Regroupement sémantique de relations pour l'extraction d'information non supervisée

Wei Wang[1] Romaric Besançon[1] Olivier Ferret[1] Brigitte Grau[2]
(1) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, Gif-sur-Yvette, F-91191 France.
(2) LIMSI, UPR-3251 CNRS-DR4, Bât. 508, BP 133, 91403 Orsay Cedex.
{wei.wang,romaric.besancon,olivier.ferret}@cea.fr brigitte.grau@limsi.fr

RÉSUMÉ
Beaucoup des recherches menées en extraction d'information non supervisée se concentrent sur l'extraction des relations et peu de travaux proposent des méthodes pour organiser les relations extraites. Nous présentons dans cet article une méthode de clustering en deux étapes pour regrouper des relations sémantiquement équivalentes : la première étape regroupe des relations proches par leur expression tandis que la seconde fusionne les premiers clusters obtenus sur la base d'une mesure de similarité sémantique. Nos expériences montrent en particulier que les mesures distributionnelles permettent d'obtenir pour cette tâche de meilleurs résultats que les mesures utilisant WordNet. Nous montrons également qu'un clustering à deux niveaux permet non seulement de limiter le nombre de similarités sémantiques à calculer mais aussi d'améliorer la qualité des résultats du clustering.

ABSTRACT
Semantic relation clustering for unsupervised information extraction

Most studies in unsupervised information extraction concentrate on the relation extraction and few work has been proposed on the organization of the extracted relations. We present in this paper a two-step clustering procedure to group semantically equivalent relations : a first step clusters relations with similar expressions while a second step groups these first clusters into larger semantic clusters, using different semantic similarities. Our experiments show the stability of distributional similarities over WordNet-based similarities for semantic clustering. We also demonstrate that the use of a multi-level clustering not only reduces the calculations from all relation pairs to basic clusters pairs, but it also improves the clustering results.

MOTS-CLÉS : Extraction d'Information Non Supervisée, Similarité Sémantique, Clustering.

KEYWORDS: Unsupervised Information Extraction, Semantic Similarity, Relation Clustering.

1 Introduction

Dans le domaine de l'Extraction d'Information (EI), les problématiques ont évolué sous l'impulsion d'une série de campagnes d'évaluation allant de MUC (*Message Understanding Conference*) à TAC (*Text Analysis Conference*) en passant par ACE (*Automatic Content Extraction*). Les tâches définies dans les campagnes MUC et ACE 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 développés manuellement ou par

apprentissage automatique peuvent être développés. Les approches semi-supervisées peuvent s'affranchir partiellement des contraintes de disponibilité de telles données. Par exemple, pour la tâche KBP (*Knowledge Base Population*) de la campagne TAC, 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. Les méthodes semi-supervisées incluent également des techniques d'amorçage (*bootstrapping*) (Grishman et Min, 2010) permettant de partir d'un nombre limité d'exemples pour en extraire d'autres.

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 en général des relations binaires, 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 était 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 pour représenter les relations au sein 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).

Des systèmes tels que TextRunner (Banko *et al.*, 2007) ou ReVerb (Fader *et al.*, 2011) se focalisent quant à eux sur l'extraction de relations à partir de phrases en s'appuyant sur un modèle d'apprentissage statistique pour garantir la validité des relations extraites. 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 restant 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.

Une part notable des travaux menés en EI non supervisée se focalisent sur l'extraction des relations. Le problème de leur regroupement a été en revanche moins abordé, en particulier pour rassembler des relations équivalentes mais exprimées de façon différente. Nous présentons dans cet article une méthode pour réaliser de tels regroupements efficacement en se fondant sur deux étapes de clustering : un premier niveau de regroupement des relations sur la forme, utilisant une mesure de similarité simple, et un second niveau permettant de rapprocher les premiers clusters obtenus en utilisant une mesure de similarité sémantique plus sophistiquée. Nos expériences montrent que ce clustering à deux niveaux permet d'améliorer le regroupement des relations.

2 Extraction de relations non supervisée

La première étape de notre processus d'EI non supervisée est l'extraction de relations entre entités. Nous avons défini pour ce faire un module d'extraction et de filtrage de relations entraîné pour la découverte de relations entre entités nommées. Plus formellement, une relation entre

entités nommées se caractérise par un couple d'entités (E1 et E2) et la caractérisation linguistique de la relation, elle-même formée des trois éléments du contexte phrastique autour de ces entités (*cf.* figure 1) : la caractérisation linguistique principale de la relation est en général portée par la partie de texte entre les entités (*Cmid*), alors que les éléments de chaque côté des entités (*Cpre* et *Cpost*) apportent en général des précisions de contexte.

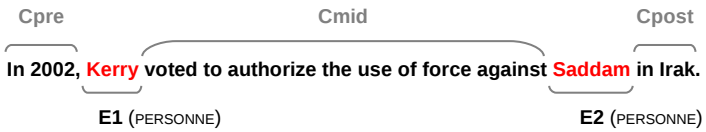


FIGURE 1 – Exemple de relation extraite

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 système aux relations entre entités nommées, à la fois pour faciliter l'organisation des relations trouvées et pour répondre au besoin le plus généralement répandu en contexte applicatif de veille. L'extraction des relations se fait alors selon les étapes suivantes :

– **Analyse linguistique** : un traitement linguistique est tout d'abord appliqué aux textes du corpus considéré pour extraire les éléments pouvant caractériser les relations candidates. Ce traitement inclut une reconnaissance des entités nommées pour les types impliqués dans les relations recherchées, mais aussi une désambiguïsation morpho-syntaxique et une lemmatisation pour normaliser les contextes linguistiques des relations. Ce traitement a été réalisé avec les outils OpenNLP ;
– **Extraction de relations candidates** : une première extraction simple est réalisée avec peu de contraintes pour permettre la collecte d'une grande variété de relations. Toutes les phrases contenant deux entités nommées sont donc extraites, avec la seule condition qu'au moins un verbe existe entre ces entités ;
– **Filtrage de relations** : à l'issue de l'extraction initiale, beaucoup de relations candidates ne sont pas des instances réelles de relations. Une étape de filtrage est alors appliquée, comprenant une première passe de filtrage heuristique, pour supprimer efficacement les relations les plus probablement fausses (discours rapporté, phrases complexes), et une seconde passe de filtrage par apprentissage statistique, entraîné sur un corpus annoté de 1 000 exemples positifs et négatifs de relations, et s'appuyant sur un modèle de Champs Conditionnels Aléatoires (CRF). Ce filtrage statistique permet d'obtenir une précision de 76,2% et un rappel de 78,2% sur les relations extraites (Wang *et al.*, 2011).

Pour nos expériences, nous avons utilisé une sous-partie du corpus AQUAINT-2 contenant 18 mois d'articles de presse du journal *New York Times*. Les relations candidates ont été extraites et filtrées selon la méthode présentée pour six types de relations fondées sur trois types d'entités nommées faisant consensus : les organisation (ORG), les lieux (LOC) et les personnes (PER). Le nombre des relations restant après filtrage, présenté dans le tableau 1, montre la nécessité de mettre en œuvre un regroupement de ces relations pour aider un utilisateur à appréhender les informations extraites.

Total	ORG-LOC	ORG-ORG	ORG-PER	PER-LOC	PER-ORG	PER-PER
165 708	15 226	13 704	10 054	47 700	40 238	38 786

TABLE 1 – Nombre de relations extraites

3 Regroupement de relations

Dans cette section, nous présentons plus spécifiquement notre méthode de clustering multi-niveau définie afin de regrouper les relations extraites en fonction de leur similarité sémantique. Cette méthode s'organise en deux étapes, à l'instar 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 ; une seconde étape de clustering est ensuite appliquée pour rassembler ces clusters initiaux sur la base d'une similarité sémantique entre relations plus complexe.

3.1 Regroupement de base

3.1.1 Principe

En EI non supervisée, le nombre de relations extraites est rapidement important comme le montre le tableau 1. De ce fait, il est quasiment impossible d'appliquer des mesures de similarité sémantique élaborées entre toutes les relations extraites. Le tableau 2 illustre cependant le fait que certaines variabilités d'expression sont très limitées et peuvent être détectées facilement.

Type de relation	**Clusters de base**
ORG – ORG	create the, who create ...
	establish the, who establish the ...
ORG – LOC	base in, a company base in ...
	locate in, which be locate in ...
ORG – PER	found by, a group found by, which be found by ...
PER – ORG	who be the head of, become head of ...
PER – LOC	work in, who work in ...
PER – PER	who call, who call his manager ...

TABLE 2 – Illustration de la variabilité linguistique des relations

Cette observation nous a conduit à mettre 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. Outre son compromis intéressant entre simplicité et efficacité, ce choix a été motivé par la possibilité d'appliquer cette similarité aux larges ensembles de relations extraites dans notre contexte par une utilisation de l'algorithme *All Pairs Similarity Search* (APSS) (Bayardo *et al.*, 2007). Moyennant la fixation *a priori* d'un seuil de similarité minimale, celui-ci permet en effet de construire de façon optimisée la matrice de similarité d'un ensemble de vecteurs suivant la mesure *Cosinus*. Cette matrice étant calculée et transformée en graphe de

similarité, nous appliquons ensuite l'algorithme *Markov Clustering* (Dongen, 2000) pour former les regroupements de relations. Cet algorithme identifie les zones d'un graphe de similarité les plus densément connectées en réalisant des marches aléatoires dans ce graphe. Outre son efficacité, il présente l'avantage, du point de vue de l'IE non supervisée, de ne pas nécessiter la fixation préalable d'un nombre de clusters.

3.1.2 Pondération des termes

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 bonne stratégie de pondération pour établir une bonne mesure de similarité entre phrases. Trois types de pondération sont considérés ici :
– pondération binaire : tous les mots de $Cmid$ ont le même poids (1,0) ;
– pondération *tf-idf* : un poids *tf-idf* est attribué à chaque mot en 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 morpho-syntaxique.

La pondération binaire est la plus simple et forme une *baseline*, qui a été utilisée dans nos premières expériences, en particulier en raison de l'efficacité de l'implémentation de l'APSS avec un poids binaire. 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 nécessairement 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 n'en sera pas moins représentatif de la relation BUY(ORG-ORG). C'est pourquoi nous avons décidé d'introduire une pondération grammaticale.

Classe	Catégories morpho-syntaxiques
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

TABLE 3 – Pondération grammaticale : distribution des poids selon la catégorie morpho-syntaxique

Une analyse des catégories morpho-syntaxiques nous a amené à les séparer en plusieurs classes selon leur importance dans la contribution à l'expression d'une relation. Plus précisément, nous considérons quatre classes :
– **(A) contribution directe**, de poids élevé : les mots de cette classe contribuent directement au sens de la relation et incluent les verbes, noms, adjectifs et 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. C'est le cas des noms propres, qui sont souvent discriminants d'un point de vue thématique mais ont plutôt à introduire des associations inadéquates sur le plan sémantique ;
– **(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 3 une configuration de pondération grammaticale. La liste des catégories morpho-syntaxiques est fondée sur les catégories du *Penn Treebank*. Les poids 1,0, 0,75, 0,5 et 0 sont respectivement attribués aux classes A, B, C, D. Pour les catégories non présentes dans cette liste, un poids par défaut de 0,5 est utilisé. Compte tenu des problèmes posés par l'évaluation de la tâche considérée (cf. section 4), ces poids n'ont pas fait l'objet d'une optimisation telle qu'elle pourrait être menée avec une procédure de type validation croisée.

3.1.3 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 relations. Au sein d'un cluster de base, la forme linguistique de ces dernières est 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, à l'instar de (Hasegawa *et al.*, 2004), nous considérons le nom ou le verbe le plus fréquent au sein d'un cluster de base comme son représentant et nous fusionnons les clusters ayant le même terme dominant, appelé *mot-clé* dans ce qui suit, pour former des clusters de base plus larges.

3.2 Regroupement sémantique

Le premier niveau de clustering ne peut clairement pas regrouper des relations exprimées avec des termes complètement différents. Dans l'exemple *a company based in* et *which is located in* présenté dans le tableau 2, les deux formes linguistiques ont peu en commun. Nous avons donc considéré l'ajout d'un second 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 second 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. 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 de relations.

3.2.1 Évaluation de la 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 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 (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 (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

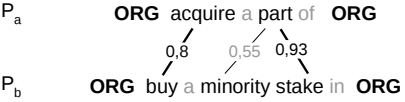
dans un corpus de référence. Ces mesures étant définies entre synsets, pour se ramener à une mesure entre mots, nous avons adopté la stratégie utilisée notamment dans (Mihalcea *et al.*, 2006) consistant à prendre comme valeur de similarité entre deux mots la plus forte valeur de similarité entre les 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 à l'ensemble des mots cooccurrant avec le mot cible dans un corpus. Cette cooccurrence peut être graphique, au sein d'une fenêtre de taille fixe, ou bien reposer sur des relations syntaxiques. Nous avons testé ici les deux types de cooccurrents, les termes au sein des contextes ainsi formés étant pondérés grâce à la mesure d'*Information Mutuelle* et les contextes eux-mêmes étant comparés grâce à la mesure *Cosinus* pour évaluer la similarité de deux mots. Ces choix résultent d'un processus d'optimisation décrit dans (Ferret, 2010) dont nous avons utilisé les thésaurus distributionnels 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 morpho-syntaxique en nous fondant sur le fait que les relations extraites se définissent généralement autour d'un verbe (*e.g. ORG found by PER*, *ORG establish by PER*) 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 relations, sans doute à cause de la focalisation sur la partie *Cmid* des relations.

3.2.2 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 3.2.1, 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 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, avec :

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

où W_k est un mot d'une phrase et f_k, sa fréquence dans la phrase, cette similarité s'écrit :

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

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 ($w_i = f_i, w_j = f_j$).

3.2.3 Similarité sémantique des clusters

Le principe adopté pour la similarité de deux relations est trop coûteux à transposer à l'échelle des clusters car il nécessiterait, pour un cluster C_a de cardinalité A et un cluster C_b de cardinalité B, de calculer $A \cdot B$ 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, obtenue en fusionnant les représentations de leurs relations. Au sein de la représentation d'un cluster, chaque mot se voit associer 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.

Concernant l'évaluation de la similarité entre les clusters, nous avons donc repris la définition de la similarité entre les relations mais avec une légère adaptation destinée à pallier le biais pouvant être induit par une trop grande différence d'effectifs entre les deux clusters. Ainsi, dans l'exemple ci-dessous, les clusters C_a et C_b ne sont pas sémantiquement similaires mais leur similarité serait élevée avec une mesure telle que $S_{P_{a,b}}$ du fait du poids élevé du mot *actor* dans C_a. Même si dans un tel cas, $sim(P_b, P_a)$ serait plus faible que $sim(P_a, P_b)$, $sim(P_b, P_a)$ influencerait fortement la moyenne des deux et conduirait à une similarité globale assez forte.

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

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 $w_{ij} = f_i \cdot f_j$.

3.2.4 Algorithme de clustering

Pour la construction de nos clusters de base, nous avons fait appel à l'association d'un seuillage sur les valeurs de similarité entre relations au travers de l'utilisation de l'APSS et de l'algorithme Markov Clustering. 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. Le cas du regroupement sémantique des clusters de base est quelque peu différent. Dans le cas des relations, 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 4.1 pour une illustration).

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 richesse 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[1].

4 Évaluation

Nous avons mené l'évaluation de ce clustering de relations multi-niveau selon une approche externe en utilisant les mesures standard de *précision* et *rappel* (combinés 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* and *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).

4.1 Évaluation du clustering de base

Le seuil de similarité utilisé pour le clustering de base (utilisé 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 paraphrase. Pour la pondération grammaticale, qui est moins stricte, un seuil de 0,60 est utilisé. Les résultats obtenus pour le clustering de base sont présentés dans le tableau 4.

	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

TABLE 4 – Résultats du clustering de base pour plusieurs pondérations en utilisant le Markov Clustering (MCL) et un premier regroupement par mots-clés

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

[1]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.

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[2], sur laquelle l'étape de regroupement par mots-clés amène 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 4.

4.2 Évaluation du clustering sémantique

Pour évaluer l'amélioration apportée par le clustering sémantique, nous comparons les approches proposées à un clustering idéal (*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 pratique, pour les mesures fondées sur WordNet, la mesure de Wu-Palmer donne de bons résultats pour les similarités entre noms alors que la mesure de Lin donne de meilleurs résultats pour les verbes. La première est calculée grâce à NLTK (`nltk.org`) tandis que pour la seconde, nous utilisons les similarités précalculées entre les verbes de WordNet de (Pedersen, 2010). Les similarités distributionnelles sont quant à elles évaluées 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}). Pour l'algorithme SNN, le voisinage de chaque instance de relation est limité aux 100 plus proches relations. Les résultats obtenus sont présentés dans le tableau 5.

	Préc.	**Rappel**	**F-score**	**Pur.**	**Pur. inv.**	**NMI**	**Nb**	**Taille**
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

TABLE 5 – Résultats du clustering sémantique

La similarité distributionnelle syntaxique donne les meilleurs résultats, bien que comparables à deux 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. Comparés au clustering de base, toutes les méthodes de clustering sémantique montrent une augmentation notable sur toutes les mesures (le F-score passe de 57,3% à 77,3%).

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

[2]Plusieurs seuils et configurations de pondérations grammaticales ont été testés. La version présentée (seuil de 0,60 et poids donnés dans le tableau 3) est celle donnant les meilleurs résultats.

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 similarités entre mots de catégories grammaticales différentes ont été effectués, sans apporter d'amélioration.

Exemples de clusters sémantiques 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 6, 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).

Type de relation	Clustering sémantique
ORG – ORG	purchase, buy, acquire, trade, own, be purchased by
ORG – LOC	start in, inaugurate service to, open in, initiate flights to
ORG – PER	sign, hire, employ, interview, rehire, receive, affiliate
PER – ORG	take over, take control of
PER – LOC	grab gold in, win the race at, reign
PER – PER	win over, defeat, beat, oust, topple, defend

TABLE 6 – Exemples de mots regroupés dans les clusters sémantiques

4.3 Étude des avantages du clustering multi-niveau

Comme indiqué au début de la section 3.1, 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 1), alors que le nombre de clusters de base n'est que de 11 726 (cf. tableau 4). Un premier avantage du clustering multi-niveau est donc d'éviter de calculer un trop grand nombre de similarités coûteuses. Mais, parallèlement, il permet également d'améliorer la qualité de l'organisation sémantique des relations, en exploitant 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 similarités entre les relations initiales et entre les clusters de base. Dans un premier temps, nous avons examiné toutes les similarités entre deux instances de relations appartenant au même cluster de référence (distribution intra-cluster D_{intra}) et les similarités entre deux instances appartenant à des clusters différents (distribution intra-cluster 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 similarités 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 2 pour la similarité Dist_{cooc}, la même tendance étant observée pour les autres similarités.

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

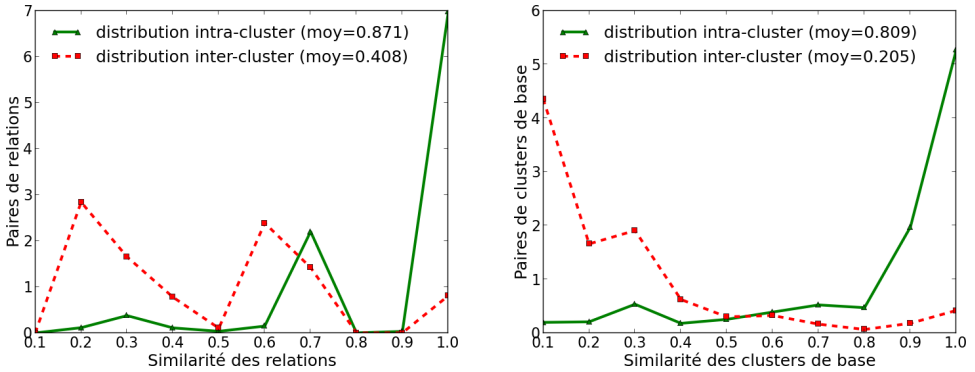


FIGURE 2 – Distribution des similarités entre les relations et entre les clusters de base

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.

5 Travaux liés au clustering sémantique 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, cette équivalence étant située 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) retrouve ainsi des variations sémantiques comme (*offer to buy – acquisition of*) au sein des clusters de relations entre entités nommées qu'il forme 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'appuie pour sa part sur WordNet pour détecter les relations de synonymie entre verbes. La démarche se rapproche d'une partie de ce que nous avons expérimenté, même si nous avons également inclus les noms dans notre champ d'étude, car ceux-ci sont dominants pour exprimer certaines relations, que nous avons appliqué cette recherche au niveau des clusters de base, et non des relations individuelles, et qu'avec les similarités distributionnelles, nous ne sommes pas restreints aux seules relations de synonymie.

La notion de clustering multiple apparaît quant à elle dans quelques travaux. (Kok et Domingos, 2008) propose ainsi 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 simulta-

nément des classes d'arguments et des classes de relations. (Min *et al.*, 2012) fait quant à lui apparaître deux niveaux de clustering mais avec une optique plus proche de (Kok et Domingos, 2008) que de la nôtre. Son 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.

Conclusion et perspectives

Nous avons présenté dans cet article une méthode de clustering à plusieurs niveaux pour regrouper des relations extraites dans un contexte d'EI non supervisée. Une première étape est appliquée pour regrouper des relations ayant des expressions linguistiques proches de façon efficace et avec une bonne précision. Une seconde étape permet d'améliorer 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 met également en évidence l'intérêt d'un clustering à deux niveaux. Parmi les perspectives envisagées, nous envisageons d'exploiter le contexte des relations, que ce soit de façon locale au niveau de la phrase au travers des parties *Cpre* et *Cpost* ou plus globalement en prenant en compte les contextes thématiques des relations pour améliorer le regroupement des relations et pouvoir les présenter de façon plus pertinente à un utilisateur.

Références

Akbik, A. et Bross, J. (2009). Extracting semantic relations from natural language text using dependency grammar patterns. *In SemSearch 2009 workshop of WWW 2009*.

Amigó, E., Gonzalo, J., Artiles, J. et Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 12(4):461–486.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M. et Etzioni, O. (2007). Open information extraction from the web. *In IJCAI'07*, pages 2670–2676.

Bayardo, R. J., Ma, Y. et Srikant, R. (2007). Scaling up all pairs similarity search. *In WWW'07*, pages 131–140.

Chen, J., Ji, D., Tan, C. et Niu, Z. (2005). Unsupervised feature selection for relation extraction. *In IJCNLP-2005*, pages 262–267.

Cheu, E., Keongg, C. et Zhou, Z. (2004). On the two-level hybrid clustering algorithm. *In International conference on artificial intelligence in science and technology*, pages 138–142.

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

DONGEN, S. V. (2000). *Graph Clustering by Flow Simulation*. Thèse de doctorat, University of Utrecht.

EICHLER, K., HEMSEN, H. et NEUMANN, G. (2008). Unsupervised relation extraction from web documents. *In LREC'08*.

ERTÖZ, L., STEINBACH, M. et KUMAR, V. (2002). A new shared nearest neighbor clustering algorithm and its applications. *In Workshop on Clustering High Dimensional Data and its Applications of SIAM ICDM 2002*.

FADER, A., SODERLAND, S. et ETZIONI, O. (2011). Identifying relations for open information extraction. *In EMNLP'11*, pages 1535–1545.

FERRET, O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. *In LREC'10*.

GAMALLO, P., GARCIA, M. et FERNÁNDEZ-LANZA, S. (2012). Dependency-based open information extraction. *In Joint Workshop on Unsupervised and Semi-Supervised Learning in NLP*.

GRISHMAN, R. et MIN, B. (2010). New York University KBP 2010 Slot-Filling System. *In Text Analysis Conference (TAC)*. NIST.

HASEGAWA, T., SEKINE, S. et GRISHMAN, R. (2004). Discovering relations among named entities from large corpora. *In ACL'04*.

KOK, S. et DOMINGOS, P. (2008). Extracting Semantic Networks from Text Via Relational Clustering. *In ECML PKDD'08*, pages 624–639.

LIN, D. (1998). An information-theoretic definition of similarity. *In ICML'98*, pages 296–304.

MIHALCEA, R., CORLEY, C. et STRAPPARAVA, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *In AAAI'06*, pages 775–780.

MIN, B., SHI, S., GRISHMAN, R. et LIN, C.-Y. (2012). Ensemble semantics for large-scale unsupervised relation extraction. *In EMNLP'12*, pages 1027–1037.

MINTZ, M., BILLS, S., SNOW, R. et JURAFSKY, D. (2009). Distant supervision for relation extraction without labeled data. *In ACL-IJCNLP 2009*, pages 1003–1011.

PEDERSEN, T. (2010). Information content measures of semantic similarity perform better without sense-tagged text. *In HLT-NAACL'10*, pages 329–332.

RINK, B. et HARABAGIU, S. (2011). A generative model for unsupervised discovery of relations and argument classes from clinical texts. *In EMNLP'11*, pages 519–528.

ROZENFELD, B. et FELDMAN, R. (2006). High-performance unsupervised relation extraction from large corpora. *In ICDM'06*, pages 1032–1037.

SEKINE, S. (2006). On-demand information extraction. *In COLING-ACL'06*, pages 731–738.

SHINYAMA, Y. et SEKINE, S. (2006). Preemptive information extraction using unrestricted relation discovery. *In HLT-NAACL'06*, pages 304–311.

WANG, W., BESANÇON, R., FERRET, O. et GRAU, B. (2011). Filtering and clustering relations for unsupervised information extraction in open domain. *In CIKM 2011*, pages 1405–1414.

WANG, W., BESANÇON, R., FERRET, O. et GRAU, B. (2012). Evaluation of unsupervised information extraction. *In LREC'12*.

WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. *In ACL'94*, pages 133–138.

YAO, L., HAGHIGHI, A., RIEDEL, S. et MCCALLUM, A. (2011). Structured relation discovery using generative models. *In EMNLP'11*, pages 1456–1466.