe et ces stratégies de rattachement, nous allons présenter la notion de graphe d’entités et la forme précise qu’elle revêt dans notre cas.

¹Les relations organisme-location sont écartées.

4.2 Graphes d'entités nommées

La structure de graphe, bien que définie depuis longtemps et largement exploitée depuis lors dans des domaines comme celui des réseaux de communication, a connu depuis quelques dizaines d'années un grand succès en tant que modèle de représentation. Le traitement automatique des langues (TAL) n'a pas échappé à cette tendance comme le prouvent en particulier les ateliers TextGraph¹. Les graphes ont été ainsi utilisés pour plusieurs tâches allant, sans être exhaustif, de la résolution de coréférence [Chen and Ji, 2009; Nicolae and Nicolae, 2006] à la désambiguation des sens de mots [Dorow and Widdows, 2003] en passant par le résumé automatique [Mihalcea, 2004] et la tâche de question-réponse [Aceves-Pérez et al., 2007; Mollá, 2006].

Nous appliquons ici les graphes au cadre de l'extraction d'événements et plus particulièrement à la représentation des *templates*. Ce choix est justifié par le fait que les événements que nous cherchons à caractériser au travers des *templates* sont constitués d'entités et de relations entre ces entités. Les *templates* peuvent ainsi être considérés comme des graphes où les nœuds représentent des entités (ce qui inclut ici des événements) et les arcs représentent les relations entre ces entités. Cette structure offre en outre une grande souplesse de représentation puisqu'elle permet à la fois de représenter la structure finale désirée, un événement lié à un ensemble d'entités, et ses versions préliminaires au cours du processus d'extraction dans lesquelles plusieurs mentions d'événements ou plusieurs occurrences d'une même entité peuvent apparaître. De façon simplificatrice, nous nommerons dans ce qui suit ces graphes «graphes d'entités nommées» ou «graphes d'entités».

Ces graphes d'entités sont plus précisément des graphes pondérés, non orientés, dont les arcs symbolisent l'existence ou l'absence d'une relation entre deux entités. Le poids associé à chaque arc correspond quant à lui à un score de confiance (w_i) et a pour objet de refléter le niveau de confiance quant à l'existence d'une relation entre deux entités. Il est à noter qu'un graphe d'entités n'est pas nécessairement connexe.

La figure 4.2 montre deux exemples de graphes d'entités, en l'occurrence ceux produits pour chacune des phrases de la même figure. Il s'agit de graphes d'entités

¹<http://lit.csci.unt.edu/~textgraphs/ws11/>

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

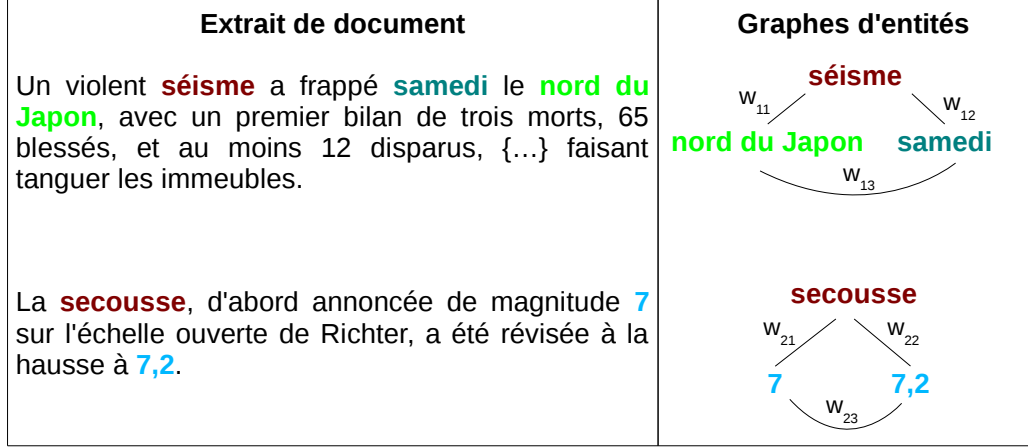


FIG. 4.2 – Exemple de graphes d'entités nommées au niveau des phrases

associés à des *templates* intermédiaires qu'il faudra fusionner pour produire le *template* final. On peut noter que dans le premier cas toutes les entités sont effectivement liées alors que dans le second graphe, les scores de confiance w_{21} et w_{23} devraient être proche de zéro puisque la meilleure valeur pour le rôle MAGNITUDE est 7,2.

L'intérêt d'adopter une structure de représentation abstraite telle que la structure de graphe est de pouvoir réutiliser les méthodes de manipulation associées. Ainsi, une manière générique d'envisager le remplissage de *templates* est de le considérer comme un problème de partitionnement d'un graphe d'entités tel que nous l'avons décrit ci-dessus. Le partitionnement de graphe (ou *clustering* de graphe) [Chen and Ji, 2010; Schaeffer, 2007] est en effet un problème connu, défini comme une tâche visant à regrouper les nœuds d'un graphe sous forme de *clusters* en tenant compte de la structure du graphe de telle façon que le nombre d'arcs à l'intérieur des *clusters* soit plus important que celui entre les *clusters*. Le partitionnement de graphe produit donc des *clusters* assimilables à des sous-graphes fortement connectés. Dans notre cas, ces sous-graphes correspondraient à des instances de *templates*.

On peut noter que la problématique du *clustering* de graphe d'entités est proche de celle du partitionnement présentée à la section 2.4.2.3 mais que compte tenu de l'absence de structure *a priori* dans ce dernier cas, le clustering s'effectue

plutôt au niveau des relations que des entités. L'application du *clustering* de graphe au remplissage de *templates* est en revanche présentée dans [Wick et al., 2006] comme nous avons pu le voir à la section 4.1.1. Cette application s'effectue plus précisément pour l'extraction d'enregistrements de base de données à partir de textes : un graphe est d'abord construit à partir de toutes les entités trouvées dans un document, puis un *clustering* de ce graphe est réalisé afin de reconstituer les enregistrements.

Les méthodes génériques de partitionnement de graphe ne sont toutefois pas très adaptées à notre problématique de remplissage de *templates*. Lors d'un tel *clustering*, les entités sont en effet regroupées sans tenir compte de la structure du *template*. En particulier, il est difficile pour ces méthodes d'intégrer des contraintes visant à exclure d'un *cluster* la présence de plusieurs entités ayant le même rôle vis-à-vis de l'événement, alors que dans notre contexte, le processus de remplissage des *templates* doit ne retenir qu'une seule entité pour chaque rôle. Dans [Wick et al., 2006], le *clustering* de graphe est en revanche plus adapté dans la mesure où les champs des *templates* peuvent être multi-valués (pluralité possible des adresses postales ou des numéros de téléphone pour une personne dans le cas présent). Nous détaillons donc dans la section suivante la méthode spécifique que nous avons définie pour la sélection de la valeur d'un champ parmi plusieurs entités de même type pour le remplissage d'un *template*. Au préalable, la méthode de construction du graphe d'entités servant de point de départ à cette sélection est elle-même précisée.

4.3 Application du rattachement à l'extraction des événements

L'approche pour l'extraction d'événements que nous proposons repose sur l'exploitation d'un graphe d'entités nommées afin de sélectionner les entités les plus pertinentes en rapport avec l'événement principal d'un texte. Dans cette section nous décrivons en détail la phase de construction du graphe d'entités (section 4.3.1) et la phase de sélection des entités à partir de ce même graphe (section 4.3.2).

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

4.3.1 Construction du graphe d'entités

La finalité du graphe d'entités que nous construisons est de caractériser, à l'échelle du document, la présence ou l'absence entre chaque paire d'entités (incluant les mentions d'événement¹) d'une relation sémantique définissant le type d'événements considéré. La construction de ce graphe s'effectue en deux temps : elle commence par la mise en évidence des relations existant à un niveau local (intra-phrastique) et se poursuit par un processus de fusion de ces relations locales permettant l'établissement de relations à une échelle plus globale. Le premier temps est l'équivalent de la construction de *templates* intermédiaires évoquée précédemment. La fusion est quant à elle guidée par la segmentation préalable des textes en événements : seuls les segments faisant référence à l'événement visé, dans notre cas l'*événement principal*, sont considérés à ce stade. Les différentes entités figurant dans ces segments sont donc supposées faire référence au même événement. Les entités partagées par les relations locales, en premier lieu les mentions d'événement, font ainsi office de points de jointure entre ces relations et constituent l'épine dorsale du processus de fusion.

La figure 4.3 montre un exemple de graphe d'entités tel que nous cherchons à le construire et illustre le processus de fusion : l'heure du séisme est localement liée à la mention d'événement *secousse* tandis que sa magnitude est localement associée à la mention d'événement *tremblement de terre*. La réunion de ces deux mentions d'événement, réalisée par le biais de la segmentation événementielle, permet ainsi de rattacher à l'événement principal du texte son heure et sa magnitude. Ce type de jointure s'effectue le plus souvent au niveau des mentions d'événement mais peut faire intervenir d'autres types d'entités lorsqu'une entité est reprise sous une même forme (cf. par exemple le cas de l'entité coordonnées géographiques de la figure 4.3) ou sous des formes pouvant être appariées. issues

Le premier point de notre démarche est la mise en évidence des relations à un niveau local, en l'occurrence au niveau des phrases. Nous reprenons en cela l'approche adoptée par la plupart des travaux comparables et justifiée par la richesse des indices utilisables au niveau phrastique. Cette richesse permet en effet d'obtenir une fiabilité de détection des relations élevée donnant une assise

¹Il n'est pas obligatoire que les relations soient entre une mention d'événement et une autre entité.

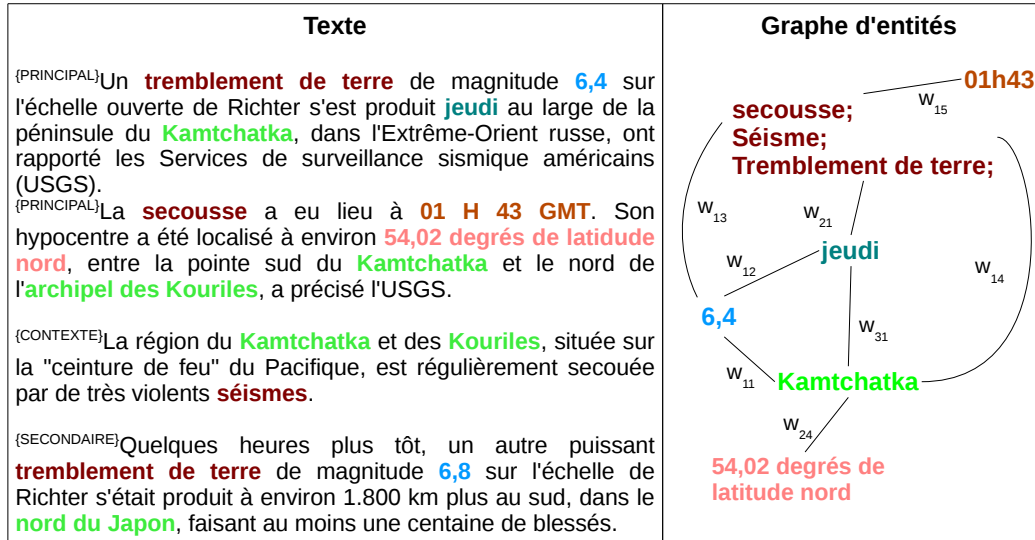


FIG. 4.3 – Exemple de graphe d’entités nommées au niveau du document

solide à l’étape suivante de fusion.

Cette mise en évidence des relations locales est effectuée par l’utilisation d’un classifieur statistique. Celle-ci permet de répondre à un double besoin de la construction du graphe d’entités : d’une part décider si deux entités d’une phrase sont liées par une relation relevant du type d’événements considéré ; d’autre part, évaluer le score de confiance de cette relation, ce qui se traduit au niveau du graphe d’entités par les poids associés aux relations. Les classifieurs statistiques asseyant généralement leur décision de classification sur un score qu’il est possible de transformer, s’il ne l’est pas déjà, en score de confiance, ils permettent de satisfaire les deux besoins. Dans cette perspective, il faut noter que les travaux existants [Afzal, 2009; Gu and Cercone, 2006; Wick et al., 2006] utilisent des ensembles de *features* très dépendants d’informations lexicales telles que les formes de surface des mots. À l’exception de [Liu et al., 2007], ils excluent en revanche les *features* issus d’informations syntaxiques. Globalement, les modèles de relations qu’ils construisent sont donc très dépendants du domaine considéré puisque leurs constituants sont fortement lexicalisés.

À l’inverse, notre objectif est d’obtenir un modèle de relation le plus générique possible, capable de s’adapter facilement à de nouveaux domaines. Un tel modèle

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

doit de notre point de vue s'appuyer principalement sur des informations syntaxiques et avoir un recours limité aux informations lexicales. Pour mesurer l'impact d'un tel choix, nous avons cherché à évaluer la contribution des *features* lexicaux par rapport aux *features* syntaxiques. Pour ce faire, nous avons entraîné plusieurs types de classifieurs en utilisant les ensembles de *features* suivants :

- *FEAT-BASE* : *features* identiques à ceux proposés dans [Afzal, 2009]. Ce sont pour l'essentiel des *features* lexicaux ;
- *FEAT-LEX* : un ensemble de *features* inspiré de [Liu et al., 2007] mêlant *features* lexicaux et syntaxiques¹ ;
- *FEAT-NOLEX* : même *features* que l'ensemble *FEAT-LEX* en supprimant les *features* lexicaux.

Le détail des éléments composant chaque ensemble de *features* est présenté dans le tableau 1.

Description des <i>features</i>	FEAT-BASE	FEAT-LEX	FEAT-NOLEX
Types des entités E1 et E2	✓	✓	✓
POS des entités E1 et E2	✓	✓	✓
Mots des entités E1 et E2	✓		
Bigrammes de mots de E1 et E2	✓	✓	✓
Mots entre E1 et E2	✓	✓	
Bigrammes de mots entre E1 et E2	✓	✓	
POS des mots entre E1 et E2	✓	✓	✓
Nb mots entre E1 et E2	✓	✓	✓
Bigrammes de POS de mots entre E1 et E2		✓	✓
Nb relations syntaxiques entre E1 et E2		✓	✓
Chemin syntaxique entre E1 et E2		✓	✓
Position relative et POS ²		✓	✓
Nb d'entités entre E1 et E2		✓	✓
Nb mentions d'événements entre E1 et E2		✓	✓
POS des deux mots avant/après E1		✓	✓
POS des deux mots avant/après E2		✓	✓

TAB. 1 – *Features* pour la classification de relations. *POS* désigne les catégories morpho-syntaxiques (*Part-Of-Speech tags*) ; E1 et E2 représentent les deux entités nommées de la relation.

¹Certains de leurs *features* ne sont pas pertinents dans notre contexte. Ces derniers ne sont applicables que dans le domaine biomédical.

²Si E1 ou E2 est une mention d'événement alors la position avant/après l'autre entité et son POS.

4.3.2 Sélection des entités et remplissage des *templates*

Compte tenu de l’approche adoptée, le remplissage des *templates* consiste à sélectionner au sein du graphe d’entités les entités effectivement liées à l’événement considéré en précisant le rôle qu’elles viennent y occuper. Ce problème peut être également exprimé sous une autre forme où il s’agirait, pour chaque rôle du *template*, de classer (ordonner selon un critère objectif) les entités et par la suite, associer à chaque rôle l’entité la mieux classée. Comme nous l’avons souligné précédemment, notre travail s’inscrit dans un contexte de *templates* figés : chaque *template* est associé à un domaine spécifique et comporte un ensemble défini *a priori* de rôles.

Cependant, certains rôles peuvent ne pas être renseignés, soit parce que les entités correspondantes n’ont pas été correctement identifiées, soit plus simplement parce que l’information n’est pas mentionnée dans le document. Voir le problème de la sélection des entités comme un problème d’ordonnancement d’une liste d’items permet de se ramener à une problématique très générale. Une façon de l’aborder, utilisée fréquemment en TAL ou en Recherche d’Information, est l’utilisation du vote, avec toutes les variantes que cette option permet (vote majoritaire, vote préférentiel, etc.). Par exemple, [MacDonald, 2009; Santos et al., 2010] proposent d’appliquer les systèmes de vote pour ordonner les entités nommées dans le cadre d’une tâche de recherche d’informations à propos de personnes. [Santos et al., 2010] s’intéresse plus précisément à l’ordonnancement d’une liste d’entités liées à une entité particulière : typiquement, si l’on se pose la question «*quels sont les joueurs de tennis ayant remporté Wimbledon ?*», on espère retrouver une liste de noms de personnes, liste qui peut être ordonnée selon différents critères. Plus généralement, l’approche par vote permet d’intégrer facilement les résultats de plusieurs systèmes reposant sur des méthodes différentes. C’est la perspective dans laquelle nous l’abordons dans notre travail.

Sous un angle différent, les chercheurs se sont intéressés à traiter le problème de l’ordonnancement des éléments d’une liste en adaptant les approches existantes de classification. Cette thématique de recherche, appelée ordonnancement par classification ou «*learning to rank*», vise ainsi à apprendre un modèle capable d’ordonner ou réordonner les items d’une liste. Les travaux de [Liu, 2011, 2009]

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

décrivent plus en détail les méthodes d'apprentissage en matière de «*learning to rank*» en se focalisant sur la recherche d'information¹. Concernant le remplissage des *templates*, nous avons choisi de ne pas retenir ce type d'algorithmes d'ordonnancement. La raison principale est que ce type de modèles nécessite un apprentissage supervisé impliquant la constitution et l'annotation d'un ensemble d'exemples d'une taille conséquente, ce qui est difficile dans le cas présent : il serait en effet nécessaire de constituer un ensemble de listes d'ambiguïtés de rattachement assez large, ce qui implique en pratique d'annoter beaucoup de textes ou d'adopter des stratégies complexes de sélection d'exemples. Par ailleurs, ce type d'algorithmes semblent plus adaptés à l'ordonnancement de longues listes d'items, comme les résultats d'un moteur de recherche, qu'à celui de très courtes listes, comme dans notre cas de figure.

Une façon différente d'aborder l'ordonnancement des entités est d'exploiter directement le graphe d'entités, en particulier au travers de sa structure. Depuis l'avènement du Web, l'étude de l'analyse des liens (*link analysis*) dans les graphes a connu un essor important et s'est notamment attachée à développer des méthodes permettant d'évaluer l'importance d'un nœud d'un graphe par rapport aux autres en se fondant sur les liens entre ces différents nœuds. Ces méthodes ont été par ailleurs largement appliquées dans le domaine du TAL et font maintenant partie des outils fréquemment exploités dans ce contexte [Mihalcea and Radev, 2006]. Nous proposons donc d'appliquer ces méthodes à la sélection des entités dans notre graphe d'entités en nous appuyant sur une adaptation d'un algorithme classique en la matière, l'algorithme *PageRank* [Page et al., 1999]. Cet algorithme est généralement utilisé sur des graphes dirigés et non pondérés. À l'origine il a été appliqué afin d'ordonner des pages Web. L'idée étant de représenter chacune des pages comme un nœud du graphe, et d'exploiter les liens de redirection (entrants et sortants) entre les pages afin de calculer un score de popularité pour chacune d'entre elles. Dans [Mihalcea and Radev, 2006], cet algorithme est appliqué au résumé automatique. Plus précisément, l'algorithme est adapté afin de calculer un score de popularité pour les nœuds mais à partir d'un graphe non dirigé et pondéré. Dans ce cadre du résumé automatique, le fait de prendre en

¹Ces méthodes ne sont pas abordées dans ce manuscrit mais nous renvoyons le lecteur vers les approches décrites dans [Burges et al., 2006; Herbrich et al., 2000; Sculley, 2010].

considération les pondérations des arcs permet d'améliorer la sélection des phrases pertinentes.

Pour notre processus de sélection des entités, nous avons choisi de tester plusieurs méthodes inspirées des méthodes à base de vote et d'analyse de liens :

Position est une approche simple mais efficace dans notre contexte applicatif qui choisit pour chaque rôle, l'entité apparaissant en première position dans le segment de l'événement principal. Cette approche nous sert de référence (*baseline*).

Confiance est une approche se contentant de sélectionner pour chaque rôle, l'entité connectée à une mention d'événements ayant le plus fort poids selon les pondérations sur les arcs du graphe d'entité.

PageRank est approche issue de l'analyse des liens, plus directement de l'algorithme PageRank. La démarche consiste à appliquer l'algorithme afin d'attribuer un score à chaque nœud du graphe. Par la suite, pour chaque rôle on retient l'entité qui a le meilleur score retourné par l'algorithme.

Vote est une approche de sélection par vote. Il s'agit d'un vote majoritaire, utilisant comme point de départ les sorties fournies par les approches *Confiance*, *PageRank* et *Position*. Un vote est organisé pour chaque rôle et celle ayant le plus grand nombre de voix est retenue.

Hybride est une approche de sélection ayant elle aussi pour objet est de combiner les résultats des approches précédentes (*Confiance*, *PageRank* et *Position*). Elle part du constat qu'une approche de sélection peut être pertinente pour un ou plusieurs rôles et plus inadaptée pour d'autres. Son principe est donc d'améliorer les performances en appliquant des stratégies de sélection d'entités adaptées en fonction des rôles dans le *template*. Plus précisément, elle associe à chaque rôle l'approche de sélection obtenant les meilleurs résultats pour ce rôle.

Par ailleurs, les résultats des approches *Confiance*, *PageRank*, *Vote* et *Hybride* sont complétés par une heuristique de sélection faisant office de mécanisme de rattrapage (*back-off*), c'est-à-dire gérant les cas où aucune entité n'est choisie pour un rôle donné. Par exemple, lorsqu'une entité apparaît de façon isolée dans une phrase, l'étape de classification des relations binaires ne peut s'appliquer. Par

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

conséquent, cette entité ne peut être sélectionnée puisque non présente dans le graphe d'entités. Ce mécanisme de *back-off* apporte donc une solution dans de tels cas de figure.

Notre contexte applicatif nous a conduit enfin à laisser de côté deux problématiques en lien avec le remplissage des *templates*. La première concerne l'opportunité de remplir un rôle. Dans le cas présent, chaque rôle pouvant être rempli par une entité compatible est effectivement occupé par cette entité. Dans le domaine et pour le type de textes considérés, cette stratégie s'est en effet avérée la plus efficace et justifie notre mécanisme de back-off. Toutefois, nous ne prétendons pas qu'il en est nécessairement ainsi pour tous les domaines. La seconde problématique a trait aux relations entre les rôles des *templates* et le type des entités. Dans l'application liée au domaine sismique qui nous a servi de cadre expérimental, cette relation est de nature bijective : chaque rôle est associé à un type d'entités spécifique, ce qui constitue un cas favorable. Lorsqu'une distinction entre rôle et type d'entités existent, il est nécessaire d'introduire des contraintes d'appariement supplémentaires. La solution la plus évidente de ce point de vue consiste sans doute à s'appuyer sur le résultat de classifieurs locaux plus spécifiques intégrant la notion de rôle.

4.4 Application et évaluation de l'approche de rattachement

Dans cette section, nous évaluons notre approche de rattachement des entités nommées en vue de la construction des *templates*. Nous détaillons ici les résultats des expérimentations pour les deux phases du processus, à la section 4.4.1 pour la construction du graphe d'entités et à la section 4.4.2 pour la sélection des entités. De plus, nous démontrons l'impact de la segmentation sur le résultat final à la section 4.4.3. Enfin, la section 4.4.4 propose une analyse des erreurs restantes suite à la construction des *templates*.

4.4.1 Construction du graphe d'entités

La construction du graphe d'entités dépend de la détection à un niveau intra-phrastique de relations sémantiques binaires entre les entités. Comme nous l'avons indiqué précédemment, cette identification est réalisée par un classifieur statistique. Plus précisément, plusieurs modèles de classification ont été évalués en utilisant les ensembles de *features* *FEAT-BASE*, *FEAT-LEX*, *FEAT-NOLEX* présentés à la section 4.3.1.

Pour l'entraînement de ces modèles, un corpus de 5 000 relations binaires¹ a été constitué à partir de 44 dépêches issues du corpus d'évaluation présenté à la section 3.7.1. Ces relations ont été manuellement annotées et incluent à la fois des relations inter-phrastiques (3 231) et des relations intra-phrastiques (969). Cette annotation consistait à attribuer une des deux classes possibles à chaque relation : classe *POSITIVE* pour indiquer que les deux entités jouent un rôle dans la description d'un même événement et classe *NEGATIVE* dans le cas où les ne sont pas liées par le même événement. Pour illustration, les extraits ci-dessous montrent des exemples de relations annotées. Pour l'entraînement des modèles de détection des relations, seules les relations intra-phrastiques ont été utilisées.

Relations intra-phrastiques

[*POSITIVE*] : Un séisme de magnitude 5,6 sur l'échelle de Richter s'est produit jeudi dans la région de Trinidad, ...

[*NEGATIVE*] : Ce tremblement de terre avait été précédé deux heures plus tôt par une secousse de magnitude 5,0 dans la même région, selon l'USGS.

Relations inter-phrastiques

[*POSITIVE*] : Un séisme de magnitude 5,1 a secoué lundi une région isolée du sud-ouest de la Chine, a annoncé {...} après le séisme du 12 mai. La secousse a été enregistrée à 01 H 56 lundi (dimanche 17 H 56 GMT) dans la province ...

[*NEGATIVE*] : Un séisme de magnitude 5,1 a secoué lundi une région isolée du sud-ouest de la Chine, a annoncé {...} après le séisme du 12 mai. La secousse a été enregistrée à 01 H 56 lundi (dimanche 17 H 56 GMT) dans la province ...

¹Ces relations incluent des relations entre deux entités uniquement ainsi que des relations entre une entité et une mention d'événement.

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

Étant donné que le corpus annoté contient un nombre significatif d’instances négatives de relations, c’est-à-dire de couples d’entités au sein d’une phrase n’entretenant pas de relation, nous avons volontairement écarté une part importante de ces exemples négatifs afin de ne pas déséquilibrer le corpus d’apprentissage en faveur de ceux-ci. Au final, nous avons retenu 690 instances positives et 236 instances négatives. Pour l’apprentissage, nous avons repris les implémentations des modèles Bayésien Naïf (*NB*), Maximum d’Entropie (*ME*) et Arbres de décision (*DT*) disponibles au sein de la boîte à outils Mallet¹. Les performances de ces modèles en termes de rappel (*R*), précision (*P*) et F1-mesure (*F*) en fonction de nos différents ensembles de *features* sont données dans le tableau 2. Ces résultats sont obtenus par le biais d’une validation croisée, en exploitant 4/5 des données pour la phase d’apprentissage et 1/5 pour celle de test. Ces résultats sont également comparés à une approche de référence simpliste (*baseline*) consistant à attribuer la classe *POSITIVE* à toutes les relations.

Ensemble de <i>features</i>	Algo.	R(%)	P(%)	F(%)
FEAT-LEX	ME	96,3	95,9	96,1
FEAT-BASE	ME	91,2	96,1	93,6
FEAT-NOLEX	ME	91,7	95,0	93,3
FEAT-LEX	DT	89,0	96,5	92,6
FEAT-LEX	NB	93,4	90,7	92,0
FEAT-NOLEX	DT	91,2	88,7	89,8
FEAT-NOLEX	NB	89,6	89,2	89,4
FEAT-BASE	DT	84,4	94,7	89,2
FEAT-BASE	NB	86,7	87,9	87,3
Baseline	—	100,0	25,5	40,5

TAB. 2 – Résultats de l’évaluation de la classification binaire des relations (avec ME : Maximum d’Entropie, DT : arbres de décision, NB : Bayésien Naïf).

Les résultats du tableau 2 montrent que l’ensemble de *features* FEAT-LEX améliore les résultats obtenus par l’ensemble FEAT-BASE pour l’ensemble des modèles utilisés, ce qui conduit à penser que les *features* supplémentaires de nature syntaxique de FEAT-LEX sont particulièrement intéressants pour l’identification des relations entre entités au sein des phrases. De plus, le tableau montre

¹<http://mallet.cs.umass.edu/>

également que l'ensemble FEAT-NOLEX, qui n'est pas lexicalisé, obtient des résultats équivalents à ceux de FEAT-BASE, qui comprend pour sa part des *features* lexicaux. Nous pouvons donc en conclure que non seulement les *features* syntaxiques utilisés ici apportent un plus par rapport à des *feature* lexicaux mais qu'ils peuvent même leur être substitués, point intéressant dans l'optique d'une adaptation à un autre domaine applicatif.

Au niveau des modèles, les résultats permettent d'établir la hiérarchie : ME > DT > NB. [Afzal, 2009] obtient pour sa part une hiérarchie différente : DT > ME > NB. Cependant les deux évaluations ne sont pas directement comparables puisque [Afzal, 2009] traite une autre langue et un corpus différent, ce qui peut influencer notablement sur les performances des différents modèles. Globalement, nos résultats sont néanmoins assez proches de ceux obtenus par [Afzal, 2009]. Les meilleurs scores de ce dernier, obtenus avec des arbres de décision, sont en effet R=95 %|P=87 %|F=91 %.

Pour la suite des traitements de notre approche, nous avons retenu le modèle de type Maximum d'Entropie entraîné à partir de l'ensemble de *features* FEAT-NOLEX. Bien que FEAT-LEX obtienne un niveau de performance un peu supérieur, notre choix est motivé par le fait que la différence entre les deux ensembles de *features* est faible et compensée à nos yeux par le fait de ne pas avoir recours à des informations fortement dépendantes d'un domaine, en l'occurrence issues des *features* lexicaux.

Dans le processus de construction du graphe d'entités, la mise en évidence des relations au niveau phrastique est suivie d'une étape de fusion. Celle-ci permet d'abord d'identifier tous les nœuds faisant référence à la même valeur d'entité (y compris les mentions d'événements) et donc, de supprimer les doublons. Ensuite et surtout, elle permet ainsi d'établir des relations inter-phrastiques entre les entités. Pour ce qui concerne les événements sismiques, l'étape de fusion est appliquée, en dehors des mentions d'événements, aux entités de type date et lieu : toutes les mentions de date ayant la même normalisation sont considérées équivalentes, de même pour toutes les mentions de lieux ayant la même forme de surface. Plus généralement, ce mécanisme de fusion peut être appliqué à tous les types d'entités en regroupant les entités ayant les mêmes valeurs.

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

4.4.2 Sélection des entités et remplissage des *templates*

Nous avons vu à la section 4.3.2 que notre approche du remplissage des *templates* est fondée sur la sélection d'entités à partir du graphe d'entités. La démarche générale consiste à déterminer, pour chaque entité, un score évaluant son importance dans le graphe, ce qui permet en particulier d'établir un ordre sur les différentes entités de même type. L'hypothèse sous-jacente est que les valeurs des champs des *templates* se trouvent parmi les entités les mieux classées.

Pour l'évaluation des stratégies de sélection des entités, tous les documents de notre corpus initial sont utilisés. Nous faisons apparaître dans le tableau 3 les résultats de l'évaluation du remplissage des *templates* en termes de rappel (R), précision (P) et F1-mesure (F), tous les rôles étant confondus.

Approche	R(%)	P(%)	F(%)
Hybride	77,6	76,9	77,2
Vote	74,9	74,3	74,5
Confiance	74,9	74,2	74,5
Position	73,4	73,1	73,2
PageRank	72,4	71,7	72,0

TAB. 3 – Évaluation du rattachement des entités aux événements

Ces résultats montrent en premier lieu que notre méthode de référence *Position* est caractérisée par un niveau déjà très élevé. En particulier, elle permet d'obtenir des performances légèrement supérieures à la stratégie *PageRank*. Une explication possible de ce constat est que la stratégie *PageRank* ne repose que sur la structure du graphe, sans considérer les poids sur les arcs. De ce fait, les entités les mieux classées sont les entités fortement connectées indépendamment du poids sur les arcs. Par conséquent, si plusieurs entités non valides sont densément connectées, l'algorithme PageRank leur attribue un bon score. Ce problème pourrait être dans une certaine mesure minimisé en adoptant la version pondérée de l'algorithme PageRank proposée dans [Mihalcea, 2004]. Le tableau 3 laisse néanmoins apparaître que la meilleure méthode non agrégative, en l'occurrence la méthode *Confiance*, dépasse nettement la méthode *Position*. Parmi les méthodes agrégatives, l'approche *Hybride* se révèle nettement supérieure à l'approche *Vote*.

En outre, elle s'avère être globalement la meilleure stratégie de sélection des entités, ce qui n'est pas complètement surprenant compte tenu de sa définition. On notera par ailleurs que tous les résultats obtenus sont assez équilibrés pour ce qui est du rapport entre la précision et le rappel.

4.4.3 Impact de la segmentation sur le rattachement

Afin de compléter l'évaluation précédente, nous proposons d'évaluer l'impact de la segmentation en événements sur le rattachement des entités aux événements. La segmentation est utilisée pour identifier les régions textuelles liée à un type d'événements particulier. Cependant, les documents ne font pas tous référence à plusieurs événements. Lorsqu'un document ne fait référence qu'à un seul événement, il peut donc sembler moins pertinent d'appliquer la segmentation puisque toutes les phrases font référence au même événement. Celle-ci n'est dès lors susceptible que d'apporter des perturbations dans la mesure où ses résultats ne sont nécessairement pas parfaits.

De ce fait, nous cherchons à évaluer dans cette section l'impact de la segmentation sur les documents ne faisant référence qu'à un seul événement en comparaison avec ceux faisant référence à plusieurs événements. Notre intuition est que la segmentation devrait avoir un effet limité sur les documents mono-événements et devrait améliorer les résultats pour les documents multi-événements. Afin de vérifier cette hypothèse, nous avons manuellement divisé le corpus initial en deux parties en fonction du nombre d'événements sismiques qu'ils relatent. Après découpage du corpus, nous avons ainsi obtenu 227 documents multi-événements (M) et 274 documents mono-événements (S). Les différentes stratégies de construction des *templates* évoquées précédemment ont été appliquées à ces deux ensembles de documents en tenant compte ou non de la segmentation. Les résultats sont présentés dans le tableau 4 en termes de F1-mesure en regroupant tous les rôles.

En ce qui concerne les documents mono-événements, les résultats du tableau 4 montrent que les meilleures approches de sélection sont sans surprise celles sans segmentation préalable, même si de façon globale la différence avec l'utilisation de la segmentation n'est pas très marquée (en moyenne +0,7 %). À l'inverse, la sélection des entités avec segmentation des dépêches obtient de meilleurs

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

	Sans segmentation		Avec segmentation	
	S(%)	M(%)	S(%)	M(%)
<i>Approche</i>				
Hybride	79,2	73,6	78,3	75,6
Vote	77,7	68,7	76,9	71,8
Confiance	72,6	66,1	71,8	69,1
Position	74,0	73,2	73,1	73,1
PageRank	70,9	59,7	70,7	65,3

TAB. 4 – Impact de la segmentation selon le nombre d’événements (mono-événements (S)/multi-événements (M)) dans les documents (F1-mesure)

résultats sur les documents faisant référence à plusieurs événements (en moyenne +2,7 %). On peut noter que la stratégie la plus efficace (*Hybride* avec segmentation) dépasse l’approche «*Baseline*», à savoir *Position* sans segmentation, pour les deux catégories de documents. Plus généralement, ces résultats démontrent que la segmentation événementielle a globalement un impact positif sur le processus de remplissage des *templates* en ne dégradant pas trop le niveau des résultats sur les documents mono-événements et en l’améliorant pour les documents multi-événements.

4.4.4 Analyse d’erreurs

Afin d’obtenir une meilleure connaissance des performances des approches proposées pour le remplissage des *templates*, nous avons poussé un peu plus loin l’analyse des erreurs en cherchant à identifier précisément les causes de la présence d’une valeur erronée (sélection d’une mauvaise entité pour un rôle) ou de l’absence de valeur (pas d’entité sélectionnée pour un rôle) au niveau d’un champ du *template* à remplir. Nous avons ainsi identifié trois types d’erreurs prépondérants, correspondant aux trois premiers états de la valeur d’un champ dans la liste suivante :

- erreur de reconnaissance des entités nommées (*NE-err*) : l’entité n’est pas reconnue lors du pré-traitement linguistique ;
- erreur de segmentation en événements (*Seg-err*) : l’entité est identifiée lors du pré-traitement mais elle appartient à une phrase qui n’est pas associée

-
- à l'événement principal ;
 - erreur de sélection des entités (*Fill-err*) : l'entité se trouve dans le segment principal mais une autre entité a été retenue comme valeur pour le rôle dans le *template* ;
 - *Correcte* : l'entité a été correctement identifiée puis sélectionnée comme rôle dans le *template*.

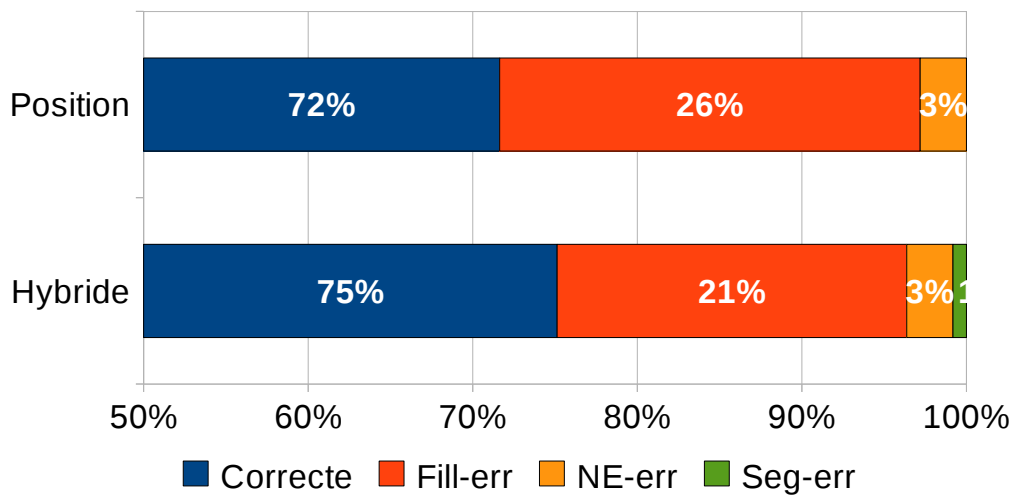


FIG. 4.4 – Répartition des erreurs, par type

La figure 4.4 présente la répartition de chaque type d'erreurs pour deux approches de construction des *templates* en utilisant l'intégralité du corpus¹ : la première correspond à une sélection à base d'heuristique sans segmentation en événements (*Position*) ; la seconde utilise une stratégie hybride avec la segmentation en événements (*Hybride*).

Le graphe montre que la stratégie *Position* permet d'identifier correctement une part importante des entités (72 %) bien qu'un nombre conséquent d'erreurs de rattachement des entités subsiste (26 %). La stratégie la plus efficace que nous avons proposée réduit ce type d'erreurs tout en augmentant le pourcentage de valeurs d'entités correctes dans le *template*. Cette double amélioration se fait au prix de l'introduction d'un nombre très limité d'erreurs dues à la segmentation

¹Les pourcentages sur le graphe sont arrondis au supérieur, ce qui explique une somme différente de 100 %.

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

en événements (1 %).

4.5 Conclusions

Ce chapitre a abordé le problème du rattachement des entités aux événements, c'est-à-dire de l'attribution des rôles des entités vis-à-vis des événements, dans la perspective du remplissage de *templates*. Nous avons vu que l'essentiel des approches dans ce domaine reposent sur des approches heuristiques exploitant surtout des informations locales. À l'inverse, nous avons proposé une approche d'extraction visant à prendre en compte les informations allant au-delà de la phrase, en particulier en utilisant les résultats d'une segmentation événementielle des textes plutôt que les phrases une à une. Les expérimentations sur l'extraction d'événements sismiques ont montré que ce choix était intéressant et qu'il s'avérait particulièrement pertinent pour des textes décrivant plusieurs événements de même nature sans entraver le traitement de textes mono-événements (décrivant un seul événement).

Dans notre approche, nous avons par ailleurs considéré les *templates* comme des relations complexes exprimant des relations de degré n entre n entités. Leur identification à une échelle textuelle s'effectue en deux temps en considérant que ces relations complexes peuvent se décomposer en un ensemble de relations binaires : le premier temps réalise l'identification de ces relations élémentaires tandis que le second sélectionne les entités jouant un rôle dans la relation complexe, c'est-à-dire sélectionne les valeurs des champs du *template* considéré.

Pour l'identification des relations sémantiques élémentaires entre les entités, plusieurs classifieurs statistiques ont été testés en s'appuyant sur différents ensembles de *features*. Les résultats ont montré que l'utilisation de *features* lexicaux permet d'obtenir une bonne classification des couples d'entités. En revanche, ce type de *features* implique une forte dépendance des modèles par rapport au domaine et au type de textes considérés. Les résultats obtenus ont aussi montré que la substitution de *features* syntaxiques aux *features* lexicaux entraîne certes une légère dégradation des résultats mais que cette dégradation est suffisamment minime par rapport au gain de généralité des modèles qui en résulte.

À la suite de cette première phase, les relations mises en évidence servent à

construire un graphe d'entités au sein duquel les entités potentiellement pertinentes sont à sélectionner. Pour réaliser cette sélection, nous avons mis en œuvre et testé trois types de stratégies, avec le souci de rendre ce processus le moins dépendant que possible du domaine. La première, qui sert aussi de *baseline*, est purement heuristique et ne repose que sur l'ordre d'apparition des entités. La deuxième s'appuie sur les pondérations des arcs du graphe tandis que la dernière exploite la structure du graphe. De plus, nous avons également proposé deux méthodes pour combiner les sorties des trois stratégies précédentes, l'une inspirée du vote majoritaire, l'autre étant fondée sur une combinaison de stratégies tenant compte de leurs performances différenciées en fonction de chaque rôle spécifique du *template*. Les expériences ont montré que les meilleurs résultats sont obtenus de façon générale par la combinaison de nos trois stratégies de base, et plus particulièrement par la combinaison que l'on peut qualifier d'informée, c'est-à-dire adaptant la stratégie à utiliser en fonction du rôle à sélectionner.

4. LE RATTACHEMENT DES ENTITÉS AUX ÉVÉNEMENTS

Chapitre 5

Peuplement de bases de connaissances

Dans les deux chapitres précédents, nous nous sommes intéressés à un processus de construction de *templates* appliqué à l'extraction d'information pour les événements sismiques. Pour reprendre la terminologie des conférences MUC, cette tâche correspond à la tâche de *scenario template extraction*. Ainsi, les *template* que nous avons extraits contenaient des relations liées à un domaine particulier et en nombre restreint. Dans une perspective différente, mais complémentaire, ce chapitre porte sur l'extraction de relations entre entités nommées à une plus large échelle et pour un domaine plus général. Par rapport à la terminologie MUC, la tâche la plus proche correspond à celle de *template element extraction* ou *slot filling*, dont le but est d'extraire des renseignements (ou informations complémentaires) concernant des entités nommées. Typiquement, il s'agit d'extraire des propriétés caractéristiques associées à un type d'entité donné, par exemple une entité de type personne est toujours liée à une date de naissance.

5.1 Introduction

Comme nous l'avons mentionné dans le préambule, les *templates* se rapportent à des scénarios servant à décrire des événements. À l'inverse, dans ce chapitre, nous nous concentrons sur des *templates* plus généraux, qui ne sont plus focalisés

5. PEUPLEMENT DE BASES DE CONNAISSANCES

sur des événements mais sur des entités : l'idée principale est de décrire une entité en se servant des relations qu'elle partage avec les autres entités. Si l'on peut considérer que toutes les relations liées à une même entité constituent une forme de *template*, la différence réside dans le fait que dans le cas des *template* d'entités, une relation est établie avec l'entité pour chaque champ, ce qui forme un ensemble de relations binaire souvent, mais pas toujours indépendantes, alors que dans le cas des *templates* d'événements, les différents champs sont en général inter-dépendants et forment une relation n-aire globale.

Plus généralement, notre objectif est d'obtenir des informations complémentaires (et indépendantes d'un domaine) concernant les entités. Notre motivation est de se servir de ces informations complémentaires afin de situer un événement dans un cadre plus général que celui du document qui en fait mention. Pour cela, notre idée est d'apporter des connaissances sur chaque entité selon son type et indépendamment de l'événement auquel elle appartient. Pour illustration, lorsqu'un événement se produit dans un lieu donné, il peut être intéressant de connaître le nombre d'habitants associé à ce lieu, ou encore le nom de la capitale lorsqu'il s'agit d'un pays, etc.

Dans l'ensemble, les informations complémentaires que nous cherchons à extraire sont des connaissances encyclopédiques sur les entités. Ce type de connaissances peut se trouver dans des sources d'informations ouvertes, en particulier dans le contexte du Web sémantique. Un grand nombre d'informations sont par exemple disponibles sous forme semi-structurée dans le contexte de l'encyclopédie collaborative Wikipédia, sous la forme d'*infobox*, c'est-à-dire de tables formatées qui contiennent des informations factuelles liées à l'entité d'intérêt de la page. Ces données semi-structurées peuvent être structurées automatiquement sous forme d'une base de données, comme le montre le projet DBpedia¹ [Bizer et al., 2009]. Malheureusement, ces ressources sont parfois incomplètes : dans Wikipedia, les entités populaires sont bien renseignées, les autres le sont beaucoup moins. Pour pallier ce problème, une alternative consiste à utiliser le contenu textuel non structuré issu des articles de l'encyclopédie pour enrichir des bases de connaissances (*Knowledge Base*, ou *KB*) incomplètes. Ici, nous faisons l'hypothèse qu'enrichir une base de connaissances revient à extraire des relations entre des entités

¹<http://dbpedia.org/About>

nommées : on considère que chacune des entrées de la KB décrit une entité et par conséquent, les différents champs de la KB définissent des relations entre cette entité et les valeurs de ces champs.

Dans le chapitre 2, nous avons mis l'accent sur les approches d'extraction de relations supervisées. Néanmoins, les relations peuvent également être extraites à partir d'approches non supervisées, dont but est d'extraire des relations sans *a priori* sur les types de relations [Banko and Etzioni, 2008; Shinyama and Sekine, 2006; Wang et al., 2011a; Yan et al., 2009], ou faiblement supervisées [Bunescu and Mooney, 2007; Mintz et al., 2009]. Le principe de ces dernières est comparable à celle des approche semi-supervisées : il s'agit d'entraîner un système en utilisant un ensemble d'exemples de relations annotés. Pour les méthodes semi-supervisées, cet ensemble est obtenu en utilisant un nombre d'exemples restreint dont on est sûr de la pertinence [Agichtein and Gravano, 2000; Brin, 1999]. À l'inverse, dans le contexte faiblement supervisé, ces exemples sont obtenus de façon automatique en exploitant des ressources extérieures (KB, corpus non annotés, etc.) [Mintz et al., 2009; Suchanek et al., 2006]. Par conséquent l'ensemble d'exemples obtenu est plus important, en revanche il peut contenir des exemples non pertinents. Plus généralement, cette démarche consistant à constituer un ensemble d'exemples de relations est aussi appelée supervision distante [Mintz et al., 2009]. La motivation de ce type d'approches est d'utiliser un ensemble de ressources plutôt que des annotateurs humains comme source de supervision.

Dans ce chapitre, nous présentons un système d'extraction d'information à large échelle fondé sur un apprentissage faiblement supervisé de patrons d'extraction de relations. Ce système a été évalué dans le cadre de la tâche de peuplement automatique d'une base de connaissances (*Knowledge Base Population – KBP*), au sein de la campagne d'évaluation TAC (*Text Analysis Conference*) organisée par le NIST *National Institute of Standards and Technology*. La section 5.2 présente l'objet du peuplement des bases de connaissances. La section 5.4 aborde l'approche d'extraction de relations que nous proposons pour cette tâche. La campagne d'évaluation de référence dans le domaine de la population de base de données est présentée en section 5.5. Les sections 5.6 et 5.7 présentent respectivement les résultats de l'évaluation de notre approche sur les données TAC-KBP 2010 et un aperçu d'autres approches utilisées pour cette tâche. Enfin, les sec-

5. PEUPLEMENT DE BASES DE CONNAISSANCES

tions 5.8 et 5.9 présentent une discussion sur les résultats ainsi que quelques conclusions.

5.2 Le peuplement de bases de connaissances

Cette thématique vise à compléter de façon automatique (éventuellement autonome) une base de connaissances à partir de faits ou d'informations collectées dans des textes non structurés. La motivation vient du fait que collecter puis compiler manuellement des informations non structurées est un processus coûteux en terme de temps et de main d'œuvre. Pour accélérer ce processus, de nombreux travaux ont proposé d'utiliser les approches d'extraction de relations sur des documents issus soit du Web [Banko et al., 2007; Etzioni et al., 2004; Pantel and Pennacchiotti, 2006; Yates et al., 2007], soit de corpus existants [Agichtein and Gravano, 2000] afin d'acquérir des connaissances sur ces faits. Les informations collectées de cette façon sont ensuite sauvegardées dans des bases de données [Etzioni et al., 2004] ou dans des ontologies [Suchanek, 2009; Suchanek et al., 2007].

5.3 Lien entre peuplement de KB et question-réponse

Les domaines du question-réponse (*question-answering (QA)*) et du peuplement de base de connaissances tel qu'il est envisagé dans *KBP* sont très proches. On peut considérer le second comme un sous-problème du premier. Le point commun entre les deux domaines est qu'il est question de trouver un ou plusieurs éléments d'information spécifiques en rapport avec la formulation d'un besoin d'information donné.

Dans la perspective QA, la formulation de la demande passe par des questions écrites en langue naturelle (Qui a fondé Apple?) alors que, pour le domaine *KBP*, cette formulation passe par des requêtes formelles exprimées par des prédicats incomplets contenant : une valeur d'entité, un type de question, ainsi que le type de réponse que l'on attend. Par exemple, dans le cadre de *KBP*, la question «Qui

a fondé Apple?» prend la forme *créateur(Apple, PERSONNE)*. Il reste ensuite à déterminer la valeur de l'entité PERSONNE manquante pour compléter le prédicat.

Une distinction entre les deux domaines est qu'un processus de QA nécessite une étape de reformulation de la question qui n'est pas nécessaire pour le domaine KBP. Cette étape de reformulation passe par une analyse de la question qui sert à extraire les éléments formels sur lesquels portent la question. Plus précisément elle consiste à : faire la détection des entités nommées dans la question (*Apple* est de type ORGANISATION) ; reconnaître le type de question (question factuelle, définition, etc.) ; et enfin, apporter des précisions sur le type de réponse attendu : une personne, une date, un lieu, etc. (ici le «*créateur*» recherché est une personne, mais il pourrait être un groupe de personne, ou une organisation).

Pour le peuplement des bases de connaissances, la nature des questions et le type des réponses sont prédéterminés par la structure de la base de connaissances.

5.4 Vue d'ensemble de l'approche pour l'extraction de relations

Nous nous concentrons dans notre approche sur l'extraction de relations à large échelle en supposant la préexistence d'une base de connaissances partiellement remplie¹, par exemple par extraction automatique à partir de données semi-structurées.

La notion de «large échelle» se décline quant à elle selon plusieurs dimensions. La première réside dans le grand nombre de relations considéré (plusieurs dizaines de relations), induisant une mise en œuvre difficile pour une approche à base de règles écrites manuellement. La deuxième est liée à la prise en compte initiale d'un grand nombre de relations existantes dans la KB (c'est-à-dire l'association de deux valeurs d'entités à un type de relations) ; ces relations fournissent un bon ensemble de départ pour l'apprentissage automatique d'un modèle de ces types de relations. Enfin, le corpus dans lequel de nouvelles relations sont recherchées

¹Nous faisons l'hypothèse que chaque entité fait l'objet d'une entrée dans la base de connaissances et est associée à un ensemble de champs. Les champs de cet ensemble ne sont néanmoins pas toujours renseignés.

5. PEUPLEMENT DE BASES DE CONNAISSANCES

est lui-même de taille importante (près de deux millions de documents), ce qui implique l'utilisation de techniques de recherche d'information (indexation des documents et formulation d'une requête) pour extraire des bons candidats car on ne peut pas envisager l'application directe de patrons sur toutes les phrases du corpus.

Dans la suite, nous appelons *type de relation* un champ R de la base de connaissance associant deux types d'entité (un pour l'entité de référence, l'autre pour la valeur du champ associé à cette entité dans la base de connaissances), *relation* le triplet $R(E1, E2)$ associant un type de relation et deux entités (qui correspond donc à une instance de relation), et *occurrence de relation* l'expression d'une relation dans un texte : cette expression est donc une phrase ou un passage contenant les deux entités et une formulation linguistique de la relation entre les deux entités.

Cette approche, illustrée par la figure 5.1, s'articule en deux phases : une phase d'*apprentissage de patrons* à partir d'occurrences de relations connues et une phase d'*extraction de relations* pour la découverte de nouvelles relations. La première phase part des relations connues $R(E1, E2)$ pour trouver des occurrences de ces relations dans un corpus, c'est-à-dire les différentes expressions de cette relation dans les textes. Cette recherche d'occurrence de relations se fait par l'interrogation d'un moteur de recherche sur une structure indexée du corpus de documents¹. Ces occurrences de relations connues sont alors utilisées pour induire² des patrons de reconnaissance pour le type de relations concerné.

La seconde phase part de relations incomplètes $R(E1, x)$, où l'entité source $E1$ est connue et l'entité cible x est à trouver, cherche des occurrences de relations impliquant $E1$ dans un corpus, puis extrait l'entité x en utilisant les patrons induits dans la première phase. Ces deux phases sont détaillées dans les sections suivantes.

¹La construction de cette structure indexée est décrite à la section 5.4.3.

²Le processus d'induction est décrit à la section 5.4.1.

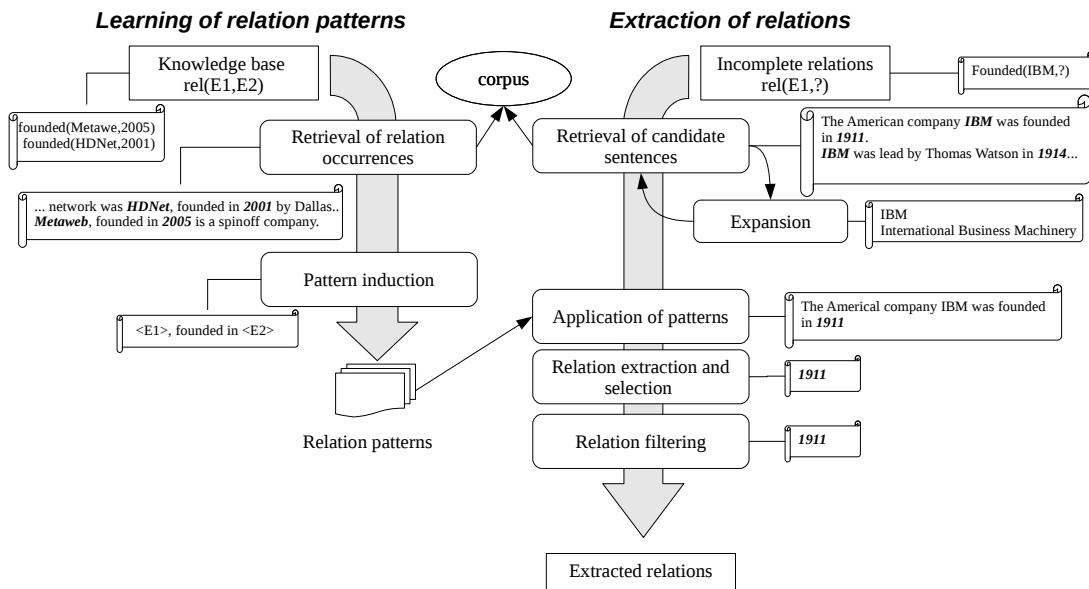


FIG. 5.1 – Architecture générale du système

5.4.1 Apprentissage des patrons de relations

L'apprentissage des patrons de relations repose sur l'induction (ou généralisation) de patrons lexicaux à partir de phrases exemples contenant des occurrences des relations considérées. L'objectif de cet apprentissage est de capturer les différentes expressions d'une relation sémantique entre deux entités. Par exemple, les deux extraits de phrases ci-dessous contiennent des occurrences de relations pour le type *founded_by*, instancié pour les couples d'entités (Charles Revson ; Revlon Cosmetics) et (Mayer Lehman ; Lehman Brothers investment).

The glamorous cabaret chanteuse reportedly had had a romantic liaison with <source>Charles Revson</source>, the founder of <cible>Revlon Cosmetics</cible> ...

Lehman was a great-grandson of <source>Mayer Lehman</source>, a founder of the <cible>Lehman Brothers investment</cible> house ...

Plusieurs travaux présentent des algorithmes de généralisation de patrons lexicaux [Ravichandran, 2005; Ruiz-Casado et al., 2007; Schlaefer et al., 2006]. Notre approche est similaire à celle de [Pantel et al., 2004] et reprend plus di-

5. PEUPLEMENT DE BASES DE CONNAISSANCES

rectement encore la méthode de [Embarek and Ferret, 2008]. L'idée générale de l'approche est de trouver, dans le contexte entre les entités cible et source, des points communs entre deux phrases exprimant la relation que l'on veut capturer. Ici, nous cherchons ces points communs parmi trois niveaux d'information linguistique : forme de surface (les mots), lemme (forme normalisée d'un mot) et catégorie morpho-syntaxique (catégorie grammaticale du mot). Ces informations linguistiques sont mises en évidence grâce à l'outil OpenNLP¹, qui est également utilisé pour la reconnaissance des entités nommées. La présence de ces trois niveaux d'information donne une plus grande expressivité aux patrons construits et permet ainsi de trouver un compromis intéressant en termes de niveau de généralisation entre la spécificité des éléments lexicalisés et le caractère plus général des catégories morpho-syntaxiques.

L'induction d'un patron à partir de deux occurrences de relation est plus précisément composée des trois étapes suivantes :

- le calcul de la distance d'édition entre les deux phrases exemples, c'est-à-dire le nombre minimal d'opérations d'éditions (insertion, suppression, substitution) à effectuer pour passer d'une phrase à l'autre. Toutes les opérations ont ici le même poids ;
- l'alignement optimal des phrases exemples à partir de la matrice des distances entre sous-séquences issue du calcul de la distance d'édition. L'algorithme classique pour trouver un tel alignement est ici étendu en permettant la mise en correspondance de deux mots lors d'une substitution selon les trois niveaux d'information possibles (forme de surface, lemme, catégorie morpho-syntaxique) ;
- la construction des patrons, en complétant si nécessaire les alignements par des opérateurs *jokers* ($*s*$), représentant 0 ou 1 mot quelconque, et ($*g*$), représentant exactement un mot quelconque.

Le tableau 1 montre un exemple d'induction de patron pour le type de relation *founded_by* à partir des deux extraits de phrases ci-dessus. On peut noter la présence de la catégorie *DET* (déterminant) comme généralisation pour (*a—the*), ce qui rend le patron pertinent pour d'autres extraits tels que « *Charles Kettering, another founder of DELCO ...* ».

¹<http://opennlp.sourceforge.net/index.html>

Charles Revson	,	the	founder	of		Revlon Cosmetics
Mayer Lehman	,	a	founder	of	the	Lehman Brothers investment
<source>	,	DET	founder	of	(*s*)	<cible>

TAB. 1 – Exemple d’induction de patron de relation

Cet exemple illustre également le fait que la généralisation peut aller jusqu’à l’utilisation de jokers pouvant se substituer à n’importe quel mot. Comme il est toujours possible de généraliser deux phrases en un patron ne contenant que des jokers, il est nécessaire de fixer une limite supérieure au nombre de jokers pouvant être introduits dans une opération de généralisation pour conserver un niveau de spécificité raisonnable des patrons, limite pouvant prendre la forme d’une valeur absolue ou d’un pourcentage appliqué aux jokers ou aux mots selon qu’il s’agit d’une limite supérieure ou inférieure.

Par ailleurs, travaillant en domaine ouvert et avec des entités nommées assez générales, nous souhaitons plutôt induire un nombre important de patrons spécifiques qu’un ensemble restreint de patrons très généraux, ceci afin de privilégier la précision. Dans l’évaluation présentée en section 5.6, le nombre maximal de jokers dans un patron a été fixé à 1.

5.4.2 Filtrage pour l’apprentissage des patrons de relations

Dans le contexte de supervision distante dans lequel nous nous plaçons, les phrases exemples ne sont pas directement fournies comme des exemples annotés d’occurrences de relations mais résultent de la projection dans un corpus de relations se présentant sous la forme de couples d’entités (par exemple le couple (Ray Charles, Albany) pour le type de relation *city_of_birth*). Plus concrètement dans notre cas, elles sont récupérées en soumettant à un moteur de recherche des requêtes contenant des couples d’entités pour un type de relations donné et en restreignant les résultats du moteur aux phrases contenant effectivement les deux valeurs des entités.

On peut souligner que la nature des restrictions appliquées a un impact direct sur la quantité et la précision des patrons induits. Plus on impose de contraintes,

5. PEUPLEMENT DE BASES DE CONNAISSANCES

moins on obtient de phrases exemples, mais meilleurs seront les patrons induits. Par exemple, les auteurs de [Agirre et al., 2009] ne retiennent que les phrases exemples dans lesquelles les paires d'entités apparaissent dans un voisinage de zéro à dix mots.

Il est important de noter que le processus d'induction de patrons s'effectue en comparant les phrases exemples deux à deux. Il peut donc être coûteux (en temps de calcul) lorsque le nombre de phrases exemples est important : pour 10 000 exemples, on a environ 50 millions de couples distincts de phrases à comparer ($n(n-1)/2$ exactement). Pour traiter ce problème, une première solution consiste à réduire de façon drastique le nombre de phrases exemples en amont du processus d'induction, la conséquence étant une réduction de la couverture des différentes formes d'expression des types de relations.

Une autre solution consiste à faire une réduction sélective du nombre de couples de phrases exemples à généraliser pour éviter de considérer les couples de phrases dont la distance est visiblement trop grande pour induire des patrons intéressants. Même si la distance utilisée pour cette induction est une distance d'édition, donc tenant compte de l'ordre des mots, il est évident qu'un faible recoupement des phrases en termes de mots conduira à une valeur élevée de la distance d'édition. Le filtrage *a priori* des couples de phrases peut donc se fonder sur une mesure s'appliquant à une représentation de type «sac de mots», telle que la mesure *cosinus*, en fixant une valeur minimale en dessous de laquelle la généralisation des couples de phrases n'est pas réalisée.

Or, la mesure *cosinus* peut être évaluée de manière efficace, soit avec une bonne approximation, comme dans le cas du *Local Sensitive Hashing* [Gionis et al., 1999], soit de manière exacte mais en fixant un seuil de similarité minimal, ce qui correspond à notre cas de figure. Nous avons donc retenu pour notre filtrage l'algorithme *All Pairs Similarity Search* (APSS), proposé dans [Bayardo et al., 2007], qui calcule la mesure *cosinus* pour les seules paires d'objets considérés – ici, les phrases exemples – dont la similarité est supérieure ou égale à un seuil fixé *a priori*. Cet algorithme se fonde plus précisément sur une série d'optimisations dans l'indexation des objets tenant compte des informations recueillies sur leurs caractéristiques et d'un tri appliqué à ces objets en fonction de ces mêmes caractéristiques.

De plus, le filtrage par l'algorithme APSS que nous proposons est combiné à une étape de *clustering*. En effet, si l'algorithme APSS nous permet de retrouver *les paires* de phrases exemples partageant des similitudes, il est potentiellement plus intéressant, dans la perspective de la réduction des comparaisons, de retrouver *les groupes* de phrases partageant des similitudes. L'idée est donc d'utiliser les scores de similarité calculés par l'algorithme APSS pour faire ce regroupement en fournissant la matrice de similarité ainsi construite en entrée d'un algorithme de *clustering*. Par la suite le processus d'induction de patrons est opéré au niveau de chaque *cluster* de phrases exemples au lieu de l'ensemble des phrases deux à deux. De cette façon, on réduit de façon très significative les comparaisons entre phrases. Pour le *clustering* des phrases candidates, l'algorithme Markov Clustering (MCL) [Dongen, 2000] est utilisé.

En pratique, lors de l'induction de patrons à partir d'un grand volume de phrases exemples, on retrouve de nombreux doublons : soit parce que la même phrase exemple se trouve dans plusieurs documents, soit parce que l'on retrouve la même forme d'expression d'un type de relations avec des valeurs différentes (*Obama's height is 1.87m* ; *Sarkozy's height is 1.65m*). Nous proposons donc de filtrer les phrases exemples à deux niveaux : d'abord avec un seuil de similarité 1 afin d'identifier et éliminer les phrases identiques ; puis avec un seuil de similarité faible pour s'assurer d'un niveau minimal de similarité entre les phrases en vue du processus d'induction. Dans nos expériences, le seuil de similarité a été fixé à 0,25.

Notons que dans le contexte de l'évaluation de la similarité entre les relations, [Bollegala et al., 2009] mettent en œuvre une étape de *clustering* proche de celle que nous proposons. Dans leur cas, il s'agit de regrouper des patrons lexicaux (qui sont quasi identiques à nos phrases exemples) afin de caractériser le type de relation sémantique exprimé par chaque *cluster*. Dans notre cas, le type de relation sémantique est connu à l'avance et l'objectif est de regrouper à travers les *clusters* les différentes formulations de la même relation sémantique.

5. PEUPLEMENT DE BASES DE CONNAISSANCES

5.4.3 Extraction des relations

L'extraction de nouvelles relations se fait à partir des types de relations existants et d'entités connues : on cherche à compléter une base de connaissances existante en complétant les informations concernant les entités qu'elle contient. La première étape de l'extraction de relations est la recherche de phrases candidates pouvant contenir l'expression d'une relation. Elle prend comme point de départ des requêtes contenant une entité nommée associée à son type et le type de l'information recherchée.

La recherche proprement dite est réalisée, comme dans le cas de l'apprentissage de patrons, grâce à un moteur de recherche ayant préalablement indexé le corpus cible pour l'extraction des relations. Nous nous sommes appuyés dans notre cas sur le moteur Lucène¹, avec une indexation adaptée aux caractéristiques de notre recherche : les documents initiaux sont découpés en unités d'indexation de petite taille, trois phrases, grâce à une fenêtre glissante et au sein de ces unités, sont indexés les mots pleins sous leur forme normalisée et les entités nommées, avec leur type.

L'interrogation du corpus présente en outre la particularité d'inclure une phase d'expansion de l'entité source. En effet, on retrouve souvent dans les documents des formes plus ou moins développées des entités nommées : par exemple *Bill Clinton* est généralement utilisé au lieu de *William Jefferson Blythe III Clinton*. Il est donc intéressant de savoir que ces deux mentions d'entités sont équivalentes et associées à la même entité, en particulier lors de la recherche de documents. Nous utilisons donc une étape d'expansion des entités visant à associer à une entité donnée les formes alternatives lui faisant référence. Pour l'entité «Barack Obama», on a ainsi : {*B. Hussein Obama*, *Barack H. Obama Junior*, *Barack Obama Jr*, *Barack Hussein Obama Jr.*, etc.}. L'intérêt est de pouvoir augmenter les chances de retrouver des phrases candidates liées à l'entité puisque l'on considère tous les documents dans lesquels apparaissent ses différentes expressions.

Une base d'expansion des entités a été constituée de façon automatique à partir du corpus Wikipédia² en collectant pour chaque entité les formulations

¹<http://lucene.apache.org/java/docs/index.html>

²Plus précisément, la version mise à disposition par l'université de New York : <http://nlp>.

extraites des pages de redirection de Wikipédia vers cette entité. Au total, la base d'expansion contient des formes étendues pour environ 2,4 millions d'entrées.

Nous appliquons ensuite sur les phrases candidates sélectionnées les patrons induits lors de la phase d'apprentissage. Les entités cibles extraites par ces patrons sont cumulées pour ne retenir finalement que les plus fréquentes : notre hypothèse est que les entités cibles les plus pertinentes apparaissent plus souvent dans les documents que les moins pertinentes. Pour les relations mono-valuées (ex. : date de naissance), une seule valeur est conservée. Pour les relations multi-valuées (ex. : lieux de résidence), un nombre arbitraire de trois valeurs sont conservées à défaut de connaissances fournies *a priori* ou extraites des textes sur le nombre de valeurs attendu.

Enfin, un dernier filtre est appliqué sur les entités cibles pour vérifier la compatibilité des valeurs obtenues avec les contraintes relatives au type d'information recherché qu'elles représentent, définies par des listes de valeurs ou d'expressions régulières : on vérifie, par exemple, que le pays de naissance d'une personne fait bien partie de la liste des pays connus.

5.4.4 Amélioration par l'utilisation d'un filtrage générique de relations

Dans cette section nous décrivons des améliorations apportées à notre système d'extractions de relations, par l'intégration d'un module de filtrage de relations dans la chaîne de traitement. Ces améliorations ont été effectuées dans le cadre de notre participation à l'édition 2011 de la campagne TAC-KBP et ne sont pas intégrés dans l'évaluation quantitative de notre approche présentée dans les sections suivantes, effectuées sur les données de la campagne TAC-KBP 2010. Nous en présentons néanmoins les principes généraux dans cette section. Les améliorations concernent essentiellement deux points, le filtrage des phrases exemples servant à la génération des patrons et le ré-ordonnancement des entités cibles extraites.

L'hypothèse sur laquelle repose notre sélection des phrases exemples, à savoir qu'un couple d'entités contenu dans une même phrase induit une relation

5. PEUPLEMENT DE BASES DE CONNAISSANCES

entre les entités, est une hypothèse simplificatrice et n'est pas toujours vérifiée. Nous avons donc cherché à améliorer la génération des patrons en éliminant les phrases exemples qui n'expriment aucune relation entre les entités. Notre motivation est qu'un tel filtrage devrait améliorer la qualité (précision) des patrons générés. Pour ce faire, nous avons intégré la méthode de filtrage de relations développée par [Wang et al., 2011b] dans le cadre de l'extraction d'information non supervisée. Plus précisément, [Wang et al., 2011b] proposent d'utiliser des critères non liés à un type de relations (nombre de mots entre les entités, types des entités, catégories morpho-syntaxiques des mots entre les entités, etc.) afin de déterminer si deux entités contenues dans une même phrase sont liées par une relation sémantique. Concernant la mise en œuvre de leur approche, les auteurs s'appuient sur un modèle statistique à base de CRF. En pratique, il s'agit pour nous d'appliquer ce classifieur (CRF) à l'ensemble des phrases exemples collectées pour un type de relation. Ce classifieur renvoie une réponse binaire qui exprime si la phrase contient ou non une relation sémantique, sans *a priori* sur le type de relation. Les phrases exemples qui contiennent effectivement une relation selon le classifieur sont conservées pour le processus d'induction de patrons décrit dans la section 5.4.1¹. Dans un cadre idéal, suite au processus d'induction des patrons on devrait observer un meilleur score de précision et une baisse limitée du score de rappel.

La deuxième amélioration que nous avons expérimentée concerne la phase d'extraction de relations, suite à l'induction des patrons, et vise plus directement le ré-ordonnancement des entités cibles extraites. Dans sa version actuelle, notre système se fonde sur un critère de redondance pour sélectionner la meilleure réponse. Il s'agit de considérer la réponse la plus fréquente comme étant la plus pertinente, sans pour cela se préoccuper de savoir si les phrases candidates (les phrases contenant une entité connue et une entité pouvant potentiellement compléter la relation) retenues pour faire ce décompte contiennent effectivement des relations. Par conséquent, nous nous sommes servis du filtrage des phrases

¹Notons que nous avons expérimenté plusieurs configurations pour l'induction des patrons. Dans la première configuration, celle décrite à la section 5.4.1, nous avons utilisé l'ensemble des occurrences de relations sans appliquer de filtrage pour le processus d'induction de patrons (configuration TAC-KBP 2010). Dans un second temps, nous avons appliqué le filtrage des occurrences de relations (configuration TAC-KBP 2011)

candidates comme critère d’ordonnancement des réponses. Notre motivation est de privilégier, lors de la sélection des réponses, les entités cibles potentielles issues des phrases candidates ayant été identifiées comme «positives» par le classifieur. Par conséquent, nous proposons d’ajouter un bonus pour chaque entité cible potentielle lorsque celle-ci est trouvée dans une phrase candidate contenant effectivement une relation selon le classifieur, équivalent à une occurrence supplémentaire de l’entité cible.

5.5 La campagne d’évaluation TAC-KBP

La tâche *KBP* de la campagne d’évaluation TAC a pour objet la découverte d’informations concernant des entités nommées à partir d’un corpus de taille importante (plus d’un million de documents), pour intégrer ces informations dans une base de connaissance existante (*KB*). Cette KB est dérivée de l’encyclopédie en ligne Wikipedia. La campagne concerne principalement la langue anglaise¹ et n’a pas, pour le moment, d’équivalent en langue française. À ce jour, trois éditions de la campagne ont été organisées entre 2009 et 2011 [TAC-KBP, 2009, 2010, 2011].

5.5.1 TAC-KBP 2009 – 2010

Durant les deux premières éditions de la campagne *KBP*, deux tâches étaient proposées aux participants.

La première tâche, appelée *entity linking*, aborde une limitation des systèmes classiques d’extraction d’information concernant l’ambiguïté sur les entités nommées. Plus concrètement, il s’agit pour un système, lorsqu’il reconnaît un nom de personne comme «*Michael Jordan*», de déterminer s’il fait référence au joueur de basket ou s’il fait référence au chercheur en intelligence artificielle du même nom². Ainsi, cette tâche est dans le domaine de l’identification de la coréférence entre entités, avec la mise en correspondance d’une entité dans un texte avec une des entités possibles dans une base de connaissances : la KB est composée d’un

¹TAC-KBP 2011 propose des tâches translingues (anglais-chinois).

²http://en.wikipedia.org/wiki/Michael_I._Jordan

5. PEUPLEMENT DE BASES DE CONNAISSANCES

ensemble d'entités ayant chacune un identifiant unique et d'autres informations associées (issues des infobox Wikipedia). Pour cette tâche, les requêtes sont composées d'une valeur d'entité et d'un document qui accompagne l'entité. Il s'agit alors de déterminer, pour la valeur d'entité donnée, l'identifiant qui lui correspond dans la KB. S'il n'y en a pas, une réponse NIL doit être renvoyée. Notons qu'il existe un autre type d'ambiguïté sur les entités nommées qui n'est pas directement abordé dans la campagne : l'ambiguïté sur la détection du type d'entité. Par exemple, «*Obama*» peut être un nom de personne ou un nom de lieu (une ville du Japon). Ces problématiques de désambiguïsation de type relèvent plus de la reconnaissance des entités nommées. Nous n'avons pas traité ces problèmes de désambiguïsation des entités nommées dans nos travaux.

La deuxième tâche, appelée *Slot Filling*, aborde la problématique qui nous intéresse, à savoir l'extraction de relations entre les entités nommées. Il s'agit pour un système d'apprendre à reconnaître un ensemble de relations à partir d'un corpus de documents. L'objectif étant de compléter une KB, les relations qui doivent être apprises sont définies en fonction de la KB : chaque champ (ou *slot*) de la KB est lié à un type de relation. De plus, pour cette tâche, les requêtes sont composées d'une valeur d'entité (et son type), d'un identifiant dans la KB, et d'une liste de champs à ignorer : il s'agit des champs déjà renseignés dans la KB et donc qu'il ne faut pas traiter. Notons que, selon les champs, une seule ou plusieurs réponses sont attendues : une date de naissance prend par exemple une seule valeur, alors que des lieux de résidences peuvent être multiples. Si le système n'est pas capable de trouver une réponse dans le corpus, il doit renvoyer une réponse NIL. Enfin, pour chacune des relations trouvées, le système doit retourner un identifiant correspondant au document dans lequel la relation a été retrouvée.

Au cours des deux éditions 2009 et 2010 de TAC-KBP la description de la tâche de *slot filling* qui nous concerne est restée inchangée. En revanche on peut relever quelques distinctions significatives sur quelques points :

Type d'entités à traiter : dans TAC-KBP 2009 il est question d'extraire des relations concernant trois types d'entités nommées génériques PERSONNE, ORGANISATION, et ENTITÉ-GÉO-POLITIQUE. Dans TAC-KBP 2010 le type ENTITÉ-GÉO-POLITIQUE a été supprimé, car les documents du

corpus contiennent très peu d'informations utiles pour l'apprentissage des relations concernant les champs associés aux entités géo-politiques (par exemple devise d'un état). De plus, un champ comme celui de la population d'une entité géo-politique est très souvent mis à jour dans des bases de connaissances existantes. Malgré ce changement, on peut relever que le nombre de relations à extraire est resté constant sur les deux éditions (42 relations) : pour TAC-KBP 2010, les relations impliquant des noms de lieux ont été dissociées pour être plus précis, par exemple la relations *PER :place_of_birth* est devenue *PER :country_of_birth*, *PER :city_of_birth*, *PER :stateOrProvince_of_birth*. La liste exhaustive des différents champs dans les campagnes TAC-KBP 2009 et 2010 est présentée dans le tableau 2

Métrique d'évaluation : dans TAC-KBP 2009, les performances des systèmes étaient évaluées en fonction de la multiplicité des réponses attendues (une métrique pour les champs mono-valués et une autre pour les champs multi-valués), puis les résultats étaient combinés. De plus, les résultats étaient fortement influencés par les réponses NIL induites par les requêtes ne trouvant pas de réponses dans le corpus : un système renvoyant uniquement la réponse NIL obtiendrait une réponse correcte sur 80 % des requêtes. En conséquence, pour TAC-KBP 2010, la métrique a été changée afin d'évaluer davantage la pertinence des réponses non-NIL renvoyées par les systèmes.

Corpus : dans TAC-KBP 2009, le corpus utilisé est constitué en grande partie d'articles de presses (environ 1 million de documents) alors que celui de TAC-KBP 2010 intègre, en plus du corpus initial, une part significative de documents issus du web (blogs, etc.) (plus de 300 000).

Adaptation des systèmes : cette problématique est récurrente en matière d'extraction d'information. Ainsi, TAC-KBP 2010 introduit la tâche «*Surprise slot-filling*» dont l'objet est de mesurer le temps et la quantité d'efforts nécessaires pour le paramétrage des systèmes à de nouvelles relations. Cette tâche permet d'avoir un aperçu des performances globales d'un système qui serait très peu optimisé sur des relations nouvelles à l'inverse de la tâche principale.

TAC-KBP 2009			TAC-KBP 2010-2011	
Person	Organization	Geo-Political Entity	Person	Organization
per :alternate_names	org :alternate_names	gpe :alternate_names	per :alternate_names	org :alternate_names
per :date_of_birth	org :political/religious_affiliation	gpe :capital	per :date_of_birth	org :political/religious_affiliation
per :age	org :top_members/employees	gpe :subsidiary_orgs	per :age	org :top_members/employees
per :place_of_birth	org :number_of_employees/members	gpe :top_employees	per :country_of_birth	org :number_of_employees/members
per :origin	org :members	gpe :political_parties	per :stateorprovince_of_birth	org :members
per :date_of_death	org :member_of	gpe :established	per :city_of_birth	org :member_of
per :place_of_death	org :subsidiaries	gpe :population	per :origin	org :subsidiaries
per :cause_of_death	org :parents	gpe :currency	per :date_of_death	org :parents
per :residences	org :founded_by		per :country_of_death	org :founded_by
per :schools_attended	org :founded		per :stateorprovince_of_death	org :founded
per :title	org :dissolved		per :city_of_death	org :dissolved
per :member_of	org :headquarters		per :cause_of_death	org :country_of_headquarters
per :employee_of	org :shareholders		per :countries_of_residence	org :stateorprovince_of_headquarters
per :religion	org :website		per :stateorprovinces_of_residence	org :city_of_headquarters
per :spouse			per :cities_of_residence	org :shareholders
per :children			per :schools_attended	org :website
per :parents			per :title	
per :siblings			per :member_of	
per :other_family			per :employee_of	
per :charges			per :religion	
			per :spouse	
			per :children	
			per :parents	
			per :siblings	
			per :other_family	
			per :charges	

TAB. 2 – Liste des relations à extraire pour TAC-KBP 2009-2010

5.5.2 TAC-KBP 2011

La campagne TAC-KBP 2011 est la plus récente des évaluations au moment de la rédaction de ce manuscrit. Dans TAC-KBP 2011, plus singulièrement la tâche *Slot Filling*, deux nouveaux aspects du processus d'extraction sont introduits :

le traitement cross-lingue des données : il s'agit d'une tâche expérimentale¹ similaire à la tâche *Slot Filling* précédente, la distinction est au niveau des documents qui peuvent être en anglais ou en chinois. De plus, les systèmes doivent renvoyer des réponses uniquement en langue anglaise même si celles-ci sont trouvées dans des documents en chinois. Compte tenu de la difficulté de la tâche, seuls 8 champs (5 pour les personnes, 3 pour les organisations) sont concernés au lieu des 42 de la tâche initiale.

la temporalité : les informations concernant les entités varient généralement dans le temps, et sont donc plus ou moins pertinentes en fonction de la fenêtre de temps dans laquelle on se place. Par exemple lorsque l'on cherche le lieu de résidence d'une personne, il semble plus correct de considérer le dernier lieu connu même si les lieux précédents ne peuvent être considérés comme incorrects. Ainsi, le but de cette tâche est de compléter les réponses extraites par des informations temporelles (intervalle temporel dans lequel la réponse est valide, etc.) Certains champs étant invariants dans le temps (date de naissance), un nombre limité de champs sont concernés (lieux de résidences, épouses, etc).

Notons que la description de la tâche *Slot Filling* est identique à la version («*monolingue*») précédente et que le corpus est identique au corpus précédent. De plus, les mêmes métriques d'évaluations que TAC-KBP 2010 sont utilisées pour cette tâche.

¹La tâche devrait être étendue pour TAC-KBP 2012.

5.6 Évaluation de l'approche dans le cadre de TAC-KBP

Nous présentons dans cette section les résultats de l'évaluation de notre système en utilisant les données de la tâche *Slot Filling* de la campagne d'évaluation TAC-KBP 2010 [TAC-KBP, 2010]. Cette évaluation est faite en plusieurs parties, pour évaluer de façon différenciée les différentes étapes de notre approche : l'induction de patrons et l'extraction des relations. Nous précisons que les expériences ont été réalisées sur un cluster de 24 nœuds (4 processeurs/nœud) avec une parallélisation par type de relations.

5.6.1 Les données

Les données d'évaluation issues de TAC-KBP sont les suivantes :

- un corpus de textes composé de 1 780 980 documents répartis en 0,1 % de transcriptions (conversations téléphoniques, journaux radio, conversations radio), 72,3 % d'articles de presse et 27,7 % de pages Web ;
- une base de connaissances (*KB*) reposant sur une image de Wikipédia d'octobre 2008. Un identifiant unique et un type d'entité sont attribués à chaque page contenant des *infobox*. Le type d'entité *personne*, *organisation*, *entité géopolitique* ou *inconnu* est associé à chaque page en fonction des champs contenus dans les *infobox*. Typiquement, les *infobox* *Infobox_Actor* sont ainsi liées à des personnes. Au final 818 741 entités ont été retenues pour former la KB, chacune d'elles étant associée à un ensemble de propriétés (champs des *infobox*) ainsi qu'à un texte la décrivant. Ainsi les relations sont représentées dans la KB par des tuples (identifiant, type *infobox*, nom, type, propriété, valeurs), ex. : (E0000437 ; *Infobox_Actor* ; Julia Roberts ; PER ; birthplace ; Atlanta) ;
- une table de correspondance entre les propriétés issues de Wikipédia et les types de relations retenus pour l'évaluation. Par exemple, *Infobox_Actor:-birthplace* est convertie en *per:city_of_birth*. Cette correspondance permet de prendre en compte une certaine hétérogénéité de désignation des propriétés dans Wikipédia ;

-
- une liste de 100 entités sources pour lesquelles on cherche à extraire toutes les entités en relation pour tous les types de relations considérés. On dénombre parmi ces entités 15 entités présentes dans la KB et 85 inconnues de la KB. Par ailleurs, toutes les relations considérées ne trouvent pas d’entités cibles dans le corpus pour ces 100 entités. Dans le cadre de cette étude, nous nous focalisons uniquement sur les relations pour lesquelles il existe une entité cible dans le corpus¹, ce qui représente au total 2069 relations. Le détail par type de relations est présenté dans la colonne *Nb Ref.* du tableau 3.

5.6.2 Métriques d’évaluation TAC-KBP 2010

Une difficulté récurrente pour l’évaluation des systèmes de questions-réponses ou d’extraction de relations se situe au niveau de la construction d’une référence. Dans le cas idéal, cette référence serait issue d’une annotation manuelle des relations dans chacun des documents du corpus. Malheureusement, ce type de référence est très coûteuse à obtenir, en particulier lorsque le nombre de types de relations à annoter est important.

En pratique, les organisateurs de campagnes d’évaluation constituent une référence à partir des résultats fournis par les participants. Cette référence est donc partielle puisqu’elle n’évalue qu’une partie des relations du corpus. Cette phase vise à concaténer l’ensemble des relations extraites par les participants puis à les faire juger par des annotateurs humains pour s’assurer de leur pertinence. Dans le cadre de TAC-KBP 2010, les annotateurs ont le choix entre trois jugements pour chaque relation :

Correct la relation extraite par le système est jugée valide par un annotateur.

Inexact la relation extraite par le système n’est pas jugée exacte par l’annotateur, puisque non complète ou inexacte. Cette notion de relation inexacte est à mi-chemin entre une réponse correcte et une réponse incorrecte. Plus précisément, il s’agit de juger le niveau de spécificité des relations extraites.

Par exemple, les titres généraux comme *docteur* sont jugés inexacte si le

¹Les entités cibles existantes dans le corpus sont établies par la référence fournie par les organisateurs de la campagne, construite à partir des résultats des participants.

Types de relations	Type de cible	Couv. Doc.	Couv. Rel.	Nb Appr.	Nb Test	Nb Induc.	Nb Patrons	Couv. Patrons	Nb Ref.
org:alternate_names	ORG	89,2	33,3	20 013	10 006	214	6 007	66,1	120
org:city_of_headquarters	LOC + liste	90,1	59,3	6 847	3 423	4 553	2 010 749	65,5	81
org:country_of_headquarters	LOC + liste	91,0	55,2	18 401	9 200	2 110	185 158	69,6	67
org:dissolved	DATE	100,0	25,0	532	266	87	775	0,0	4
org:founded_by	ORG/PER	95,5	31,8	1 954	977	197	4 385	77,9	28
org:founded	DATE	92,9	53,6	13 688	6 844	127	22 482	77,3	22
org:member_of	ORG	100,0	100,0	7 951	3 976	102	103	70,0	2
org:members	ORG	77,8	11,1	531	265	183	552	86,0	9
org:number_of_employees_members	regex+liste	90,5	23,8	7 173	3 586	216	3 109	100,0	21
org:parents	ORG	96,7	43,3	22 361	11 181	3 013	485 947	69,0	30
org:political_religious_affiliation	ORG	78,6	64,3	3 427	1 713	406	3 250	55,4	14
org:shareholders	ORG/PER	66,7	33,3	3	2	0	0	0,0	3
org:stateorprovince_of_headquarters	LOC + liste	92,7	63,2	9 672	4 836	1 422	148 610	69,9	68
org:subsidiaries	ORG	82,7	28,9	5 588	2 794	498	3 764	56,5	52
org:top_members_employees	PER	91,5	37,2	40 929	20 464	108	1 010	70,6	223
org:website	regex	78,3	30,4	30 813	15 407	32	28	0,0	23
per:age	regex+liste	85,3	32,1	157	79	3	1	0,0	109
per:alternate_names	PER	61,6	11,6	18 115	9 057	68	2 818	82,6	86
per:cause_of_death	liste	100,0	0,0	1	1	0	0	0,0	2
per:charges	liste	61,5	0,0	184	92	0	0	0,0	13
per:children	PER	72,0	16,0	2 010	1 005	147	238	0,0	25
per:cities_of_residence	LOC + liste	77,6	34,5	3 631	1 815	722	14 297	77,9	58
per:city_of_birth	LOC + liste	69,2	15,4	4 745	2 373	2 252	62 455	63,3	13
per:city_of_death	LOC + liste	100,0	100,0	1 631	816	505	2 860	70,3	1
per:countries_of_residence	LOC + liste	73,5	20,6	8 098	4 049	2 181	205 344	80,1	34
per:country_of_birth	LOC + liste	82,4	5,9	11 085	5 542	11 192	9 145 385	65,0	17
per:country_of_death	LOC + liste		2 873	1 436	1 068	22 374	62,9	0	0
per:date_of_birth	DATE	90,0	20,0	11 689	5 845	30	22	0,0	20
per:date_of_death	DATE	100,0	0,0	4 692	2 346	54	63	33,3	1
per:employee_of	ORG	84,2	29,3	24 762	12 381	2 435	704 833	71,1	133
per:member_of	ORG	82,4	36,3	27 523	13 761	3 901	740 999	57,3	91
per:origin	liste	81,6	42,1	37 626	18 813	2 710	276 653	74,4	76
per:other_family	PER	86,7	33,3	4	2	0	0	0,0	30
per:parents	PER	78,1	9,4	1 314	657	37	604	77,8	64
per:religion	liste	85,7	57,1	1 468	734	515	1 575	80,0	7
per:schools_attended	ORG+liste	87,5	37,5	2 246	1 123	67	170	4,2	16
per:siblings	PER	78,3	20,3	4	2	0	0	0,0	69
per:spouse	PER	80,0	35,6	5 385	2 693	3 094	314 329	80,0	45
per:stateorprovince_of_birth	LOC + liste	80,0	50,0	7 047	3 523	2 097	60 782	75,4	10
per:stateorprovince_of_death	LOC + liste	100,0	100,0	1 616	808	278	911	66,7	1
per:states_or_provinces_of_residence	LOC + liste	84,2	50,0	4 980	2 490	1 166	115 418	77,9	38
per:title	liste	84,6	52,8	31 574	15 787	8 797	1 573 512	49,1	343

TAB. 3 – Résultats des différentes étapes, pour tous les types de relations

Type de cible : mécanisme utilisé pour retrouver l'entité cible. *Couv. Doc.* : couverture des documents de référence dans les résultats de la recherche de phrases. *Couv. Rel.* : couverture des phrases candidates de référence. *Nb Appr.* : nombre de relations pour l'apprentissage des patrons. *Nb Test* : nombre de relations pour l'évaluation des patrons. *Nb Induc.* : nombre de phrases contenant des occurrences de relations pour l'induction des patrons. *Nb Patrons* : nombre de patrons induits à partir des occurrences de relations. *Couv. Patrons* : couverture des patrons induits. *Nb Ref.* : nombre de relations de référence.

document mentionne un titre plus spécifique comme *dentiste*. Cette notion juge aussi la pertinence du passage dans lequel la relation issu, il s'agit de s'assurer que la relation est suffisamment motivée dans le document.

Wrong la relation extraite par le système est jugée comme incorrecte par l'annotateur.

Redundant la relation extraite par le système est correcte mais redondante, c'est-à-dire que la base de connaissance contient déjà cette information ou qu'elle a déjà été retournée par le système.

Les phrases suivantes sont des illustrations de jugements effectués par les annotateurs :

- Relation à compléter : *per:countries_of_residence(Hugo Chavez, ?)*
Extrait du document : *President Hugo Chavez was granted free rein to accelerate changes in broad areas of society by presidential decree – a move critics said propels Venezuela toward dictatorship.*
Jugement : *Correct*. Il y a une relation implicite entre le fait de présider un pays et d'y habiter.
- Relation à compléter : *per:cities_of_residence(Hugo Chavez, ?)*
Extrait du document : *Born of schoolteacher parents in the western town of Sabaneta on July 28, 1954, Chavez studied at the Military Academy of Venezuela in Caracas.*
Jugement : *Inexact*. Le fait d'être né dans la ville de Sabaneta n'a pas été jugé suffisant pour justifier que Hugo Chavez y a résidé.
- Relation à compléter : *per:age(Hugo Chavez, ?)*
Extrait du document : *The 80-year-old revolutionary leader, who was shown standing and chatting with Chavez, looked far fitter than he did when images of him were last broadcast here three months ago.*
Jugement : *Wrong*. L'âge de 80 ans correspond à celui de Fidel Castro et non celui de Hugo Chavez.
- Relation à compléter : *per:title(Beyonce Knowles, ?)*
Extrait du document : *Newspapers said the guest list includes Korean actors and singers as well as American pop star Beyonce Knowles, who arrived Thursday for her first concert here.*

5. PEUPLEMENT DE BASES DE CONNAISSANCES

Jugement : *Redundant*. La relation extraite figure déjà dans la KB ou est redondante par rapport à une réponse retournée par le système.

Il faut souligner que, dans le cadre de TAC-KBP 2010, les réponses NIL (la relation n'existe pas dans le corpus) identifiées comme telles par les systèmes ne sont pas prises en compte. La raison principale en est que le but de la tâche est d'évaluer la pertinence des relations qui sont effectivement dans le corpus, et non celles qui ne le sont pas. Par conséquent, seules les réponses non-NIL (accompagnées des identifiants de documents) retournées par les participants sont utilisées pour construire la référence.

La conséquence est que si tous les participants ont renvoyés NIL pour une requête donnée, alors la référence ne contient aucune valeur pour cette relation. Cela ne garantit pas que la relation ne puisse être ailleurs dans un autre document du corpus. Par exemple, [Chen et al., 2010a] ont étendu la référence initiale de TAC-KBP 2009 fournie par les organisateurs : ils ont demandé à des annotateurs humains de juger les sorties de leur système qui n'étaient pas trouvées par les autres participants. Les relations extraites jugées comme correctes ont été ajoutées à celles des autres participants afin d'augmenter le nombre de relations dans la référence.

Afin évaluer les performances des systèmes les organisateurs proposent d'utiliser les métriques classiques en extraction d'information : rappel, précision et F1-mesure. Toutes ces métriques sont calculées à partir des jugements réalisées par les annotateurs [TAC-KBP, 2010] :

$$\begin{aligned} \text{Rappel} &= \frac{\text{Correct}}{\text{Reference}} \\ \text{Précision} &= \frac{\text{Correct}}{\text{System}} \\ \text{F1 - Mesure} &= \frac{2 * (\text{Rappel} * \text{Précision})}{\text{Rappel} + \text{Précision}} \\ \text{Correct} &= \text{nombre de champs non-NIL retournés par le système et jugés corrects} \\ \text{System} &= \text{nombre de champs non-NIL retournés par le système} \\ \text{Reference} &= (\text{nombre de réponses mono-valuées non-NIL correctes}) + \\ &\quad (\text{nombre de classes d'équivalences pour les réponses multi-valuées}) \end{aligned}$$

où les *classes d'équivalences* sont utilisées pour regrouper les réponses multi-valuées non-NIL similaires. Pour les réponses multi-valuées, les réponses renvoyées par le système doivent appartenir à des classes d'équivalences distinctes afin d'être jugées correctes, sinon elles sont jugées redondantes. De plus, pour le nombre *Reference*, les nombres de réponses (mono/multi-valués) sont calculés à partir des réponses correctes extraites de l'adjudication de tous les systèmes.

5.6.3 Évaluation de l'apprentissage des patrons

Les patrons servent à confirmer/infirmier la présence d'une relation entre deux entités. Il est donc important de vérifier que les patrons appris aient une couverture suffisamment large pour retrouver le plus possible de variantes parmi les occurrences de relations. Pour évaluer la qualité des patrons, nous avons séparé les relations connues en deux ensembles : un ensemble d'apprentissage E_A (2/3 des relations) et un ensemble de test E_T (1/3 des relations).

Nous mesurons la qualité de la couverture des patrons en calculant le pourcentage des occurrences de relations de l'ensemble de test que l'on retrouve en appliquant les patrons appris à partir des occurrences de relations de l'ensemble d'apprentissage. La couverture des patrons est calculée à partir de la formule 5.1 : occ_{ref} désigne l'ensemble des occurrences de relations issues de E_T , occ désigne l'ensemble des relations de E_T pour lesquelles au moins un patron a été appliqué.

$$Couv. Patrons = \frac{|occ_{test} \cap occ|}{|occ_{test}|} \quad (5.1)$$

Le corpus utilisé pour réaliser cette évaluation est le corpus TAC-KBP 2010 décrit ci-dessus. Précisons que l'utilisation de ce corpus pour évaluer l'extraction des relations n'empêche pas son utilisation pour l'apprentissage des patrons, les relations étant différentes pour les deux tâches.

Nous indiquons dans le tableau 3 le nombre de relations des ensembles E_A et E_T respectivement dans les colonnes *Nb. Appr* et *Nb. Test*. Le nombre de phrases trouvées contenant des occurrences des relations du corpus d'entraînement, qui ont donc servi pour l'induction des patrons, est indiqué dans la colonne *Nb. Induc*. Le nombre de patrons générés à partir de ces phrases candidates est indiqué dans

5. PEUPLEMENT DE BASES DE CONNAISSANCES

la colonne *Nb. Patrons* de ce même tableau.

Par exemple, pour le type de relation *org:alternate_names*, à partir des 20 013 relations de l'ensemble E_A , seules 214 phrases candidates contenant l'expression d'une de ces relations sont sélectionnées. Ces 214 phrases servent à générer 6 007 patrons, qui ont une couverture de 66,1 % (*i.e.* on retrouve 66,1 % des phrases contenant des occurrences des 10 006 relations de test). L'écart conséquent entre les 20 013 relations et les 214 phrases trouvées est dû à deux facteurs :

- une contrainte réductrice imposée lors de la sélection des phrases candidates. Seules les phrases dont tous les mots des entités nommées sont correctement identifiés sont en effet conservées. Or, les entités peuvent être partiellement (ou mal) reconnues lors des traitements linguistiques ;
- la nature des documents du corpus : 72 % des documents sont des articles de presse édités entre janvier 2008 et août 2009, ce qui explique le peu de documents, voir aucun, concernant certaines personnes ou organisations présentes dans la KB.

Les résultats de la couverture des patrons sont présentés dans le tableau 3 pour chaque type de relations dans la colonne *Couv. Patrons*. À titre indicatif, le temps passé pour l'induction des patrons pour le type de relations *per:country_of_birth* (11 192 phrases exemples à comparer) est de plus de 11 heures pour la version sans filtrage et passe à moins d'une minute pour la version avec filtrage¹, ce qui illustre l'intérêt de celui-ci en termes de temps de calcul.

5.6.4 Évaluation de l'extraction des relations

L'extraction des relations comprenant plusieurs étapes, chacune d'entre elles peut influencer sur le résultat global. Nous proposons donc de faire une évaluation séparée de la recherche des phrases candidates et de l'extraction des relations proprement dite.

¹La version avec filtrage étant parallélisée, le temps donné est une somme des temps comptabilisés au niveau de chaque processeur.

5.6.4.1 Recherche des phrases candidates

Une condition nécessaire pour extraire des relations pertinentes est de s'assurer que le moteur de recherche renvoie suffisamment de documents pertinents pour nous permettre de retrouver des entités cibles. Nous avons donc mesuré la couverture en documents de notre recherche de phrases candidates, à savoir le pourcentage de documents renvoyés par l'index que l'on retrouve effectivement dans la référence. Cette couverture est calculée à partir de la formule 5.2 : *ref* désigne l'ensemble des documents de la référence, c'est-à-dire des documents contenant des occurrences de relations jugées correctes par les annotateurs, *doc* désigne l'ensemble des documents renvoyés par le moteur de recherche.

$$Couv. Doc = \frac{|ref \cap doc|}{|ref|} \quad (5.2)$$

Nous avons testé de ce point de vue différentes stratégies en faisant varier des paramètres comme le nombre de résultats retournés et l'utilisation ou non de l'expansion pour la requête. Les résultats de cette évaluation nous ont ainsi conduit à utiliser les entités sources et leurs formes étendues pour interrogation de l'index et prendre en compte les 1000 premiers résultats retournés : ces paramètres permettent de retrouver 84,2 % des documents de référence. Le résultat détaillé par type de relations est donné par la colonne *Couv. Doc* du tableau 3.

À partir des documents ainsi sélectionnés, les phrases candidates à l'extraction d'une relation pour un type donné sont extraites en retenant les phrases contenant à la fois l'entité source et le type de l'entité cible. La qualité et la quantité des phrases candidates sont largement influencées par la qualité de la reconnaissance des entités nommées. Comme nous ne disposons pas d'annotation de référence pour les entités nommées du corpus, il n'est pas possible de mesurer les pertes causées par la mauvaise reconnaissance des entités.

En revanche, nous avons évalué la proportion de documents de référence dans lesquels nous retrouvons des phrases candidates. Cette seconde couverture est calculée à partir de la formule 5.3 : *ref* désigne l'ensemble des documents de la référence, *seg* désigne l'ensemble des documents dans lesquelles des phrases

5. PEUPLEMENT DE BASES DE CONNAISSANCES

candidates ont été sélectionnées.

$$Couv. Rel = \frac{|ref \cap seg|}{|ref|} \quad (5.3)$$

Cette donnée permet de fixer une borne maximale pour le pourcentage de relations qu'il serait possible d'extraire si les étapes à la suite se déroulaient idéalement. Nous obtenons au total une couverture de 37,5 % des phrases appartenant aux documents de référence. Le détail par type de relations est présenté à la colonne *Couv. Rel* du tableau 3.

Il faut souligner que la recherche des phrases candidates est un processus assez coûteux en temps de calcul : dans notre cas, il faudrait parcourir les 1000 premiers documents pour chacune des 2069 relations à compléter. En pratique, nous nous sommes restreints à rechercher des phrases candidates dans les 300 premiers documents afin d'accélérer le processus. Un autre aspect important est que seules les relations non-NIL (2069) sont considérées pour cette évaluation. Néanmoins en situation réelle, il faut aussi tenir compte des relations NIL (bien plus nombreuses) ce qui augmente de façon significative la proportion de documents dans laquelle il faut rechercher les phrases candidates.

5.6.4.2 Extraction de relations

Pour évaluer les relations extraites, nous avons réutilisé les mesures et les outils d'évaluation fournis par la campagne TAC-KBP¹ sans nous limiter aux seuls documents présents dans la référence pour accepter une relation correcte, c'est-à-dire que les relations extraites par notre système sont jugées seulement en fonction de la valeur de l'entité retournée et non en fonction du document justificatif contenant l'occurrence de la relation. En effet, la référence n'étant constituée qu'à partir des résultats des participants à l'évaluation TAC-KBP, elle n'est pas forcément complète. Le tableau 4 fournit les résultats de cette évaluation en agglomérant tous les types de relations et en caractérisant l'impact du filtrage *a posteriori* des entités cibles sur les relations extraites en termes de rappel (*R.*), précision (*P.*) et f1-mesure (*F1.*). Pour mémoire, ce filtrage consiste à s'assurer que l'entité cible valide des expressions régulières et/ou une liste fermée de valeurs.

¹<http://nlp.cs.qc.cuny.edu/kbp/2010/scoring.html>

Nous indiquons dans la colonne *Type de cible* du tableau 3 le mécanisme utilisé pour chaque type de relations.

Les résultats du tableau 4 montrent d’une part, que ce filtrage améliore les performances (en moyenne +2,7 % de f1-mesure) et d’autre part, valident l’hypothèse que les patrons induits à partir des exemples filtrés par l’algorithme APSS sont aussi pertinents que ceux induits en considérant tous les exemples de relations deux à deux (dans ce cas, il y a même une amélioration de +1,7 % de la f1-mesure en moyenne).

	Avant filtrage des entités cibles			Après filtrage des entités cibles		
	R. (%)	P. (%)	F1. (%)	R. (%)	P. (%)	F1. (%)
Tous les couples d’entités	16,3	11,2	13,3	18,1	13,7	15,6
APSS	16,9	12,8	14,5	18,7	16,9	17,7

TAB. 4 – Évaluation de l’impact du filtrage des réponses

Le tableau 5 présente les résultats de différents systèmes sur deux corpus très similaires, les corpus TAC-KBP 2009 et TAC-KBP 2010, ce dernier ajoutant au premier des documents Web et des transcriptions, *a priori* plus difficiles. Bien que ces chiffres ne portent que sur les relations effectivement présentes dans le corpus, ils intègrent la contrainte pour les systèmes ayant participé à la tâche *Slot Filling* de devoir décider si la relation existe ou non dans le corpus, ce que notre système, développé en dehors du contexte de ces campagnes, ne fait pas. Dans ce tableau, les colonnes 2009 et 2010 désignent les scores des trois systèmes les plus et les moins performants de TAC-KBP 2009 et 2010. [Ji et al., 2010] ont montré que sur 492 relations de référence, 60,4 % se trouvaient dans la même phrase tandis que les 39,6 % restantes dépassaient l’espace phrastique dans leur expression et nécessitaient pour leur extraction la résolution de coréférences ou l’application de mécanismes d’inférence impliquant par exemple la composition de plusieurs relations ou l’utilisation de connaissances *a priori* sur les types de relations. De ce fait, nous avons distingué dans la colonne 2010 (a) du tableau 5 les scores des systèmes qui nous sont les plus directement comparables, c’est-à-dire ceux se limitant à l’extraction de relations au niveau phrastique.

On peut noter que le meilleur système de TAC-KBP 2010 [Chada et al., 2010] se détache très nettement : +36,6 % par rapport au deuxième et +4,7 % par

5. PEUPLEMENT DE BASES DE CONNAISSANCES

rapport à un annotateur humain. Cette prédominance s’appuie à la fois sur l’utilisation d’un corpus annoté manuellement (différent du corpus TAC-KBP) de 3 millions de documents et la présence de plusieurs mécanismes d’extraction de relations au niveau inter-phrastique : coréférence pronominale, métonymie entre entités, résolution de dépendances sémantiques entre les mots et les entités, etc. L’utilisation du corpus supplémentaire semble être l’élément déterminant par rapport aux systèmes venant à la suite immédiate, ceux-ci se distinguant de systèmes plus médians par la prise en compte des relations inter-phrastiques. Les plus mauvais résultats, plus faibles en 2010, sont dûs pour une bonne part à des systèmes en cours de développement.

Systèmes TAC KBP	2009	2010	2010 (a)
Nb. soumissions (N) / participants	N=16 / 8	N=31 / 15	N=18
Annotateur humain	59,0	61,1	61,1
1 ^{er} score	34,4	65,8	29,2
2 ^{ème} score	25,1	29,2	14,2
3 ^{ème} score	18,0	28,3	14,1
(N-2) ^{ème} score	5,9	0,6	0,6
(N-1) ^{ème} score	2,6	0,2	0,2
N ^{ème} score	1,8	0,1	0,1
Notre système	–	17,7	17,7
Moyenne	13,4	17,5	9,7
Médiane	13,9	14,1	12,3

TAB. 5 – Résultats sur les données TAC-KBP (f1-mesure)

Concernant notre système, le tableau 5 permet de situer nos résultats dans la moyenne des résultats obtenus par les participants de l’évaluation TAC-KBP 2010 et parmi les trois premiers systèmes pour les approches faisant de l’extraction de relations au niveau de la phrase. Dans ce dernier cas, l’approche la plus performante (29,1 % de f1-mesure) [Byrne and Dunnion, 2010] utilise des règles construites manuellement permettant d’atteindre un score de précision (66,5 %) équivalent au meilleur score de la campagne (66,8 %) et un score de rappel (18,7 %) se situant dans la moyenne de la campagne (15,3 %).

5.6.5 Vue d'ensemble des résultats pour TAC-KBP 2011

Dans cette section nous présentons brièvement les résultats de notre participation à la campagne TAC-KBP 2011. Nous soulignons que tous les résultats officiels ainsi que les descriptions des systèmes ne sont pas disponibles au moment de la rédaction de ce manuscrit.

Lors de cette édition nous avons présenté trois soumissions représentant chacune une configuration différente de notre système. La première configuration, notée LVIC1, est directement inspirée de l'approche que nous avons présenté précédemment sur le jeu de donnée TAC-KBP 2010. Les deux autres soumissions, notées LVIC2 et LVIC3, sont comparables à LVIC1 sauf qu'elles intègrent les améliorations présentés dans la section 5.4.4.

Plus précisément, LVIC2 intègre par rapport à LVIC1, une phase de filtrage des exemples de relations en amont du processus d'induction des patrons. La motivation est de détecter puis d'éliminer du processus d'induction des patrons, les phrases qui n'expriment pas de relations sémantiques et donc qui ne sont pas pertinentes,

LVIC3 quant à elle, ajoute à LVIC1, une phase de ré-ordonnancement des entités cibles extraites. La motivation dans ce dernier cas est de s'assurer que les phrases candidates contenant les relations expriment biens des relations sémantiques entre les entités. Les entités cibles issues de ces phrases sont privilégiées par rapport à celles issues des autres phrases.

Le tableau 6 présente les résultats globaux pour tous les participants à la tâche *slot-filling* pour l'édition 2011. Dans le tableau, la colonne *Top-1* désigne les systèmes ayant une étape faisant intervenir des ressources se trouvant sur le Web et donc extérieure aux ressources initialement fournies. La colonne *Top-2* concerne les systèmes n'utilisant pas ce type de ressources. Plus généralement, les résultats sont moins élevées que lors de l'édition 2010 : le meilleur score était de 65,8 % en F1-mesure ce qui reste plus élevé que celui des deux catégories *Top-1* et *Top-2*. Concernant la médiane, elle est inférieure aux médianes des deux éditions précédentes 13,9 % et 14,1 % en 2009 et 2010. Une hypothèse qui peut expliquer cette différence, mais qui reste à vérifier, est que les relations à extraire pour cette dernière édition sont plus complexes. Malheureusement, il nous manque pour le

5. PEUPLEMENT DE BASES DE CONNAISSANCES

moment une analyse d'erreur pour en être sûr.

Systèmes TAC KBP	LDC	Top-1	Top-2	Médiane
Nb. soumission (N) / participants	N=31	N=31 / 14	N=31 / 13	N=31 / 13
Accès Web	Non	Oui	Non	Non
Précision	86,2	35,0	49,2	10,3
Rappel	72,6	25,5	12,6	16,5
F1-mesure	78,8	29,5	20,1	12,7

TAB. 6 – Résultats sur les données TAC-KBP 2011

Le tableau 7 présente les résultats, agrégés pour toutes les catégories, des trois configurations précédentes. Les résultats sont calculés à partir des mesures officiels de la campagne. Le premier constat est que le niveau de résultat est peu élevé par rapport à la médiane des participants et aussi par rapport aux précédents résultats obtenus sur les données TAC-KBP 2010. Un point important à relever est notre évaluation sur les données TAC-KBP 2010 ne tenait compte que des relations non-NIL, c'est-à-dire que l'entité cible est effectivement dans le corpus. En conséquence ce point n'est pas abordée dans notre approche initiale. Pour traiter cette aspect pour TAC-KBP 2011 la stratégie que nous avons adoptée visaient à considérer une relation comme NIL lorsqu'aucune entité cible était retrouvée dans le corpus.

Il semble que cette stratégie nous ait conduit à renvoyer beaucoup plus de réponses que nécessaire. Aussi on peut relever que les résultats sont très proches pour les configurations LVIC1 et LVIC3, cependant les entités cibles extraites ne sont pas identiques. Précisément les relations extraites entre LVIC1 et LVIC3 sont à 93 % identiques, à 85 % identiques entre LVIC2 et LVIC3 et enfin à 90 % identiques entre LVIC1 et LVIC2. Au moment de la rédaction du manuscrit l'analyse plus détaillée des erreurs qui expliquerait cette baisse important des résultats n'est pas terminée. Une explication possible pourrait aussi venir du traitement des relations inter-phrastiques qui ne sont pas non plus traitées par notre approche. D'après le score de précision (86,2 %) atteint par les annotateurs humains cela reste plausible.

	LVIC1	LVIC2	LVIC3
Nb. rel. Réf.	945	945	945
Nb. réponses Sys.	2103	2070	2103
Nb. Correctes (non-NIL)	97	94	96
Nb. redondantes	10	10	14
Nb. incorrectes	1952	1921	1950
Nb. inexactes	44	45	43
Précision	4,6	4,5	4,6
Rappel	10,3	10,0	10,2
F1-mesure	6,4	6,2	6,3

TAB. 7 – Résultats de notre système lors de TAC-KBP 2011

5.7 Aperçu des systèmes utilisés pour TAC-KBP

Dans cette section, nous présentons brièvement les approches des systèmes ayant obtenu les meilleurs résultats aux campagnes TAC-KBP 2009 et TAC-KBP 2010.

Le système de l'équipe *THU QUANTA* [Li et al., 2009a] a été développé pour participer à plusieurs tâches de la campagne TAC-KBP 2009 et en particulier celle de *slot filling*. Concernant l'apprentissage des relations le système repose sur des patrons lexico-syntaxiques. La démarche utilisée pour générer les patrons est très proche de notre approche : (i) les phrases exemples sont collectées en injectant des paires d'entités dans le corpus TAC-KBP pour chaque type de relation ; (ii) les valeurs des entités cibles et sources dans les phrases exemples sont remplacées par des balises («<source>, <target>») ; (iii) une analyse morpho-syntaxique est effectuée sur les phrases exemples et les mots du contexte entre les entités sont remplacés par leur catégorie morpho-syntaxique ; (iv) enfin les n phrases exemples les plus fréquentes sont conservées afin servir de patrons pour le type de relation concerné. À la différence de notre approche, [Li et al., 2009a] n'utilise pas de processus de généralisation entre les phrases exemples.

Pour ce qui est de la recherche de documents, les auteurs considèrent les 700 premiers documents renvoyés par un moteur de recherche. Par la suite, une chaîne de traitement linguistique (découpage en phrase et étiquetage des catégories morpho-syntaxiques) est appliquée à chacun de ces documents. Dans

5. PEUPLEMENT DE BASES DE CONNAISSANCES

notre cas, l'ensemble des pré-traitements linguistiques sont appliqués sur l'ensemble du corpus.

Concernant l'extraction des relations, le système est fondé sur une démarche en deux phases, premièrement l'application des patrons et deuxièmement la validation des entités cibles extraites. Les entités cibles sont repérées en appliquant des expressions régulières, qui sont soit dérivées des patrons de relations, soit des expressions spécifiques (cas des nombres et des dates). La sélection des entités cibles se fait sur un critère de redondance : les réponses les plus fréquentes sont retenues. La validation des entités cibles vise à vérifier que celles-ci sont incluses dans des listes fermées de valeurs. Par exemple, pour le champ *per :alternate_names* la liste est constituée à partir des liens de redirection issus de Wikipedia.

Selon la description ci-dessus, le système [Li et al., 2009a] est assez similaire à notre approche. Néanmoins les auteurs l'ont complété en ajoutant une étape permettant d'acquérir, *a posteriori*, des entités cibles. L'idée générale de cette étape est de rechercher des entités cibles dans un autre corpus que celui de la campagne (en l'occurrence des articles de Wikipedia) et de les ajouter aux entités cibles déjà extraites. Afin de trouver ces entités cibles, deux cas de figure sont possibles : l'entité cible recherchée fait l'objet d'un article Wikipédia (elle forme le titre de l'article) ou l'entité cible recherchée se trouve dans un article Wikipedia. Dans le second cas, des patrons de relations sont appliquées aux articles Wikipedia afin de retrouver les entités cibles potentielles. Dans les deux cas, il faut s'assurer que ces entités cibles supplémentaires sont bien mentionnées dans un document du corpus initial afin qu'elles soient pertinentes. Si c'est le cas, les entités cibles supplémentaires viennent se rajouter aux autres entités. Ce processus de validation de réponses est repris par d'autres participants dans TAC-KBP 2010, à l'instar de [Chen et al., 2010b; Lehmann et al., 2010]. La distinction est que des bases de connaissances dérivées de Wikipedia sont utilisées au lieu de la version originelle, [Lehmann et al., 2010] se servent de DBpedia¹ et [Chen et al., 2010b] se servent de Freebase² [Bollacker et al., 2007].

Le système de l'équipe *Cortex* [Chada et al., 2010] a été évalué seulement sur la

¹<http://dbpedia.org/About>

²<http://www.freebase.com/>

tâche *Slot Filling*. Ce système est en fait composé de trois systèmes indépendants bien que très liés les uns aux autres : (i) *HERBE* (*Heuristic Evidence-and-Resource-Based Extraction*) est responsable des traitements linguistiques de base (détection des entités nommées, découpage en phrases/paragraphes, normalisation, etc.) ; (ii) *Summon System* s'occupe des analyses syntaxiques/sémantiques et de la résolution des anaphores ; (iii) *Noun Tree* est chargé de représenter les relations sémantiques entre les entités nommées sous forme de graphes hiérarchiques.

Plus précisément, les relations entre les entités du graphe sont dirigées, le graphe est acyclique et les entités peuvent être connectés à plusieurs parents (ce qui explique le nom d'arbre). Le système *Noun Tree* se sert des sorties des deux autres systèmes pour construire une représentation globale cohérente¹ de toutes les connaissances dans les textes à partir des relations collectées. Il est important de noter que dans le cadre de la campagne TAC-KBP 2010, un corpus de plus de 3 millions de documents (articles de presses et blogs) à été manuellement afin de construire l'ensemble de règles (qui sont appliquées au système *Noun Tree*) permettant d'extraire les relations spécifiques à la campagne.

Au niveau des pré-traitements linguistiques, ils sont effectués en amont de la phase d'extraction de relations et sur l'ensemble des documents du corpus. Globalement, ces traitements sont effectués en appliquant les trois systèmes précédents auxquels on rajoute une phase d'indexation de toutes les entités et de leurs formes alternatives. L'indexation ayant pour but de retrouver les documents mentionnant une entité donnée.

À l'opposé de notre système et de celui de [Li et al., 2009a], le système de l'équipe Cortex n'utilise pas de patrons lexico-syntaxiques pour extraire les relations. À la place, il extrait les relations en vérifiant qu'un certain nombre de règles (contraintes) sémantiques sont applicables sur la représentation sous forme de graphe d'un document. Ces règles viennent s'ajouter aux relations syntaxiques et sémantiques déjà détectées par le *Summon System*. Précisément, ces règles² correspondent à des sous-arbres dont on cherche à vérifier la présence dans le *Noun Tree*.

¹Avant d'établir une nouvelle relation entre deux entités le système vérifie qu'il n'y ait pas d'incompatibilités avec les relations déjà présentes dans le graphe.

²Il y a environ 200 règles au total pour les 42 types de relations.

5. PEUPLEMENT DE BASES DE CONNAISSANCES

Par rapport au système de l'équipe Thu Quanta celui de l'équipe Cortex ne se sert ni d'étape de validation des réponses, ni d'étape d'acquisition de réponse supplémentaire. Pourtant le second obtient de bien meilleurs résultats¹ sur lesquels nous reviendrons dans la section suivante. Une des raisons est que l'approche de l'équipe Thu Quanta ne semble utiliser aucun mécanisme permettant d'extraire des relations inter-phrastiques (résolution de coréférence, etc.).

5.8 Discussion sur les résultats de TAC-KBP

L'extraction de relations à large échelle, au sens où nous l'avons définie à la section 5.4, est une problématique encore récente. Néanmoins, au travers notamment des campagnes d'évaluation TAC-KBP, elle a été l'objet d'un certain nombre de travaux proposant différentes approches. Concernant spécifiquement l'extraction des relations, les travaux se répartissent entre l'utilisation de l'apprentissage statistique [Agirre et al., 2009; Chen et al., 2010b; Li et al., 2009b], l'induction de patrons lexicaux [Chen et al., 2010b; de Pablo-Sánchez et al., 2009; Li et al., 2009a; McNamee et al., 2009] et enfin, l'adaptation de systèmes existants pour la détection de relations [Bikel et al., 2009; Schone et al., 2009]. On note pour TAC-KBP 2010 l'introduction d'approches à base de règles développées manuellement, par exemple [Byrne and Dunnion, 2010], et d'approches reposant sur le principe de «Distant supervision» à partir de classifieurs, dont celle de [Surdeanu et al., 2010].

Notre approche relève de l'induction de patrons lexicaux et fait l'hypothèse, comme [Mintz et al., 2009], que la seule présence d'un couple d'entités dans une phrase est suffisante pour marquer la présence effective d'une relation entre ces entités. En pratique ce n'est pas toujours le cas et on peut distinguer au moins trois configurations qui posent problème :

1. Les entités ne sont pas en relation même si elles sont dans la même phrase
- «*Larry Page has employed Eric Schmidt, he will be working as the new chairman at Google.*»
2. Les entités sont en relation et le type de la relation ne correspond à aucun

¹Nous soulignons que les corpus, les requêtes et les métriques d'évaluations sont différentes.

type défini dans la campagne TAC-KBP

- «*Even Larry Page and Sergey Brin wanted to sell Google to Yahoo.*»

3. Les entités sont en relation et le type de relation correspond à un où plusieurs types définis dans la campagne TAC-KBP. Ici il s'agit de *org :founded_by* et *org :top_members_employees*

- «*Larry Page co-founded Google, along with Sergey Brin*»

- «*... Larry Page is also the Chief Executive Officer of Google ...*»

Nous pensons ainsi qu'il est important de filtrer en amont les exemples utilisés pour l'induction des patrons, à l'instar de ce que propose [Banko and Etzioni, 2008; Wang et al., 2011b] (problème 1), [Riedel et al., 2010] (problèmes 1 et 2) et [Bunescu and Mooney, 2007; Hoffmann et al., 2011] (problème 3).

Notre approche, comme la plupart de celles des participants à TAC-KBP, utilise comme source de supervision distante les infobox issues de Wikipedia [Li et al., 2009a; Surdeanu et al., 2010]. Un inconvénient de cette démarche est que Wikipedia est un outil communautaire et donc on ne peut garantir l'absence d'informations fausses, incorrectes ou inexacts. Ce type d'informations influence directement les phrases exemples servant à l'induction des patrons, cependant il est difficile de mesurer cet impact puisque l'on ne connaît pas la proportion de couples d'entités incorrectes «*a priori*». Dans notre approche, des heuristiques simples ont été utilisées pour filtrer les couples d'entités présents dans la base de connaissances et utilisées pour la sélection des phrases exemples. Par exemple, le champ *per :age* de la KB doit être une valeur numérique inférieure à 100. Une alternative, consiste à utiliser une autre source de supervision de grande taille et contenant moins d'informations erronées. Dans cette perspective [Nguyen and Moschitti, 2011] utilise YAGO [Suchanek et al., 2007] comme source de supervision pour obtenir des couples d'entités qui sont ensuite projetés dans un corpus d'articles issus de Freebase afin d'obtenir des phrases exemples. Les auteurs se servent d'un modèle SVM pour l'apprentissage des relations. Il faut souligner qu'ils n'utilisent pas les données de la campagne TAC-KBP mais qu'ils cherchent à repérer un ensemble de 52 types de relations, comparable à celui de la campagne. Concernant les performances, ils reportent un score de global de 67 % de F1-mesure pour la détection des entités nommées et l'extraction des relations.

Comme notre système, ceux élaborés pour TAC-KBP 2009 n'exploitent pas

5. PEUPLEMENT DE BASES DE CONNAISSANCES

les liens de dépendance entre les types de relations, à l'image du lien entre la date de naissance et l'âge par exemple. Dans [Chen et al., 2010a], les auteurs montrent que les résultats obtenus dans [Li et al., 2009a] (32 % de f1-mesure) peuvent être améliorés (ils obtiennent 34,8 % de f1-mesure) par l'intégration des dépendances entre les relations en utilisant des règles d'inférence fondées sur une extension de la logique du premier ordre. Plus généralement, [Chada et al., 2010] ont montré dans le cadre de TAC-KBP 2010 une augmentation très significative des performances en intégrant des mécanismes qui d'une part permettent d'extraire des relations au-delà de la phrase et d'autre part intègrent des informations syntaxiques et sémantiques.

Sur un autre plan, [Li et al., 2009a] se distinguent dans TAC-KBP 2009 en utilisant deux étapes d'extraction de relations : la première vise à retrouver dans les documents du corpus des entités cibles potentielles ; la seconde se sert d'une version récente de Wikipédia pour trouver des entités cibles potentielles supplémentaires qui n'auraient pas été identifiées lors de la première étape. Cette récupération d'entités améliore les performances de façon significative (+9 % de f1-mesure par rapport à [Bikel et al., 2009]) mais ajoute l'utilisation d'un corpus externe que l'on peut considérer comme trop lié à la KB. Les résultats sur TAC-KBP 2010 ont d'ailleurs montré que les performances globales pouvaient être améliorées sans cette ressource supplémentaire et que son impact sur les résultats est plus limité que pour TAC-KBP 2009 (une baisse des résultats a même été observée).

5.9 Conclusions

Dans ce chapitre, nous avons présenté un système d'extraction d'information à large échelle permettant d'extraire des relations de nature attributive entre entités nommées. Le qualificatif «à large échelle» recouvre à la fois la prise en compte d'un grand nombre de types de relations et la recherche de ces relations dans un large corpus. Ce système se fonde sur une approche faiblement supervisée dans laquelle les exemples se limitent à des couples d'entités en relation.

L'extraction des relations s'effectue par l'application de patrons lexico-syntaxiques caractéristiques des types de relations considérés, ces patrons étant appris

à partir de phrases issues de la projection des couples d'entités exemples dans un corpus. Nous avons évalué les résultats de cette approche en utilisant le cadre d'évaluation offert par la tâche *Slot Filling* de la campagne d'évaluation TAC-KBP en nous concentrant sur la problématique de l'extraction des relations proprement dite, sans nous attacher à la détection de l'absence d'une relation dans un corpus.

Les résultats obtenus dans ce contexte se situent dans la moyenne des résultats obtenus par les participants de l'édition 2010, ce que nous pouvons considérer comme un point de départ intéressant dans la mesure où notre système repose sur une approche volontairement générique et n'exploite que très faiblement les spécificités des types de relations traités. Nous avons aussi pu montrer que des techniques permettant de prendre en compte certains aspects d'un passage à une «large échelle», comme le filtrage des couples de phrases exemples utilisées pour l'induction des patrons par l'utilisation d'un algorithme optimisé pour le calcul de similarités entre relations (APSS), ne dégradent pas les performances et peuvent même contribuer à les améliorer.

Ces résultats sont également à mettre en parallèle avec ceux obtenus avec des techniques similaires dans un domaine de spécialité, en l'occurrence le domaine médical [Embarek and Ferret, 2008]. Dans ce dernier cas, l'extraction de relations obtient une f1-mesure moyenne aux alentours de 60 %¹. Même en tenant compte des pertes issues du processus initial de recherche d'information (de l'ordre de 15 % dans notre cas au niveau des documents), on obtient une performance de 50 % environ, à comparer aux 30 % obtenus par le meilleur système de TAC-KBP 2009. On constate donc que le fait de travailler en domaine ouvert, c'est-à-dire avec des types d'entités et de relations peu spécifiques, constitue une difficulté certaine à prendre en compte.

Concernant les perspectives d'amélioration, plusieurs voies sont envisageables. La première vise le traitement de la coréférence (entre un pronom et une entité) ou plus généralement des phénomènes de référence (entre les entités nommées) pour permettre la détection de relations entre des entités présentes dans des

¹Dans [Embarek and Ferret, 2008], une f1-mesure de 74 % est obtenue pour quatre types de relations mais avec une sélection manuelle des phrases exemples. Après analyse, l'impact de ce processus de sélection sur les résultats est estimé à 15 %.

5. PEUPLEMENT DE BASES DE CONNAISSANCES

phrases différentes. Pour le moment, les phrases candidates sur lesquelles les patrons sont appliquées sont seulement celles qui contiennent l'entité source (ou une forme étendue de l'entité source). Notre idée est d'ajouter à ces phrases candidates celles qui contiennent une anaphore (pronom ou groupe nominal) de l'entité cible. Par la suite, cette anaphore sera remplacée par la valeur initiale de l'entité cible afin d'appliquer les patrons. En résumé l'alternative que nous proposons vise à intégrer la résolution de la coréférence afin d'appliquer les patrons sur un ensemble plus important de phrases candidates. Le système pourra ainsi tenir compte des cas ci-dessous, qui ne sont pas pris en charge actuellement.

- L'entité source et la cible potentielle ne sont pas dans la même phrase. Aussi, l'entité source est reprise par un pronom :

«<source>Barack Hussein Obama II</source> (born August 4, 1961) is the 44th and current President of the United States. **He** is the first African American to hold the office. **He** is a graduate of <cible>Columbia University</cible> and ...».

- L'entité source et la cible potentielle sont dans la même phrase. L'entité source est reprise par un groupe nominal :

«<source>SAP</source> was founded in Mannheim, Germany, in 1972 by five former IBM employees, **the company** now has over <cible>55,000 employees</cible> around the world.»

La deuxième perspective concerne la construction des listes fermées de valeurs (listes de noms de pays, etc.) servant à la recherche des phrases candidates. Actuellement, ces listes sont utiles à notre système pour contourner les erreurs de reconnaissances d'entités nommées et/ou se substituer au processus de détection des entités (en particulier pour les types d'entités nommées spécifiques, ex. : *org :political_regligious_affiliation*). Si ces listes permettent d'améliorer la pertinence des relations extraites, elles influencent les performances de la sélection des phrases candidates : le temps de traitement augmente avec la taille de la liste. Notre motivation est de diminuer la taille de ces listes en fonction de différents facteurs : le type de la relation à traiter, la langue du corpus, etc. Pour illustration, si on considère les noms de lieux, plus spécifiquement les noms de villes ils

sont souvent différents en fonction de la langue considérée (ex. : Pékin, Beijing, Pekin). L'idée est de ne pas considérer les valeurs dans une langue différente de celle du corpus. Dans notre cas les listes sont très génériques et peuvent contenir plusieurs valeurs équivalentes. Nous avons, dans un premier temps, effectué une revue manuelle pour éliminer certaines redondances, mais cette démarche devrait être étendue de façon automatique. Ce problème de constitution de ressources pour l'extraction de relations à large échelle a été abordé dans [Hoffmann et al., 2010], où les listes constituées servent à construire des dictionnaires utilisés pour enrichir les *features* d'un modèle CRF. Ces listes de valeurs pourraient également être utilisées pour entraîner des modèles capables de reconnaître directement les entités spécifiques et améliorer ainsi de façon directe la reconnaissance des entités spécifiques. Il s'agirait, par exemple, de construire ces modèles sur le même principe de faible supervision que nous avons présenté pour les relations entre entités.

La troisième perspective concerne la prise en compte des dépendances entre les types de relations. La piste que nous envisageons vise à ajouter des contraintes sur les valeurs extraites pour les entités cibles. Ces contraintes peuvent porter sur la cohérence entre deux valeurs qui sont liées : par exemple pour les champs *per :date_of_birth* et *per :age*, on peut détecter une erreur si la valeur de l'âge que l'on extrait et celle calculée à partir de la date de naissance ne sont pas cohérentes. Ces contraintes peuvent également porter sur des incompatibilités : par exemple, les champs *per :spouse* et *per :children* ne peuvent pas contenir de valeurs identiques, ou sur des déductions par transition : par exemple, on peut établir des relations de parentés entre les personnes (champ *per :siblings*) s'ils ont des valeurs se recouvrant dans le champ *per :parents*.

5. PEUPLEMENT DE BASES DE CONNAISSANCES

Chapitre 6

Conclusion

Dans ce chapitre de conclusion, nous proposons un résumé et une analyse de notre approche d'extraction d'information, ainsi que plusieurs perspectives pour des travaux futurs.

6.1 Bilan des résultats

Les travaux présentés dans cette thèse portent sur l'extraction d'information, premièrement dans un cadre contraint par un domaine d'application spécifique et deuxièmement dans un cadre plus général et à plus large échelle. En ce qui concerne l'extraction d'information dans un cadre contraint, nous nous sommes intéressés au domaine sismique avec pour objectif de construire des *templates* sur des événements à partir de dépêches de presse. Nous avons proposé une approche d'extraction d'information en deux étapes que nous avons détaillée dans les chapitres précédents.

La première étape de notre démarche (chapitre 3) réalise une segmentation événementielle des documents à partir d'indices temporels trouvés au niveau de chaque phrase d'un document. L'objet de cette segmentation est d'isoler les phrases pertinentes en rapport avec le séisme principal du document et d'écarter les phrases concernant les autres événements afin qu'elles ne viennent pas perturber le processus d'extraction. Dans la section 3.2, nous avons présenté différentes approches pour la segmentation en événements des textes. Parmi toutes ces approches, celle que nous avons retenue est fondée sur un modèle d'appren-

6. CONCLUSION

tissage statistique (CRF) qui permet d’obtenir des résultats satisfaisants pour l’identification des phrases en rapport avec le séisme principal d’un document. Plus généralement, les expérimentations sur la segmentation en événements ont montrées que, dans le cadre des mentions d’événements sismiques dans des articles de presse, il est possible de mettre en avant des structures discursives liées aux événements en suivant les changements d’indices temporels.

La construction des *templates*, qui est la deuxième étape de notre démarche (chapitre 4), vient s’inscrire dans la continuité de la segmentation événementielle. Son objectif est de sélectionner, à partir des segments pertinents (ceux identifiés lors de la segmentation événementielle), les entités qui serviront à l’élaboration des *templates* sur les événements. Pour la sélection des entités, nous avons proposé différentes méthodes génériques qui s’appuient sur un graphe d’entités. Plus globalement, la finalité de ce graphe d’entités est de représenter les relations entre les entités au niveau du document. Dans la section 4.3.1 nous avons proposé plusieurs approches à base d’apprentissage statistique, ainsi que différents ensembles de *features* afin de construire un tel graphe. Par la suite, ce graphe d’entités sert de support pour le processus de sélection des entités pertinentes. Concernant ce processus de sélection, nous avons suggéré différentes méthodes génériques et non dépendantes d’un domaine particulier (ou d’une langue spécifique) (section 4.3.2). Par ailleurs, suite aux expérimentations sur les méthodes de sélection des entités, nous avons opté pour une approche hybride (qui exploite les propriétés de plusieurs approches) qui obtient des résultats encourageants pour la construction des *templates*. Plus globalement, les expérimentations ont montrés que pour les événements sismiques, la combinaison de la segmentation en événements et de la sélection des entités à partir d’un graphe permet d’obtenir des résultats satisfaisants, avec un score de F-mesure de 77 % pour les entités rattachées aux événements.

Sous un autre angle, nous avons abordé l’extraction d’information dans un cadre plus général dans le chapitre 5. À la différence des deux étapes précédentes où le but était d’élaborer des relations complexes entre les entités afin de constituer des événements, cette partie des travaux concerne l’extraction de relations entre des entités à large échelle (corpus d’environ 2 millions de documents, 42 types de relations à extraire). En particulier, nous nous sommes intéressés à l’ex-

traction de relations dans le but de compléter une base de connaissance existante. Notre motivation est de se servir des connaissances issues d’une telle base afin de proposer aux utilisateurs des informations génériques concernant les entités participant à des événements. L’intérêt de telles informations est qu’elles ne sont pas liées à un domaine particulier et restent pertinentes quelque soit le domaine des événements. Pour ce faire, nous avons proposé une approche faiblement supervisée qui s’appuie sur des patrons lexico-syntaxiques. Cette approche, décrite dans la section 5.2, a été évaluée dans le cadre de la campagne TAC-KBP. Plus précisément, notre approche est composée de deux étapes principales, la première est chargée de l’apprentissage des patrons de relations, la seconde de l’application des patrons et de l’extraction des relations. Par rapport aux systèmes ayant participé à cette campagne, notre système se distingue principalement par les patrons de relations utilisés. Il s’agit de patrons multi-niveaux qui sont induits à partir d’un ensemble (conséquent, mais bruité) de phrases exemples. Afin d’améliorer la qualité de ces patrons, un filtrage des phrases exemples a été proposé (*cf.* section 5.4.1). Les expérimentations ont montré que ce filtrage des phrases exemples permettait d’améliorer les performances en terme de F1-mesure pour l’extraction des relations. Plus généralement, en faisant une évaluation postérieure sur les données TAC-KBP 2010, notre système se situe dans la moyenne des résultats des participants à cette évaluation.

6.2 Analyse de notre contribution

Dans cette section, nous proposons une analyse de notre contribution selon les cinq familles de critères d’évaluations proposées dans [Burstein and Gregor, 1999], qui fournissent des directives génériques pour évaluer le développement de systèmes d’informations, en les adaptant au contexte de l’extraction d’information et plus singulièrement à nos travaux.

Le premier critère, de *significativité* (*Significance*), vise à mesurer l’importance de l’impact théorique et/ou pratique du système ou de la méthode proposée par rapport aux travaux existants (au cas où des travaux similaires existent). Par exemple, on cherche à comparer les différences du point de vue des performances

6. CONCLUSION

(système plus rapide, etc.) ou du point de vue des fonctionnalités (fonctions supplémentaires, etc.). Pour notre travail, ce critère se traduit par les questions suivantes :

- Quels sont les principales contributions du travail de recherche pour l'extraction d'information ?
- Est-ce que les approches proposées améliorent les performances par rapport aux approches existantes ?
- Quels sont les distinctions entre les approches proposées et les travaux précédents ?

Vis à vis de la première question, nos résultats, résumés dans la section 6.1, mettent en avant les contributions suivantes : la première concerne une segmentation en événements des articles de presses à partir d'indices temporels ; la seconde concerne une approche de construction de *templates* utilisant un graphe d'entités nommées ainsi que le résultat de la segmentation événementielle ; la troisième concerne une approche faiblement supervisée pour l'extraction de relations entre entités nommées.

Concernant nos performances (deuxième question), il n'est pas simple de juger de façon systématique les résultats, en tout cas, pour les étapes de segmentation en événements et de construction des *templates*. Nos expérimentations ont été effectuées sur un corpus de petite taille en comparaison des corpus utilisés lors des campagnes MUC ou encore TAC-KBP. Aussi, peu de travaux similaires ont été menés pour la langue française. En revanche, concernant l'extraction des relations à large échelle, nos résultats se situent dans la moyenne des systèmes pour TAC-KBP 2010, et dans les 5 premiers pour l'extraction de relations à l'intérieur des phrases.

Pour ce qui est de la dernière question, le processus d'extraction d'information en deux étapes que nous avons proposé est relativement différent des processus existants : l'utilisation d'informations temporelles n'avait jamais, à notre connaissance, été exploitée pour distinguer les événements pertinents des documents. Par ailleurs, les relations complexes entre entités n'avaient pas été utilisées pour constituer des événements au niveau du document (elles ont été appliquées seulement au niveau des phrases).

Pour le second critère, de *validité interne* (*Internal validity*), il est question de

mesurer la crédibilité des arguments qui sont avancés, et d'évaluer si les résultats avancés sont logiques ou cohérents avec ces arguments. Ce critère peut être associé aux questions suivantes :

- Est-ce que la démarche proposée est adaptée à la problématique ? Atteint-elle son objectif ?
- Si des hypothèses ont été faites sur les résultats, ont-elles été validées ?
- Des approches similaires (concurrentes) ont-elles été considérées ?

Pour la première question, notre objectif premier concerne la construction de *templates* pour des événements dont les caractéristiques sont dispersées dans des documents. Nous avons vu que notre démarche en deux étapes était adaptée à ce type de problème : la segmentation cible les passages pertinents sur les événements pour l'événement principal, puis les informations liées à l'événement principal sont regroupées pour construire un *template*. Nous avons vu que la segmentation en événements semble plus particulièrement convenir pour les documents mentionnant plusieurs événements, ce qui est fréquemment le cas pour notre cas d'application (plus de la moitié des dépêches de presse mentionnent plusieurs événements dans notre corpus). Par ailleurs, pour le regroupement des informations pertinentes sous forme de *templates*, nous avons montré que l'utilisation des relations complexes entre les entités, exprimées au niveau du document, apportaient une solution. Les évaluations ont montré que ce dernier problème n'était pas totalement résolu et que des erreurs d'associations entre entités et événement subsistaient. Enfin, pour l'extraction de relations à large échelle, notre approche couvre bien le problème. En revanche des améliorations restent à apporter à plusieurs niveaux (résolution de la coréférence, prise en compte des relations intra-phrastiques, etc.).

Au niveau des hypothèses que nous avons faites (deuxième question), nous avons cherché à les valider au fur et à mesure des expérimentations. Concernant la prise en compte des approches similaires (troisième question), si l'on considère la problématique d'extraction de relations entre entités nommées notre démarche s'inspire du principe de supervision distante proposé par [Mintz et al., 2009]. En revanche, concernant la problématique de construction de *templates*, la démarche globale est différente des approches existantes sur plusieurs points, comme cela a été mis en avant précédemment. Il faut tout de même signaler que

6. CONCLUSION

cette démarche a des similitudes avec des travaux existants. Par exemple, lors de l'étape de construction des *templates*, la phase de construction du graphe d'entités nommées repose sur une étape de classification d'entités nommées qui est inspirée des travaux de [McDonald et al., 2005].

Le troisième critère, de *validité externe* (*External validity*) concerne le niveau de généralisation associé aux résultats trouvés. Par exemple, il peut s'agir d'une abstraction d'une approche existante ou l'application d'une même démarche dans des configurations différentes. Les questions liées à ce critère sont :

- Si les résultats sont obtenus à partir d'approches existantes, est-ce qu'ils sont cohérents avec ceux obtenus par d'autres travaux reposant sur ces mêmes approches ?
- Est-ce que les méthodes ou processus produits sont suffisamment génériques pour être transférés à d'autres domaines ?
- Quels sont les limitations à considérer afin d'appliquer les approches proposées à d'autres domaines ?

Pour ce qui est de la comparaison de nos résultats vis à vis des approches que nous avons réutilisées, nous avons fourni, dans la mesure du possible, des évaluations comparatives. Par exemple, pour l'étape de construction des relations complexes, nous avons obtenu des résultats comparables à ceux de [Afzal, 2009; McDonald et al., 2005] concernant la classification des relations binaires. Aussi, lorsque c'était possible nous avons présenté des résultats pour des systèmes servant de «*baseline*». En ce qui concerne l'extraction de relations entre entités nommées, nous nous sommes comparés aux participants à la campagne TAC-KBP.

Concernant la transposition de notre approche à d'autres domaines (deuxième question), différents axes sont envisageables selon les problématiques. Pour la problématique de construction de *templates* concernant les événements, il s'agit de la langue et du genre des documents, du type d'événement ainsi que du domaine visé. Nous avons essayé de conserver un niveau de généralité important pour notre démarche de construction de *template* : la phase de segmentation se sert d'un classifieur statistique qui repose en grande partie sur les indices temporels présents dans les documents (*a priori* non dépendants d'un domaine ou d'un genre, mais dépendant de la langue) ; la phase de construction des *templates* s'ap-

puie sur un graphe d'entités qui est construit à partir d'un classifieur statistique qui se sert de *features* non lexicalisés. Enfin, le processus de sélection des entités dont on se sert pour l'élaboration des *templates* exploite les propriétés du graphe d'entités, qui sont donc indépendantes de la langue, du genre des documents ou du domaine. Nous avons mené quelques expérimentations pour transposer notre démarche à l'extraction d'événements sismiques pour des dépêches de presse en langue anglaise.

Pour la seconde problématique d'extraction de relations à large échelle, la transposition à un autre domaine vise surtout la langue des documents. L'approche faiblement supervisée que nous avons présentée pour l'apprentissage et l'extraction des relations est très général, par conséquent elle devrait être transférable à une autre langue que l'anglais. Nous n'avons pas menés d'expérimentations dans ce sens, même si nous sommes optimistes sur la facilité de la transposition.

Pour répondre à la troisième question, nous listons quelques points dont il faut tenir compte pour faire une transposition de notre démarche à d'autres domaines, d'abord pour la construction de *templates* sur des événements, puis pour l'extraction faiblement supervisée de relations.

Construction de *templates* sur les événements

- Notre segmentation en événements ne distingue pas les événements secondaires les uns des autres. Cela n'a pas de conséquence dans une perspective de veille événementielle, dans laquelle les utilisateurs ne sont intéressés que par l'événement le plus récent, et dans laquelle les événements mentionnés sont dans des cadres temporels différents. En revanche cela peut impacter d'autres domaines, en particulier si plusieurs événements de même nature peuvent être identifiées dans le même cadre temporel (par exemple, pour des événements de type «*changement de personnel dans une entreprise*», les annonces évoquent souvent une série de changements qui concernent plusieurs personnes et plusieurs postes).
- Il faut tenir compte des événements à considérer dans les *templates* : dans nos travaux, seules les entités associées à l'événement principal d'un texte sont considérées par l'approche de construction des *templates*. Pour intégrer les informations des autres événements, il faudrait considérer

6. CONCLUSION

de façon séparée les segments associés aux autres événements (qui sont identifiés lors de la segmentation en événements).

- La segmentation en événements est particulièrement intéressante lorsque les documents ont une structure événementielle clairement identifiable, par exemple s'ils discutent d'un événement central et de plusieurs événements qui lui sont comparables (qui le précèdent). Pour des types de documents qui n'ont pas cette structure, par exemple pour les dialogues, la segmentation en événements peut ne pas être pertinente. Dans ce cas, cette étape peut être supprimée du processus d'extraction des *templates* sans perturber la construction des *templates*.
- Les deux étapes de notre démarche reposent sur des classifieurs statistiques qui ont été entraînés à partir de *features* très généraux, mais sur un corpus particulier, lié au domaine sismique et au genre journalistique. Sur un corpus correspondant à un autre domaine ou un autre genre de document, il conviendrait de tester les modèles appris sur notre corpus et les comparer à des modèles appris sur ces nouveaux corpus, pour vérifier la généralité de notre approche.

Extraction faiblement supervisée de relations

Concernant la transposition à une autre langue, les phases d'apprentissage et d'extraction de relations restent identiques. Les processus sont identiques à condition de changer les documents du corpus pour qu'ils soient dans la langue cible qui nous intéresse.

Concernant les autres critères de [Burstein and Gregor, 1999], de *confirmabilité* (*Objectivity/Confirmability*) et de *fiabilité* (*Reliability/Dependability/Audibility*), nous avons choisi de les regrouper car ils ne concernent pas directement les approches que nous proposons (ou les résultats obtenus), mais plutôt la manière dont les ressources et les procédures sont décrites afin de pouvoir reproduire les résultats. Les questions liées à ces deux critères sont :

- Les problématiques de recherche ont-elles été clairement définies ?
- Les approches proposées sont-elles décrites avec suffisamment de détails ?
- Les expérimentations sont-elles décrites avec suffisamment de détails ?
- Les données servant à l'entraînement (apprentissage) et à l'évaluation (test) des modèles sont-elles clairement spécifiées ?

Les problématiques auxquelles nous avons essayé de répondre ont été détaillées dans la section 2.9 du manuscrit. Chacune de ces problématiques est reprise dans un chapitre et nous avons proposé une démarche pour répondre à une partie des problèmes qui se posent. Les hypothèses, ainsi que les détails concernant les approches de segmentation en événements et de construction de *template*, sont présentés dans les chapitres 3 et 4. L'extraction de relations à large échelle est discutée dans le chapitre 5. Les *features* utilisées par les modèles statistiques et les données utilisées sont aussi présentés lors de chaque phase d'expérimentation. En revanche, tous les corpus utilisés ne sont pas ouverts à la communauté : par exemple, le corpus annoté de dépêches concernant les événements sismiques a été construit dans le cadre d'un projet qui limite les possibilités de diffusion de ces données. Le corpus de la campagne TAC-KBP est quant à lui disponible après avoir fait la demande auprès des organisateurs. Lorsque nous avons utilisé des outils existants, nous avons fourni des liens vers les implémentations que nous avons utilisées.

6.3 Perspectives

Cette section discute de quelques perspectives pour continuer les travaux concernant les problématiques de construction des *templates* et d'extraction de relations à large échelle.

En ce qui concerne la *construction des templates*, une première piste concerne l'élaboration et l'enrichissement de *templates* en considérant plusieurs documents à la fois sur le même événement. Les informations sur les événements sont en effet souvent complétées ou corrigées, de façon incrémentale, au fur et à mesure que les documents sont publiés. Actuellement, la segmentation et la construction des *templates* ne s'appliquent qu'à un seul document indépendamment des autres, mais il serait pertinent d'étendre nos approches à plusieurs documents.

Par ailleurs, dans notre approche, les événements visés ainsi que les informations qui leur sont associées sont portées par des entités nommées nominales. Une extension possible peut être d'enrichir la seconde étape de construction des

6. CONCLUSION

templates en prenant en compte les événements portés par des formes verbales, qui sont relativement courants.

Enfin, les phases d'extraction d'information sur les événements et d'extraction faiblement supervisée de relations ne sont pas directement liées. Une voie intéressante pour la suite des travaux serait de combiner ces deux approches afin d'adapter le système à de nouveaux événements.

En ce qui concerne l'*extraction des relations à large échelle*, notre approche se sert de patrons lexicaux-syntaxiques multi-niveaux pour les phases d'apprentissage et d'extraction de relations. Il serait intéressant de tester le remplacement de ces patrons par des classifieurs statistiques. Dans cette perspective, un classifieur serait attribué à une relation au lieu d'un ensemble de patrons. Une difficulté liée à ce choix réside dans le corpus d'apprentissage, constitué par les phrases exemples annotées pour chaque type de relations, qui est actuellement construit de façon automatique et contient donc des erreurs. Il faudrait donc soit améliorer la qualité de ce corpus d'apprentissage, soit utiliser un modèle d'apprentissage plus robuste au bruit.

D'autre part, pour le moment, aucune phase de notre approche faiblement supervisée ne se sert des informations syntaxiques. Or, de telles informations sont souvent utilisées pour l'extraction de relations, et peuvent effectivement être utiles à plusieurs niveaux :

- Pour l'apprentissage des relations, des patrons syntaxiques peuvent se substituer aux patrons multi-niveaux ou les compléter. Ces mêmes patrons syntaxiques seraient appliqués pour l'extraction des relations.
- Concernant la collecte des phrases exemples servant à la construction des patrons, les phrases actuellement utilisées sont obtenues à partir de couples d'entités projetées dans un corpus. Le problème lié à cette collecte est qu'une même phrase peut servir d'exemple pour différentes relations. Notre idée est de faire un *clustering* à partir des chemins syntaxiques contenus dans les phrases exemples. De cette façon, on espère que les clusters ainsi formés fourniront une meilleure représentation de la relation sémantique. Par la suite, on pourrait supprimer des phrases exemples en fonction de certains critères : taille des clusters, présence de mots ou de verbes dans la phrase,

etc.

Enfin, concernant les phrases exemples qui servent à la construction de patrons, ces dernières sont actuellement collectées uniquement à partir d'un corpus unique. Il serait intéressant de pouvoir collecter d'autres phrases à partir d'autres ressources comme Wikipedia et plus généralement du web. L'utilisation de corpus plus importants permettrait d'avoir une plus grande variation d'expressions pour les relations, une plus grande redondance, grâce à laquelle on pourrait pondérer de façon fiable les patrons appris (par exemple selon leur fréquence), ainsi qu'un plus grand nombre d'occurrences candidates, sur lesquelles on pourrait appliquer des critères de filtrage plus stricts, ce qui améliorerait l'ensemble des phrases exemples pour l'induction de patrons tout en en gardant un nombre suffisant.

Plus généralement, une perspective importante pour la suite des travaux concerne l'adaptation des approches à des domaines différents de ceux que nous avons testés pour nos travaux. Les systèmes sont le plus souvent conçus et optimisés pour un domaine spécifique, par conséquent les performances sont impactées lorsque l'on change de domaine. Très récemment, quelques travaux se sont intéressés à cette problématique dans l'optique de rendre la démarche de transfert d'un domaine vers un autre plus générique [Freedman et al., 2011; Surdeanu et al., 2011]. La finalité est de pouvoir transposer le travail capitalisé sur un domaine vers un autre, avec un minimum d'effort, tout en conservant un niveau de résultat satisfaisant. Dans notre démarche de construction de *templates*, quelques composants s'appuient sur des classifieurs, qui ont une certaine capacité de généralisation, mais qui nécessitent néanmoins une annotation manuelle de corpus. En pratique, cette annotation devrait être reproduite pour chaque domaine. Aussi, afin que notre approche soit adaptable facilement à de nouveaux domaines il conviendrait d'intégrer des méthodes semi-supervisée (de type auto-apprentissage), à l'instar de [Freedman et al., 2011]. Sous un autre angle, l'adaptation à de nouveaux domaines pourrait passer par la combinaison de notre approche avec des approches à base de règles, qui peuvent capter de façon simple certaines adaptations, ou par l'utilisation d'un processus semi-automatique, intégrant les retours fait par des utilisateurs : typiquement pour l'extraction de relations, le système pourrait demander une validation à un utilisateur pour les relations extraites. Par la suite

6. CONCLUSION

les relations correctement extraites serviraient à trouver de nouveaux exemples (sur le modèle du *bootstrap*).

Liste des publications

1. Jean-Louis L., Besançon R., Ferret O. (2011). Text segmentation and graph-based method for template filling in information extraction. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, IJCNLP 2011. Chiang Mai, Thailand.
2. Jean-Louis L., Besançon R., Ferret O., Durand A. (2011). A weakly supervised approach for large-scale relation extraction. In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, KDIR 2011. Paris, France. (*Best Student Papers Award*)
3. Jean-Louis L., Besançon R., Ferret O. (2011). Une approche faiblement supervisée pour l'extraction de relations à large échelle. In *18ème Conférence annuelle de Traitement Automatique des Langues Naturelles*, TALN 2011. Montpellier, France.
4. Besançon R., Ferret O., Jean-Louis L. (2011). Construire et évaluer une application de veille sur les événements sismiques. In *8ème Conférence en Recherche d'Information et Applications*, CORIA 2011, article court. Avignon, France.
5. Jean-Louis L., Besançon R., Ferret O., Wang W. (2011). Using a weakly supervised approach and lexical patterns for the KBP slot filling task. In *Proceedings of the Text Analysis Conference Workshop*, TAC 2011. Gaithersburg, Maryland, USA.
6. Jean-Louis L., Besançon R., Ferret O. (2010). Using temporal cues for segmenting texts into events. In *Proceedings of the 7th International Conference on Natural Language Processing*, IceTAL 2010. Reykjavik, Iceland.

Liste des publications

7. Jean-Louis L., Besançon R., Ferret O. (2010). Utilisation d'indices temporels pour la segmentation événementielle de textes. In *17ème Conférence annuelle de Traitement Automatique des Langues Naturelles*, TALN 2010. Montréal, Canada.
8. Jean-Louis L., Besançon R., Ferret O. (2010). Using Conditional Random Fields for segmenting texts into events. In *Séminaire "CRF pour le TAL"*. Paris, France.

References

- Aceves-Pérez, R. M., Montes-y Gómez, M., and Villaseñor Pineda, L. (2007). Graph-Based Answer Fusion in Multilingual Question Answering. In *Proceedings of the 10th international conference on Text, speech and dialogue*, TSD'07, pages 621–629, Berlin, Heidelberg. Springer-Verlag. [81](#)
- Afzal, N. (2009). Complex Relations Extraction. In *Conference on Language & Technology 2009 (CLT09)*, Lahore, Pakistan. [34](#), [78](#), [79](#), [80](#), [85](#), [86](#), [93](#), [148](#)
- Agichtein, E. and Gravano, L. (2000). Snowball : Extracting Relations from Large Plain-Text Collections. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 85–94, New York, NY, USA. ACM. [38](#), [103](#), [104](#)
- Agirre, E., Chang, A., Jurafsky, D., Manning, C., Spitkovsky, V., and Yeh, E. (2009). Stanford-UBC at TAC-KBP. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA. [110](#), [136](#)
- Aone, C., Halverson, L., Hampton, T., and Ramos-Santacruz, M. (1998). SRA : Description of the IE2 system used for MUC. In *Proceedings of the 7th Message Understanding Conference, MUC-7*, Fairfax, Virginia. [42](#), [47](#), [50](#)
- Aone, C. and Ramos-Santacruz, M. (2000). REES : a Large-Scale Relation and Event Extraction System. In *Proceedings of the sixth conference on Applied natural language processing*, ANLC '00, pages 76–83, Stroudsburg, PA, USA. Association for Computational Linguistics. [4](#), [42](#), [47](#), [50](#), [76](#)
- Appelt, D. E., Hobbs, J. R., Bear, J., Israel, D., Kameyama, M., Martin, D., Myers, K., and Tyson, M. (1995). SRI International FASTUS System : MUC-6

REFERENCES

- Test Results and Analysis. In *Proceedings of the 6th conference on Message understanding*, MUC6 '95, pages 237–248, Stroudsburg, PA, USA. Association for Computational Linguistics. [42](#), [47](#)
- Appelt, D. E. and Onyshkevych, B. (1998). The Common Pattern Specification Language. In *Proceedings of a workshop on held at Baltimore, Maryland : October 13-15, 1998*, TIPSTER '98, pages 23–30, Stroudsburg, PA, USA. Association for Computational Linguistics. [17](#)
- Atzmüller, M., Klügl, P., and Puppe, F. (2008). Rule-Based Information Extraction for Structured Data Acquisition using TextMarker. In *LWA '08*, pages 1–7. [26](#)
- Bach, N. and Badaskar, S. (2007). A Survey on Relation Extraction. Technical report, Language Technologies Institute, Carnegie Mellon University. [39](#)
- Bagga, A. (1998). Evaluation of Coreferences and Coreference Resolution Systems. In *Language Resources and Evaluation*. [29](#)
- Baldwin, B. (1997). CogNIAC : High Precision Coreference with Limited Knowledge and Linguistic Resources. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, ANARESOLUTION '97, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics. [32](#)
- Banko, M., Cafarella, M. J., Soderl, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction from the Web. In *In IJCAI*, pages 2670–2676. [104](#)
- Banko, M. and Etzioni, O. (2008). The Tradeoffs Between Open and Traditional Relation Extraction. In *Proceedings of ACL-08 : HLT*, pages 28–36, Columbus, Ohio. Association for Computational Linguistics. [59](#), [103](#), [137](#)
- Bayardo, R., Ma, Y., and Srikant, R. (2007). Scaling Up All Pairs Similarity Search. In *16th International Conference on World Wide Web (WWW'07)*, pages 131–140, Banff, Alberta, Canada. [110](#)

REFERENCES

- Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A Maximum Entropy Approach to Natural Language Processing. *Comput. Linguist.*, 22 :39–71. [20](#)
- Berkhin, P. (2002). Survey Of Clustering Data Mining Techniques. Technical report, Accrue Software, San Jose, CA. [22](#)
- Besançon, R., Ferret, O., and Jean-Louis, L. (2011). Construire et Évaluer une Application de Veille pour l’Information sur les Événements Sismiques. In *CORIA*, pages 287–294. [63](#)
- Besançon, R., de Chalendar, G., Ferret, O., Gara, F., and Semmar, N. (2010). LIMA : A Multilingual Framework for Linguistic Analysis and Linguistic Resources Development and Evaluation . In *7th Conference on Language Resources and Evaluation (LREC 2010)*, Malta. [64](#)
- Bestgen, Y. and Vonk, W. (2000). Temporal Adverbials as Segmentation Markers in Discourse Comprehension. *Journal of Memory and Language*, 42(1) :74–87. [55](#)
- Bikel, D., Castelli, V., Radu, F., and jung Han, D. (2009). Entity Linking and Slot Filling through Statistical Processing and Inference Rules. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA. [136](#), [138](#)
- Bikel, D. M., Schwartz, R., and Weischedel, R. M. (1999). An Algorithm that Learns What’s in a Name. *Machine Learning*, 34 :211–231. 10.1023/A :1007558221122. [28](#)
- Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O’Reilly Media, 1 edition. [19](#)
- Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia - A Crystallization Point for the Web of Data. *J. Web Sem.*, 7(3) :154–165. [102](#)
- Bollacker, K., Cook, R., and Tufts, P. (2007). Freebase : a Shared Database of Structured General Human Knowledge. In *Proceedings of the 22nd national conference on Artificial intelligence - Volume 2*, pages 1962–1963. AAAI Press. [134](#)

REFERENCES

- Bollegala, D. T., Matsuo, Y., and Ishizuka, M. (2009). Measuring the Similarity Between Implicit Semantic Relations from the Web. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 651–660, New York, NY, USA. ACM. [111](#)
- Borthwick, A. E. (1999). *A Maximum Entropy Approach to Named Entity Recognition*. PhD thesis, Computer Science Department New York University, New York, NY, USA. AAI9945252. [28](#)
- Brin, S. (1999). Extracting Patterns and Relations from the World Wide Web. In Atzeni, P., Mendelzon, A., and Mecca, G., editors, *The World Wide Web and Databases*, volume 1590 of *Lecture Notes in Computer Science*, pages 172–183. Springer Berlin Heidelberg. [37](#), [38](#), [103](#)
- Bunescu, R. and Mooney, R. (2006). Subsequence Kernels for Relation Extraction. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems 18*, pages 171–178. MIT Press, Cambridge, MA. [41](#)
- Bunescu, R. C. and Mooney, R. J. (2005). A Shortest Path Dependency Kernel for Relation Extraction. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 724–731, Stroudsburg, PA, USA. Association for Computational Linguistics. [41](#)
- Bunescu, R. C. and Mooney, R. J. (2007). Learning to Extract Relations from the Web using Minimal Supervision. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, Czech Republic. [103](#), [137](#)
- Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2 :121–167. 10.1023/A :1009715923555. [20](#)
- Burges, C. J. C., Ragno, R., and Le, Q. V. (2006). Learning to Rank with Nonsmooth Cost Functions. In Schölkopf, B., Platt, J. C., Hoffman, T., Schölkopf,

REFERENCES

- B., Platt, J. C., and Hoffman, T., editors, *NIPS*, pages 193–200. MIT Press. [88](#)
- Burstein, F. and Gregor, S. (1999). The Systems Development or Engineering Approach to Research in Information Systems : An Action Research Perspective. *System*, pages 122–134. [145](#), [150](#)
- Byrne, L. and Dunnion, J. (2010). UCD IIRG at TAC 2010 KBP Slot Filling Task. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA. [130](#), [136](#)
- Califf, M. E. and Mooney, R. J. (1997). Relational Learning of Pattern-Match Rules for Information Extraction. In *Proceedings of the ACL Workshop on Natural Language Learning*, pages 9–15, Madrid, Spain. [26](#)
- Cardie, C. and Wagstaff, K. (1999). Noun Phrase Coreference as Clustering. In *Empirical Methods in Natural Language Processing*. [32](#)
- Chada, D., Aranha, C., and Monte, C. (2010). An Analysis of The Cortex Method at TAC 2010 KBP Slot-Filling. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA. [129](#), [134](#), [138](#)
- Chambers, N. and Jurafsky, D. (2011). Template-Based Information Extraction Without the Templates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, pages 976–986, Stroudsburg, PA, USA. Association for Computational Linguistics. [76](#)
- Chan, Y. S. and Roth, D. (2010). Exploiting Background Knowledge for Relation Extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 152–160, Stroudsburg, PA, USA. Association for Computational Linguistics. [35](#)
- Chang, C., Kaye, M., Girgis, M., and Shaalan, K. (2006). A Survey of Web Information Extraction Systems. *IEEE transactions on knowledge and data engineering*, pages 1411–1428. [8](#)

REFERENCES

- Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA. [19](#)
- Charolles, M. (1995). Cohésion, Cohérence et Pertinence du Discours. *Travaux de Linguistique*, (29) :125–151. [55](#)
- Chen, Z. and Ji, H. (2009). Graph-based event coreference resolution. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-4, pages 54–57, Stroudsburg, PA, USA. Association for Computational Linguistics. [81](#)
- Chen, Z. and Ji, H. (2010). Graph-Based Clustering for Computational Linguistics : A Survey. In *Proceedings of TextGraphs-5 - 2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 1–9, Uppsala, Sweden. Association for Computational Linguistics. [82](#)
- Chen, Z., Tamang, S., Lee, A., Li, X., Passantino, M., and Ji, H. (2010a). Top-down and Bottom-up : A Combined Approach to Slot Filling. In *6th Asia Information Retrieval Symposium on Information Retrieval Technology*, Gaithersburg, Maryland, USA. Springer-Verlag. [124](#), [138](#)
- Chen, Z., Tamang, S., Lee, A., Li, X., Snover, M., Passantino, M., Lin, W.-P., and Ji, H. (2010b). CUNY-BLENDER TAC-KBP2010 Slot Filling System Description. In *Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA. [134](#), [136](#)
- Chieu, H. and Ng, H. (2002). A maximum entropy approach to information extraction from semi-structured and free text. In *PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 786–791. Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999. [42](#)
- Chieu, H. L., Ng, H. T., and Lee, Y. K. (2003). Closing the Gap : Learning-Based Information Extraction Rivaling Knowledge-Engineering Methods. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 216–223, Stroudsburg, PA, USA. Association for Computational Linguistics. [43](#), [50](#), [76](#)

REFERENCES

- Cicurel, F. (1993). Pré-visibilité des Discours Journalistiques : À Propos d'un Événement-Catastrophe. Les Carnets du Cediscor [En ligne], mis en ligne le 28 août 2009, consulté le 03 août 2011. URL : <http://cediscor.revues.org/603>. [51](#), [56](#), [58](#)
- Ciravegna, F. (2001). (LP)², an Adaptive Algorithm for Information Extraction from Web-related Texts. In *International Joint Conference on Artificial Intelligence*. [26](#)
- Connolly, D., Burger, J. D., and Day, D. S. (1994). A Machine Learning Approach to Anaphoric Reference. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*. ACL. [32](#)
- Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20 :273–297. [10.1007/BF00994018](#). [20](#)
- Cowie, J. and Lehnert, W. (1996). Information Extraction. *Special natural language processing issue of the communications of the ACM*, 39 :80–91. [3](#), [8](#)
- Crowe, J. (1995). Constraint-Based Event Recognition for Information Extraction. In *In ACL 33*, pages 296–298. [50](#), [53](#)
- Cunningham, H. (1997). Information Extraction - A User Guide. Research memorandum, Department of Computer Science, University of Sheffield. [11](#)
- Cunningham, H. (2005). Information Extraction, Automatic. *Encyclopedia of Language and Linguistics, 2nd Edition*. [3](#), [8](#), [10](#), [12](#)
- Cunningham, H., Cunningham, H., Maynard, D., Maynard, D., Tablan, V., and Tablan, V. (2000). JAPE : a Java Annotation Patterns Engine. Technical report, Department of Computer Science, University of Sheffield. [17](#)
- Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*. [26](#)

REFERENCES

- de Pablo-Sánchez, C., Perea, J., Segura-Bedmar, I., and Martínez, P. (2009). The UC3M Team at the Knowledge Base Population Task. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA. [136](#)
- Dongen, S. (2000). A Cluster Algorithm for Graphs. Technical report, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, The Netherlands, The Netherlands. [22](#), [111](#)
- Dorow, B. and Widdows, D. (2003). Discovering Corpus-Specific Word Senses. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 2*, EACL '03, pages 79–82, Stroudsburg, PA, USA. Association for Computational Linguistics. [81](#)
- Ehrmann, M. (2008). *Les Entités Nommées de la Linguistique au TAL : Statut Théorique et Méthodes de Désambiguïsation*. PhD thesis, Université Paris VII - Denis Diderot, LaTTICe Langues, Textes, Traitement Informatique, Cognition. [22](#)
- Embarek, M. and Ferret, O. (2008). Learning Patterns for Building Resources about Semantic Relations in the Medical Domain. In *6th Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. [108](#), [139](#)
- Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S., and Yates, A. (2004). Web-scale information extraction in knowitall : (preliminary results). In *Proceedings of the 13th international conference on World Wide Web, WWW '04*, pages 100–110, New York, NY, USA. ACM. [104](#)
- Feng, D., Burns, G., and Hovy, E. H. (2007). Extracting Data Records from Unstructured Biomedical Full Text. In *EMNLP-CoNLL'07*, pages 837–846. [77](#)
- Filatova, E., Hatzivassiloglou, V., and McKeown, K. (2006). Automatic Creation of Domain Templates. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 207–214, Stroudsburg, PA, USA. Association for Computational Linguistics. [76](#)

REFERENCES

- Freedman, M., Ramshaw, L., Boschee, E., Gabbard, R., Kratkiewicz, G., Ward, N., and Weischedel, R. (2011). Extreme Extraction – Machine Reading in a Week. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1446, Edinburgh, Scotland, UK. Association for Computational Linguistics. [153](#)
- Fundel, K., Küffner, R., and Zimmer, R. (2007). RelEx—Relation Extraction Using Dependency Parse Trees. *Bioinformatics*, 23 :365–371. [36](#), [41](#)
- Gionis, A., Indyk, P., and Motwani, R. (1999). Similarity Search in High Dimensions via Hashing. In *25th International Conference on Very Large Data Bases (VLDB’99)*, pages 518–529, Edinburgh, Scotland, UK. [110](#)
- Gonzàlez, E. and Turmo, J. (2009). Unsupervised Relation Extraction by Massive Clustering. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, ICDM ’09*, pages 782–787, Washington, DC, USA. IEEE Computer Society. [36](#)
- Grishman, R. (1997). Information Extraction : Techniques and Challenges. In *International Summer School on Information Extraction : A Multidisciplinary Approach to an Emerging Information Technology*, pages 10–27, London, UK. Springer-Verlag. [8](#), [14](#), [15](#)
- Gu, Z. and Cercone, N. (2006). Segment-Based Hidden Markov Models for Information Extraction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44*, pages 481–488, Stroudsburg, PA, USA. Association for Computational Linguistics. [4](#), [53](#), [85](#)
- GuoDong, Z., Jian, S., Jie, Z., and Min, Z. (2005). Exploring Various Knowledge in Relation Extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL ’05*, pages 427–434, Stroudsburg, PA, USA. Association for Computational Linguistics. [39](#)
- Hasegawa, T., Sekine, S., and Grishman, R. (2004). Discovering Relations among Named Entities from Large Corpora. In *Proceedings of the 42nd Annual Meeting*

REFERENCES

- ting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics. [36](#)
- Hassan, H., Hassan, A., and Emam, O. (2006). Unsupervised Information Extraction Approach Using Graph Mutual Reinforcement. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 501–508, Stroudsburg, PA, USA. Association for Computational Linguistics. [36](#)
- Hearst, M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *14th International Conference on Computational linguistics (COLING'92)*, pages 539–545, Nantes, France. Association for Computational Linguistics. [37](#)
- Herbrich, R., Graepel, T., and Obermayer, K. (2000). Large Margin Rank Boundaries for Ordinal Regression. In Bartlett, P. J., Schölkopf, B., Schuurmans, D., and Smola, A. J., editors, *Advances in Large Margin Classifiers*, pages 115–132. MIT Press. [88](#)
- Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M. (2008). Identifying Sections in Scientific Abstracts using Conditional Random Fields. In *Third International Joint Conference on Natural Language Processing (IJCNLP 2008)*, pages 381–388, Hyderabad, India. [59](#)
- Ho-Dac, L.-M. and Péry-Woodley, M.-P. (2008). Temporal Adverbials and Discourse Segmentation Revisited. In *Multidisciplinary Approaches to Discourse 2008 (MAD 08)*, pages 65–77, Oslo. W.Ramm & C. Fabricius-Hansen (eds.). [55](#)
- Hobbs, J. and Riloff, E. (2010). Information Extraction. *Handbook of Natural Language Processing*. [15](#), [16](#)
- Hobbs, J. R. (1978). Resolving Pronoun References. *Lingua*, 44(4) :311 – 338. [31](#)
- Hobbs, J. R. (1993). The Generic Information Extraction System. In *Proceedings of the 5th conference on Message understanding*, MUC5 '93, pages 87–91, Stroudsburg, PA, USA. Association for Computational Linguistics. [4](#), [14](#), [15](#)

REFERENCES

- Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. (2011). Knowledge-Based Weak Supervision for Information Extraction of Overlapping Relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics. [137](#)
- Hoffmann, R., Zhang, C., and Weld, D. S. (2010). Learning 5000 Relational Extractors. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 286–295, Stroudsburg, PA, USA. Association for Computational Linguistics. [141](#)
- Humphreys, K., Gaizauskas, R., and Azzam, S. (1997). Event Coreference for Information Extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts, ANARESOLUTION '97*, pages 75–81, Stroudsburg, PA, USA. Association for Computational Linguistics. [48](#)
- Ji, H., Grishman, R., and Trang Dang, H. (2010). Overview of the TAC 2010 Knowledge Base Population Track. In *Third Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA. [129](#)
- Jiang, J. and Zhai, C. (2007). A Systematic Exploration of the Feature Space for Relation Extraction. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 113–120. Association for Computational Linguistics. [40](#)
- Kambhatla, N. (2004). Combining Lexical, Syntactic, and Semantic Features with Maximum Entropy Models for Extracting Relations. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, ACLdemo '04*, Stroudsburg, PA, USA. Association for Computational Linguistics. [39](#), [40](#)
- Kennedy, C. and Boguraev, B. (1996). Anaphora for Everyone : Pronominal Anaphora Resolution Without a Parser. In *Proceedings of the 16th conference on Computational linguistics - Volume 1, COLING '96*, pages 113–118, Stroudsburg, PA, USA. Association for Computational Linguistics. [32](#)

REFERENCES

- Khuller, S. (1997). *Approximation Algorithms for Finding Highly Connected Subgraphs*, pages 236–265. PWS Publishing Co., Boston, MA, USA. [22](#)
- Kitani, T., Eriguchi, Y., and Hara, M. (1994). Pattern Matching and Discourse Processing in Information Extraction from Japanese Text. *J. Artif. Int. Res.*, 2 :89–110. [50](#), [51](#), [56](#)
- Klinger, R. and Tomanek, K. (2007). Classical Probabilistic Models and Conditional Random Fields. Technical Report TR07-2-013, Department of Computer Science, Dortmund University of Technology. [19](#), [61](#)
- Kotsiantis, S., Zaharakis, I., and Pintelas, P. (2007). Supervised Machine Learning : A Review of Classification Techniques. *Emerging artificial intelligence applications in computer engineering : real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*, 160 :3. [19](#)
- Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML '01 : Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. [20](#), [28](#), [59](#)
- Laporte, E., Nakamura, T., and Voyatzi, S. (2008). A French Corpus Annotated for Multiword Expressions with Adverbial Function. In *6th Conference on Language Resources and Evaluation (LREC'08)*, pages 48–51, Marrakech, Maroc. [61](#)
- Lappin, S. and Leass, H. J. (1994). An Algorithm for Pronominal Anaphora Resolution. *Comput. Linguist.*, 20 :535–561. [32](#)
- Lehmann, J., Monahan, S., Nezda, L., Jung, A., and Shi, Y. (2010). LCC Approaches to Knowledge Base Population at TAC 2010. In *Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA. [134](#)
- Li, F., Zheng, Z., Bu, F., Tang, Y., Zhu, X., and Huang, M. (2009a). THU QUANTA at TAC 2009 KBP and RTE Track. In *Second Text Analysis Confe-*

REFERENCES

- rence (*TAC 2009*), Gaithersburg, Maryland, USA. [133](#), [134](#), [135](#), [136](#), [137](#), [138](#)
- Li, P., Jiang, J., and Wang, Y. (2010). Generating Templates of Entity Summaries with an Entity-Aspect Model and Pattern Mining. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 640–649, Stroudsburg, PA, USA. Association for Computational Linguistics. [76](#)
- Li, S., Gao, S., Zhang, Z., Li, X., Guan, J., Xu, W., and Guo, J. (2009b). PRIS at TAC 2009 : Experiments in KBP Track. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA. [136](#)
- Liu, T. (2011). *Learning to Rank for Information Retrieval*. Springer. [87](#)
- Liu, T.-Y. (2009). Learning to Rank for Information Retrieval. *Found. Trends Inf. Retr.*, 3 :225–331. [87](#)
- Liu, Y., Shi, Z., and Sarkar, A. (2007). Exploiting Rich Syntactic Information for Relation Extraction from Biomedical Articles. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers*, NAACL-Short '07, pages 97–100, Stroudsburg, PA, USA. Association for Computational Linguistics. [34](#), [78](#), [79](#), [80](#), [85](#), [86](#)
- Lucas, N. (2004). La rhétorique des dépêches de presse à travers les marques énonciatives du temps, du lieu et de la personne. In *Actes de la Semaine du Document Numérique (SDN 2004)*, Journée ATALA. [44](#), [51](#)
- MacDonald, C. (2009). *The Voting Model for People Search*. PhD thesis, Department of Computing Science, University of Glasgow. [87](#)
- Malouf, R. (2002). Markov Models for Language-Independent Named Entity Recognition. In *proceedings of the 6th conference on Natural language learning - Volume 20*, COLING-02, pages 1–4, Stroudsburg, PA, USA. Association for Computational Linguistics. [28](#)

REFERENCES

- Mansuri, I. R. and Sarawagi, S. (2006). Integrating Unstructured Data into Relational Databases. In *Proceedings of the 22nd International Conference on Data Engineering, ICDE '06*, pages 29–, Washington, DC, USA. IEEE Computer Society. [77](#)
- McCallum, A. (2005). Information Extraction : Distilling Structured Data from Unstructured Text. *Queue*, 3 :48–57. [8](#), [10](#), [15](#)
- McCallum, A., Freitag, D., and Pereira, F. C. N. (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML '00*, pages 591–598, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. [28](#)
- McDonald, R., Pereira, F., Kulick, S., Winters, S., Jin, Y., and White, P. (2005). Simple Algorithms for Complex Relation Extraction with Applications to Bio-medical IE. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pages 491–498, Stroudsburg, PA, USA. Association for Computational Linguistics. [13](#), [34](#), [76](#), [78](#), [79](#), [80](#), [148](#)
- McNamee, P., Dredze, M., Gerber, A., Garera, N., Finin, T., Mayfield, J., Piatko, C., Rao, D., Yarowsky, D., and Dreyer, M. (2009). HLTCOE Approaches to Knowledge Base Population at TAC 2009. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA. [136](#)
- Mihalcea, R. (2004). Graph-Based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions, ACLdemo '04*, Stroudsburg, PA, USA. Association for Computational Linguistics. [81](#), [94](#)
- Mihalcea, R. and Radev, D. (2006). Graph-based algorithms for natural language processing and information retrieval. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Tutorial Abstracts, NAACL-Tutorials '06*, pages 303–304, Stroudsburg, PA, USA. Association for Computational Linguistics. [88](#)

REFERENCES

- Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics. [103](#), [136](#), [147](#)
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York. [19](#)
- Mitkov, R. (1998). Robust Pronoun Resolution With Limited Knowledge. In *ACL-36 : Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 869–875, Montreal, Quebec, Canada. Association for Computational Linguistics. [32](#)
- Mitkov, R. (1999). Anaphora Resolution : The State Of The Art. Technical report, School of Languages and European Studies. University of Wolverhampton. [28](#), [31](#)
- Mollá, D. (2006). Learning of Graph-Based Question Answering Rules. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, TextGraphs-1*, pages 37–44, Stroudsburg, PA, USA. Association for Computational Linguistics. [81](#)
- Moncecchi, G., Minel, J.-L., and Wonsever, D. (2010). A Survey of Kernel Methods for Relation Extraction. In *Workshop on NLP and Web-based technologies*, Bahía Blanca, Argentine. [41](#)
- Mooney, R. J. and Bunescu, R. (2005). Mining Knowledge From Text Using Information Extraction. *SIGKDD Explorations*, 7 :3–10. [8](#)
- Moschitti, A. (2006). Making Tree Kernels Practical for Natural Language Learning. In *EACL*. [41](#)
- Muslea, I. (1999). Extraction Patterns for Information Extraction Tasks : A Survey. In *In AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 1–6. [17](#)

REFERENCES

- Nadeau, D. and Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes*, 30(1) :3–26. Publisher : John Benjamins Publishing Company. [24](#), [28](#)
- Naughton, M. (2007). Exploiting Structure for Event Discovery Using the MDI Algorithm. In *45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 31–36, Prague. [53](#), [54](#)
- Ng, V. (2008). Unsupervised Models for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08*, pages 640–649, Stroudsburg, PA, USA. Association for Computational Linguistics. [32](#), [33](#)
- Ng, V. (2010). Supervised Noun Phrase Coreference Research : the First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10*, pages 1396–1411, Stroudsburg, PA, USA. Association for Computational Linguistics. [28](#), [31](#), [33](#)
- Ng, V. and Cardie, C. (2002). Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics. [32](#)
- Nguyen, T.-V. T. and Moschitti, A. (2011). End-to-End Relation Extraction Using Distant Supervision from External Semantic Repositories. In *ACL (Short Papers)'11*, pages 277–282. [137](#)
- Nicolae, C. and Nicolae, G. (2006). BESTCUT : A Graph Algorithm for Coreference Resolution. In *EMNLP'06*, pages 275–283. [81](#)
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking : Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab. [88](#)
- Pantel, P. and Pennacchiotti, M. (2006). Espresso : Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual*

REFERENCES

- meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120, Stroudsburg, PA, USA. Association for Computational Linguistics. [104](#)
- Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards Terascale Knowledge Acquisition. In *20th International Conference on Computational Linguistics (COLING'04)*, pages 771–777, Geneva, Switzerland. [107](#)
- Patwardhan, S. (2008). Combining Global Relevance Information with Local Contextual Clues for Event-Oriented Information Extraction. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 3*, pages 1863–1864. AAAI Press. [43](#)
- Patwardhan, S. and Riloff, E. (2007). Effective information extraction with semantic affinity patterns and relevant regions. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 717–727. [5](#), [43](#), [53](#), [54](#), [69](#)
- Patwardhan, S. and Riloff, E. (2009). A Unified Model of Phrasal and Sentential Evidence for Information Extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing : Volume 1*, EMNLP '09, pages 151–160, Stroudsburg, PA, USA. Association for Computational Linguistics. [43](#)
- Poesio, M., Paolo Ponzetto, S., and Versley, Y. (2010). Computational Models of Anaphora Resolution : A Survey. To be published; available at <http://clic.cimec.unitn.it/massimo/Publications/lilt.pdf>. [31](#)
- Pustejovsky, James, Knippen, Robert, Littman, Jessica, Sauri, and Roser (2005). Temporal and Event Information in Natural Language Text. *Computers and the Humanities*, 39(2-3) :123–164. [54](#)
- Quinlan, J. R. (1996). Learning Decision Tree Classifiers. *ACM Comput. Surv.*, 28 :71–72. [20](#)
- Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77 :257–286. [20](#), [58](#)

REFERENCES

- Ratnaparkhi, A. (1998). *Maximum Entropy Models for Natural Language Ambiguity Resolution*. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA. [61](#)
- Ravichandran, D. (2005). *Terascale Knowledge Acquisition*. PhD thesis, Faculty of the Graduate School University of Southern California, Los Angeles, CA, USA. [107](#)
- Reiss, F., Raghavan, S., Krishnamurthy, R., Zhu, H., and Vaithyanathan, S. (2008). An Algebraic Approach to Rule-Based Information Extraction. In *International Conference on Data Engineering*, pages 933–942. [26](#)
- Riedel, S., Yao, L., and McCallum, A. (2010). Modeling Relations and Their Mentions without Labeled Text. In Balcázar, J., Bonchi, F., Gionis, A., and Sebag, M., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 6323 of *Lecture Notes in Computer Science*, pages 148–163. Springer Berlin / Heidelberg. [137](#)
- Riloff, E. (1993). Automatically Constructing a Dictionary for Information Extraction Tasks. In *In Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 811–816. MIT Press. [4](#), [17](#), [26](#), [43](#)
- Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI-01 workshop on "Empirical Methods in AI"*. [20](#)
- Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2007). Automatising the Learning of Lexical Patterns : An Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia. *Data Knowledge Engineering*, 61 :484–499. [107](#)
- Santos, R. L. T., Macdonald, C., and Ounis, I. (2010). Voting for Related Entities. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 1–8, Paris, France, France. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. [87](#)
- Sarawagi, S. (2008). Information Extraction. *Foundations and Trends in Databases*, 1 :261–377. [8](#), [9](#), [16](#), [17](#), [24](#), [26](#), [28](#)

REFERENCES

- Schaeffer, S. (2007). Graph Clustering. *Computer Science Review*, 1(1) :27–64. [82](#)
- Schlaefter, N., Giesermann, P., Schaaf, T., and Waibel, A. (2006). A Pattern Learning Approach to Question Answering Within the Ephyra Framework. In Sojka, P., Kopecek, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 4188 of *Lecture Notes in Computer Science*, pages 687–694. Springer Berlin / Heidelberg. [107](#)
- Schone, P., Goldschen, A., Langley, C., Lewis, S., Onyshkevych, B., Cutts, R., Dawson, B., MacBride, J., Matrangola, G., McDonough, C., Pfeifer, C., and Ursiak, M. (2009). TCAR at TAC-KBP 2009. In *Second Text Analysis Conference (TAC 2009)*, Gaithersburg, Maryland, USA. [136](#)
- Sculley, D. (2010). Combined Regression and Ranking. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 979–988, New York, NY, USA. ACM. [88](#)
- Seymore, K., McCallum, A., and Rosenfeld, R. (1999). Learning Hidden Markov Model Structure for Information Extraction. In *AAAI-99 Workshop on Machine Learning for Information Extraction*, pages 37–42. [28](#)
- Shinyama, Y. and Sekine, S. (2006). Preemptive Information Extraction using Unrestricted Relation Discovery. In *HLT-NAACL 2006*, pages 304–311, New York City, USA. [36](#), [103](#)
- Simões, G. F., Galhardas, H., and Coheur, L. (2009). Information Extraction Tasks : a Survey (short paper). *Inforum*. [8](#)
- Soderland, S. (1999). Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning*, 34 :233–272. 10.1023/A :1007562322031. [17](#)
- Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine Learning Approach to Coreference Resolution of Noun Phrases. *Comput. Linguist.*, 27 :521–544. [32](#)
- Stevenson, M. (2006). Fact Distribution in Information Extraction. *Language resources and evaluation*, 40(2) :183–201. [45](#)

REFERENCES

- Suchanek, F. M. (2009). *Automated Construction and Growth of a Large Ontology*. PhD thesis, Saarland University. [104](#)
- Suchanek, F. M., Ifrim, G., and Weikum, G. (2006). Combining Linguistic and Statistical Analysis to Extract Relations from Web Documents. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 712–717, New York, NY, USA. ACM. [103](#)
- Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago : a Core of Semantic Knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM. [104](#), [137](#)
- Surdeanu, M., McClosky, D., Smith, M. R., Gusev, A., and Manning, C. D. (2011). Customizing an Information Extraction System to a New Domain. In *Proceedings of the Workshop on Relational Models of Semantics*. [153](#)
- Surdeanu, M., McClosky, D., Tibshirani, J., Bauer, J., Chang, A., Spitzkovsky, V., and Manning, C. (2010). A Simple Distant Supervision Approach for the TAC-KBP Slot Filling Task. In *Text Analysis Conference (TAC 2010)*, Gaithersburg, Maryland, USA. [136](#), [137](#)
- Sutton, C. and McCallum, A. (2007). An Introduction to Conditional Random Fields for Relational Learning. In Getoor, L. and Taskar, B., editors, *Introduction to statistical relational learning*, pages 93–129. MIT Press. [20](#)
- TAC-KBP (2009). Task Description for Knowledge-Base Population at TAC 2009. <http://apl.jhu.edu/~paulmac/kbp/090601-KBPTaskGuidelines.pdf>. [115](#)
- TAC-KBP (2010). Preliminary Task Description for Knowledge-Base Population at TAC 2010. http://nlp.cs.qc.cuny.edu/kbp/2010/KBP2010_TaskDefinition.pdf. [115](#), [120](#), [124](#)
- TAC-KBP (2011). Proposed Task Description for Knowledge-Base Population at TAC 2011. http://nlp.cs.qc.cuny.edu/kbp/2011/KBP2011_TaskDefinition.pdf. [115](#)

REFERENCES

- Tikk, D., Thomas, P., Palaga, P., Hakenberg, J., and Leser, U. (2010). A Comprehensive Benchmark of Kernel Methods to Extract Protein–Protein Interactions from Literature. *PLoS Comput Biol*, 6(7). [41](#)
- Turmo, J., Ageno, A., and Català, N. (2006). Adaptive Information Extraction. *ACM Comput. Surv.*, 38. [8](#), [12](#), [15](#)
- Uren, V., Cimiano, P., Iria, J., Hanndschuh, S., Vargas-Vera, M., Motta, E., and Ciravegna, F. (2006). Semantic Annotation for Knowledge Management : Requirements and a Survey of the State of the Art. *Journal of Web Semantics*, 4(1) :14–28. [8](#)
- Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1 :1–305. [19](#)
- Wang, W., Besançon, R., Ferret, O., and Grau, B. (2011a). Filtering and clustering relations for unsupervised information extraction in open domain. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, Scotland, UK, 24th-28th October*. [103](#)
- Wang, W., Besançon, R., Ferret, O., and Grau, B. (2011b). Filtrage de Relations pour l’Extraction d’Information Non Supervisée. In *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2011, article court)*, Montpellier, France. [114](#), [137](#)
- Wick, M., Culotta, A., and McCallum, A. (2006). Learning Field Compatibilities to Extract Database Records from Unstructured Text. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP ’06*, pages 603–611, Stroudsburg, PA, USA. Association for Computational Linguistics. [77](#), [78](#), [83](#), [85](#)
- Xu, R. and Wunsch, D., I. (2005). Survey of Clustering Algorithms. *Neural Networks, IEEE Transactions on*, 16(3) :645 –678. [21](#), [22](#)
- Xu, S. (2011). Discovering and Tracking Events From News, Blogs and Microblogs on the Web. In *Doctoral Colloquium, COSIT 2011*. [55](#)

REFERENCES

- Yamron, J. P., Carp, I., Gillick, L., Lowe, S., and Van Mulbregt, P. (1998). A Hidden Markov Model Approach to Text Segmentation and Event Tracking. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 333–336. [58](#)
- Yan, Y., Matsuo, Y., and Ishizuka, M. (2009). An Integrated Approach for Relation Extraction from Wikipedia Texts. In *WWW 2009 workshop on Content Analysis in the WEB 2.0 (CAW2.0)*, Madrid, Spain. [103](#)
- Yang, X., Zhou, G., Su, J., and Tan, C. L. (2003). Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 176–183, Stroudsburg, PA, USA. Association for Computational Linguistics. [32](#)
- Yates, A., Cafarella, M., Banko, M., Etzioni, O., Broadhead, M., and Soderland, S. (2007). TextRunner : Open Information Extraction on the Web. In *Proceedings of Human Language Technologies : The Annual Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations*, NAACL-Demonstrations '07, pages 25–26, Stroudsburg, PA, USA. Association for Computational Linguistics. [104](#)
- Zelenko, D., Aone, C., and Richardella, A. (2003). Kernel Methods for Relation Extraction. *J. Mach. Learn. Res.*, 3 :1083–1106. [39](#), [41](#)
- Zhu, X. (2005). Semi-Supervised Learning Literature Survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison. [19](#), [21](#)