Extraction et regroupement de relations entre entités pour l'extraction d'information non supervisée

Wei Wang* — **Romaric Besançon*** — **Olivier Ferret*** — **Brigitte Grau****

* *CEA, LIST, Laboratoire Vision et Ingénierie des Contenus*
91191 Gif-sur-Yvette Cedex, France
wei.wang@lip6.fr,{romaric.besancon,olivier.ferret}@cea.fr
** *LIMSI, UPR-3251 CNRS-DR4, Bât. 508, BP 133, 91403 Orsay Cedex*
brigitte.grau@limsi.fr

RÉSUMÉ. *Cet article se situe dans le cadre de l'extraction d'information non supervisée en domaine ouvert en se concentrant sur l'extraction et le regroupement à large échelle de relations entre entités nommées de type non défini a priori. L'étape d'extraction combine l'utilisation de critères simples mais efficaces et une procédure de filtrage à base d'apprentissage. L'étape de regroupement organise quant à elle les relations extraites pour en caractériser le type selon une stratégie multiniveau permettant de prendre en compte à la fois un volume important et des critères de regroupement élaborés. Les évaluations menées montrent que cette approche a la capacité d'extraire des relations avec une bonne précision et de les grouper selon leurs similarités sémantique et thématique.*

ABSTRACT. *This article takes place in the context of unsupervised information extraction in open domain and focuses on the extraction and the clustering at a large scale of relations between named entities without defining their type a priori. The extraction step combines the use of basic but efficient criteria and a filtering procedure based on machine learning. The clustering step organizes extracted relations into clusters to characterize their type according to a multi-level strategy that takes into account both large volumes of relations and sophisticated clustering criteria. Experiments show that our approach is able to extract relations with a good precision and to organize them according to their semantic and topical similarity.*

MOTS-CLÉS : *extraction d'information non supervisée, extraction de relations entre entités nommées,* clustering *de relations.*

KEYWORDS: *unsupervised information extraction, relation extraction, relation clustering.*

1. Introduction

Le domaine de l'extraction d'information (EI) s'est longtemps inscrit dans le paradigme établi par les conférences d'évaluation MUC (*Message Understanding Conference*) et poursuivi par des campagnes telles que ACE (*Automatic Content Extraction*). Les tâches définies par ces campagnes concernent l'extraction d'information supervisée, pour laquelle le type d'information à extraire est prédéfini et des instances sont annotées dans des corpus représentatifs. À partir de ces données, des systèmes conçus manuellement ou par apprentissage automatique peuvent être développés. Des approches semi-supervisées ont été définies plus récemment pour s'affranchir partiellement des contraintes de disponibilité de telles données. Par exemple, dans le cadre de la tâche KBP (*Knowledge Base Population*) de la campagne TAC (*Text Analysis Conference*), l'extraction de relations s'appuie sur une base de connaissances existante (construite à partir des infoboxes de Wikipédia), mais sans données annotées. Dans ce cas, des techniques de supervision distante (Mintz *et al.*, 2009) peuvent être appliquées. Ces méthodes semi-supervisées incluent également des techniques d'amorçage (*bootstrapping*) (Grishman et Min, 2010) permettant de s'appuyer sur un nombre limité d'exemples pour en extraire d'autres, comme par exemple dans (Brin, 1998) pour extraire une relation entre un livre et son auteur.

L'extraction d'information non supervisée diffère de ces tâches en ouvrant la problématique de l'extraction de relations à des relations de type inconnu *a priori*, ce qui permet de faire face à l'hétérogénéité des relations rencontrées en domaine ouvert, notamment sur le Web. Le type de ces relations doit alors être découvert de façon automatique à partir des textes. Dans ce cadre, les structures d'information considérées sont fréquemment des relations binaires intervenant entre des entités nommées, à l'instar de Hasegawa *et al.* (2004). Ce travail, parmi les premiers sur cette problématique, a avancé l'hypothèse que les relations les plus intéressantes entre entités nommées sont aussi les plus fréquentes dans une collection de textes, de sorte que les instances de relations susceptibles de former des clusters de grande taille peuvent être distinguées des autres. Pour opérer cette distinction, un seuil de similarité minimale appliqué à une représentation des relations de type sac de mots a été établi pour défavoriser les clusters de petite taille. Des améliorations ont par la suite été apportées à cette approche initiale par l'adoption de patrons comme éléments de représentation des relations pour former des clusters (Shinyama et Sekine, 2006) ou l'usage d'un algorithme d'ordonnancement de ces patrons pour la sélection de relations candidates (Chen *et al.*, 2005).

Une part notable des travaux menés en EI non supervisée se focalisent sur l'extraction des relations à partir de phrases, avec des approches variées. Des systèmes tels que TEXTRUNNER (Banko *et al.*, 2007) ou WOE (Wu et Weld, 2010) reposent ainsi sur des modèles d'apprentissage statistique supervisés tandis que les systèmes REVERB (Fader *et al.*, 2011) ou OLLIE (Mausam *et al.*, 2012) exploitent des contraintes ou des patrons lexico-syntaxiques. Des approches à base de règles (Akbik et Broß, 2009 ; Gamallo *et al.*, 2012) ou des modèles génératifs (Rink et Harabagiu, 2011 ; Yao *et al.*, 2011) ont également été proposés pour ce faire. Tout en res-

tant pour l'essentiel non supervisées, d'autres approches font appel à un utilisateur pour délimiter un domaine d'extraction de façon peu contrainte. Ainsi, le système *On-Demand Information Extraction* (Sekine, 2006) initie le processus d'extraction par des requêtes de moteur de recherche. Le problème du regroupement des relations a été en revanche moins abordé, en particulier pour rassembler des relations équivalentes mais exprimées de façon différente. En dehors des premiers travaux dans ce domaine (Hasegawa *et al.*, 2004 ; Rozenfeld et Feldman, 2006), qui se limitaient à l'application du modèle vectoriel saltonien aux relations, beaucoup des travaux sur cette question (Kok et Domingos, 2008 ; Yao *et al.*, 2011 ; Min *et al.*, 2012) se sont concentrés sur la problématique du *co-clustering* des relations et de leurs arguments.

Le travail que nous présentons dans cet article s'inscrit dans un contexte applicatif plus large visant à répondre à des problématiques de veille telle que « suivre tous les événements faisant intervenir les sociétés X et Y ». Dans un tel contexte, les sources d'information et les entités intéressantes sont connues *a priori* mais la nature des relations entretenues par ces entités est assez ouverte et constitue précisément l'objet de la recherche. Un moteur de recherche se focalisant sur les relations intervenant entre les entités du domaine considéré au sein d'un ensemble de documents collectés constitue ainsi un outil de fouille intéressant pour les analystes. Un tel moteur doit, en amont de son processus d'indexation, s'appuyer sur un processus d'EI non supervisée afin de mettre au jour les relations au sein des documents et les regrouper pour en faciliter l'appréhension. Ce processus peut s'appuyer, dans notre cas, sur le fait que les relations font essentiellement intervenir des entités nommées. L'intérêt de former des classes d'arguments pour faciliter le rapprochement des relations est donc beaucoup plus faible que pour certains travaux évoqués précédemment car les types d'entités représentent déjà de telles classes. Pour aller plus avant dans le rapprochement des relations, il est alors nécessaire de se focaliser sur la prise en compte de la variabilité de leur expression même, et donc de leur similarité à un niveau plus sémantique. Le défi dans ce contexte est de conjuguer la mise en œuvre d'une telle similarité et son application pour le regroupement d'un grand nombre de relations. Pour ce faire, nous présentons dans cet article une méthode efficace pour à la fois extraire et regrouper des relations entre entités nommées à une large échelle.

L'étape d'extraction se fonde sur l'identification de couples d'entités nommées co-occurrentes à un niveau phrastique, combinée à une procédure de filtrage pour éliminer les fausses relations (Wang *et al.*, 2011). L'étape de regroupement s'appuie, quant à elle, sur une approche nouvelle s'articulant autour de deux niveaux de regroupement des relations. Le premier niveau s'effectue sur la forme, en utilisant une mesure de similarité et un algorithme de *clustering* efficaces sur le plan calculatoire, permettant ainsi une application à une large échelle. Le second niveau repose sur une mesure de similarité sémantique plus élaborée, donc plus exigeante en termes de temps de traitement mais compatible avec le nombre plus restreint, par rapport aux relations initiales, des clusters de premier niveau auxquels elle est appliquée (Wang *et al.*, 2013). Outre son intérêt pour le passage à l'échelle, cette approche à deux niveaux permet, comme nous l'illustrerons dans ce qui suit, d'obtenir de meilleurs résultats en mettant d'abord en évidence les éléments discriminants des relations pour ensuite prendre en compte

la variabilité d'expression de ces éléments. Nous montrons également que la similarité sémantique intégrant cette variabilité peut exploiter avec avantage un thésaurus distributionnel construit automatiquement à partir d'un corpus, en comparaison avec l'utilisation de réseaux lexicaux construits manuellement comme WordNet. Finalement, ces deux niveaux se complètent d'un regroupement réalisé suivant un axe thématique, ajoutant ainsi une dimension de regroupement nouvelle et intéressante du point de vue applicatif par rapport aux travaux existants.

Dans la suite de cet article, nous introduisons d'abord la notion de relation sous-tendant notre travail à la section 2 puis nous donnons une vue d'ensemble de l'approche proposée à la section 3. Les sections 4 et 5 détaillent respectivement les méthodes d'extraction et de regroupement des relations. Enfin, les sections 6 et 7 rendent compte de l'évaluation de cette méthode de regroupement sous plusieurs angles et la mettent en perspective.

2. Notion de relation

À la base du processus d'EI non supervisée considéré ici se trouve une notion de relation reprenant pour l'essentiel les hypothèses des travaux mentionnés ci-dessus : une relation est définie par la cooccurrence de deux entités nommées dans une phrase. Compte tenu du caractère non supervisé de la démarche, l'idée sous-jacente à ces restrictions est de se focaliser en premier lieu sur des cas simples, autrement dit des relations s'appuyant sur des arguments facilement identifiables dans un espace textuel suffisamment limité pour rendre leur caractérisation synthétique et s'affranchir des problèmes de coréférence au niveau de leurs arguments.

Dans les systèmes d'EI non supervisée, les entités en relation peuvent être des entités nommées (Hasegawa *et al.*, 2004) ou, de façon plus ouverte, des syntagmes nominaux (Rozenfeld et Feldman, 2006). Les entités nommées permettent en général d'avoir une meilleure séparation des différents types de relations alors que l'utilisation de syntagmes nominaux permet d'avoir un plus grand nombre de candidats. Nous nous intéressons dans notre cas aux relations entre entités nommées, à la fois pour faciliter l'organisation des relations trouvées et parce qu'il s'agit du besoin le plus généralement répandu en contexte applicatif de veille.

Plus formellement, comme illustré par la figure 1[1], les relations extraites des textes, que l'on devrait en toute rigueur appeler instances de relations, même si leur type n'est pas explicitement défini, sont caractérisées par trois grandes catégories d'information permettant tout à la fois de les définir et de fournir les éléments nécessaires à leur regroupement :

– un couple d'entités nommées (E1 et E2). Dans les expérimentations menées, nous nous sommes restreints aux entités de type personne (PERS), organisation (ORG) et lieu (LIEU) ;

1. L'exemple est donné en anglais car nos expérimentations ont été réalisées dans cette langue.

Figure 1. *Exemple de relation extraite*

– une caractérisation linguistique de la relation. Il s'agit de la façon dont la relation est exprimée linguistiquement. Chaque relation étant extraite sur la base de la présence dans une phrase d'un couple d'entités nommées correspondant aux types ci-dessus, sa caractérisation linguistique comporte trois parties :

- *Cpre* : la partie de la phrase précédant la première entité (E1),
- *Cmid* : la partie de la phrase se situant entre les deux entités,
- *Cpost* : la partie de phrase suivant la seconde entité (E2).

Le plus souvent *Cmid* exprime la relation proprement dite tandis que *Cpre* et *Cpost* fournissent plutôt des éléments de contexte pouvant être utiles dans la perspective de son regroupement avec d'autres relations. Dans notre cas, nous accentuons quelque peu cette tendance naturelle en faisant porter les contraintes d'extraction des relations exclusivement sur la partie *Cmid* comme nous le verrons dans ce qui suit. Cette focalisation conduit par ailleurs à favoriser très fortement une forme d'expression verbale des relations extraites, au détriment d'une forme nominale ;

– un contexte thématique : ce contexte est formé des mots pleins du segment de texte thématiquement homogène environnant l'instance de relation extraite.

On notera qu'une telle relation revêt une forme que l'on peut qualifier de semi-structurée dans la mesure où une partie de sa définition – le couple d'entités – renvoie à des éléments d'une ontologie prédéfinie tandis que son autre partie n'apparaît que sous une forme linguistique.

3. Vue d'ensemble du processus d'extraction d'information non supervisée

Le processus d'EI non supervisée défini autour de la notion de relation présentée à la section précédente se décompose, comme l'illustre la figure 2, autour de deux grandes étapes : l'extraction des relations et leur regroupement. Alors que ce processus est non supervisé dans sa finalité – le type des relations n'est pas connu *a priori* – l'étape d'extraction est centrée sur un modèle d'apprentissage supervisé tandis que celle du regroupement relève de l'apprentissage non supervisé.

Chacune de ces deux grandes étapes se définit elle-même autour de deux processus principaux selon le même principe : un premier processus adopte une approche simple

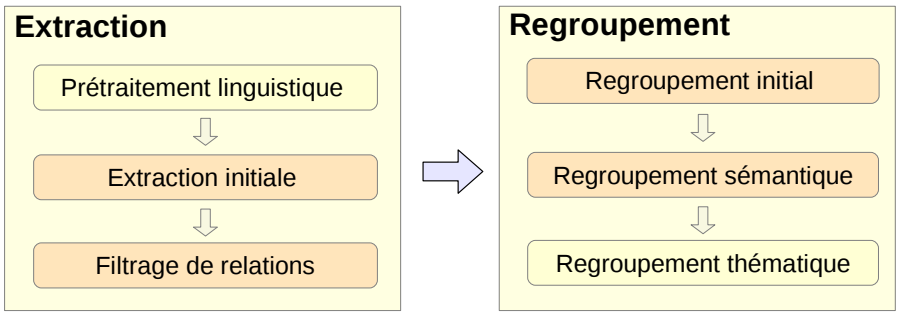


Figure 2. *Processus d'extraction d'information non supervisée*

mais efficace en termes calculatoires pour mettre en œuvre la fonctionnalité visée à une large échelle ; un second processus plus élaboré mais également plus coûteux est ensuite appliqué pour raffiner les résultats du premier. Le fait qu'une partie du travail ait déjà été effectuée par un premier processus permet de compenser la complexité calculatoire plus élevée du second. Concrètement, après le *prétraitement linguistique* des documents, l'étape d'extraction de relations se divise ainsi entre un processus d'*extraction initiale* de relations candidates et un processus de *filtrage* de ces relations. Le premier met en œuvre les critères minimaux de définition d'une relation et assure ainsi l'extraction rapide d'un ensemble de relations potentielles tandis que le second s'appuie sur des moyens plus élaborés pour faire le tri parmi ces relations candidates. De même, l'étape de regroupement repose sur deux processus principaux : un *regroupement initial* des relations extraites fondé sur une similarité de forme, qui peut de ce fait bénéficier de techniques d'optimisation intéressantes et permet de construire des regroupements très homogènes ; un *regroupement sémantique* des premiers clusters formés exploitant une mesure de similarité faisant appel à des ressources sémantiques pour réunir les formes d'expression diverses d'un même type de relation. Le processus de *regroupement thématique* vient compléter ce diptyque en ajoutant une dimension supplémentaire de raffinement des clusters sémantiques.

Dans ce qui suit, nous commencerons par décrire assez rapidement l'étape d'extraction des relations avant de présenter plus en détail l'étape de regroupement, qui constitue plus particulièrement le cœur de cet article.

4. Extraction et filtrage de relations

4.1. *Prétraitement linguistique des documents*

Le premier processus constituant l'étape d'extraction des relations est le prétraitement linguistique des documents constituant le corpus considéré. Ce prétraitement permet de mettre en évidence dans les textes les informations nécessaires à

la définition des relations. Ce prétraitement comporte donc une reconnaissance des entités nommées pour les types d'entités visés, une désambiguïsation morphosyntaxique des mots ainsi que leur normalisation. Ces traitements s'appuient sur les outils d'OpenNLP (`http://opennlp.apache.org`)[2]. En outre, chaque document fait l'objet d'une segmentation thématique linéaire de sorte que chaque instance de relation est associée à un segment thématique à partir duquel est construit son contexte thématique. Cette segmentation est réalisée par l'outil LCseg (Galley *et al.*, 2003).

4.2. *Extraction initiale des relations candidates*

Lors de l'extraction initiale des relations candidates, les contraintes sont très limitées. Sont ainsi extraites les relations correspondant à tout couple d'entités nommées dont les types correspondent aux types ciblés, avec pour seules restrictions la cooccurrence de ces entités dans une même phrase et la présence d'au moins un verbe entre les deux. Pour la sous-partie du corpus AQUAINT-2, constituée de dix-huit mois du journal *New York Times*, que nous avons utilisée pour toutes les expérimentations présentées dans cet article, cette stratégie conduit à extraire près de 929 404 relations couvrant les neuf types de relations possibles à partir des trois types d'entités nommées retenus (PERS, ORG et LIEU).

Un examen de ces relations candidates montre cependant qu'un nombre très significatif des relations ainsi extraites ne correspond pas à de véritables relations entre les entités impliquées. Il semble donc que cette stratégie basique d'extraction, qui peut donner des résultats intéressants dans des domaines de spécialité[3], ne soit pas suffisamment sélective en domaine ouvert. Nous avons donc cherché à la compléter par un processus de filtrage spécifique visant à déterminer si deux entités dans une phrase sont ou ne sont pas liées par une relation, sans *a priori* sur la nature de cette relation.

4.3. *Filtrage heuristique*

Dans une perspective exploratoire, nous avons défini un nombre restreint d'heuristiques de filtrage et analysé leur impact. Ces heuristiques sont au nombre de trois :

– la suppression des relations comportant entre leurs deux entités un verbe exprimant un discours rapporté (dans le cas présent, la liste se limite aux verbes *to say* et *to present*). Ceci vise à éviter d'extraire une relation entre les entités *Homgren* et *Allen* dans l'exemple suivant :

Holmgren said Allen was more involved with the team [...]

2. Bien qu'ayant travaillé avec des textes océrisés, Rodriquez *et al.* (2012) donnent quelques éléments d'évaluation comparative d'OpenNLP concernant la reconnaissance des trois types d'entités nommées considérés ici.

3. Le travail rapporté dans (Embarek et Ferret, 2008) montre que dans le domaine médical, les relations extraites sur la base de cette stratégie sont correctes dans 79 % des cas.

– nombre de mots entre les deux entités limité à dix. Au-delà de cette limite empirique, le nombre des relations effectives devient en effet très faible ;

– limitation à un du nombre de verbes entre les deux entités, sauf si ces verbes ont valeur d'auxiliaire (*be*, *have* et *do*).

L'application de ces heuristiques aux relations extraites a globalement pour conséquence de réduire leur volume d'environ 50 %. Une analyse plus détaillée montre que l'heuristique la plus filtrante est clairement celle de la distance entre entités mais celle limitant le nombre de verbes a également un impact très significatif.

Ce ratio de filtrage doit être mis en parallèle avec une évaluation de l'efficacité de ces heuristiques en termes de sélection des relations correctes. Pour ce faire, nous avons choisi au hasard cinquante instances pour chaque catégorie et nous avons procédé à une annotation manuelle de leur validité. Le tableau 1 donne le résultat de cette évaluation montrant que globalement, le taux de fausses relations parmi les relations filtrées est assez élevé pour toutes les catégories de relations mais que parmi les relations conservées, certaines catégories de relations, en particulier toutes les relations ayant un lieu comme première entité nommée, se caractérisent par un taux de fausses relations encore très important.

Catégories	**Filtrées**		**Gardées**	
	correctes	**fausses**	**correctes**	**fausses**
LIEU – LIEU	1	49 (98 %)	9 (18 %)	41
LIEU – ORG	4	46 (92 %)	8 (16 %)	42
LIEU – PERS	3	47 (94 %)	2 (4 %)	48
ORG – LIEU	7	43 (86 %)	14 (28 %)	36
ORG – ORG	6	44 (88 %)	20 (40 %)	30
ORG – PERS	4	46 (92 %)	20 (40 %)	30
PERS – LIEU	13	37 (74 %)	40 (80 %)	10
PERS – ORG	12	38 (76 %)	40 (80 %)	10
PERS – PERS	5	45 (90 %)	14 (28 %)	36

Tableau 1. *Évaluation du filtrage par les heuristiques*

Ce constat n'est d'ailleurs pas surprenant dans la mesure où la première entité d'une relation occupe souvent un rôle d'agent alors que les lieux apparaissent le plus fréquemment comme des circonstants. Compte tenu de cette observation, nous avons choisi d'écarter les relations ayant un lieu comme première entité dans ce qui suit.

4.4. *Modèle de séquence pour le filtrage par apprentissage*

L'évaluation précédente a mis en évidence l'intérêt des heuristiques testées pour écarter les mauvaises relations mais a également montré leur insuffisance pour conserver une proportion significative des relations correctes. Nous avons de ce fait choisi

d'adjoindre à ces heuristiques un module de filtrage de type apprentissage statistique pour décider si une relation extraite est véritablement sous-tendue par une relation effective entre ses entités. À l'instar de Li *et al.* (2011) pour l'extraction de relations de type connu *a priori* et de Banko et Etzioni (2008) pour des relations sans type prédéfini, nous avons abordé ce problème comme une tâche d'annotation : ce module annote la partie d'une phrase exprimant une relation entre deux entités nommées, lorsqu'une telle relation existe. En adéquation avec ce point de vue, nous avons développé un modèle fondé sur les champs conditionnels aléatoires (CRF), un paradigme d'apprentissage statistique utilisé avec succès pour de telles tâches d'annotation, en particulier grâce à son exploitation conjointe de la notion de séquence et de traits associés aux unités de cette séquence. Outre que, comme l'illustre Wang *et al.* (2011), un modèle de type CRF linéaires obtient des résultats supérieurs à ceux de classifieurs de type machine à vecteurs de support (SVM), maximum d'entropie (MaxEnt), arbre de décision ou bayésien naïf, il présente aussi l'avantage de prendre en compte les dépendances de nature séquentielle de façon beaucoup plus souple que ces classifieurs (pas de taille arbitraire des relations, par exemple).

La mise en œuvre de ce filtrage des relations a commencé par la construction d'un corpus de référence en annotant manuellement un ensemble de relations. Plus précisément, 200 relations ont été sélectionnées au hasard et annotées pour chacune des six catégories de relations finalement considérées. L'annotation distinguait les relations correctes, les relations incorrectes du fait d'un problème de reconnaissance des entités nommées et les relations fausses du fait de l'absence de relation effective. Les résultats de cette annotation sont présentés dans le tableau 2.

Catégories	**Correctes**	**Erreurs EN**	**Fausses**
ORG – LIEU	38 % (77)	18 % (35)	44 % (88)
ORG – ORG	39 % (78)	14 % (28)	47 % (94)
ORG – PERS	36 % (72)	18 % (36)	46 % (92)
PERS – LIEU	51 % (102)	31 % (62)	18 % (36)
PERS – ORG	60 % (120)	18 % (36)	22 % (44)
PERS – PERS	41 % (82)	20 % (39)	40 % (79)
Tous	44 % (531)	20 % (236)	36 % (433)

Tableau 2. *Résultat de l'annotation manuelle des relations*

Les relations incorrectes du fait des entités nommées représentent environ 20 % de l'ensemble et ont été laissées de côté pour l'entraînement et le test du modèle CRF. Le corpus résultant se compose donc de 964 relations, 531 étant correctes et 433 étant fausses, ce qui constitue un ensemble suffisamment équilibré pour ne pas poser de problème spécifique pour l'apprentissage du modèle.

Comme nous l'avons indiqué en préambule, la tâche considérée prend la forme d'une annotation, et plus précisément d'un étiquetage, illustré par la figure 3. Plus précisément, il s'agit d'étiqueter chaque mot d'une phrase par l'une des quatre éti-

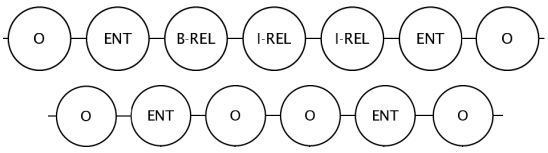


Figure 3. *Étiquetage des relations par un modèle CRF*

quettes suivantes, suivant en cela le modèle IOB introduit par Ramshaw et Marcus (1995) :

– O : mot de la phrase en dehors d'une relation ;

– ENT : mot d'une entité nommée impliquée dans une relation (E1 ou E2) ;

– B-REL : premier mot d'une relation suivant E1 ;

– I-REL : mot faisant partie d'une relation.

Dans ce schéma, une relation est jugée correcte lorsque l'étiquetage suit la première configuration de la figure 3 (avec un nombre de I-REL variable selon la relation) tandis qu'elle est jugée fausse lorsque l'étiquetage produit la deuxième configuration [4].

Ce modèle CRF s'appuie sur l'ensemble des caractéristiques suivantes, présentant la particularité de ne pas être lexicalisées[5], comme dans le cas de Banko et Etzioni (2008) :

– la catégorie morphosyntaxique du mot courant, du mot précédent et du mot suivant ;

– les bigrammes de catégories morphosyntaxiques $<\text{cat}_{i-1}, \text{cat}_i>$, avec $i = -1, 0, 1$ (0 : mot courant ; -1 : mot précédent ; 1 : mot suivant) ;

– le type d'entité nommée du mot courant et de chacun des six mots le précédant et le suivant. Ce type peut avoir une valeur nulle (notée NIL) lorsque le mot ne fait pas partie d'une entité nommée.

Compte tenu de la taille relativement réduite du corpus pour chaque catégorie de relations, nous avons choisi d'évaluer notre modèle CRF, implémenté au moyen de l'outil Wapiti (Lavergne *et al.*, 2010), en faisant appel à la technique classique de la validation croisée à k-plis (k-fold), k étant égal à 10 dans le cas présent. Le tableau 3 donne le résultat de cette évaluation pour les mesures standard d'exactitude (*accuracy*), de précision, rappel et F-score (F_1-mesure).

4. D'autres séquences marquant l'absence de relation seraient en principe possibles (comme O – ENT – B-REL – O – O – ENT – O) mais seule la deuxième est observée en pratique, sans doute du fait de la présence des deux seuls types de séquences de la figure 3 dans le corpus d'apprentissage, issues d'une transformation automatique des annotations manuelles.

5. C'est-à-dire n'intégrant pas la forme fléchie ou normalisée des mots non grammaticaux.

Modèle	Exactitude	Précision	Rappel	F-score
CRF	0,745	0,762	0,782	0,771
(Banko et Etzioni, 2008)	/	0,883	0,452	0,598

Tableau 3. *Évaluation du modèle CRF de filtrage des relations*

Outre le niveau globalement assez élevé des résultats obtenus, ce tableau laisse apparaître un équilibre intéressant entre la précision et le rappel. Il montre également que notre modèle CRF se compare favorablement à celui de Banko et Etzioni (2008) pour une tâche similaire. Dans ce dernier cas, le profil des résultats est un peu différent puisque la précision est plus forte que la nôtre mais le rappel très largement inférieur. Il faut néanmoins préciser que dans (Banko et Etzioni, 2008), les relations extraites peuvent faire intervenir des entités plus générales que des entités nommées, ce qui est *a priori* un facteur de difficulté. En revanche, Banko et Etzioni (2008) construisent automatiquement leurs exemples d'apprentissage à partir des résultats d'une analyse syntaxique, ce qui pose plus de contraintes sur la forme des relations considérées que dans notre cas.

4.5. *Comparaison avec le système* REVERB

Les résultats de Banko et Etzioni (2008) donne un premier point de repère pour situer les performances de notre système d'extraction de relations mais le fait d'opérer sur des corpus différents avec des types de relations différents oblige à ne considérer la comparaison avec cette référence qu'avec prudence. C'est pour cette raison que nous avons voulu mener une comparaison plus directe avec le système REVERB, un successeur de TEXTRUNNER (Banko et Etzioni, 2008) librement disponible et accompagné de son corpus d'évaluation. Celui-ci est constitué de 500 phrases recueillies grâce au service Yahoo random link, avec l'annotation de 2 474 relations candidates, 621 ayant été jugées comme de véritables relations.

Pour avoir une comparaison équilibrée, nous avons procédé à une évaluation croisée en appliquant le système REVERB sur notre corpus d'évaluation et notre propre système sur celui de REVERB. Compte tenu de la différence entre REVERB et notre approche concernant les arguments des relations – REVERB s'appuie sur des groupes nominaux tandis que nous nous limitons aux entités nommées – quelques adaptations ont dû être réalisées. Ainsi pour l'application de notre système au corpus REVERB, nous avons considéré les groupes nominaux identifiés par REVERB comme des entités nommées et réentraîné notre CRF sans prendre en compte le type des entités. Par ailleurs, compte tenu des problèmes d'alignement du résultat du prétraitement linguistique avec la référence, seules 2 412 relations candidates ont été considérées, parmi lesquelles 606 relations sont jugées correctes. À l'inverse, pour l'application de REVERB à notre corpus d'évaluation, ses arguments de relations ont dû être mis en

correspondance avec les entités nommées extraites par OpenNLP, ce qui a été fait de façon large sur la base du partage au minimum d'un mot plein. Les résultats de cette double évaluation sont donnés par le tableau 4.

Corpus	Système	Précision	Rappel	F-score
Corpus REVERB	REVERB	0,505	0,578	0,539
	Notre système	0,357	0,627	0,455
Notre corpus	REVERB	0,810	0,363	0,501
	Notre système	0,762	0,782	0,771

Tableau 4. *Comparaison avec le système* REVERB

Le premier constat évident que laisse apparaître ce tableau est que chaque système obtient à la fois ses meilleurs résultats et ses résultats les plus équilibrés sur son propre corpus d'évaluation, ce qui illustre la dépendance de ces corpus d'évaluation par rapport aux hypothèses faites sur les relations à extraire. Au-delà de cette observation, l'évaluation montre aussi clairement la différence de stratégie des deux systèmes : REVERB pose des contraintes fortes, notamment en termes syntaxiques, sur la forme des relations à extraire, ce qui lui permet d'obtenir une bonne précision, en particulier sur notre corpus. Notre approche est inverse : nous nous appuyons sur un extracteur statistique guidé par des exemples annotés sans restriction particulière autre que la nature des arguments, ce qui peut expliquer notre meilleur rappel. Cette différence de stratégie est d'ailleurs plus globale : notre extraction est une étape préalable au regroupement des relations, ce qui n'est pas la perspective de REVERB. Or, ce regroupement constitue aussi une forme indirecte de filtrage : les relations non valides sont en effet plus susceptibles de former des singletons ou de très petits clusters jugés comme inintéressants.

4.6. *Application du filtrage des relations*

L'extraction des relations telle que nous l'avons envisagée précédemment se compose des quatre étapes suivantes, appliquées successivement :

1) une extraction initiale ne posant comme contraintes que la cooccurrence dans une phrase d'entités nommées relevant d'un ensemble donné de types et la présence d'au moins un verbe entre les deux ;

2) l'application des heuristiques permettant d'écarter avec une bonne précision un grand nombre de relations fausses ;

3) l'application d'un modèle de filtrage à base de CRF permettant de discriminer plus finement les relations correctes ;

4) l'élimination des relations redondantes.

Le constat de la présence, dans nos relations filtrées, d'un certain nombre de relations identiques, pour une part issues d'articles sur un même sujet ou d'articles correspondant à des rubriques très formatées, nous a conduits à compléter le processus

de filtrage constitué par les trois premières étapes par un dédoublonnage final visant à éliminer ces relations redondantes. Pour ce faire, nous reprenons les outils utilisés par le processus de regroupement de relations de la section 5 pour évaluer la similarité entre les relations et détecter celles dont la similarité est maximale ce qui, compte tenu de l'existence d'une borne supérieure pour la mesure utilisée, signifie que les relations sont identiques. Pour chaque ensemble de relations identiques, un représentant est alors choisi. Il est à noter que cette opération de dédoublonnage vient en dernière position à la fois parce que son coût est le plus important mais également parce qu'elle repose sur l'évaluation de la similarité entre les relations, exploitée ensuite directement pour le regroupement des relations.

	Initial	**Heuristiques**	**Classifieur CRF**	**Dédoublonnage**
ORG – LIEU	71 858	33 505 (47 %)	16 700 (23 %)	15 226 (21 %)
ORG – ORG	77 025	37 061 (48 %)	17 025 (22 %)	13 704 (18 %)
ORG – PERS	73 895	32 033 (43 %)	12 098 (16 %)	10 054 (14 %)
PERS – LIEU	152 514	72 221 (47 %)	55 174 (36 %)	47 700 (31 %)
PERS – ORG	126 281	66 035 (52 %)	50 487 (40 %)	40 238 (32 %)
PERS – PERS	175 802	78 530 (45 %)	42 463 (24 %)	38 786 (22 %)
TOTAL	677 375	319 385 (47 %)	193 947 (29 %)	165 708 (24 %)

Tableau 5. *Niveau de filtrage des relations à l'issue de chacune des étapes*

Le tableau 5 illustre l'application des quatre étapes de filtrage aux relations extraites de notre corpus de travail. On constate que ce filtrage laisse de côté un grand nombre des relations extraites initialement mais que le volume des relations restantes est *a priori* suffisant pour alimenter efficacement les étapes suivantes de notre processus d'extraction d'information non supervisée. Par ailleurs, comme Banko et Etzioni (2008), nous nous situons dans un contexte de traitement de volumes textuels importants caractérisés par une certaine redondance informationnelle où la perte d'une certaine quantité d'instances de relations n'est pas un obstacle pour appliquer notre approche.

5. Regroupement de relations

5.1. *Principe du regroupement des relations*

À l'instar de travaux dans le domaine de l'EI non supervisée comme (Shinyama et Sekine, 2006) ou (Rosenfeld et Feldman, 2007), notre objectif final est le regroupement des relations selon leur similarité, en particulier pour en faciliter l'exploration. La méthode de regroupement visée doit à la fois être capable de traiter le volume important de relations issu de leur filtrage et la variété de la forme de ces relations, inhérente au fait de travailler en domaine ouvert. Pour ce faire, nous proposons une méthode, illustrée par la figure 4, s'organisant en deux étapes principales, à l'image de l'approche multiniveau de (Cheu *et al.*, 2004) : un premier *clustering* de base est

réalisé en s'appuyant sur la similarité des formes de surface des relations, ce qui permet de former de manière efficace de petits clusters homogènes en regroupant des instances de relations définies autour d'un même mot-clé principal, comme pour les formes {*create, create the, that create, who create the, etc*}. ; une seconde étape de *clustering* est ensuite appliquée pour rassembler ces clusters initiaux sur la base d'une similarité sémantique entre relations plus complexe. Cette similarité permet de prendre en compte des phénomènes tels que la synonymie, voire la paraphrase, pour rassembler des formes telles que {*create, establish, found, launch, inaugurate, etc*}.

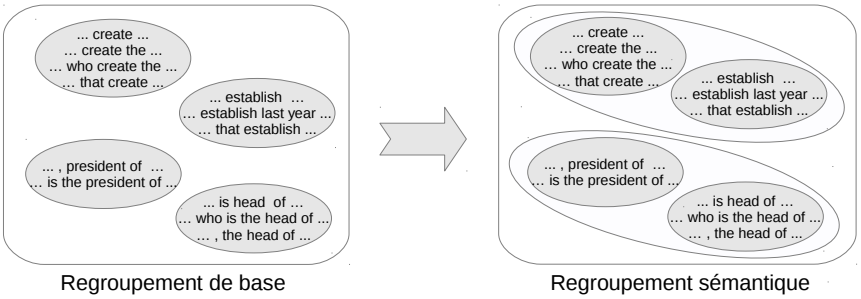


Figure 4. *Regroupement des relations en niveaux*

En plus de ces deux niveaux de *clustering* pour opérer un regroupement sémantique de qualité, nous proposons d'intégrer un troisième type de *clustering*, dont l'objet est différent mais complémentaire des deux premiers : son objectif est de rassembler les instances de relations dont les contextes font référence au même thème. Cette autre dimension de structuration des instances de relations possède à la fois un intérêt applicatif et un intérêt du point de vue sémantique : le fait de se situer dans un contexte thématique homogène tend en effet à réduire le problème de l'ambiguïté sémantique des mots, ce qui permet des rapprochements plus sûrs.

5.2. *Regroupement de base*

En EI non supervisée, le nombre de relations extraites est rapidement important. De ce fait, il est quasiment impossible d'appliquer des mesures de similarité sémantique élaborées entre toutes les relations extraites. Nous mettons en œuvre un premier niveau de *clustering* afin de former des regroupements de relations proches les unes des autres sur le plan de leur expression linguistique, comme le fait de regrouper *create the* et *who create*. Pour ce faire, nous nous sommes appuyés sur une similarité Cosinus appliquée à une représentation de type « sac de mots » de la partie *Cmid* des relations.

5.2.1. *Pondération des termes*

Une représentation en sac de mots d'un contenu textuel s'appuie sur une pondération choisie de chacun des mots du texte. Si l'on considère que tous les mots d'une

phrase n'apportent pas la même contribution au sens général de la phrase, il est nécessaire d'établir une stratégie de pondération adéquate pour établir une bonne mesure de similarité entre phrases. Trois types de pondérations sont considérés ici :

– pondération binaire : tous les mots de $Cmid$ ont le même poids (1,0) ;

– pondération *tf-idf* : un poids *tf-idf* standard est attribué à chaque mot (prenant en compte la fréquence du mot dans la relation et la fréquence inverse du mot dans l'ensemble des relations) ;

– pondération grammaticale : des poids spécifiques sont donnés aux mots en fonction de leur catégorie morphosyntaxique.

La pondération binaire est la plus simple et forme une *baseline*, qui a été utilisée dans nos premières expériences. La pondération *tf-idf* prend en compte, par le biais du facteur *idf*, une mesure de l'importance du terme dans le corpus. Néanmoins, la fréquence des mots dans un corpus n'est pas forcément corrélée à leur rôle dans la caractérisation d'une relation. Par exemple, le verbe *buy* peut être fréquent dans un corpus de documents financiers (et donc avoir un poids faible), mais il n'en sera pas moins représentatif de la relation BUY(ORG – ORG). C'est pourquoi nous avons décidé d'introduire une pondération grammaticale.

Une analyse des catégories morphosyntaxiques nous a amenés à les séparer en plusieurs classes selon leur importance dans la contribution à l'expression d'une relation, selon un schéma de séparation différent et plus fin que les simples filtrages appliqués usuellement sur les catégories grammaticales pour distinguer les mots pleins des mots vides (par exemple, les noms propres ne sont pas sémantiquement représentatifs d'un type de relation). Plus précisément, nous considérons les quatre classes suivantes :

– **(A) contribution directe**, de poids élevé : les mots de cette classe contribuent directement au sens de la relation et incluent les verbes, noms, adjectifs, prépositions ;

– **(B) contribution indirecte**, de poids moyen : les mots de la classe B ne sont pas directement liés au sens de la relation mais sont pertinents dans l'expression de la phrase, comme les adverbes et les pronoms ;

– **(C) information complémentaire**, de poids faible : cette classe contient des mots fournissant une information complémentaire sur la relation, comme les noms propres ;

– **(D) pas d'information**, de poids nul : cette classe contient les mots vides que l'on veut ignorer (symboles, nombres, déterminants etc.).

Nous présentons dans le tableau 6 une configuration de pondération grammaticale. La liste des catégories morphosyntaxiques est fondée sur les catégories du *Penn Treebank*. Des poids de 1,0, 0,75, 0,5 et 0 sont attribués aux classes A, B, C et D. Pour les catégories non présentes dans cette liste, un poids par défaut de 0,5 est utilisé.

5.2.2. *Regroupement par mots-clés représentatifs*

Pour renforcer ce premier niveau de *clustering*, la stratégie généraliste présentée ci-dessus a été complétée par une heuristique tenant compte de la spécificité des rela-

Classe	Catégories morphosyntaxiques
A ($w = 1,0$)	VB VBD VBG VBN VBP VBZ NN NNS JJ JJR JJS IN TO RP
B ($w = 0,75$)	RB RBR RBS WDT WP WP$ WRB PDT POS PRP PRP$
C ($w = 0,5$)	NNP NNPS UH
D ($w = 0,0$)	SYM CC CD DT MD

Tableau 6. *Pondération grammaticale : distribution des poids selon la catégorie morphosyntaxique*

tions. Au sein d'un cluster de base, la forme linguistique de ces dernières est en effet souvent dominée par un verbe (*founded* pour *a group founded by* ou *which is founded by*) ou par un nom (*head* pour *who is the head of*, *becomes head of*), ce terme dominant possédant une fréquence élevée dans le cluster. De ce fait, nous considérons le nom ou le verbe le plus fréquent au sein d'un cluster de base comme son représentant, à l'instar de travaux comme (Hasegawa *et al.*, 2004), et nous fusionnons les clusters partageant le même terme dominant, appelé « mot-clé » dans ce qui suit, pour former des clusters de base plus larges.

5.3. *Regroupement sémantique*

Le premier niveau de *clustering* ne peut clairement pas regrouper des relations exprimées avec des termes complètement différents. Dans l'exemple *who create the* et *that establish*, présenté à la figure 4, les deux formes linguistiques ont peu en commun. Nous avons donc considéré l'ajout d'un deuxième niveau de *clustering* ayant pour objectif de regrouper les clusters formés précédemment sur des bases plus sémantiques, plus précisément en intégrant les similarités sémantiques au niveau lexical. Contrairement au premier, ce deuxième niveau bénéficie en outre du fait de travailler à partir de clusters et non de relations individuelles, ce qui permet d'exploiter une information plus riche : en effet, la redondance d'information dans les clusters de premier niveau permet de mettre en évidence les mots les plus importants des relations et offrir ainsi un meilleur socle sur lequel appuyer les mesures de similarité. Il nécessite de ce fait de définir trois niveaux de similarité sémantique : similarité entre les mots, entre les relations et entre les clusters de base des relations.

5.3.1. *Similarité sémantique entre les mots*

Les mesures de similarité sémantique au niveau lexical se répartissent en deux grandes catégories aux caractéristiques souvent complémentaires : la première rassemble les mesures fondées sur des connaissances élaborées manuellement prenant typiquement la forme de réseaux lexicaux de type WordNet ; la seconde recouvre les mesures de nature distributionnelle, construites à partir de corpus. Pour évaluer la similarité sémantique entre les relations, nous avons choisi de tester des mesures relevant de ces deux catégories afin de juger de leur intérêt respectif.

Concernant le premier type de mesures, le fait de travailler avec des textes en anglais ouvre le champ des différentes mesures définies à partir de WordNet. Nous en avons retenu deux caractéristiques : la mesure de Wu et Palmer (1994), qui évalue la proximité de deux synsets en fonction de leur profondeur dans la hiérarchie de WordNet et de la profondeur de leur plus petit ancêtre commun ; la mesure de Lin (1998), qui associe le même type de critère que la mesure de Wu et Palmer et des informations de fréquence d'usage des synsets dans un corpus de référence. Ces mesures étant définies entre synsets, pour se ramener à une mesure entre mots, nous avons adopté la stratégie utilisée notamment dans (Mihalcea *et al.*, 2006) consistant à prendre comme valeur de similarité entre deux mots la plus forte valeur de similarité entre les synsets dont ils font partie.

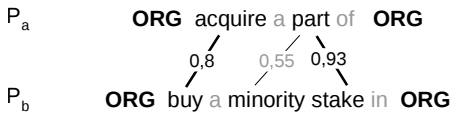
Les mesures de similarité distributionnelles sont, quant à elles, fondées sur l'hypothèse que les mots apparaissant dans les mêmes contextes tendent à avoir le même sens. La notion de contexte renvoie ici traditionnellement à l'ensemble des mots cooccurrents avec le mot cible dans un corpus. Cette cooccurrence peut être purement graphique, au sein d'une fenêtre de taille fixe, ou bien reposer sur des relations syntaxiques. Nous avons testé ici les deux types de cooccurrences, les termes au sein des contextes ainsi formés étant pondérés grâce à la mesure d'information mutuelle (*Pointwise Mutual Information*) et les contextes eux-mêmes étant comparés grâce à la mesure Cosinus pour évaluer la similarité de deux mots. Plus précisément, nous avons utilisé les thésaurus distributionnels présentés dans (Ferret, 2010) pour disposer de ces similarités sous une forme précalculée.

Dans le cadre de la comparaison de relations, nous nous sommes intéressés essentiellement à la similarité sémantique entre des mots appartenant à la même catégorie morphosyntaxique en nous fondant sur le fait que les relations extraites se définissent généralement autour d'un verbe (*e.g.* ORG *found by* PERS, ORG *establish by* PERS) ou d'un nom (*e.g.* ORG *be partner of* ORG, ORG *have cooperation with* ORG), mais pas sous les deux formes pour un même type de relation, sans doute à cause de la focalisation sur la partie *Cmid* des relations.

5.3.2. *Similarité sémantique des relations*

La similarité s'applique ici à l'échelle de la définition linguistique des relations, *i.e.* leur partie *Cmid*, ce qui s'apparente à la problématique de la détection de paraphrases. De ce fait, nous avons repris le principe expérimenté dans (Mihalcea *et al.*, 2006) pour cette tâche : chaque phrase (ici relation) à comparer est représentée sous la forme d'un sac de mots et lors de l'évaluation de la similarité $sim(P_a, P_b)$ d'une phrase P_b par rapport à une phrase P_a, chaque mot de P_a est apparié au mot de P_b avec lequel sa similarité sémantique, au sens de la section 5.3.1, est la plus forte. Ainsi, dans l'exemple ci-dessous, *acquire* est apparié à la seule possibilité, *buy*, tandis que *part* est apparié à *stake*, avec lequel il partage la plus grande similarité selon la mesure de Wu-Palmer.

Un mot d'une phrase peut éventuellement ne pas être apparié si sa similarité avec tous les autres mots de l'autre phrase est nulle. Cette mesure de similarité n'étant



pas symétrique, la similarité complète est égale à la moyenne de $sim(P_a, P_b)$ et $sim(P_b, P_a)$. Plus formellement, si l'on définit P_a et P_b comme :

$$P_a = W_1 : f_1,\ W_2 : f_2,\ \ldots, W_i : f_i, \ldots,\ W_M : f_M$$
$$P_b = W_1 : f_1,\ W_2 : f_2,\ \ldots, W_j : f_j, \ldots,\ W_N : f_N$$

où W_i est un mot d'une phrase et f_i, sa fréquence dans la phrase, cette similarité complète s'écrit :

$$S_{P_{a,b}} = \frac{1}{2}\Big(\frac{1}{\sum_{i\in[1,M]} w_i} \sum_{i\in[1,M]} \max_{j\in[1,N]} \{S_{W_{i,j}}\} \cdot w_i + \frac{1}{\sum_{j\in[1,N]} w_j} \sum_{j\in[1,N]} \max_{i\in[1,M]} \{S_{W_{i,j}}\} \cdot w_j\Big) \quad [1]$$

où $S_{W_{i,j}}$ est la similarité sémantique entre les mots W_i et W_j, qu'elle soit fondée sur WordNet ou sur un thésaurus distributionnel et w_i et w_j sont les poids de ces mots respectivement dans P_a et P_b, définis par leur fréquence ($w_i = f_i, w_j = f_j$).

5.3.3. *Similarité sémantique des clusters*

Le principe guidant la similarité de deux relations est trop coûteux à transposer au niveau des clusters car il nécessiterait, pour deux clusters C_a et C_b, de calculer $|C_a|\cdot|C_b|$ similarités, lesquelles ne peuvent pas être précalculées comme pour les mots. La similarité à l'échelle des relations étant fondée sur une représentation de type sac de mots, nous avons choisi de construire pour les clusters une représentation de même type en fusionnant les représentations de leurs relations. Au sein de la représentation d'un cluster, chaque mot se voit associé sa fréquence parmi les relations du cluster, les mots de plus fortes fréquences étant supposés les plus représentatifs du type de relation sous-jacent au cluster. Ainsi, l'importance des mots des relations est évaluée sur les regroupements effectués et non pas sur des critères externes aux relations. Nous avons aussi vu que le calcul du poids des mots sur des critères propres aux relations prises individuellement (exemple de *tf-idf*) n'apportait pas d'informations intéressantes.

Pour le calcul de la similarité entre les clusters, nous proposons une adaptation de la similarité entre les relations, destinée à pallier un biais possible lorsque les deux clusters sont de tailles différentes. Par exemple, les clusters C_a et C_b définis ci-dessous ne sont pas sémantiquement similaires mais ont une valeur $S_{P_{a,b}}$, telle que définie dans l'équation (1), élevée du fait du poids élevé du mot *actor* dans C_a.

C_a = found:3, actor:3 … {*i.e.* PERS *an actor who found* ORG}
C_b = study:9, actor:1 … {*i.e.* PERS *study at* ORG, PERS *an actor study at* ORG}

Même si dans un tel cas, $sim(P_b, P_a)$ serait plus faible que $sim(P_a, P_b)$, cette dernière influencerait fortement la moyenne des deux et conduirait à une similarité globale assez forte. Pour contrecarrer cet effet, nous introduisons la fréquence des mots dans les deux clusters et non dans celui servant de référence seulement, en remplaçant, dans l'équation (1), les poids w_i et w_j par w_{ij}, défini par $w_{ij} = f_i \cdot f_j$.

5.4. *Regroupement thématique des relations*

Les deux niveaux de regroupement de relations (regroupement de base et regroupement sémantique) présentés ci-dessus ont pour objectif de regrouper des instances de relations équivalentes sur le plan sémantique et ce, en s'appuyant uniquement sur des informations locales des phrases les contenant, en l'occurrence leur partie *Cmid*. Mais chaque relation s'inscrit également dans un contexte plus large, faisant référence à des thèmes tels que la politique, l'économie ou le sport par exemple. En proposant de regrouper les instances de relations suivant cette dimension thématique, nous poursuivons deux objectifs : sur un plan applicatif, proposer une autre dimension de regroupement, complémentaire de la dimension sémantique, qui permet de mettre en contexte les relations extraites ; sur le plan du regroupement même des instances de relations, former des clusters sémantiques plus précis en désambiguïsant indirectement les mots des relations sur lesquels ils reposent. Deux instances de relations peuvent en effet avoir été regroupées sur la base d'un mot utilisé avec des sens différents car faisant référence à des contextes thématiques différents, à l'instar par exemple du mot *title* qui possède un sens particulier dans le domaine du sport et un autre dans le domaine des arts.

Ce regroupement thématique est plus précisément effectué de façon indirecte : il ne s'applique pas en premier lieu aux instances de relations mais aux seuls contextes (*i.e.* les segments thématiques) dans lesquels elles apparaissent. Tous les segments extraits du corpus considéré sont ainsi regroupés selon leur similarité en adoptant les mêmes modalités que pour le *clustering* de base des instances de relations : une représentation de type « sac de mots » pour chaque segment avec une pondération *tf.idf* des mots et l'utilisation de la mesure de similarité Cosinus. La figure 5 illustre le processus de regroupement thématique des relations, qui se fait par l'intégration des clusters thématiques ainsi formés et des clusters sémantiques de relations. Plus précisément, chaque cluster thématique CT_k, qui incarne un thème du corpus par le regroupement des segments thématiques S_j, est également un regroupement d'instances des relations R_{ij} qui apparaissent dans les segments regroupés. La fusion des deux informations se fait alors par l'intersection des clusters de relations CR_i et des clusters de segments thématiques CT_k : au sein de chaque cluster sémantique, les instances de relations faisant partie d'un même cluster thématique sont regroupées pour former un cluster thématique de relations (CTR_{ik}, défini formellement par $CTR_{ik} = CR_i \cap CT_k$). Les instances de relations d'un cluster sémantique non regroupées à l'issue de ce processus forment elles-mêmes un cluster thématique.

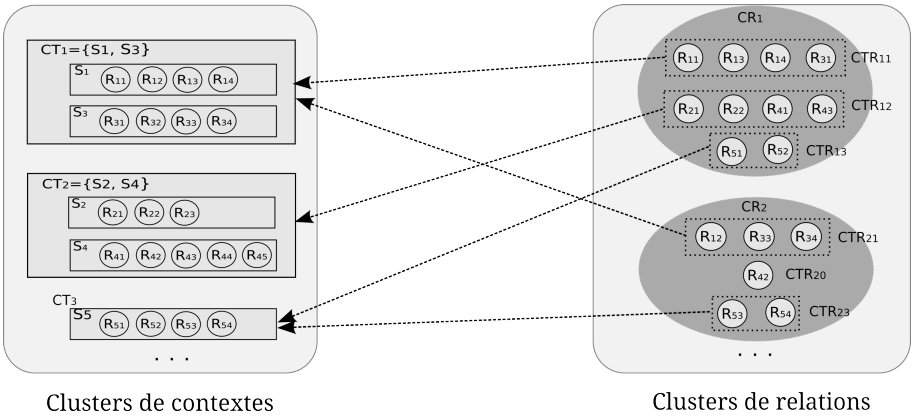


Figure 5. *Regroupement thématique des relations*

5.5. *Algorithmes de regroupement*

Le choix de la mesure de similarité Cosinus, outre son compromis intéressant entre simplicité et efficacité, a été motivé par la possibilité de calculer les similarités deux à deux pour un grand nombre d'éléments par une utilisation de l'algorithme *All Pairs Similarity Search* (APSS) (Bayardo *et al.*, 2007), qui permet de construire de façon optimisée la matrice de similarité d'un ensemble de vecteurs, pour des valeurs de similarité supérieures à un seuil minimal fixé *a priori*. Pour la construction de nos clusters de base, nous avons mis en œuvre l'association d'un tel seuillage sur les valeurs de similarité entre les relations avec un algorithme *Markov Clustering* (Dongen, 2000), qui s'appuie sur une représentation en graphe de la matrice de similarité, et identifie les zones du graphe les plus densément connectées en réalisant des marches aléatoires dans ce graphe. Cet algorithme présente l'avantage, du point de vue de l'IE non supervisée, de ne pas nécessiter la fixation préalable d'un nombre de clusters. Par ailleurs, le seuillage réalisé conduit à éclaircir le graphe de similarité et rend possible l'application du *Markov Clustering* qui, en dépit de son efficacité, ne pourrait gérer la matrice complète de similarité des relations. Par ailleurs, la taille des clusters à former peut être assez variable selon le contenu du corpus considéré mais la valeur de similarité de deux relations est assez facile à étalonner à partir de résultats de référence (cf. section 6.1 pour une illustration), ce qui justifie le fait de se focaliser sur la similarité entre relations. La problématique est similaire pour le regroupement thématique des contextes : la taille des clusters formés peut être assez variable selon le niveau de représentation d'un thème dans le corpus considéré mais la similarité de deux segments thématiques est suffisamment indicative pour fixer des seuils.

Le cas du *clustering* sémantique est assez différent. Le fait d'utiliser des ressources de natures assez diverses rend difficile la fixation *a priori* d'un seuil de similarité car les intervalles de valeurs ne sont pas les mêmes selon les cas. En revanche, la ri-

chesse des ressources sémantiques utilisées permet d'avoir une idée approximative du nombre de voisins d'un cluster de base. Un tel cluster se définissant souvent autour d'un terme clé, ce nombre de voisins est assez directement en rapport avec le nombre de synonymes ou de mots sémantiquement liés à ce terme. De ce fait, pour le *clustering* sémantique, nous avons adopté l'algorithme *Shared Nearest Neighbor* (SNN) proposé dans (Ertöz *et al.*, 2002) plutôt que le *Markov Clustering* utilisé initialement. Cet algorithme définit en effet implicitement la taille des clusters qu'il forme en seuillant le nombre de voisins possibles pour chaque élément à regrouper[6].

6. Résultats et évaluations

Nous avons mené l'évaluation de ce *clustering* de relations multiniveau selon une approche externe en utilisant les mesures standard de précision et rappel (combinées par la F-mesure). Ces mesures sont appliquées à des paires de relations en considérant que les relations peuvent être regroupées dans le même cluster ou séparées dans des clusters différents et ce, de façon correcte ou incorrecte par rapport à la référence. Nous utilisons également les mesures standard pour le *clustering* de pureté, pureté inverse et information mutuelle normalisée (NMI) (Amigó *et al.*, 2009). Le *clustering* de référence utilisé a été construit manuellement à partir d'un sous-ensemble de relations provenant de l'extraction initiale. Il est formé de 80 clusters couvrant 4 420 relations : une douzaine de clusters sont construits pour chaque paire de types d'entités en relation, avec des tailles variant entre 4 et 280 relations. De plus amples détails sur la construction de cette référence et les mesures d'évaluation utilisées sont donnés dans (Wang *et al.*, 2012).

Les évaluations suivantes ont été mises en œuvre : en premier lieu, une expérimentation a été menée pour évaluer l'impact de la qualité de l'extraction de relations sur leur regroupement, et en particulier l'impact de l'utilisation de l'étape de filtrage présentée dans la section 4. Nous proposons ensuite une évaluation du *clustering* des relations, en évaluant dans un premier temps le *clustering* de base et l'impact du choix de la pondération des termes pour ce *clustering* et, dans un deuxième temps, l'apport du second niveau de *clustering* sémantique et la comparaison des différentes mesures de similarité sémantique. Une étude supplémentaire permettant de mettre en évidence l'intérêt d'un *clustering* multiniveau par rapport à un *clustering* direct est également présentée. Enfin, une évaluation du *clustering* thématique des relations est proposée dans la section 6.3.

6. Les hypothèses faites sur l'adéquation entre le type d'éléments à regrouper et les algorithmes de regroupement ont été confirmées expérimentalement : l'algorithme SNN donne de moins bons résultats que le *Markov Clustering* pour le premier niveau de *clustering* mais l'ordre s'inverse pour le *clustering* sémantique.

6.1. *Évaluation de l'impact du filtrage sur le regroupement des relations*

Nous avons en premier lieu évalué l'impact de la procédure de filtrage sur les résultats du regroupement des relations. Pour ce faire, nous avons appliqué le *clustering* de base sur l'ensemble des instances de relations extraites avant la procédure de filtrage et sur celui obtenu après le filtrage. Le seuil de similarité utilisé pour ce *clustering* de base (pour élaguer la matrice de similarité grâce à l'algorithme APSS) a été fixé à 0,45. Ce seuil a été choisi empiriquement en étudiant le comportement de l'algorithme de *clustering* sur les phrases du corpus *Microsoft Research Paraphrase* (Dolan *et al.*, 2004) et couvre les trois quarts des valeurs de similarité de ses phrases en état de paraphrases. Les performances de ces deux applications du *clustering* de base sont données dans le tableau 7.

	Préc.	**Rappel**	**F-score**	**Pur.**	**Pur. inv.**	**NMI**	**Nb**	**Taille**
Sans filtrage	0,708	0,282	0,403	**0,915**	0,381	0,743	82 338	5,54
Avec filtrage	**0,756**	**0,312**	**0,442**	0,902	**0,407**	**0,750**	15 833	**7,50**

Tableau 7. *Impact du filtrage sur les résultats du regroupement des relations*

Les colonnes de ce tableau correspondent respectivement aux mesures de précision, rappel, F-mesure, pureté, pureté inverse et information mutuelle normalisée, auxquelles s'ajoutent le nombre de clusters et la taille moyenne des clusters. À l'exception de la pureté, toutes ces mesures montrent l'impact positif du filtrage des relations sur leur regroupement. Ce comportement de la pureté est d'ailleurs compensé d'une certaine façon par une augmentation plus importante de la pureté inverse. En outre, la réduction du bruit au niveau des instances de relations résultant de leur filtrage se manifeste aussi par la tendance à former des clusters plus grands, la taille moyenne de ceux-ci passant de 5,54 à 7,50 instances de relations. Le filtrage favorise donc le rapprochement des instances de relations. La suite des expérimentations est réalisée à partir du résultat produit après filtrage.

6.2. *Évaluation du* clustering *sémantique des relations*

6.2.1. *Évaluation du* clustering *de base*

La phase suivante de notre évaluation du *clustering* a porté sur les différents schémas de pondération décrits à la section 5.2.1 pour le *clustering* de base. Le même seuil (0,45) présenté à la section précédente est utilisé pour la pondération binaire et celle par *tf-idf*. La pondération grammaticale est moins stricte et les similarités avec ces pondérations sont en général plus élevées : dans ce cas, un seuil plus élevé de 0,60 a été utilisé. Les résultats obtenus pour le *clustering* de base avec les différentes pondérations proposées sont présentés dans le tableau 8.

Le regroupement sur la base de la similarité utilisant une pondération grammaticale donne les meilleurs résultats, avec une meilleure précision et un rappel satisfai-

	Préc.	Rappel	F-score	Pur.	Pur. inv.	NMI	Nb	Taille
Binaire	0,756	0,312	0,442	0,902	0,407	0,750	15 833	7,50
Tf-idf	0,203	**0,445**	0,279	0,646	**0,573**	0,722	11 911	11,44
Gramm.	**0,810**	0,402	**0,537**	**0,963**	0,513	**0,812**	13 648	7,56
Mots-clés	**0,812**	**0,443**	**0,573**	0,953	0,552	**0,825**	11 726	8,80

Tableau 8. *Résultats du* clustering *de base pour différentes pondérations des termes en utilisant le* Markov Clustering *(MCL) et le regroupement par mots-clés*

sant. Cette pondération utilise en effet plus de connaissances pour mettre en évidence le rôle des verbes, noms ou adjectifs et diminuer l'influence des mots qui ne contribuent qu'à des variations linguistiques légères (*who* + verbe, *the one that* + verbe). La pondération *tf-idf* donne, quant à elle, de moins bons résultats. Cette pondération favorise en effet les mots rares. Or, les noms communs et les verbes, qui supportent le plus souvent les relations, sont plus fréquents que des noms propres ou des occurrences de nombres, par exemple, qui se verront attribuer un score important avec cette pondération alors qu'ils n'apportent pas d'information sur la relation.

Les résultats utilisés par la suite pour le *clustering* sémantique sont ceux obtenus avec la pondération grammaticale. Plusieurs seuils et configurations des pondérations grammaticales ont été testés : la version présentée (seuil de 0,60 et poids du tableau 6) est celle donnant les meilleurs résultats. L'étape de regroupement par mots-clés amène, avec cette pondération, une amélioration légère de la *F-mesure*, due à un accroissement du rappel ; mais cette étape permet surtout de réduire le nombre de clusters et d'augmenter leur taille moyenne, comme illustré par les deux dernières colonnes du tableau 8.

6.2.2. *Évaluation du* clustering *fondé sur les similarités sémantiques*

Pour évaluer l'apport du *clustering* sémantique décrit à la section 5.3 sur les clusters de base, nous comparons les résultats obtenus avec les différentes mesures de similarité à un *clustering* idéal donnant le meilleur regroupement possible des clusters de base obtenus par la première étape : chaque cluster de base est associé au cluster de référence avec lequel il partage le plus de relations ; puis les clusters associés aux mêmes clusters de référence sont regroupés.

En ce qui concerne les mesures fondées sur WordNet, la mesure de Wu-Palmer est calculée grâce à NLTK (`nltk.org`) tandis que pour la mesure de Lin, nous utilisons les similarités précalculées entre les verbes de WordNet de Pedersen (2010). Dans les résultats présentés, la première a été utilisée pour mesurer la similarité entre les noms et la deuxième entre les verbes (cette configuration étant celle qui donne les meilleurs résultats). Les similarités distributionnelles sont, quant à elles, calculées à partir du corpus AQUAINT-2, sur la base d'une mesure Cosinus entre des vecteurs de contexte obtenus soit avec une fenêtre glissante de taille 3 (Dist_{cooc}), soit en suivant les liens syntaxiques entre les mots (Dist_{syn}). L'algorithme de *clustering* SNN nécessite de

fixer le nombre de voisins considérés pour chaque relation : dans nos expériences, ce nombre a été fixé à 100. Les résultats obtenus sont présentés dans le tableau 9.

	Préc.	**Rappel**	**F-score**	**Pur.**	**Pur. inv.**	**NMI**	**Nb**	**Taille**
Base	0,812	0,443	0,573	0,953	0,552	0,825	11 726	8,80
WordNet	0,821	0,507	0,627	0,942	0,622	0,839	9 403	10,98
Dist_{cooc}	0,814	0,540	0,649	0,932	0,634	0,841	10 161	10,16
Dist_{syn}	**0,831**	**0,549**	**0,661**	**0,950**	**0,645**	**0,847**	10 116	10,20
Idéal	0,847	0,788	0,816	0,957	0,831	0,899	13 468	7,66

Tableau 9. *Résultats du* clustering *sémantique, comparés au* clustering *de base*

La similarité distributionnelle syntaxique donne les meilleurs résultats, devant ceux de la similarité distributionnelle graphique. Les deux approches distributionnelles sont meilleures pour cette tâche que celle fondée sur WordNet, ce qui signifie que la méthode pourra plus facilement être adaptée à d'autres langues. Comparées au *clustering* de base, toutes les méthodes de *clustering* sémantique montrent une augmentation notable pour toutes les mesures (le F-score passe de 57,3 % à 66,1 %).

Pour les similarités WordNet, d'autres tests ont été effectués pour vérifier l'importance relative des différentes catégories grammaticales dans ce regroupement. Par exemple, si l'on ne considère que les verbes, les résultats sont un peu inférieurs, en particulier en termes de rappel. Nous avons également expérimenté l'intégration des adjectifs dans la mesure de similarité, mais les résultats ont montré que ces mots n'ont pas d'influence notable sur le regroupement des relations. D'autres tests intégrant des mesures de similarité entre des mots de catégories grammaticales différentes ont été effectués, sans apporter d'améliorations.

Pour donner une idée qualitative des résultats du *clustering* sémantique, nous présentons quelques exemples de clusters sémantiques, créés en utilisant la mesure Dist_{cooc}. Un exemple de cluster sémantique obtenu pour chaque type de relation est présenté dans le tableau 10, où chaque mot représente un cluster. Il est clair avec ces exemples que des mots différents mais sémantiquement similaires sont regroupés. Néanmoins, des erreurs subsistent : le fait de ne pas différencier les voies active et passive conduit ainsi à certaines erreurs de regroupement pour les relations entre des entités de même type (par exemple, *purchase* et *be purchased by* pour des relations ORG – ORG).

6.2.3. *Étude des avantages du* clustering *multiniveau*

Comme indiqué au début de la section 5.2, le calcul des similarités sémantiques est beaucoup plus coûteux que le calcul d'une simple mesure Cosinus. Le nombre total de relations atteint 165 708 (cf. tableau 5), alors que le nombre de clusters de base n'est que de 11 726 (cf. tableau 8). Un premier avantage du *clustering* multiniveau est donc d'éviter de calculer un trop grand nombre de similarités coûteuses et de permettre ainsi le passage à l'échelle. Mais, parallèlement, nous faisons également l'hypothèse qu'il permet d'améliorer la qualité de l'organisation sémantique des relations, en exploi-

Catégories	Clusters sémantiques
ORG – ORG	*purchase, buy, acquire, trade, own, be purchased by*
ORG – LIEU	*start in, inaugurate service to, open in, initiate flights to*
ORG – PERS	*sign, hire, employ, interview, rehire, receive, affiliate*
PERS – ORG	*take over, take control of*
PERS – LIEU	*grab gold in, win the race at, reign*
PERS – PERS	*win over, defeat, beat, oust, topple, defend*

Tableau 10. *Exemples de mots regroupés dans les clusters sémantiques*

tant la redondance d'information présente dans les clusters de base. Pour vérifier cette hypothèse, nous avons comparé, en nous appuyant sur notre référence, la distribution des mesures de similarité appliquées aux relations unitaires initiales et aux clusters de base. Dans un premier temps, nous avons calculé les valeurs de similarité entre chaque paire d'instances de relations appartenant au même cluster de référence (distribution intracluster D_{intra}) et les similarités entre les paires d'instances appartenant à des clusters différents (distribution intercluster D_{inter}), avec l'hypothèse que ces distributions sont bien séparées (avec une moyenne élevée pour D_{intra} et basse pour D_{inter}). Dans un second temps, nous établissons les mêmes distributions de similarité pour les clusters de base, en associant à chaque cluster de référence l'ensemble des clusters de base qu'il recouvre. Les distributions de similarité obtenues sont présentées à la figure 6 pour la similarité Dist_{cooc}, la même tendance étant observée pour les autres similarités.

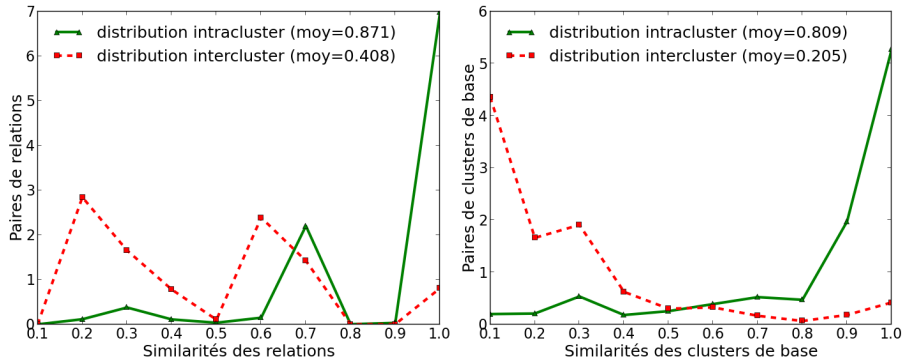


Figure 6. *Distribution des similarités entre les relations et entre les clusters de base*

On voit clairement sur ces figures que le *clustering* sémantique effectué à partir des clusters de base peut obtenir de meilleurs résultats parce que les distributions de similarité à l'intérieur des clusters de référence ou entre clusters sont mieux séparées et que la moyenne des similarités pour des relations entre des clusters différents est relativement basse. Ceci confirme notre hypothèse que l'information redondante dans

les clusters de base peut être utilisée pour diminuer le bruit causé par les mots non représentatifs de la relation.

6.3. *Évaluation du* clustering *thématique de relations*

Le dernier volet de notre évaluation a bien entendu porté sur le *clustering* thématique des relations de la section 5.4. Pour la mise en œuvre du *clustering* des contextes thématiques des relations, l'algorithme MCL a été appliqué avec un seuil empirique pour la mesure Cosinus égal à 0,15. L'évaluation proprement dite du regroupement thématique des relations a demandé la construction d'une référence spécifique en se focalisant sur un cluster sémantique et en répartissant ses instances de relations en fonction des différents thèmes caractérisant leur contexte d'occurrence. Cette référence a ainsi permis de juger de façon précise de l'impact de la structuration thématique du contenu des clusters sémantiques opérée par le *clustering* thématique. En pratique, nous avons annoté 65 instances de la relation *lead by* pour le type ORG – PERS. Ces instances ont été réparties manuellement en trois sous-groupes correspondant aux trois grands thèmes du contexte dans lequel elles apparaissaient : politique (30 instances), économie (21 instances) et sport (14 instances). L'évaluation, par rapport à cette référence, du résultat de l'application de la procédure de regroupement thématique au cluster sémantique *lead by* est donnée par le tableau 11.

	Préc.	**Rappel**	**F-score**	**Pur.**	**Pur. inv.**	**NMI**
Sémantique	0,362	**0,842**	**0,507**	0,477	**0,908**	0,127
Thématique	**0,400**	0,219	0,283	**0,723**	0,431	**0,348**

Tableau 11. *Résultats du* clustering *thématique de relations*

Ce tableau fait nettement apparaître une amélioration de la précision du regroupement des instances de relations, accompagnée d'une chute du rappel. La pureté, qui mesure la précision au niveau des clusters, est, quant à elle, significativement améliorée (passant de 0,477 à 0,723), de même que la mesure NMI (de 0,127 à 0,348). Cette amélioration globale des mesures de précision tend à confirmer l'intérêt de l'utilisation de l'information thématique pour invalider certains rapprochements opérés sur la base de sens différents de certains mots. Parallèlement, la chute des mesures de rappel suggère néanmoins que les clusters thématiques formés sont trop petits et opèrent certains distinguos trop spécifiques. Nous avons vérifié ce dernier point de manière plus qualitative en examinant comment les trois clusters thématiques de notre référence se répartissaient parmi les clusters formés par notre procédure de regroupement thématique. Le tableau 12 donne, pour chaque cluster de référence, quelques-uns de ces clusters formés, caractérisés par leurs mots les plus fréquents.

Ce tableau montre clairement que chaque grand thème se retrouve divisé en plusieurs sous-thèmes. Ainsi, un thème comme le sport se retrouve en pratique éclaté en sous-thèmes renvoyant à des sports particuliers, comme le base-ball, le basket-ball ou la boxe. La présence de mots partagés entre ces différents sports comme *game*, *play*

Thème	Mots caractéristiques
Politique	
1	***iraq american*** *official baghdad sunni* ***force military*** *kill bush* ***government*** *police*
2	***oil*** *price* ***energy*** *company* ***gas*** *state gasoline production bill saudi barrels* ***government***
3	***palestinian israel*** *gaza sharon* ***hamas*** *bank minister state* ***government security***
Économique	
1	*share* ***company*** *quarter oracle* ***earnings revenue*** *analyst report rise sales* ***stock*** *profit business*
2	***china japan trade company american*** *government world taiwan beijing market dollar* ***export***
3	*cell cancer* ***research human disease patient*** *study university breast* ***treatment*** *drug* ***medical health***
Sports	
1	*Sox* ***game*** *Yankee red* ***team season*** *run series play hit boston win pitch angel start world* ***league*** *manager* ***player***
2	*Bryant Lakers* ***game play*** *point* ***season*** *team O'Neal Odom jackson Kobe player coach quarter shot* ***NBA***
3	*stone Ruiz Toney* ***fight show*** *jagger world rolling* ***play win*** *title* ***champion*** *game heavyweight* ***boxing***

Tableau 12. *Mots caractéristiques des clusters thématiques au niveau de la référence pour le type de relation* lead by

ou *season* ne suffit pas en effet à les rassembler. De ce point de vue, on peut noter en particulier l'influence du nom des joueurs ou des équipes, comme les *Sox* et les *Yankee* pour le base-ball, les *Lakers* et *Bryant* pour le basket-ball ou *Ruiz* et *Toney* pour la boxe entre autres. Les différents sports impliquent également des actions et donc des verbes particuliers comme *hit* et *pitch* pour le base-ball, *shot* pour le basket-ball ou *fight* pour la boxe. L'ajout d'une information thématique permet donc de différencier ces différents sports mais la structurer de manière plus hiérarchique conduirait à gagner la capacité à opérer des regroupements plus larges de relations.

7. Travaux liés au *clustering* de relations

Le *clustering* de relations occupe des positions diverses dans le domaine de l'EI non supervisée. En premier lieu, il est absent des travaux se concentrant essentiellement sur la découverte et l'extraction de relations, à l'instar du système TEXTRUNNER dans lequel les relations extraites sont directement indexées pour être interrogées. Dans la plupart des autres travaux, la finalité du *clustering* de relations peut être qualifiée de sémantique dans la mesure où son objectif est de regrouper des relations équivalentes, ces équivalences étant situées plus ou moins explicitement sur le plan sémantique. Enfin, quelques travaux plus marginaux, à l'image de Sekine (2006), intègrent également une dimension plus thématique dans les regroupements réalisés.

Même lorsque le *clustering* de relations possède une vocation sémantique, les moyens pour le mettre en œuvre ne sont pas nécessairement eux-mêmes sémantiques. À l'image de notre premier niveau de *clustering*, Hasegawa *et al.* (2004) retrouvent ainsi des variations sémantiques comme *offer to buy* et *acquisition of* au sein des clusters de relations entre entités nommées qu'ils forment en appliquant une simple mesure Cosinus au contexte immédiat de ces relations. Sekine (2006) va, quant à lui, un peu plus loin en exploitant un ensemble de paraphrases constitué *a priori* sur la base de cooccurrences d'entités nommées pour faciliter l'appariement de phrases issues de plusieurs articles journalistiques relatant un même événement. Concernant toujours l'évaluation de la similarité entre les relations, Eichler *et al.* (2008) s'appuient, pour leur part, sur WordNet pour détecter les relations de synonymie entre verbes. La démarche se rapproche d'une partie de ce que nous avons expérimenté, avec toutefois certaines différences significatives : nous avons aussi inclus les noms dans notre champ d'étude car ceux-ci sont dominants pour exprimer certaines relations ; nous avons appliqué cette recherche au niveau des clusters de base, et non des relations individuelles ; enfin, avec les similarités distributionnelles, nous ne nous sommes pas restreints aux seules relations de synonymie.

La notion de *clustering* multiple apparaît, quant à elle, dans quelques travaux, avec généralement l'idée d'associer le *clustering* des différentes formes d'expression d'une relation et celui de ses arguments. Le fait de se concentrer sur des relations entre entités nommées dans notre cas permet en première instance de s'affranchir de la nécessité de regrouper les arguments des relations. Néanmoins, ces types d'entités étant très généraux, il est possible qu'un tel regroupement, au sein de chaque type, puisse améliorer les résultats, même s'il tendrait plutôt à favoriser la précision alors que les insuffisances se situent au niveau du rappel. Dans cette optique de *co-clustering* entre relations et arguments, Kok et Domingos (2008) proposent de construire un réseau de relations sémantiques de haut niveau à partir des résultats du système TEXTRUNNER grâce à une méthode de *co-clustering* engendrant simultanément des classes d'arguments et des classes de relations. En s'inscrivant dans cette même tendance, Bollegala *et al.* (2010) mettent particulièrement l'accent sur une représentation duale des relations : en extension, sous la forme d'un ensemble de couples d'entités liées, ou en intension, sous la forme d'un ensemble de patrons lexico-syntaxiques d'expressions. Le travail propose ainsi un algorithme spécifique de *co-clustering* exploitant une matrice de cooccurrences entre couples d'entités liées et patrons lexico-syntaxiques. Min *et al.* (2012) font, quant à eux, apparaître deux niveaux de *clustering* mais avec une optique plus proche de Kok et Domingos (2008) que de la nôtre. Leur premier niveau de *clustering* porte en effet sur les arguments des relations tandis que le second se focalise sur les relations proprement dites. L'objectif du premier niveau de *clustering* est ainsi de regrouper des relations ayant la même expression et de trouver des arguments équivalents tandis que le second niveau de *clustering* vise à regrouper des relations ayant des expressions similaires en s'appuyant notamment sur les classes d'arguments dégagées par le premier *clustering*. Ce dernier exploite un vaste graphe de relations de similarité et d'hyperonymie entre entités, construit automatiquement à la fois sur la base de similarités distributionnelles et de patrons lexico-syntaxiques. S'y ajoute pour

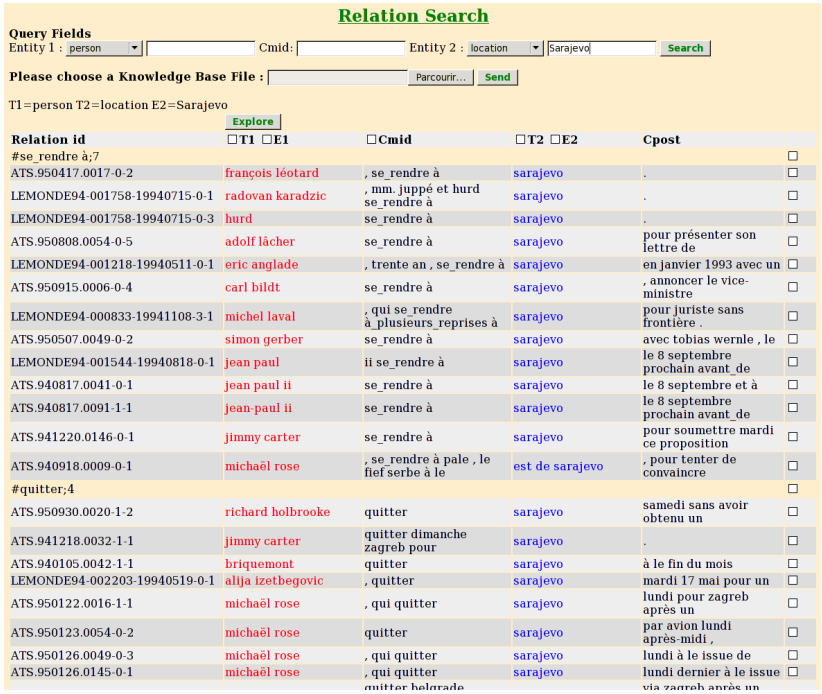
le second niveau de *clustering* une large base de paraphrases elle aussi construite automatiquement à partir de corpus. Cette exploitation de ressources distributionnelles rapproche pour partie ce travail du nôtre.

8. Conclusion et perspectives

Dans cet article, nous avons présenté un travail sur l'extraction d'information non supervisée en mettant l'accent sur une application à large échelle. Nous cherchons d'abord à déterminer si deux entités nommées apparaissant dans une même phrase sont en relation, sans *a priori* sur la nature de cette relation. Nous avons développé pour ce faire une procédure de filtrage par la combinaison d'heuristiques, pour éliminer efficacement les cas les plus simples, et d'un classifieur à base de CRF mettant en œuvre des critères plus élaborés. Concernant ce dernier, les performances obtenues, équilibrées en termes de précision et de rappel, se comparent favorablement aux résultats du système REVERB (Fader *et al.*, 2011) appliqué à notre corpus d'évaluation.

Nous avons également présenté dans cet article une méthode de *clustering* à deux niveaux pour regrouper des relations extraites dans un contexte d'EI non supervisée. Une première étape permet de regrouper des relations ayant des expressions linguistiques proches de façon efficace et avec une bonne précision. Une seconde étape améliore ce premier regroupement en utilisant des mesures de similarité sémantique plus riches afin de rassembler les clusters déjà formés et augmenter le rappel. Nos expériences montrent que dans ce contexte, des mesures de similarité distributionnelle donnent des résultats plus stables que des mesures fondées sur WordNet. Une analyse des distributions des similarités entre les relations initiales et entre les clusters de premier niveau confirme en outre l'intérêt d'un *clustering* à deux niveaux. Nous avons montré enfin que ce dernier peut être complété par un *clustering* de nature thématique, apportant à la fois un axe de structuration différent et une amélioration de la précision.

Cette amélioration de la précision se faisant néanmoins au prix d'une chute du rappel encore trop importante, des travaux complémentaires restent à mener concernant l'intégration des regroupements sémantique et thématique, notamment en considérant un *clustering* à plus gros grain des contextes thématiques des relations. Par ailleurs, la similarité sémantique des relations pourrait bénéficier de façon plus avancée des travaux menés sur l'identification des paraphrases, en intégrant notamment un ensemble plus large de critères. Enfin, le contexte applicatif de ce travail étant la veille, une évaluation utilisateur reste à mener de ce point de vue, évaluation qui pourrait se faire au travers d'un moteur de recherche sémantique orienté relation. Le prototype d'un tel moteur a déjà été développé, avec l'interface Web illustrée par la figure 7, et permet d'interroger l'ensemble des relations extraites d'un corpus en spécifiant des contraintes portant sur les entités (nom ou type) ou sur l'expression de la relation. La réponse à une telle interrogation se présente sous la forme d'une liste de relations, structurée selon les clusters sémantiques formés. Outre la nécessité d'ajouter la structuration des résultats selon la dimension thématique, un protocole d'évaluation



Relation Search

Query Fields
Entity 1 : person Cmid: Entity 2 : location Sarajevo Search

Please choose a Knowledge Base File : Parcourir... Send

T1=person T2=location E2=Sarajevo

Explore

Relation id	☐T1 ☐E1	☐Cmid	☐T2 ☐E2	Cpost	
#se_rendre à;7					☐
ATS.950417.0017-0-2	françois léotard	, se_rendre à	sarajevo	.	☐
LEMONDE94-001758-19940715-0-1	radovan karadzic	, mm. juppé et hurd se_rendre à	sarajevo	.	☐
LEMONDE94-001758-19940715-0-3	hurd	se_rendre à	sarajevo	.	☐
ATS.950808.0054-0-5	adolf lâcher	se_rendre à	sarajevo	pour présenter son lettre de	☐
LEMONDE94-001218-19940511-0-1	eric anglade	, trente an , se_rendre à	sarajevo	en janvier 1993 avec un	☐
ATS.950915.0006-0-4	carl bildt	se_rendre à	sarajevo	, annoncer le vice-ministre	☐
LEMONDE94-000833-19941108-3-1	michel laval	, qui se_rendre à_plusieurs_reprises à	sarajevo	pour juriste sans frontière .	☐
ATS.950507.0049-0-2	simon gerber	se_rendre à	sarajevo	avec tobias wernle , le	☐
LEMONDE94-001544-19940818-0-1	jean paul	ii se_rendre à	sarajevo	le 8 septembre prochain avant_de	☐
ATS.940817.0041-0-1	jean paul ii	se_rendre à	sarajevo	le 8 septembre et à	☐
ATS.940817.0091-1-1	jean-paul ii	se_rendre à	sarajevo	le 8 septembre prochain avant_de	☐
ATS.941220.0146-0-1	jimmy carter	se_rendre à	sarajevo	pour soumettre mardi ce proposition	☐
ATS.940918.0009-0-1	michaël rose	, se_rendre à pale , le fief serbe à le	est de sarajevo	, pour tenter de convaincre	☐
#quitter;4					☐
ATS.950930.0020-1-2	richard holbrooke	quitter	sarajevo	samedi sans avoir obtenu un	☐
ATS.941218.0032-1-1	jimmy carter	quitter dimanche zagreb pour	sarajevo	.	☐
ATS.940105.0042-1-1	briquemont	quitter	sarajevo	à le fin du mois	☐
LEMONDE94-002203-19940519-0-1	alija izetbegovic	, quitter	sarajevo	mardi 17 mai pour un	☐
ATS.950122.0016-1-1	michaël rose	, qui quitter	sarajevo	lundi pour zagreb après un	☐
ATS.950123.0054-0-2	michaël rose	quitter	sarajevo	par avion lundi après-midi ,	☐
ATS.950126.0049-0-3	michaël rose	, qui quitter	sarajevo	lundi à le issue de	☐
ATS.950126.0145-0-1	michaël rose	, qui quitter	sarajevo	lundi dernier à le issue	☐
		quitter belgrade		via zagreb après un	

Figure 7. *Moteur de recherche orienté relation*

spécifique à cette forme de recherche reste encore à définir en lien avec la notion de moteur de recherche sémantique (Guha *et al.*, 2003).

9. Bibliographie

Akbik A., Broß J., « Extracting Semantic Relations from Natural Language Text Using Dependency Grammar Patterns », *SemSearch 2009 workshop of WWW 2009*, 2009.

Amigó E., Gonzalo J., Artiles J., Verdejo F., « A comparison of extrinsic clustering evaluation metrics based on formal constraints », *Information Retrieval*, vol. 12, n° 4, p. 461-486, 2009.

Banko M., Cafarella M. J., Soderland S., Broadhead M., Etzioni O., « Open information extraction from the web », *IJCAI'07*, p. 2670-2676, 2007.

Banko M., Etzioni O., « The Tradeoffs Between Open and Traditional Relation Extraction », *48th Annual Meeting of the ACL : Human Language Technologies (ACL-08 : HLT)*, Columbus, Ohio, p. 28-36, 2008.

Bayardo R. J., Ma Y., Srikant R., « Scaling up all pairs similarity search », *WWW'07*, p. 131-140, 2007.

Bollegala D. T., Matsuo Y., Ishizuka M., « Relational duality : unsupervised extraction of semantic relations between entities on the web », *19th International Conference on World Wide Web (WWW2010)*, Raleigh, North Carolina, USA, p. 151-160, 2010.

Brin S., « Extracting Patterns and Relations from the World Wide Web », *WebDB'98*, p. 172-183, 1998.

Chen J., Ji D., Tan C., Niu Z., « Unsupervised Feature Selection for Relation Extraction », *IJCNLP-2005*, p. 262-267, 2005.

Cheu E., Keongg C., Zhou Z., « On the two-level hybrid clustering algorithm », *International conference on artificial intelligence in science and technology*, p. 138-142, 2004.

Dolan B., Quirk C., Brockett C., « Unsupervised construction of large paraphrase corpora : exploiting massively parallel news sources », *COLING'04*, 2004.

Dongen S. V., Graph Clustering by Flow Simulation, PhD thesis, University of Utrecht, 2000.

Eichler K., Hemsen H., Neumann G., « Unsupervised Relation Extraction From Web Documents », *LREC'08*, 2008.

Embarek M., Ferret O., « Learning patterns for building resources about semantic relations in the medical domain », *6th Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Ertöz L., Steinbach M., Kumar V., « A New Shared Nearest Neighbor Clustering Algorithm and its Applications », *Workshop on Clustering High Dimensional Data and its Applications of SIAM ICDM 2002*, 2002.

Fader A., Soderland S., Etzioni O., « Identifying relations for open information extraction », *2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, p. 1535-1545, 2011.

Ferret O., « Testing Semantic Similarity Measures for Extracting Synonyms from a Corpus », *LREC'10*, 2010.

Galley M., McKeown K., Fosler-Lussier E., Jing H., « Discourse Segmentation of Multi-party Conversation », *41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, p. 562-569, 2003.

Gamallo P., Garcia M., Fernández-Lanza S., « Dependency-Based Open Information Extraction », *Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*, 2012.

Grishman R., Min B., « New York University KBP 2010 Slot-Filling System », *Text Analysis Conference (TAC)*, NIST, 2010.

Guha R., McCool R., Miller E., « Semantic search », *12th International Conference on World Wide Web (WWW'03)*, p. 700-709, 2003.

Hasegawa T., Sekine S., Grishman R., « Discovering relations among named entities from large corpora », *ACL'04*, 2004.

Kok S., Domingos P., « Extracting Semantic Networks from Text Via Relational Clustering », *ECML PKDD'08*, p. 624-639, 2008.

Lavergne T., Cappé O., Yvon F., « Practical Very Large Scale CRFs », *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, p. 504-513, 2010.

Li Y., Jiang J., Chieu H. L., Chai K. M. A., « Extracting Relation Descriptors with Conditional Random Fields », *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, Chiang Mai, Thailand, p. 392-400, 2011.

Lin D., « An Information-Theoretic Definition of Similarity », *ICML'98*, Morgan Kaufmann Publishers Inc., p. 296-304, 1998.

Mausam, Schmitz M., Soderland S., Bart R., Etzioni O., « Open Language Learning for Information Extraction », *2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2012)*, Jeju Island, Korea, p. 523-534, 2012.

Mihalcea R., Corley C., Strapparava C., « Corpus-based and knowledge-based measures of text semantic similarity », *AAAI'06*, p. 775-780, 2006.

Min B., Shi S., Grishman R., Lin C.-Y., « Ensemble Semantics for Large-scale Unsupervised Relation Extraction », *EMNLP'12*, 2012.

Mintz M., Bills S., Snow R., Jurafsky D., « Distant supervision for relation extraction without labeled data », *ACL-IJCNLP 2009*, p. 1003-1011, 2009.

Pedersen T., « Information Content Measures of Semantic Similarity Perform Better Without Sense-Tagged Text », *HLT-NAACL'10*, p. 329-332, 2010.

Ramshaw L., Marcus M., « Text Chunking Using Transformation-Based Learning », *Third Workshop on Very Large Corpora*, Cambridge, Massachusetts, USA, p. 82-94, 1995.

Rink B., Harabagiu S., « A generative model for unsupervised discovery of relations and argument classes from clinical texts », *EMNLP'11*, p. 519-528, 2011.

Rodriquez K. J., Bryant M., Blanke T., Luszczynska M., « Comparison of Named Entity Recognition tools for raw OCR text », *in* J. Jancsary (ed.), *11^{th} Conference on Natural Language Processing (KONVENS 2012)*, Vienna, Austria, p. 410-414, 2012.

Rosenfeld B., Feldman R., « Clustering for unsupervised relation identification », *Sixteenth ACM conference on Conference on information and knowledge management (CIKM'07)*, Lisbon, Portugal, p. 411-418, 2007.

Rozenfeld B., Feldman R., « High-Performance Unsupervised Relation Extraction from Large Corpora », *ICDM'06*, p. 1032-1037, 2006.

Sekine S., « On-demand information extraction », *COLING-ACL'06*, p. 731-738, 2006.

Shinyama Y., Sekine S., « Preemptive information extraction using unrestricted relation discovery », *HLT-NAACL'06*, p. 304-311, 2006.

Wang W., Besançon R., Ferret O., Grau B., « Filtering and Clustering Relations for Unsupervised Information Extraction in Open Domain », *20^{th} ACM international Conference on Information and Knowledge Management (CIKM 2011)*, p. 1405-1414, 2011.

Wang W., Besançon R., Ferret O., Grau B., « Evaluation of Unsupervised Information Extraction », *LREC'12*, 2012.

Wang W., Besançon R., Ferret O., Grau B., « Regroupement sémantique de relations pour l'extraction d'information non supervisée », *$20^{ème}$ Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, Les Sables-d'Olonne, France, p. 353-366, 2013.

Wu F., Weld D. S., « Open Information Extraction Using Wikipedia », *48^{th} Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, Uppsala, Sweden, p. 118-127, 2010.

Wu Z., Palmer M., « Verbs semantics and lexical selection », *ACL'94*, p. 133-138, 1994.

Yao L., Haghighi A., Riedel S., McCallum A., « Structured relation discovery using generative models », *EMNLP'11*, p. 1456-1466, 2011.