Détection et classification non supervisées de relations sémantiques dans des articles scientifiques

Kata Gábor, Isabelle Tellier, Thierry Charnois, Haïfa Zargayouna, Davide Buscaldi

Détection et classification non supervisées de relations sémantiques dans des articles scientifiques

Kata Gábor[1] Isabelle Tellier[2] Thierry Charnois[1] Haïfa Zargayouna[1]
Davide Buscaldi[1]
(1) LIPN, CNRS (UMR 7030), Université Paris 13 Sorbonne Paris Cité
(2) LaTTiCe, CNRS (UMR 8094), ENS Paris, Université Sorbonne Nouvelle - Paris 3
(1) prénom.nom@lipn.univ-paris13.fr (2) isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Dans cet article, nous abordons une tâche encore peu explorée, consistant à extraire automatiquement l'état de l'art d'un domaine scientifique à partir de l'analyse d'articles de ce domaine. Nous la ramenons à deux sous-tâches élémentaires : l'identification de concepts et la reconnaissance de relations entre ces concepts. Une extraction terminologique permet d'identifier les concepts candidats, qui sont ensuite alignés à des ressources externes. Dans un deuxième temps, nous cherchons à reconnaître et classifier automatiquement les relations sémantiques entre concepts de manière non-supervisée, en nous appuyant sur différentes techniques de clustering et de biclustering. Nous mettons en œuvre ces deux étapes dans un corpus extrait de l'archive de l'ACL Anthology. Une analyse manuelle nous a permis de proposer une typologie des relations sémantiques, et de classifier un échantillon d'instances de relations. Les premières évaluations suggèrent l'intérêt du biclustering pour détecter de nouveaux types de relations dans le corpus.

ABSTRACT

Unsupervised Classification of Semantic Relations in Scientific Papers

In this article, we tackle the yet unexplored task of automatically building the "state of the art" of a scientific domain from a corpus of research papers. This task is defined as a sequence of two basic steps : finding concepts and recognizing the relations between them. First, candidate concepts are identified using terminology extraction, and subsequently linked to external resources. Second, semantic relations between entities are categorized with different clustring and biclustering algorithms. Experiences were carried out on the ACL Anthology Corpus. Results are evaluated against a hand-crafted typology of semantic relations and manually categorized examples. The first results indicate that biclustering techniques may indeed be useful for detecting new types of relations.

MOTS-CLÉS : analyse de la littérature scientifique, extraction de relations, clustering, biclustering.

KEYWORDS: analysis of scientific literature, relation extraction, clustering, biclustering.

1 Introduction

De nos jours, la production d'articles scientifiques croît à un rythme accéléré. Cette explosion d'information rend le travail des chercheurs, des relecteurs et des experts de plus en plus difficile (Larsen & von Ins, 2010). Ce problème a attiré l'attention de plusieurs chercheurs dans les domaines du web sémantique, de la scientométrie et du traitement du langage naturel, qui explorent des solutions

pour fournir un meilleur accès à la littérature scientifique. (Osborne & Motta, 2012, 2014) ont par exemple travaillé sur l'analyse des communautés scientifiques en utilisant les liens de citations et des taxinomies de topiques de recherche dans le domaine informatique. (Shotton, 2010) a proposé une ontologie pour les différents types de citations. Mais l'analyse des liens de citation n'est pas suffisante pour la compréhension profonde d'un domaine scientifique. Celle-ci requiert la maîtrise de son "état de l'art", dont les éléments figurent généralement dans les parties introductives des articles de ce domaine. Nous visons dans ce travail à extraire automatiquement et avec un minimum de supervision ces éléments. L'approche que nous préconisons prend donc appui sur l'analyse automatique d'articles. Nous l'avons mise en œuvre sur des textes extraits des archives des conférences ACL, parce que le champ qu'elle couvre, le traitement automatique des langues, nous est familier.

Par exemple, à partir de la portion de texte : "semi-automatic extensions using Bayesian Inference [...]", nous souhaitons extraire l'information structurée suivante :

```
<relation type="usedfor">
  <arg1>Bayesian Inference</arg1>
  <arg2>semi-automatic extensions</arg2>
</relation>
```

où "Bayesian Inference" et "semi-automatic extensions" sont deux entités liées par une relation d'application ou d'utilisation. Découvrir ce genre de relations permettrait de répondre à des questions comme "à quelle tâche a été appliquée l'inférence bayésienne ?".

Notre approche est spécifiquement dédiée à la littérature scientifique mais reste générique, au sens où elle est applicable sur n'importe quel domaine pour lequel une telle littérature existe. Elle repose sur deux étapes successives :

— l'identification des *entités* (concepts) cités dans les articles. Ces concepts doivent pouvoir être mis en relation avec une ontologie du domaine scientifique traité.
— la reconnaissance et la classification des *relations sémantiques* qui relient deux concepts.

Le résultat de cette démarche, serait donc une base de connaissances décrivant un domaine, constituée automatiquement à partir d'un ensemble d'articles, qui résume et synthétise les connaissances scientifiques qu'ils contiennent.

Les intérêts d'un tel système sont multiples, il recoupe et étend diverses autres tâches :

— l'extraction d'information multi-documents en domaine de spécialité : cette tâche ressemble à des traitements couramment effectués en bioinformatique, mais nous visons à couvrir la littérature scientifique en général ;
— le peuplement automatique d'ontologie (Petasis *et al.*, 2011) ou de thesauri (Ferret, 2013) ;
— l'identification de liens entre documents pour faciliter l'accès à leur contenu sémantique et la navigation intertextuelle (Presutti *et al.*, 2014) ;
— l'étude de l'évolution dans le temps d'un champ scientifique (Chavalarias & Cointet, 2013; Omodei *et al.*, 2014) ;
— la recherche d'information sémantique pour l'accès aux publications scientifiques (Sateli & Witte, 2015).

L'extraction de relations entre entités est une tâche relativement ancienne dans le domaine de l'extraction d'information. Elle se caractérise par un ensemble de relations pré-définies à extraire et par un corpus annoté. Ces données ont permis le développement d'approches supervisées lors des campagnes MUC et ACE (Hobbs & Riloff, 2010). Avec la multiplication des corpus disponibles sur le web, une nouvelle problématique *Open Information Extraction (OpenIE)* est maintenant explorée (Banko *et al.*, 2007; Del Corro & Gemulla, 2013). Ses principaux objectifs consistent à développer des approches peu supervisées afin de s'affranchir de la nécessité de disposer de corpus d'apprentissage

étiquetés et viser l'indépendance au domaine. Le type des relations à extraire ne sont par ailleurs pas limitées à un ensemble pré-défini, ce qui permet de traiter la diversité des relations existantes en domaine ouvert.

L'approche que nous proposons est proche du cadre de l'*OpenIE*. Elle est mise en œuvre sur l'anthologie des conférences ACL mais est applicable à tout domaine pour lequel il existe des textes scientifiques. Faute de données annotées disponibles, elle est le moins supervisée possible.

L'identification des concepts scientifiques est ainsi réalisée grâce à un extracteur de termes et à des lexiques spécialisés. Une typologie des différentes relations possible a ensuite été constituée manuellement. Cette typologie est destinée à servir de guide d'annotation pour créer un corpus d'évaluation, mais n'est pas utilisée pour diriger l'extraction. Comme dans un certain nombre de travaux en *OpenIE* (e.g. Wang *et al.* (2013)), la reconnaissance et la classification des relations sont traitées par des méthodes de clustering sans limiter *a priori* le nombre de clusters au nombre de relations de cet ensemble. Ceci permet d'extraire non seulement les relations connues, mais aussi potentiellement d'autres d'un type nouveau, enrichissant ainsi notre base de connaissances. Les informations disponibles pour l'extraction des relations sont les propriétés du couple d'entités que la relation relie (notamment la source qui a permis de les identifier et leur représentation distributionnelle dans le corpus) ainsi que la portion de texte qui les sépare dans le document. Nous avons testé des méthodes de clustering exploitant ces différentes informations indépendamment, ainsi que des techniques de biclustering inspirées de la recherche en génétique (mais encore à notre connaissance peu exploitées en TAL) qui essaient de les combiner. Le résultat d'un biclustering est en effet une double partition dans deux espaces simultanément : pour nous, l'espace des couples d'entités et celui des séquences de caractères les séparant. Pour évaluer nos résultats, nous avons manuellement annoté une partie des documents avec la typologie des relations. Nos clusterings sont évalués en testant si les relations annotées manuellement présentes dans un même cluster correspondent effectivement à une même relation dans cette typologie.

Dans la suite de cet article, nous décrivons d'abord nos données, les concepts scientifiques que nous y avons trouvés et la typologie des relations sémantiques que nous avons construite en les analysant. Nos différentes expériences de clustering sont ensuite exposées, ainsi que les résultats obtenus.

2 Données et ressources

Dans cette partie, nous décrivons le corpus sur lequel nous avons mené nos expériences. Nous expliquons comment nous avons identifié les entités pertinentes qu'il contient et proposons une typologie des relations sémantiques possibles entre deux entités, construite à partir de l'annotation manuelle d'un échantillon de données.

2.1 Identification des entités

Nous avons choisi pour aborder cette tâche un extrait du corpus ACL Anthology (Radev *et al.*, 2009), comportant 11 000 résumés d'articles pré-traités par Omodei *et al.* (2014), auxquels nous avons ajouté les introductions. La totalité du corpus comprend 4 200 000 mots.

Comme notre objectif est de mettre en place une chaîne d'extraction qui servira pour l'enrichissement de bases de connaissances par de relations entre concepts, il est important de pouvoir retrouver les

concepts annotés dans des ressources ontologiques. Nous avons donc décidé de n'annoter que les concepts qui peuvent être liés à des ressources externes. Notre objectif est d'obtenir une équilibre entre *précision d'annotation*, c'est-à-dire la pertinence des entités par rapport au domaine traité, et *couverture*, mesurée en termes de densité de l'annotation, celle-ci étant particulièrement importante dans un contexte de fouille de séquences. Pour une description plus détaillée de l'annotation des concepts, le lecteur est invité à se reporter à (Gábor *et al.*, 2016).

Dans un premier temps, nous avons appliqué l'extracteur terminologique TermSuite (Daille *et al.*, 2013) au corpus. Le processus d'extraction prend en entrée un ensemble de documents et ordonne les candidats termes selon leur spécificité. Plusieurs paramètres de filtrage peuvent être ajoutés (étiquettes morpho-syntaxiques, fréquences, variantes), nous avons fixé leurs valeurs empiriquement. La liste des termes candidats a été validée manuellement.

Les entités retenues pour l'annotation sont celles que l'on a pu aligner avec des ressources externes. Diverses ontologies ainsi que des ressources lexicales ont été étudiées pour jouer ce rôle. Saffron Knowledge Extraction Framework [1] propose ainsi des ressources spécifiques automatiquement extraites de corpus de domaine (Bordea, 2013; Bordea *et al.*, 2013). Nous nous sommes appuyés sur les *modèles de domaine* dédiés au traitement automatique des langues et à l'informatique : ils comprennent respectivement 200 et 120 noms simples. Les *topical hierarchies* pour le TAL, contenant 500 multi-mots ont également été incluses.

Malgré la qualité des ressources produites par Saffron (Bordea, 2013), la densité des annotations résultantes est très limitée pour la fouille de relations. En effet, un test d'annotation sur 1 100 000 mots a produit une moyenne de 13 concepts annotés/100 mots. Pour augmenter cette densité, nous avons décidé d'utiliser une ressource générique, BabelNet (Navigli & Ponzetto, 2012) pour l'aligner avec les termes extraits par TermSuite. En tant qu'ontologie générique, BabelNet tout seul ne peut pas garantir la pertinence par rapport au domaine ; cependant, en combinant l'ontologie avec l'extraction terminologique, nous nous attendons à obtenir un bon niveau de spécificité. 3 805 concepts à l'intersection de BabelNet et TermSuite ont été ainsi retenus, en plus des concepts provenant des ressources Saffron. La combinaison de ces ressources a conduit à une densité de 23 concepts annotés/100 mots, ce qui semble satisfaisant pour nos objectifs. Nous avons évalué les annotations en termes de précision en validant les concepts, abstraction faite des erreurs dues à un mauvais étiquetage morpho-syntaxique. La précision a été calculée sur un échantillon de 100 phrases contenant au total 932 entités annotées. Nous atteignons des précisions comparables pour les annotations par les ressources lexicales de Saffron ou par celles de BabelNet filtrées (0.97 et 0.98 respectivement). Ceci confirme l'intérêt de combiner des ressources spécifiques extraites du corpus à des ressources externes ayant une large couverture.

2.2 Typologie des relations sémantiques visées

Afin de construire une typologie des relations sémantiques, nous avons sélectionné un échantillon de 500 résumés (approximativement 100 mots/résumé) et nous y avons manuellement annoté les instances de relations avec Gate (Cunningham *et al.*, 2002). La typologie proposée est donnée en Table 1. Elle couvre bien le champ de la science en général, sans se restreindre au domaine du traitement automatique des langues. Trois relations génériques ont été utilisées en plus de ces relations spécifiques : l'antonymie, l'hyponymie et la co-hyponymie.

1. http ://saffron.insight-centre.org/

affects	ARG1 : *specific property of data* ARG2 : *results*
based_on :	ARG1 : *method, system* based on ARG2 : *other method*
char	ARG1 : *observed characteristics* of an observed ARG2 : *entity*
compare	ARG1 : *result (of experiment)* compared to ARG2 : *result2*
composed_of	ARG1 : *database/resource* ARG2 : *data*
datasource	ARG1 : *information* extracted from ARG2 : *kind of data*
methodapplied	ARG1 : *method* applied to ARG2 : *data*
model	ARG1 : *abstract representation* of an ARG2 : *observed entity*
phenomenon	ARG1 : *entity, a phenomenon* found in ARG2 : *context*
problem	ARG1 : *phenomenon* is a problem in a ARG2 : *field/task*
propose	ARG1 : *paper/author* presents ARG2 : *an idea*
study	ARG1 : *analysis* of a ARG2 : *phenomenon*
tag	ARG1 : *tag/meta-information* associated to an ARG2 : *entity*
taskapplied	ARG1 : *task* performed on ARG2 : *data*
usedfor	ARG1 : *method/system* ARG2 : *task*
uses_information	ARG1 : *method* relies on ARG2 : *information*
yields	ARG1 : *experiment/method* ARG2 : *result*
wrt	ARG1 *a change* in/with respect to ARG2 : *property*

TABLE 1 – Typologie des relations sémantiques

3 Extraction des relations

Dans cette partie, nous nous concentrons sur l'extraction des relations sémantiques. Nous revenons d'abord sur les spécificités de notre tâche, qui justifient l'emploi de techniques non supervisées. Nous détaillons ensuite les deux types de représentation des données que nous avons employés et présentons les différents algorithmes de clustering et de biclustering que nous avons testés.

3.1 Spécificités de l'approche

L'objectif de l'extraction des relations est double : nous cherchons d'une part à trouver de nouveaux types de relations que nous pourrions ensuite ajouter à la base de connaissances, d'autre part à identifier les couples d'entités qui instancient ces relations dans le corpus. Notre projet présente deux différences par rapport aux tâches d'apprentissage de relations, notamment celles ayant fait l'objet des campagnes Semeval (Hendrickx *et al.*, 2010; Jurgens *et al.*, 2012). D'abord, par opposition aux relations sémantiques lexicales prises en compte dans Semeval, celles que nous recherchons sont largement contextuelles. Dans notre corpus, un même couple d'entités peut ainsi représenter plusieurs relations différentes suivant le contexte d'utilisation, comme dans les exemples suivants :
Uses_information : (...) **models** extract rules using **parse trees** (...)
Usedfor : (...) **models** derives discourse **parse trees** (...)

Deuxièmement, ces tâches disposent d'un inventaire pré-défini de relations et se concentrent sur des techniques de classification supervisée. Bien que nous cherchions également à identifier des relations pré-définies, nous ne souhaitons pas nous limiter à elles. Ainsi, nous avons choisi de n'employer que des méthodes non supervisées - moins précises, par nature, que les méthodes entraînées sur des exemples étiquetés, elles permettent aussi de détecter des relations nouvelles dans le corpus. Nous

avons mené plusieurs expériences de clustering global ainsi qu'une expérience de clustering local (biclustering) qui permet d'identifier en même temps les couples d'entités qui partagent la même relation et les portions de texte qui instancient cette relation.

3.2 Représentations vectorielles

L'objectif de cette partie est de représenter chaque couple d'entités qui a des co-occurrences dans le corpus dans un espace vectoriel qui permettra de calculer une similarité. Nous avons mené des expériences avec deux types de représentations distributionnelles.

3.2.1 Représentation par séquences

La première représentation vectorielle s'appuie sur l'hypothèse que la relation sémantique entre deux entités est explicite dans le texte, dans au moins une partie des phrases qui contiennent une co-occurrence des deux entités (Lin & Pantel, 2001; Turney, 2005). Les attributs correspondent donc dans ce cas aux séquences qui relient les entités dans le corpus. Nous avons procédé à l'extraction de tous les couples d'entités présents dans une même phrase et extrait les portions de texte les séparant.

fréq	e_1 : entité 1	e_2 : entité 2	p :séquence
2	parser	sentences	which is modular in design and which processes
2	parser	sentences	is learned given a set of
1	parser	sentences	to produce an analysis of
1	parser	sentences	that was trained to minimize cost over

TABLE 2 – Exemples de couples d'entités et séquences extraits

Les séquences extraites peuvent contenir d'autres entités mais une limite de longueur (≤ 8 mots, sauf mots composés) a été appliquée. Au final, 998 000 instances ont ainsi été extraites (cf. Table 2).
Nous nous sommes servis des données de co-occurrences entre couples d'entités et portions de texte pour construire une matrice creuse M dont les lignes représentent les couples d'entités *e=*(e_1, e_2) et les colonnes correspondent aux séquences de texte $p \in P$. Après un filtrage par fréquence [2], nous avons obtenu une matrice de 1 385 couples d'entités et de 1 166 séquences. Les cellules $M_{e,p}$ contiennent une valeur d'association entre e et p. L'une des représentations testée utilise les valeurs de co-occurrences brutes, l'autre la pondération sPPMI. Cette dernière est une variante de l'information mutuelle point à point (PMI), dans laquelle les valeurs inférieures à 0 sont remplacées par 0. De plus, un lissage de contextes proposé par Levy *et al.* (2015) est appliqué qui élève les nombres de co-occurrences des contextes (des séquences p dans notre cas) à la puissance α, dont la valeur idéale serait de α = 0,75 selon (Mikolov *et al.*, 2013b). :

$$sPPMI_\alpha(e,p) = max(log_2 \frac{P(e,p)}{P(e) \times P_\alpha(p)}, 0) \quad (1)$$

$$P_\alpha(p) = \frac{freq(p)^\alpha}{\sum_p freq(p)^\alpha} \quad (2)$$

2. Les couples d'entités et les séquences avec une fréquence inférieure ou égale à 5 ont été éliminés, ainsi que les couples d'entités et les séquences dont les occurrences sont limitées à moins de 4 contextes différents.

Ce lissage, inspiré par les modèles utilisant des représentations distribuées pour des tâches sémantiques (Mikolov *et al.*, 2013a; Baroni *et al.*, 2014), pemet de compenser le biais de la PMI vers les contextes rares.

3.2.2 Représentation par vecteurs d'entités

Le deuxième espace est aussi calculé à partir des contextes mais, dans ce cas, il s'agit d'une représentation distributionnelle calculée indépendamment pour les deux entités avec word2vec. Alors que la fonctionnalité principale de word2vec est d'apprendre un modèle de langage, les représentations qu'il produit "en cours de route" sont en effet également utiles pour d'autres tâches pertinentes à la similarité distributionnelle. Ce modèle s'est montré particulièrement bien adapté aux tâches nécessitant d'évaluer une similarité sémantique, en particulier pour le calcul des analogies relationnelles (Mikolov *et al.*, 2013b). Les couples d'entités peuvent être représentés par un seul vecteur construit à partir des vecteurs des entités individuelles, par exemple par concaténation (Baroni *et al.*, 2012) ou soustraction (Weeds *et al.*, 2014).

Word2vec a été entrainé sur le corpus ACL avec le *skip-gram model* (Mikolov *et al.*, 2013a). Les vecteurs neuronaux résultants (de taille=200) sont extraits pour chaque entité. Le vecteur d'un couple d'entités *e* est construit par la concaténation des vecteurs de ses deux entités [3].

3.3 Algorithmes

3.3.1 Clustering

Deux méthodes de clustering hiérarchique ont été comparées en utilisant le logiciel Cluto (Zhao *et al.*, 2005) et la mesure de similarité cosinus. La première méthode est une bissection itérative descendante : à chaque itération, un des clusters est divisé en deux jusqu'à ce que le nombre final de clusters soit atteint. Ce nombre doit être pré-défini : nous avons testé plusieurs valeurs. La division optimale est choisie en maximisant la somme des similarités intra-clusters des clusters résultants.

La deuxième méthode est un clustering hiérarchique ascendant avec une initialisation bissective (Zhao & Karypis, 2002) : un clustering en $\sqrt{n}$ clusters (où n est le nombre de clusters choisi) est d'abord calculé par des bissections itératives. L'espace vectoriel est ensuite augmenté par $\sqrt{n}$ nouvelles dimensions qui représentent les clusters calculés à la première étape, instanciées par la distance de chaque objet par rapport au centre de ces clusters. Le clustering ascendant est ensuite effectué sur cette représentation augmentée, de façon à maximiser la somme des similarités entre chaque pair d'éléments concernés par l'unification de clusters. Cela permet d'éviter l'*effet de chaîne*.

3.3.2 Biclustering

Le biclustering, méthode populaire pour la classification des gènes (Madeira & Oliveira, 2004), est encore peu utilisé dans le domaine du TAL. Par opposition au clustering, qui cherche à regrouper les objets selon la totalité des dimensions, le biclustering a pour but de regrouper les objets et les

3. Les entités formées par des expressions multi-mots sont représentées par leur propre vecteur, issu du module de reconnaissance de mots composés dans word2vec

dimensions en même temps. Dans notre cas, nous nous intéressons à des couples d'entités qui partagent un certain nombre de contextes. Cependant, toutes nos portions de texte n'expriment pas une relation sémantique et tous les couples d'entités ne sont pas en relation.

Parmi les méthodes de biclustering, certaines approches recherchent des clusters qui correspondent à un critère d'homogénéité (Hartigan, 1972), d'autres sont basées sur un clustering itératif (Getz *et al.*, 2000) ou sur un clustering de graphes (Tanay *et al.*, 2002). Dans le domaine du TAL, les systèmes qui apprennent à classifier des objets et des relations en même temps utilisent plutôt des modèles génératifs. Ainsi, Yao *et al.* (2011) adapte l'Allocation de Dirichlet Latente (LDA) à l'apprentissage non supervisé de relations. Les variables observées sont les mots en relation et leurs attributs : la séquence qui les relie, les étiquettes morpho-syntaxiques de cette séquence et les informations sur la catégorie des entités. L'algorithme apprend en parallèle à attribuer une probabilité à la relation et une probabilité qu'un attribut indique une relation. Kok & Domingos (2008) ramènent le problème du clustering itératif de relations et d'objets à ce qu'ils nomment *invention de prédicats statistiques*, avec le but de découvrir de nouveaux concepts, relations et attributs. La méthode est basée sur l'utilisation de la logique de Markov, une extension probabiliste de la logique des prédicats. Notre approche est dans la lignée de celle de Min *et al.* (2012), dont la motivation principale est d'augmenter le rappel par rapport aux résultats de Kok & Domingos (2008) avec un clustering *soft*. Un clustering itératif en deux étapes est appliqué aux données : dans un premier temps, les entités sont regroupées dans des classes sémantiques qui seront utilisées comme arguments des séquences de relations. Ces relations sont ensuite mises dans le même cluster en utilisant une combinaison des similarités calculées sur les arguments, leurs hyperonymes partagés, et sur les séquences elles-mêmes.

Notre méthode de biclustering est également une combinaison de deux étapes. Contrairement à Min *et al.* (2012), nous utilisons les mêmes attributs, à savoir les co-occurrences, pour regrouper les relations d'abord et ensuite les couples d'entités. Notre approche consiste à regrouper les séquences selon leur distribution sur la totalité des couples d'entités, et à associer à chaque groupe de séquences résultant les couples d'entités qui y sont spécifiques. Cela permet de mettre en relation des couples d'entités même s'ils n'apparaissent qu'avec des variantes différentes de la même séquence. L'autre avantage de la méthode est de produire un clustering recouvrant (*soft clustering*), ce qui correspond mieux à la structure de données recherchée.

Nous partons d'une matrice de séquences $p \in P$ et de couples d'entités $e \in E$ et : $M_{p,e} = P \times E$. Les lignes correspondent aux séquences, les colonnes aux couples d'entités, et les valeurs des celles aux occurrences d'une séquence entre les deux entités dans le corpus. Ces valeurs sont pondérées par sPPMI (équation 1) pour refléter la spécificité des couples d'entités par rapport aux séquences. Les clusters de séquences résultants serviront ensuite à extraire des biclusters en considérant les séquences comme espace de traits et les couples d'entités comme objets.

Les séquences sont regroupés en premier selon leur distribution sur les couples d'entités en appliquant l'algorithme de clustering divisive décrit en 3.3.1. Nous avons privilégié cette méthode parce qu'elle permet de mieux contrôler la taille des clusters par rapport au clustering ascendant. Cet aspect est important puisque la deuxième étape du biclustering produira un clustering recouvrant et nous cherchons à éviter les clusters trop peuplés.

Notre première étape de clustering des séquences divise l'espace en un nombre *n* pré-défini de clusters *C* où $C_n \in P$. Ensuite, nous utiliserons chaque groupe de séquences C_n pour créer une sous-matrice $M_{e,p} = E \times C_n$ avec l'ensemble des couples d'entités comme lignes et les séquences dans le cluster C_n comme colonnes. Comme le poids sPPMI n'est pas symétrique, les valeurs des cellules $M_{e,p}$ ne seront pas identiques aux valeurs $M_{p,e}$. Pour transformer ces matrices en biclusters, nous allons

simplement éliminer les lignes *e* qui ne sont connectées à aucune séquence. Un couple d'entités *e* est connecté à une séquence *p* si $m_{e,p} > 0$, c'est-à-dire $sPPMI(e, p) > 0$. Ainsi, chaque couple d'entités qui a une valeur de spécificité supérieur au seuil avec une des séquences du cluster fera partie du bicluster.

4 Évaluation

4.1 Classification standard et mesures d'évaluation

Les couples d'entités de la matrice ont été classifiés à la main selon notre typologie de relations (Table 1). Un couple peut n'appartenir à aucune relation, aussi bien qu'à plusieurs relations. La liste complète a été examinée deux fois selon la méthode suivante : au premier tour, les 1 385 couples ont été catégorisés sans examiner leurs occurrences dans le corpus. Au deuxième tour, les couples ont été examinés par la même personne, mais en examinant cette fois leurs occurrences. Finalement, les couples d'entités qui n'avaient pas reçu exactement les mêmes catégories ont été ré-examinées et catégorisées.

Nous avons choisi d'évaluer le clustering comme une série de décisions de regrouper deux couples d'entités dans le même cluster ou dans des clusters différents. Cette évaluation est moins influencée par les différences structurelles éventuelles entre deux solutions de clustering, et est donc plus adaptée à nos expériences étendues à de nouveaux types de relations. Ainsi, les paires de couples d'entités qui sont étiquetées avec la même relation dans le corpus de référence et qui se trouvent dans le même cluster constituent les vrais positifs. Les paires qui sont dans le même cluster mais qui n'appartiennent pas à la même relation sont des faux positifs. Les paires qui sont étiquetées avec la même relation dans le corpus de référence mais qui ne sont pas dans le même cluster sont des faux négatifs. Cette évaluation peut être appliquée à des jeux de données qui contiennent des objets non présents dans la référence.

Nous avons également calculé APP *Adjusted Pairwise Precision* (Korhonen *et al.*, 2008) : une mesure de précision moyenne par cluster, pondérée par la taille des clusters. Elle nous servira à estimer la proportion de clusters pertinents.

$$APP = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \frac{nb\ de\ couples\ corrects\ dans\ k_i}{nb\ de\ couples\ dans\ k_i} \times \frac{|k_i| - 1}{|k_i| + 1} \tag{3}$$

Pour chaque expérience, une solution de clustering aléatoire a été produite pour estimer la difficulté de la tâche et l'apport de nos approches. Les clusters aléatoires ont été créés en deux étapes : d'abord, un objet aléatoire différent a été associé à chaque cluster, puis les autres objets ont été associés à un cluster aléatoire.

4.2 Résultats

Le tableau 3 montre les résultats des différentes méthodes de clustering sur les couples d'entités qui ont été étiquetés dans le corpus de référence. Comme nous ne souhaitons pas fixer *a priori* le nombre

de classes à trouver, nous avons testé différents nombres de clusters. Il faut prendre en considération que la mesure APP, par sa pondération, favorise les gros clusters.

Input	#clusters	algorithme	poids	APP	Prec	Rapp	F-mesure
baseline	100	random	N/A	0.0813	0.0955	0.0097	0.0176
baseline	50	random	N/A	0.0883	0.1036	0.0198	0.0332
baseline	25	random	N/A	**0.0979**	**0.1040**	**0.0410**	**0.0588**
sequences	100	divisive	freq	0.2498	0.3037	0.1030	0.1538
sequences	100	divisive	sPPMI	0.3112	**0.4905**	0.0462	0.0844
sequences	50	divisive	freq	0.2985	0.2805	0.1302	0.1778
sequences	50	divisive	sPPMI	0.3625	0.3789	0.0799	0.1320
sequences	25	divisive	freq	**0.3941**	0.2219	**0.1904**	**0.2050**
sequences	25	divisive	sPPMI	0.3555	0.3133	0.1400	0.1936
sequences	100	ascendant	sPPMI	0.3020	**0.4184**	0.1582	0.2296
sequences	50	ascendant	sPPMI	0.2535	0.3246	0.2142	**0.2581**
sequences	25	ascendant	sPPMI	**0.2585**	0.2898	**0.2277**	0.2550
word2vec	100	inclus	divisive	0.3396	**0.5734**	0.0527	0.0965
word2vec	50	inclus	divisive	0.3541	0.4761	0.0890	0.1499
word2vec	25	inclus	divisive	**0.3545**	0.4182	**0.1539**	**0.2250**

TABLE 3 – Évaluation du clustering avec des paramètres divers sur le standard

Nous pouvons constater que la représentation word2vec, comparée à la représentation par séquences sans pondérations avec les mêmes paramètres (algorithme divisif, même nombre de classes) arrive à une meilleure précision et aussi une meilleure APP, au prix d'un rappel bien plus faible. Bien souvent, cette méthode basée sur la similarité entre les entités individuelles trouve des couples d'entités très proches mais peine à retrouver la même relation entre des entités différentes. Par exemple, le cluster le plus fort obtenu par cette méthode est le suivant : *parsing - sentences, parsing - sentence, parses - sentence, parses - sentences, parse - sentence*.

Alors que le clustering divisif utilisant les séquences sans pondération a un meilleur rappel, il reste inférieur au niveau de la précision. [4] Il est intéressant de noter qu'en appliquant la pondération par sPPMI, les résultats sur les séquences se transforment pour ressembler à ceux atteints par word2vec (qui contient la même pondération de manière implicite) : la précision augmente au détriment du rappel. Finalement, l'algorithme ascendant combiné avec sPPMI arrive à un meilleur équilibre entre précision et rappel, pour donner systématiquement les meilleurs résultats en termes de F-mesure.

Des expériences de biclustering ont été effectuées et évaluées sur les mêmes données. Le tableau 4 donne une comparaison entre le biclustering et les méthodes de clustering qui se sont montrées les plus performantes précédemment, c'est-à-dire le clustering ascendant par séquences pour les F-mesures et word2vec pour les meilleures précisions. Nous constatons que le biclustering produit systématiquement le meilleur rappel, et ses F-mesures sont au coude à coude avec l'autre approche gagnante, le clustering ascendant.

Ces mesures sont toutefois à prendre avec précaution : le biclustering est le seul algorithme à pouvoir classifier un objet dans plusieurs clusters. Bien que ce soit un avantage pour notre tâche, il

4. Cependant, le maximum de précision moyenne par APP est atteint par cette méthode, ce qui suggère qu'elle produirait moins de classes « poubelle ».

Input	#clusters	algorithme	poids	APP	Prec	Rapp	F-mesure
sequences	100	ascendant	sPPMI	0.3020	0.4184	0.1582	0.2296
word2vec	100	divisive	inclus	**0.3396**	**0.5734**	0.0527	0.0965
sequences	100	biclustering	sPPMI	0.2168	0.1625	**0.6037**	**0.2561**
sequences	50	ascendant	sPPMI	0.2535	0.3246	0.2142	**0.2581**
word2vec	50	divisive	inclus	**0.3541**	**0.4761**	0.0890	0.1499
sequences	50	biclustering	sPPMI	0.2031	0.1524	**0.6364**	0.2460
sequences	25	ascendant	sPPMI	0.2585	0.2898	0.2277	**0.2550**
word2vec	25	divisive	inclus	**0.3545**	**0.4182**	0.1539	0.2250
sequences	25	biclustering	sPPMI	0.1806	0.1363	**0.7352**	0.2299

TABLE 4 – Comparaison des meilleurs clusterings avec le biclustering sur le standard

s'accompagne d'une baisse de précision. En effet, notre biclustering - comme le clustering ascendant - est enclin à produire un ou plusieurs clusters « poubelle » de grande taille. La mesure APP nous permet d'en estimer l'impact. Comme elle pénalise moins le biclustering que l'autre mesure de précision, cela suggère qu'un nombre limité de clusters « poubelle » sont à la source de la baisse de précision.

5 Conclusion et perspectives

Nous avons présenté une première approche pour l'analyse automatique de l'état de l'art dans des articles scientifiques. Le processus repose d'une part sur une annotation automatique des entités pertinentes du texte et leur alignement à des ressources ontologiques, d'autre part sur une classification non supervisée des relations sémantiques qui les relient. Cette deuxième étape a pour but à la fois de catégoriser les instances de relations extraites du texte et de découvrir de nouveaux types de relations. Les séquences de textes caractéristiques des nouvelles relations permettront de les qualifier.

Les expériences de classification non supervisée de relations sémantiques nous ont permis de comparer les avantages et les limites des différentes techniques testées. Alors qu'une très bonne précision peut être atteinte par les représentations sémantiques distributionnelles, le rappel de cette méthode est trop faible pour permettre de découvrir de nouveaux types de relations. Le biclustering est en revanche plus performant en rappel et présente un autre avantage important : les séquences caractéristiques des relations sont extraites en même temps que les instances d'entités qu'elles relient. Cela facilite l'interprétation et la découverte de nouveaux types de relations qui n'ont pas encore été nommées (voir tableau 5). Les séquences semblent plus faciles à interpréter et à associer à une relation que les couples d'entités hors contexte.

Pour améliorer la précision de cette méthode, une extension naturelle est d'exploiter le poids de l'association entre les couples d'entités et les séquences (Tanay *et al.*, 2002). Cela permettrait également d'associer un score de qualité à chaque bicluster résultant.

séquences	entités
show that the	results - method
show that our	experiments - approach
showed that the	results - performance
show that the proposed	results - system
of the first	results - standard
show that	results - model
indicate that our	results - approach
show that this	results - analysis
	probability - occurrence
	results - parsing
	experiments - method
	output - parser
	parse tree - sentence
that	translations - words
to automatically	techniques - extract
used to	tools - help
only	method - extract
that automatically	methods - extract
to improve the	approach - model
to obtain	information - word
to improve	information - speech
to evaluate the	techniques - parsing
	method - machine translation
	order - performance
	approach - statistical machine translation

TABLE 5 – Exemples de biclusters

Remerciements

Ce travail a bénéficié partiellement d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'Avenir » portant la référence ANR-10-LABX-0083.

Références

BANKO M., CAFARELLA J., SODERLAND S., BROADHEAD M. & ETZIONI O. (2007). Open information extraction from the web. In *IJCAI*.

BARONI M., BERNARDI R., DO N.-Q. & SHAN C.-C. (2012). Entailment above the word level in distributional semantics. In *ACL '12*.

BARONI M., DINU G. & KRUSZEWSKI G. (2014). Dont count, predict ! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*.

BORDEA G. (2013). *Domain Adaptive Extraction of Topical Hierarchies for Expertise Mining*. Phd thesis, National University of Ireland, Galway.

BORDEA G., BUITELAAR P. & POLAJNAR T. (2013). Domain-independent term extraction through domain modelling. In *10th International Conference on Terminology and Artificial Intelligence (TIA 2013)*.

CHAVALARIAS D. & COINTET J.-P. (2013). Phylomemetic patterns in science evolution - the rise and fall of scientific fields. *PLOS ONE*, **8(2)**.

CUNNINGHAM H., MAYNARD D., BONTCHEVA K. & TABLAN V. (2002). GATE : A framework and graphical development environment for robust NLP tools and applications. In *ACL*.

DAILLE B., JACQUIN C., MONCEAUX L., MORIN E. & ROCHETEAU J. (2013). TTC TermSuite : Une chaîne de traitement pour la fouille terminologique multilingue. In *TALN*.

DEL CORRO L. & GEMULLA R. (2013). Clausie : Clause-based open information extraction. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW'13.

FERRET O. (2013). Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *TALN-RÉCITAL*.

GÁBOR K., ZARGAYOUNA H., BUSCALDI D., TELLIER I. & CHARNOIS T. (2016). Semantic annotation of the acl anthology corpus for the automatic analysis of scientific literature. In *LREC '16*, Portoroz, Slovenia. in press.

GETZ G., LEVINE E. & DOMANY E. (2000). Coupled two-way clustering analysis of gene microarray data. *PNAS*, **97**.

HARTIGAN J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, **67**.

HENDRICKX I., KIM S. N., KOZAREVA Z., NAKOV, P.AND O SÉAGHDHA D., PADÓ S., PENNACCHIOTTI M., ROMANO L. & SZPAKOWICZ S. (2010). Semeval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations*.

HOBBS J. R. & RILOFF E. (2010). Information extraction. In N. INDURKHYA & F. J. DAMERAU, Eds., *Handbook of Natural Language Processing, Second Edition*. Boca Raton, FL : CRC Press, Taylor and Francis Group.

JURGENS D. A., TURNEY P., MOHAMMAD S. M. & HOLYOAK K. J. (2012). Semeval-2012 task 2 : Measuring degrees of relational similarity. In *Proceedings of the Workshop on Semantic Evaluations*.

KOK S. & DOMINGOS P. (2008). Extracting semantic networks from text via relational clustering. In *Proceedings of ECML PKDD'08*.

KORHONEN A., KRYMOLOWSKI Y. & COLLIER N. (2008). The choice of features for classification of verbs in biomedical texts. In *COLING*.

LARSEN P. O. & VON INS M. (2010). The rate of growth in scientific publication and the decline in coverage provided by science citation index.

LEVY O., GOLDBERG Y. & DAGAN I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the ACL*, **3**.

LIN D. & PANTEL P. (2001). Dirt : Discovery of inference rules from text. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

MADEIRA S. & OLIVEIRA A. (2004). Biclustering algorithms for biological data analysis : A survey. *Transactions on Computational Biology and Bioinformatics*, **1**.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*.

MIKOLOV T., YIH W. & ZWEIG G. (2013b). Linguistic regularities in continuous space word representations. In *NAACL*.

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

NAVIGLI R. & PONZETTO S. P. (2012). BabelNet : The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, **193**.

OMODEI E., COINTET J.-P. & POIBEAU T. (2014). Mapping the natural language processing domain : Experiments using the acl anthology. In *LREC*.

OSBORNE F. & MOTTA E. (2012). Mining semantic relations between reserach areas. In *International Semantic Web Conference, Boston (MA)*.

OSBORNE F. & MOTTA E. (2014). Rexplore : unveiling the dynamics of scholarly data. *Digital Libraries*, **8**(12).

PETASIS G., KARKALETSIS V., PALIOURAS G., KRITHARA A. & ZAVITSANOS E. (2011). Ontology population and enrichment : State of the art. In *Knowledge-driven multimedia information extraction and ontology evolution* : Springer-Verlag.

PRESUTTI V., CONSOLI S., NUZZOLESE A. G., RECUPERO D. R., GANGEMI A., BANNOUR I. & ZARGAYOUNA H. (2014). Uncovering the semantics of wikipedia pagelinks. In *Knowledge Engineering and Knowledge Management*, p. 413–428. Springer.

RADEV D., MUTHUKRISHNAN P. & QAZVINIAN V. (2009). The ACL Anthology Network Corpus. In *ACL Workshop on Text and Citation Analysis for Scholarly Digital Libraries*.

SATELI B. & WITTE R. (2015). What's in this paper ? : Combining rhetorical entities with linked open data for semantic literature querying. In *Proceedings of the 24th International Conference on World Wide Web*.

SHOTTON D. (2010). CiTO, the Citation Typing Ontology. *J. Biomedical Semantics*, **1**(S-1).

TANAY A., SHARAN R. & SHAMIR R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**.

TURNEY P. (2005). Measuring semantic similarity by latent relational analysis. In *IJCAI-05*.

WANG W., BESANÇON R., FERRET O. & GRAU B. (2013). Extraction et regroupement de relations entre entités pour l'extraction d'information non supervisée. *TAL*, **54**(2).

WEEDS J., CLARKE D., REFFIN J., WEIR D. & KELLER B. (2014). Learning to distinguish hypernyms and co-hyponyms. In *COLING '14*.

YAO L., HAGHIGHI A., RIEDEL S. & MCCALLUM A. (2011). Structured relation discovery using generative models. In *EMNLP'11*.

ZHAO Y. & KARYPIS G. (2002). Evaluation of hierarchical clustering algorithms for document datasets. In *CIKM*.

ZHAO Y., KARYPIS G. & FAYYAD U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining for Knowledge Discovery*, **10**.