Regroupement 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, 91191 Gif-sur-Yvette
{wei.wang, romaric.besancon, olivier.ferret}@cea.fr
(2) LIMSI, UPR-3251 CNRS-DR4, Bât. 508, BP 133, 91403 Orsay Cedex
brigitte.grau@limsi.fr

RÉSUMÉ. *En contexte de veille, l'extraction d'information non supervisée a pour but d'extraire des relations entre entités sans fournir de connaissances* a priori *sur les natures de ces relations. Dans ce contexte, nous nous intéressons à l'identification et la caractérisation de nouvelles relations entre des types d'entités fixés. Nous présentons dans cet article une procédure de filtrage de relations combinant des méthodes heuristiques et des méthodes par apprentissage, permettant d'atteindre un score de F-mesure de 77,1%. Nous présentons ensuite une méthode de regroupement des relations extraites combinant un calcul optimisé des similarités entre les relations (All Pairs Similarity Search) et un algorithme de clustering (Markov Clustering). Une évaluation de ce regroupement, effectuée grâce à des mesures internes et externes, montre que l'utilisation du filtrage permet de doubler le rappel en conservant une précision équivalente.*

ABSTRACT. *The purpose of unsupervised information extraction is to extract information from text without fixing the type of information. Our work concentrates on the task of extracting and characterizing new relations between given entity types. We first propose in this article a filtering procedure to remove false relation candidates by combining heuristics and machine learning models. Best results achieve a score of 77.1% F-measure. Similar relations are then grouped together semantically using Markov Clustering and All Pairs Similarity Search algorithm, which can efficiently identify similar candidates in large scale. Finally, evaluations of clustering results, using both internal and external measures, show that the integration of the filtering step allows to double the recall while keeping the same precision.*

MOTS-CLÉS : *Extraction d'information non supervisée, filtrage de relations, clustering*

KEYWORDS: *Unsupervised information extraction, relations filtering, clustering*

CORIA 2012, pp. 191–202, Bordeaux, 21-23 mars 2012

1. Introduction

L'extraction d'information, longtemps envisagée sous le seul angle du paradigme des conférences MUC (Message Understanding Conferences) (Grishman *et al.,* 1995), revêt depuis quelques années des formes plus diverses. À côté de la tâche consistant à extraire de textes les éléments d'information venant occuper un rôle bien défini dans une structure informationnelle donnée *a priori* (souvent appelée *template*), sont apparues des tâches moins contraintes, notamment du point de vue de la spécification des informations à extraire. Traditionnellement, celles-ci sont décrites sous la forme d'une configuration de relations entre entités, chaque relation étant spécifiée par un modèle élaboré manuellement (généralement sous la forme d'un ensemble de règles) ou par un ensemble d'exemples de relations en contexte permettant d'apprendre un modèle, souvent de nature statistique. Cette approche peut être qualifiée globalement de dirigée par les buts ou de supervisée. Sans remettre fondamentalement en cause ce dernier caractère, elle peut également prendre des formes moins contraintes. Les travaux fondés sur l'amorçage où les relations sont spécifiées initialement par un nombre très limité d'exemples ou de patrons linguistiques (Agichtein *et al.,* 2000) en sont un exemple. Plus récemment, les travaux relevant de la notion de supervision distante (*distant supervision*) (Mintz *et al.,* 2009), dans lesquels les exemples de relations se limitent à des couples d'entités, sans instanciation en corpus, en sont une autre manifestation.

Les dernières années ont vu également apparaître une approche inverse, que nous qualifierons ici d'extraction d'information non supervisée. Cette approche prend comme point de départ des entités ou des types d'entités et se fixe comme objectif de mettre en évidence les relations intervenant entre ces entités, sans connaissance *a priori* de leur type. Cette mise en évidence est éventuellement suivie d'un regroupement des relations extraites en fonction de leurs similarités pour en faire la synthèse. Les travaux effectués dans ce champ de recherche peuvent plus précisément être appréhendés selon trois grands points de vue. Le premier voit cette extraction non supervisée de relations comme un moyen d'acquérir des connaissances, que ce soit des connaissances sur le monde collectées à vaste échelle à partir du Web, comme avec le concept d'*Open Information Extraction* (Banko *et al.,* 2008), ou dans des domaines plus spécialisés, comme le domaine médical, où cette extraction est le moyen d'ajouter de nouveaux types de relations entre entités à une ontologie existante (Ciaramita *et al.,* 2005). Les deux autres points de vue sont plus directement liés à l'extraction d'information. Le principal correspond à la volonté d'offrir aux utilisateurs des modes d'extraction de l'information plus souples et plus ouverts quant à la spécification de leur besoin informationnel. L'approche *On-demand information extraction* (Sekine, 2006), préfigurée dans (Hasegawa *et al.,* 2004) et concrétisée par la *Preemptive Information Extraction* (Shinyama *et al.,* 2006), vise ainsi à induire l'équivalent d'un *template* à partir d'un ensemble de documents représentatifs des informations à extraire, documents typiquement obtenus par le biais de requêtes soumises à un moteur de recherche. Cette même perspective se retrouve pleinement dans (Eichler *et al.,* 2008) et, au travers de l'accent mis sur le regroupement des relations, dans (Rosenfeld *et al.,* 2007). Le dernier point de vue, plus minoritaire, voit l'extraction d'information non

supervisée comme un moyen d'améliorer l'extraction d'information supervisée. Cette dernière est en effet souvent tributaire de corpus annotés qui, compte tenu de la complexité des tâches considérées, ne sont généralement pas de grande taille. Les résultats d'une approche non supervisée peuvent alors être utilisés pour élargir la couverture des modèles appris. Cette conception est particulièrement développée dans (Banko *et al.,* 2008) et se retrouve dans (González *et al.,* 2009). Dans cet article, nous nous plaçons dans le cadre du deuxième point de vue exposé ci-dessus avec un travail mettant l'accent sur le regroupement de relations entre entités nommées extraites sans *a priori* sur leur type.

2. Processus d'extraction de relations non supervisée

Le travail que nous présentons dans cet article s'inscrit dans un contexte plus large visant à développer un processus d'extraction d'information non supervisée susceptible de répondre à des problématiques de veille telle que « suivre tous les événements faisant intervenir les sociétés X et Y ». À la base de ce processus 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.

Plus formellement, comme illustré par la figure 1, les relations extraites des textes sont caractérisées par deux grandes catégories d'information permettant tout à la fois de les définir et de fournir les éléments nécessaires à leur regroupement :

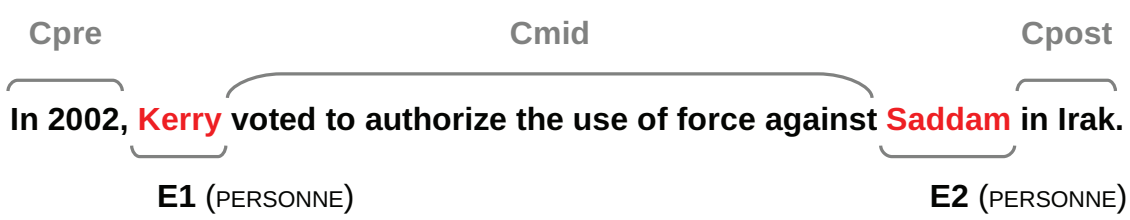

Figure 1. *Exemple de relation extraite*

– 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) ;

– une caractérisation linguistique de la relation. Il s'agit de la façon dont la relation est exprimée linguistiquement. La présence de deux entités conduit à diviser cette caractérisation linguistique en trois parties : (*Cpre*, *Cmid*, *Cpost*). 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.

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.

Le processus d'extraction d'information non supervisée défini autour de cette notion de relation s'articule quant à lui de la façon suivante :

– pré-traitement linguistique des textes ;

– extraction de relations candidates ;

– filtrage des relations candidates ;

– regroupement des relations selon leur similarité.

Le pré-traitement linguistique des textes permet de mettre en évidence dans les textes les informations nécessaires à la définition des relations. Ce pré-traitement comporte donc une reconnaissance des entités nommées pour les types d'entités visés, une désambiguïsation morpho-syntaxique des mots ainsi que leur normalisation. Ces traitements, réalisés ici pour l'anglais, s'appuient sur les outils d'OpenNLP[1]. L'extraction des relations candidates se caractérise quant à elle par des contraintes très limitées : sont ainsi extraites les relations correspondant à tout couple d'entités nommées dont les types font partie des 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. Le filtrage de ces relations candidates met en revanche en œuvre des moyens plus élaborés en combinant des heuristiques et un classifieur statistique de type Conditional Random Fields. Les expérimentations menées sur la sous-partie du corpus AQUAINT-2 constituée de 18 mois du journal *New York Times* ont crédité ce dernier d'une précision de 0,762 et d'un rappel de 0,782. Plus globalement, ces expérimentations ont montré qu'avec une extraction initiale produisant entre 71 858 et 175 802 relations candidates selon les couples de types d'entités, la procédure de filtrage ne conserve en final qu'entre 10 054 et 47 700 relations, soit de 14% à 32% d'entre elles. De plus amples détails concernant la mise en œuvre et l'évaluation de l'extraction et du filtrage des relations peuvent être trouvés dans (Wang *et al.,* 2011).

3. Regroupement de relations

À l'instar de la plupart des travaux sur l'extraction d'information non supervisée, notre objectif final est le regroupement des relations selon leur similarité, en particulier pour en faciliter l'exploration. Reprenant (Hasegawa *et al.,* 2004) et d'autres à sa suite, nous avons adopté le principe d'un seul niveau de regroupement des relations, niveau qualifié de sémantique puisque cherchant à rassembler les relations équivalentes sur le plan sémantique. Cette équivalence est ici à prendre au sens large. Elle dépasse la stricte notion de paraphrase ou d'implication et se rapproche de la notion de redondance informationnelle telle qu'elle est envisagée en résumé automatique.

1. `http://opennlp.sourceforge.net`

Concrètement, la mise en œuvre de ce regroupement des relations nécessite la donnée de deux éléments : une mesure de similarité entre les relations traduisant la notion d'équivalence évoquée ci-dessus et un algorithme de regroupement s'appuyant sur l'évaluation de cette mesure de similarité entre les relations à regrouper. Pour le premier point, nous avons repris l'adoption largement répandue, là encore à la suite de (Hasegawa *et al.,* 2004), de la mesure *cosinus* appliquée à une représentation de type « sac de mots » de la caractérisation linguistique des relations. Dans notre cas, seule la partie *Cmid* de cette caractérisation est prise en compte afin d'accentuer le caractère sémantique du regroupement en se focalisant sur le cœur de la définition des relations.

À côté du problème du choix d'une mesure pour évaluer la similarité des relations, se pose le problème, beaucoup moins abordé dans les travaux existants, de la possibilité de calculer cette mesure sur de larges ensembles de relations, en l'occurrence plusieurs dizaines de milliers. Beaucoup d'algorithmes de regroupement s'appuient en effet sur une matrice de similarité coûteuse à construire et les algorithmes les moins exigeants de ce point de vue, comme l'algorithme *k-means*, nécessitent souvent de fixer *a priori* un nombre de classes qu'il est difficile d'évaluer dans notre cas de figure. Compte tenu de la nature de notre mesure de similarité entre relations, nous avons choisi d'utiliser l'algorithme *All Pairs Similarity Search* (APSS) (Bayardo *et al.,* 2007) qui permet, moyennant la fixation d'un seuil de similarité minimale, de calculer efficacement et de manière exacte une mesure telle que *cosinus* pour tous les couples d'éléments dont la similarité est supérieure à un seuil fixé. Cet algorithme puise son efficacité dans une série d'optimisations dans l'indexation des éléments à comparer tenant compte d'informations collectées *a priori* sur les valeurs des caractéristiques de ces éléments et d'un tri de ces derniers. La valeur de similarité minimale a été fixée en s'appuyant sur les données du Microsoft Research Paraphrase Corpus (Dolan *et al.,* 2004). Ce corpus rassemble un ensemble de couples de phrases jugées quant à leur statut de paraphrases. Cependant, les phrases non paraphrases étant en pratique très proches sur le plan sémantique, elles constituent une bonne référence minimale pour notre similarité. Le calcul de la mesure *cosinus* pour tous ces couples nous a permis de retenir une valeur minimale de 0,45 pour les expérimentations de la section suivante, seuil correspondant aux trois-quarts des valeurs calculées.

Concernant l'algorithme de regroupement, nous avons opté pour l'algorithme *Markov Clustering* (Dongen, 2000). Cet algorithme partitionne un graphe de similarité[2] en clusters disjoints en réalisant une série de marches aléatoires dans ce graphe. L'hypothèse est ici qu'un cluster se caractérise par une forte densité de liens entre ses éléments et qu'en « sortir » ne peut donc se faire qu'après un nombre important de pas. Cet algorithme itératif converge en pratique rapidement et se montre capable de traiter nos graphes composés de plusieurs dizaines de milliers de nœuds. Par ailleurs, il détermine de manière intrinsèque le nombre de clusters à former et n'est dépendant pour ce faire que d'un seul paramètre – le facteur d'inflation – que nous avons laissé à sa valeur par défaut lors des expérimentations de la section suivante.

2. Le graphe de similarité est construit en créant un nœud pour chaque relation et en liant tous les nœuds dont les relations correspondantes ont une similarité évaluée par l'APSS.

4. Évaluation du regroupement des relations

La méthode de regroupement des relations présentée à la section précédente a été appliquée aux relations extraites avec ou sans l'étape de filtrage pour évaluer la qualité des regroupements formés selon ces deux conditions et de façon générale, pour évaluer la qualité globale du clustering. L'évaluation d'un regroupement à grande échelle reste une tâche difficile à cause de l'absence de ressources de référence fiable. Une approche directe consisterait à vérifier manuellement la qualité des résultats obtenus en parcourant l'ensemble des clusters produits. Mais cette évaluation est coûteuse et se limite à un résultat donné, correspondant à une configuration spécifique du clustering. Elle est donc difficile à mettre en œuvre pour comparer différents algorithmes ou paramétrages du clustering.

Nous proposons donc dans cette section deux types d'évaluation : une première évaluation utilisant des mesures internes de regroupement et une seconde évaluation utilisant des critères de qualité externes. Les mesures internes permettent d'établir si le regroupement obtenu reflète bien les valeurs de similarité dans l'espace des relations (Halkidi *et al.,* 2002). Plus précisément, nous voulons tester dans ce cas l'hypothèse que l'espace des relations après filtrage présente une meilleure distribution des similarités en ce qui concerne sa capacité à faciliter le regroupement des relations. D'un autre côté, les mesures externes permettent d'évaluer directement le résultat du regroupement par rapport à un regroupement de référence et vérifient donc si les relations regroupées sont effectivement sémantiquement similaires. Dans notre cadre, où les relations sont extraites à grand échelle, il est difficile de construire un regroupement de référence sur la totalité des relations extraites. Nous proposons donc de sélectionner un ensemble de relations pour la création de clusters de référence et d'évaluer la distribution de ces relations dans les résultats.

4.1. *Évaluation du regroupement par des mesures internes*

Parmi les différents indices internes pour l'évaluation du clustering, nous avons retenu la mesure de la densité attendue (*expected density*), qui est évaluée dans (Stein *et al.,* 2003) comme la mesure ayant la meilleure corrélation avec la mesure externe de F-mesure pour le clustering de documents (la mesure usuelle de l'indice de Dunn étant jugée moins stable). Cette mesure, notée ρ dans l'équation 1, est définie sur un graphe pondéré (V, E, w) (V est l'ensemble des nœuds, *i.e.* des relations, E est l'ensemble des arcs et w, la pondération des arcs, *i.e.* la similarité entre les relations) et un ensemble de clusters $C = \{C_i\}$, avec $C_i = (V_i, E_i, w)$, en fonction d'une mesure de densité θ de l'ensemble des objets à regrouper, définie par $\theta = \frac{ln(w(G))}{ln(|V|)}$, avec $w(G) = |V| + \sum_{e \in E} w(e)$, et de θ_i, définie de la même façon sur le graphe restreint aux nœuds du cluster C_i.

$$\rho = \sum_{i=1}^{|C|} \frac{|V_i|}{|V|} |V_i|^{\theta_i - \theta} \qquad \rho' = \sum_{i=1}^{|C|} \frac{|V_i|}{|V|} \frac{\theta_i}{\theta} \qquad [1]$$

Nous avons plus précisément utilisé dans nos expériences une version modifiée de cette mesure, notée ρ', car les deux regroupements que nous comparons (avec et sans filtrage) se font sur des ensembles de relations de tailles très différentes et la mesure ρ reste trop dépendante de la taille de l'ensemble des objets à regrouper. Nous proposons donc de limiter cette dépendance en modifiant la mesure par la suppression du facteur exponentiel. Plus cette mesure ρ' est élevée, plus le clustering est jugé de bonne qualité.

De façon complémentaire, la mesure de connectivité (*connectivity*) (Handl *et al.,* 2005) mesure dans quelle proportion les relations des plus proches voisins ne sont pas coupées par le clustering. Cette mesure est intéressante dans notre cas car elle s'appuie sur la structure du graphe de similarité que nous utilisons aussi pour la méthode de clustering. Elle se définit par :

$$c = \sum_{i=1}^{|V|} \sum_{j=1}^{p} x_{i,nn_i(j)}$$

avec p, le nombre de voisins considérés, $nn_i(j)$, le $j^{ième}$ plus proche voisin de i et $x_{i,nn_i(j)}$, égal à 0 si i et $nn_i(j)$ sont dans le même cluster et $1/j$ sinon. De la même façon que pour la mesure de densité attendue, cette mesure est dépendante de la taille du corpus. Pour pallier ce problème, nous avons choisi de calculer cette mesure pour un sous-ensemble de relations choisies aléatoirement dans la collection (5 000 relations ont été considérées dans les résultats présentés). Cette mesure évalue de façon inverse à ρ' : plus c est bas, meilleur est le clustering.

	Densité attendue		*Connectivité* ($p = 20$)	
	avant filtrage	après filtrage	avant filtrage	après filtrage
ORG – ORG	1,06	**1,13**	5335,7	**3450,8**
ORG – LIEU	**1,13**	1,02	4458,7	**2837,6**
ORG – PERS	1,09	**1,17**	3025,4	**1532,4**
PERS – ORG	1,02	**1,06**	5638,0	**4620,0**
PERS – LIEU	**1,08**	1,07	5632,5	**4571,3**
PERS – PERS	1,13	**1,15**	3892,7	**2569,2**

Tableau 1. *Résultats de l'évaluation interne du regroupement des relations. Les résultats en gras sont les meilleurs scores (la mesure Densité attendue doit être maximisée, la mesure Connectivité doit être minimisée).*

Les résultats de ces mesures sont présentés dans le tableau 1 avant et après filtrage. On constate tout d'abord que les deux mesures utilisées sont globalement meilleures pour la seconde condition, ce qui indique que le clustering est dans la plupart des cas de meilleure qualité après le filtrage des relations. Les deux couples d'entités pour lesquels cette tendance n'est pas vérifiée, ORG – LIEU et PERS – LIEU, ont en commun de faire intervenir des entités de type lieu. Ce constat s'explique peut-être par une spécificité de ces entités. En effet, lorsqu'une entité de type lieu est présente dans une

phrase et qu'elle n'est pas en relation avec l'autre entité considérée dans la phrase, il est fréquent qu'elle soit inclue dans un complément circonstanciel de lieu. Or, avec la mesure de similarité choisie, les formes des compléments circonstanciels de lieu peuvent induire une similarité entre les phrases valides du point de vue d'un regroupement global et donc donner un bon score de clustering.

4.2. *Évaluation du regroupement par des mesures externes*

Pour confirmer l'intérêt du filtrage pour le regroupement des relations et avoir une base plus solide pour l'évaluation de la qualité générale du clustering, une évaluation utilisant des mesures externes a été effectuée pour une sous-partie des relations extraites, c'est-à-dire en comparant les résultats du clustering à un regroupement de référence. Dans une telle démarche, la première tâche consiste à sélectionner les relations et à construire le regroupement de référence. Cette tâche est effectuée de manière itérative, selon le processus illustré par la figure 2. Les relations extraites sont d'abord indexées par un moteur de recherche[3], en indexant de façon différenciée le texte, les entités nommées et les types d'entités nommées. Cette différenciation permet d'interroger le moteur avec des requêtes ciblant spécifiquement la première ou la deuxième entité nommée (*E1*, *E2*) d'une relation, le type de ces entités (*T1*, *T2*) ou la caractérisation linguistique de relations *Cmid*, *Cpost* ou *Cpre* (cf. l'exemple de la figure 1).

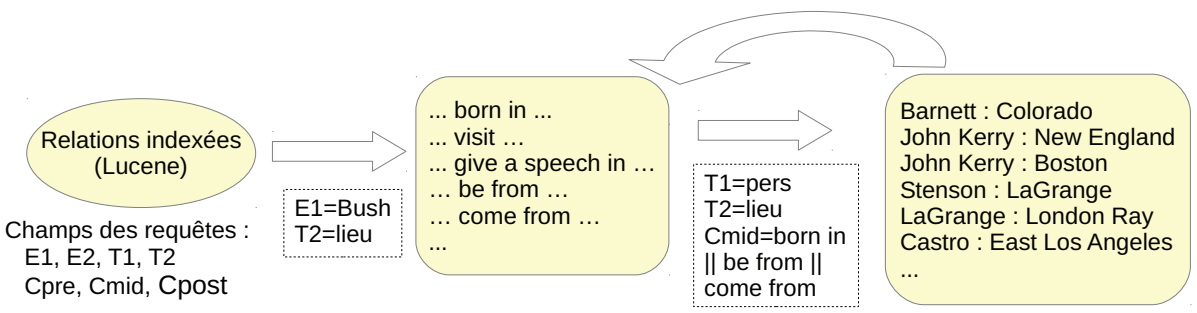

Figure 2. *Processus itératif de création de la référence*

L'index est d'abord interrogé avec une entité nommée et un type d'entités (*e.g. E1=Bush, T2*=LIEU) fixés pour explorer des relations potentielles (*e.g. born in*, *visit*, *etc*). Parmi les relations obtenues, une relation est ensuite choisie et l'index est interrogé en fixant le type des entités (*e.g. T1*=PERS, *T2*=LIEU) et les mots caractéristiques (*Cmid contient "visit"*) pour trouver d'autres couples d'entités ayant la même relation dans le corpus. Ces deux étapes sont répétées itérativement pour exploiter différents couples d'entités en relation et différentes caractérisations linguistiques des relations. Le regroupement des relations se fait manuellement pendant ce processus avec un outil d'annotation spécifique.

Actuellement, la référence construite contient 253 relations entre les entités PERS et LIEU regroupées en 17 clusters correspondant à des types de relations tels que *come*

3. Nous avons utilisé le moteur *Lucene* (http://lucene.apache.org).

from, *be going to*, *have a speech in*, *like*, *etc*. Nous présentons à la figure 3 quelques exemples d'instances de la relation *grow up in*.

Pitcher Brandon Backe, who grew up 50 miles from here in *Galveston* and dreamed of pitching for the Astros in the postseason, displayed a veteran's savvy with his varying speeds.
Chief Justice Wallace B. Jefferson, a Republican, named *Pat Priest*, a retired Democratic judge from his hometown of *San Antonio*, to hear the case – but not before Jefferson's own multiple ties to DeLay's political operation were questioned.
By the time he turned 10, *Levine* had performed as a soloist with his hometown *Cincinnati* Orchestra.
John Kerry comes from *New England*, where people don't talk about personal things like religion very easily in public.

Figure 3. *Exemples de relations regroupées dans la référence pour le type de relations "grow up in"*

De nombreuses mesures d'évaluation externes pour comparer les résultats d'un clustering à un regroupement de référence ont été proposées (Manning *et al.,* 2008), les plus utilisées étant les mesures de *Rand Index*, *Pureté* et *Information Mutuelle Normalisée*. Pour N relations, la mesure de *Rand Index* a pour but de comparer la manière dont les $N(N-1)/2$ paires de relations sont regroupées dans le clustering obtenu par rapport au clustering de référence, l'objectif étant de maximiser le nombre de relations similaires rassemblées dans le même cluster et le nombre de relations dissimilaires séparées. On mesure les regroupements de paires de relations selon quatre décisions : *Vrai Positif (VP)* si deux relations similaires sont dans un même cluster, *Vrai Négatif (VN)* si deux relations dissimilaires sont dans deux clusters différents, *Faux Positif (FP)* si deux relations dissimilaires sont dans un même cluster et *Faux Négatif (FN)* si deux relations similaires sont dans des clusters différents. VP et VN sont des décisions correctes alors que FP et FN correspondent à des erreurs de regroupement. La mesure *Rand Index* est alors définie par :

$$Rand\ Index = \frac{VP + VN}{VP + FP + FN + VN}$$

À partir de ces décisions, des scores de précision P, rappel R et F-mesure peuvent aussi être définis de la façon suivante :

$$P = \frac{VP}{VP + FP} \quad R = \frac{VP}{VP + FN} \quad F = \frac{2 \cdot P \cdot R}{P + R}$$

Le tableau 2 présente les résultats pour ces mesures en comparant le clustering sur les relations avant ou après l'étape de filtrage. On peut voir que le nombre de décisions de type VP a beaucoup augmenté après l'application de la procédure de filtrage. Plus précisément, le rappel a presque doublé, passant de 0,1271 à 0,2211, la précision restant autour de 0,53.

Étape	*VP*	*FP*	*FN*	*VN*
avant filtrage	469	404	3221	27784
après filtrage	866	774	3051	28979

Étape	*rand index*	*précision*	*rappel*	*F1-mesure*
avant filtrage	0,8863	0,5372	0,1271	0,2056
après filtrage	0,8864	0,5280	0,2211	0,3117

Tableau 2. *Évaluation par mesures externes pour des relations de type* PERS – LIEU

En complément de ces mesures, qui portent sur toutes les paires de relations, la qualité du clustering peut être aussi évaluée au niveau de chaque cluster. Les mesures de *Pureté* et d'*Information Mutuelle Normalisée* (IMN) sont souvent utilisées dans ce cas. Une condition préalable pour calculer ces mesures est d'attribuer à chaque cluster du résultat un cluster de référence. Nous avons choisi la stratégie la plus simple pour cette étape : le cluster de référence qui partage le plus grand nombre de relations avec le cluster obtenu est choisi comme cluster associé. La mesure de *Pureté* se définit alors de la façon suivante :

$$Pureté(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |w_k \cap c_j|$$

où $\Omega = \{w_1, w_2, ...w_K\}$ est l'ensemble des clusters du résultat et $\mathbb{C} = \{c_1, c_2, ..., c_J\}$ est l'ensemble des clusters de la référence. La mesure de *Pureté* peut néanmoins présenter un biais, surtout quand le nombre de clusters est grand : la *Pureté* est en effet égale à 1 s'il n'y a qu'une seule relation dans chaque cluster. La mesure d'*Information Mutuelle Normalisée* permet de faire un compromis entre le nombre de clusters et leur qualité. Elle se définit par :

$$IMN(\Omega, \mathbb{C}) = \frac{IM(\Omega, \mathbb{C})}{(H(\Omega) + H(\mathbb{C}))/2}$$

où $IM(\Omega, \mathbb{C})$ est l'information mutuelle entre Ω et $\mathbb{C}$, et $H(\Omega)$ et $H(\mathbb{C})$ les entropies respectives de Ω et $\mathbb{C}$, définies par :

$$IM(\Omega, \mathbb{C}) = \sum_k \sum_j P(w_k \cap c_j) \log \frac{P(w_k \cap c_j)}{P(w_k) * P(c_j)}$$

$$H(\Omega) = -\sum_k P(w_k) \log P(w_k)$$

avec $P(w_k)$, $P(c_j)$ et $P(w_k \cap c_j)$ les probabilités pour une relation d'être dans un cluster du résultat w_k, un cluster de la référence c_j, ou l'intersection de ces deux. Les probabilités sont estimées en comptant les cardinalités des clusters.

De la même manière que pour l'évaluation par la mesure *Rand Index*, nous avons calculé ces mesures pour les clusters obtenus à partir des relations extraites avant ou après filtrage. Les résultats sont présentés dans le tableau 3 et montrent que les mesures de *Pureté* et d'*Information Mutuelle Normalisée* sont toutes les deux améliorées par la procédure de filtrage. En particulier, l'augmentation de la *Pureté* confirme l'amélioration de la mesure de rappel dans le tableau 2.

Étape	*Pureté*	*IM*	*IMN*
avant filtrage	0,5540	1,6338	0,6008
après filtrage	0,5889	1,7216	0,6285

Tableau 3. *Évaluation par mesures externes pour des relations de type* PERS – LIEU

5. Conclusion et perspectives

Dans cet article, nous avons présenté un travail sur l'extraction d'information non supervisée pour extraire des relations entre entités, sans *a priori* sur la nature de ces relations. Une procédure de filtrage associant des heuristiques et un classifieur statistique permet d'abord de sélectionner les relations après une première extraction reposant sur des critères minimaux. Une méthode de regroupement de ces relations combinant un calcul optimisé des similarités par APSS et une phase de regroupement par Markov Clustering est ensuite appliquée pour les structurer. Une évaluation par des mesures internes et externes a été réalisée permettant en particulier de mettre en évidence l'amélioration apportée par le filtrage au travers d'un doublement du rappel pour une précision équivalente. Une évaluation plus avancée de ce travail est en cours par l'enrichissement de la référence. Un deuxième prolongement de ce travail concerne l'amélioration des méthodes de regroupement en envisageant deux niveaux de clustering : un clustering thématique et un clustering sémantique. Enfin, l'identification non supervisée de relations peut également être exploitée afin d'améliorer un système de peuplement de base de connaissances dans un contexte de supervision distante, avec pour rôle de filtrer les relations résultant de la projection dans un corpus de couples d'entités nommées en relation issues d'une base de connaissances.

Remerciements

Ce travail a été soutenu par le projet FILTRAR-S de l'appel ANR CSOSG 2008.

6. Bibliographie

Agichtein E., Gravano L., « Snowball : Extracting relations from large plain-text collections », *5^{th} ACM International Conference on Digital Libraries*, 2000.

Banko M., Etzioni O., « The Tradeoffs Between Open and Traditional Relation Extraction », *48^{th} 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 », *16th international conference on World Wide Web*, Banff, Alberta, Canada, p. 131-140, 2007.

Ciaramita M., Gangemi A., Ratsch E., Saric J., Rojas I., « Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology », *19th International Joint Conference on Artificial Intelligence (IJCAI 2005)*, Edinburgh, UK, 2005.

Dolan B., Quirk C., Brockett C., « Unsupervised construction of large paraphrase corpora : exploiting massively parallel news sources », *20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, 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 », *6th Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

González E., Turmo J., « Unsupervised Relation Extraction by Massive Clustering », *Ninth IEEE International Conference on Data Mining (ICDM 2009)*, Miami, Florida, USA, p. 782-787, 2009.

Grishman R., Sundheim B., « Design of the MUC6 evaluation », *MUC-6 (Message Understanding Conferences)*, Morgan Kauffmann Publisher, Columbia, MD, 1995.

Halkidi M., Batistakis Y., Vazirgiannis M., « Cluster validity methods : part I », *ACM SIGMOD Record (Special Interest Group on Management of Data)*, vol. 31, p. 40-45, June, 2002.

Handl J., Knowles J., Kell D. B., « Computational cluster validation in post-genomic data analysis. », *Bioinformatics (Oxford, England)*, vol. 21, n° 15, p. 3201-3212, August, 2005.

Hasegawa T., Sekine S., Grishman R., « Discovering Relations among Named Entities from Large Corpora », *42nd Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, Spain, p. 415-422, 2004.

Manning C. D., Raghavan P., Schütze H., *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

Mintz M., Bills S., Snow R., Jurafsky D., « Distant supervision for relation extraction without labeled data », *Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapore, p. 1003-1011, 2009.

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

Sekine S., « On-demand information extraction », *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, Sydney, Australia, p. 731-738, 2006.

Shinyama Y., Sekine S., « Preemptive Information Extraction using Unrestricted Relation Discovery », *HLT-NAACL 2006*, New York City, USA, p. 304-311, 2006.

Stein B., Sven, Wißbrock F., « On Cluster Validity and the Information Need of Users », *3rd IASTED International Conference on Artificial Intelligence and Applications (AIA'03)*, p. 404-413, 2003.

Wang W., Besançon R., Ferret O., Grau B., « Filtering and clustering relations for unsupervised information extraction in open domain », *20th ACM international Conference on Information and Knowledge Management (CIKM 2011)*, p. 1405-1414, 2011.