Réduction de la dispersion des données par généralisation des contextes distributionnels : application aux textes de spécialité

Auteur1[1] Auteur2[2]
(1) Affliation
Affliation
Affliation
Email
(2) Affliation
Affliation
Email

Résumé. Les modèles d'espace vectoriels mettant en œuvre l'analyse distributionnelle s'appuient sur la redondance d'informations se trouvant dans le contexte des mots à associer. Cependant, ces modèles souffrent du nombre de dimensions considérable et de la dispersion des données dans la matrice des vecteurs de contexte. Il s'agit d'un enjeu majeur sur les corpus de spécialité pour lesquels la taille est beaucoup plus petite et les informations contextuelles moins redondantes. Nous nous intéressons au problème de la dispersion des données sur des corpus de spécialité et proposons une méthode permettant de densifier la matrice en généralisant les contextes distributionnels. L'évaluation de la méthode sur un corpus médical en français montre qu'avec une petite fenêtre graphique et l'indice de Jaccard, la généralisation des contextes avec les patrons lexico-syntaxiques permet d'améliorer les résultats, alors qu'avec une large fenêtre et le cosinus, il est préférable de généraliser avec l'inclusion lexicale.

Abstract. Vector space models implement the distributional hypothesis relying on the repetition of information occurring in the contexts of words to associate. However, these models suffer from a high number of dimensions and data sparseness in the matrix of contextual vectors. This is a major issue with specialized corpora that are of much smaller size and with much lower context frequencies. We tackle the problem of data sparseness on specialized texts and we propose a method that allows to make the matrix denser, by generalizing of distributional contexts. The evaluation of the method is performed on a French medical corpus, and shows that with a small graphical window and the Jaccard Index, the context generalization with lexico-syntactic patterns improves the results, while with a large window and the cosine measure, it is better to generalize with lexical inclusion.

Mots-clés : Analyse distributionnelle, textes de spécialité, hyperonymie, dispersion des données, modèle d'espace vectoriel, méthode hybride.

Keywords: Distributional analysis, specialized texts, hypernymy, data sparseness, Vector Space Model, hybrid method.

1 Introduction

L'analyse distributionnelle (AD) s'appuie sur l'hypothèse que des mots apparaissant dans des contextes similaires ont tendance à être proches sémantiquement (Harris, 1954; Firth, 1957). Cette hypothèse est généralement mise en œuvre grâce à des modèles d'espace vectoriels (VSM) où les vecteurs représentent à la fois ces informations contextuelles mais également des données statistiques distributionnelles (Sahlgren, 2006). Chaque mot cible d'un texte est représenté comme un point dans un espace mathématique en fonction de ses propriétés distributionnelles dans le texte (Turney & Pantel, 2010; Lund & Burgess, 1996). La similarité sémantique entre deux mots est alors définie comme une proximité dans un espace à n-dimensions où chaque dimension correspond à des contextes partagés possibles. Les VSM ont ainsi l'avantage de permettre une quantification facile de la proximité sémantique entre deux mots en mesurant la distance entre deux vecteurs au sein de cet espace (par ex. le cosinus de leur angle). Cependant, outre le nombre important de dimensions (à titre d'exemple, Sahlgren (2006) manipule des VSM allant jusqu'à plusieurs millions de dimensions), les VSM souffrent également de la dispersion des données dans la matrice représentant l'espace vectoriel (Chatterjee & Mohan, 2008):

beaucoup d'éléments de la matrice sont à 0 car peu de contextes sont associés à un mot cible. Cet inconvénient est dû notamment à la distribution des mots dans le corpus (Baroni *et al.*, 2009): quelle que soit la taille du corpus, la plupart des mots ont des fréquences basses et un nombre de contextes très limité au regard du nombre de mots dans le corpus. Ces deux derniers points rendent difficile le calcul de la similarité entre deux mots. En conséquence, les méthodes basées sur l'analyse distributionnelle obtiennent de meilleures performances lorsque beaucoup d'informations sont disponibles, et notamment sur des corpus de langue générale, en général très volumineux (Weeds & Weir, 2005; van der Plas, 2008). Mais, la réduction de la dispersion des données reste un aspect important sur des corpus de langue générale. Elle est aussi un enjeu majeur lorsque l'on travaille sur des corpus de spécialité. En effet, ces corpus se caractérisent par des tailles beaucoup plus petites, avec des fréquences et un nombre de contextes différents d'autant plus faibles. Nous nous intéressons à ce dernier point en proposant l'adaptation d'une méthode d'analyse distributionnelle qui permette d'obtenir de meilleurs résultats sur des textes de spécialité. Pour cela, nous avons cherché à réduire la diversité des contextes en les généralisant. Il est alors possible d'augmenter la fréquence des contextes distributionnels qui résultent de cette généralisation et ainsi réduire la dispersion des données et la dimension de l'espace vectoriel. Nous présentons ici une méthode à base de règles permettant la généralisation des contextes distributionnels à l'aide de résultats issus de méthodes d'acquisition de relations sémantiques. Nous adaptons les paramètres de la méthode distributionnelle utilisée aux corpus de spécialité, en intégrant notamment ces contextes généralisés.

Dans la suite de l'article, nous présentons tout d'abord un état de l'art des méthodes de réduction de la dispersion des données dans les méthodes distributionnelles. Puis nous décrivons la méthode de généralisation de contextes proposée, ainsi que les expériences réalisées pour évaluer son impact sur un corpus de spécialité. Les résultats sont évalués puis analysés en terme de précision, de R-précision et de MAP.

2 État de l'art

La réduction de la dispersion des données est un enjeu majeur en analyse distributionnelle. Pour cela, les méthodes proposées visent à influer sur la sélection des contextes utiles ou à intégrer des informations sémantiques de manière à modifier la distribution des contextes. Ainsi, Broda *et al.* (2009) proposent de pondérer les contextes non pas en utilisant les fréquences des contextes à l'état brut comme il est d'usage, mais en ordonnant les contextes en fonction de leur fréquence, puis se servent du rang pour pondérer les contextes. D'autres approches s'appuient sur des modèles de langue pour déterminer les substituts les plus probables pour représenter les contextes (Baskaya *et al.*, 2013). Ces modèles assignent des probabilités à des séquences arbitraires de mots en se basant sur les fréquences de co-occurrence dans un corpus d'entraînement (Yuret, 2012). Ces mots substituts et leurs probabilités sont ensuite utilisés pour créer des paires de mots de manière à alimenter un modèle de co-occurrence, avant d'utiliser un algorithme de clustering. Ces méthodes sont limitées car leur performance est proportionnelle à la taille du vocabulaire et elles nécessitent de disposer de données d'entraînement importantes.

L'influence sur les contextes peut également être réalisée en y intégrant de l'information sémantique supplémentaire. En effet, il a été démontré que l'intégration de ce type d'information afin de modifier la méthode classique de l'AD permet d'améliorer sa performance (Tsatsaronis & Panagiotopoulou, 2009). Cette information sémantique, ou plus précisément les relations sémantiques, peuvent être calculées automatiquement ou provenir d'une ressource existante. Ainsi, avec une méthode d'amorçage, Zhitomirsky-Geffet & Dagan (2009) modifient les poids des éléments au sein des contextes en se basant sur les voisins sémantiques trouvés à l'aide d'une mesure de similarité distributionnelle. En s'appuyant sur ces travaux, Ferret (2013) s'intéresse au problème des mots de faibles fréquences. Afin de mieux prendre en compte ces informations, il propose d'utiliser un jeu d'exemples positifs et négatifs sélectionnés de manière non-supervisée à partir d'un thésaurus distributionnel, et ainsi entraîner un classifieur supervisé. Ce classifieur est ensuite appliqué pour réordonnancer les voisins sémantiques. La méthode permet ainsi d'améliorer la qualité de la relation de similarité entre des noms de faible ou moyenne fréquence.

D'autres travaux s'intéressent au problème de la dispersion des données d'un point de vue algorithmique en cherchant à limiter les dimensions de la matrice des contextes, notamment en la lissant afin de réduire le nombre de composants vectoriels (Turney & Pantel, 2010). Ainsi, l'analyse Sémantique Latente (LSA) (Landauer & Dumais, 1997; Padó & Lapata, 2007) met en œuvre une méthode de factorisation de matrice par Décomposition aux Valeurs Singulières (SVD). Les données originales de la matrice des contextes sont abstraites en composants linéaires indépendants, permettant ainsi de réduire le bruit et d'en faire ressortir les éléments essentiels. Outre la réduction du coût de traitement, la réduction de dimension améliore considérablement la précision dans les applications de la LSA. Par exemple, l'utilisation de la SVD à la similarité entre mots permet ainsi d'atteindre des scores équivalents à ceux d'un humain dans un Test du TOEFL

avec des questions de synonymie à choix multiples (Landauer & Dumais, 1997). Ceci s'explique, entre autres, par le fait qu'avec les mesures de similarité les plus fréquemment employées, les termes sont vus uniquement comme similaires s'ils apparaissent dans les mêmes contextes. En ce qui concerne les mots de faible fréquence, la SVD est une manière de simuler le texte manquant, en compensant le manque de données (Vozalis & Margaritis, 2003). Certaines méthodes, comme la factorisation en matrice non-négative (Lee & Seung, 1999), permettent de mieux modéliser la fréquence des mots. Mais, lorsqu'il s'agit d'acquérir des relations sémantiques, les performances semblent moins bonnes que celles obtenues avec la LSA (Turney & Pantel, 2010; Utsumi, 2010).

Aussi, la réduction de dimensions facilite le traitement des vecteurs de contextes, mais ne résout pas le problème initial de construction d'une matrice de co-occurrence potentiellement immense. Ainsi, l'*indexation aléatoire*, ou *Random Indexing* (RI) Kanerva *et al.* (2000), apporte une solution à ce problème en construisant incrémentalement la matrice des contextes en fonction d'un vecteur d'index du mot cible généré aléatoirement. Cette approche permet d'éviter la construction d'une trop grande matrice tout en réduisant la dimension de la matrice. Les performances obtenues avec le RI sont alors équivalentes à ceux obtenus avec la LSA lors de l'identification de synonymes de manière similaire au test du TOEFL (Karlgren & Sahlgren, 2001). Récemment, Polajnar & Clark (2014) ont montré que la sélection des meilleurs contextes combinés à une normalisation de leur poids permet d'améliorer la qualité de la matrice obtenue par SVD. Dans des cadres applicatifs comme la recherche de définition et le calcul de similarité entre syntagmes, leur impact sur les performances de modèles de sémantique compositionnelle dépend des opérateurs utilisés.

A l'instar de (Tsatsaronis & Panagiotopoulou, 2009; Zhitomirsky-Geffet & Dagan, 2009; Ferret, 2013), notre approche ajoute des informations sémantiques dans les contextes distributionnels, mais notre objectif diffère de cet ajout: il s'agit de réduire le nombre de contextes et d'augmenter leur fréquence. Et contrairement aux méthodes basées sur la SVD qui limitent les contextes en supprimant de l'information, les contextes sont regroupés en généralisant l'information en contexte grâce à l'intégration de connaissances sémantiques supplémentaires calculées sur le corpus de travail.

3 Matériel

Nous présentons dans cette section le corpus de travail ainsi que les approches mises en œuvre pour acquérir les relations sémantiques utilisées lors de la généralisation des contextes.

3.1 Corpus

Pour évaluer notre approche, nous avons utilisé le corpus Menelas (Zweigenbaum, 1994). Il s'agit d'une collection de textes du domaine médical, en français, dont la thématique est les maladies coronariennes. Le corpus comporte 84 839 mots. Il est constitué de deux grandes parties : un manuel de référence sur la coronarographie et les maladies coronariennes (environ 15 000 mots), et un ensemble de comptes rendus d'hospitalisation et de lettres de médecins hospitaliers aux médecins traitants concernant des malades atteints d'une maladie coronarienne (environ 70 000 mots).

Le corpus a été analysé à travers la plate-forme de TAL Ogmios (Hamon *et al.*, 2007). Nous avons configuré la plate-forme de manière à ce que cette analyse linguistique comprenne un étiquetage morpho-syntaxique et une lemmatisation du corpus, à l'aide de TreeTagger (Schmid, 1994), et une extraction de termes à l'aide de YaTeA (Aubin & Hamon, 2006), celle-ci permettant d'identifier dans notre corpus de travail, les groupes nominaux dénotant les notions du domaine.

3.2 Acquisition de relations sémantiques

La méthode de généralisation des contextes distributionnels s'appuie sur des relations sémantiques existantes. Pour obtenir ces relations à partir de corpus, nous avons choisi d'utiliser plusieurs approches classiques d'acquisition de relations sémantiques entre termes : utilisation de patrons lexico-syntaxiques (PLS), de l'hypothèse d'inclusion lexicale (IL), et de règles de variation terminologique (VT).

Patrons lexico-syntaxiques (PLS) Nous utilisons les patrons définis par (Morin & Jacquemin, 2004) pour acquérir des relations d'hyperonymie entre termes simples ou complexes, soit par exemple:

1. {quelques | plusieurs etc.} SN : LISTE.
2. {autre}? SN tels que LISTE.

où SN est un syntagme nominal et LISTE une liste de syntagmes.

Inclusion lexicale (IL) Cette approche s'appuie sur l'hypothèse selon laquelle si un terme (ex: *infarctus*) est inclus lexicalement dans un autre (ex: *infarctus du myocarde*) il existe généralement une relation d'hyperonymie entre ces deux termes (Grabar & Zweigenbaum, 2004). Nous contraignons l'approche en exploitant l'analyse syntaxique des termes fournie par YaTeA . Nous ne considérons ici que les relations syntaxiques entre le terme complexe et sa tête.

Variation terminologique (VT) Nous utilisons la méthode d'acquisition de variantes terminologiques proposés par (Jacquemin, 2001) et implémentée dans Faster. Cette méthode exploite des règles de transformation morpho-syntaxique, essentiellement l'insertion (*chirurgie coronarienne / chirurgie de revascularisation coronarienne*) pour identifier des relations sémantiques entre termes. Ces règles de variation terminologique comme l'insertion peuvent notamment d'acquérir des relations d'hyperonymie (*anomalie significative / anomalie coronarienne significative*).

Nous disposons ainsi de trois sources de relations sémantiques offrant principalement des relations d'hyperonymie. Les patrons lexico-syntaxiques nous fournissent le moins de relations, avec 98 relations d'hyperonymie. L'inclusion lexicale nous permet de disposer un nombre nettement plus important de relations : 7 187. Enfin, nous avons pu acquérir 171 variantes terminologiques.

4 Méthode de généralisation de contextes distributionnels

L'analyse distributionnelle appliquée à des corpus de spécialité ou des corpus de petite taille souffre d'une dispersion des données : la matrice des contextes, représentant la distribution des mots ou des termes, est très creuse (beaucoup d'éléments ont une valeur nulle). Une solution à ce problème consiste à densifier la matrice des contextes en faisant abstraction des variations superficielles ou des contextes peu significatifs statistiquement ou liés au bruit de la méthode d'identification de ces distributions. Pour cela, nous avons cherché, dans un premier temps, à filtrer les contextes de manière à sélectionner ceux qui semblent les plus pertinents, et surtout, à généraliser les contextes en exploitant des informations sémantiques extraites du corpus. En particulier, nous utilisons des relations sémantiques acquises automatiquement par des approches utilisées habituellement sur les corpus de spécialité : patrons lexico-syntaxiques, inclusion lexicale, variation terminologique.

Dans un premier temps, nous décrivons le processus d'analyse distributionnelle mis en œuvre, puis nous présentons la méthode de généralisation des contextes distributionnelle que nous proposons.

4.1 Méthode distributionnelle

Dans le contexte d'applications en langue de spécialité, l'identification de relations sémantiques entre des noms et des termes est primordiale. Pour cela, nous nous restreignons à l'analyse distributionnelle entre des noms et des termes simples ou complexes. Ces deux catégories de mots constitueront les mots cibles.

La méthode d'analyse distributionnelle que nous avons mise en œuvre suit le schéma présenté dans la figure 1. Il est d'abord nécessaire de définir les contextes distributionnels des mots cibles. Nous avons ainsi choisi d'utiliser des fenêtres graphiques d'une largeur donnée. Ainsi, les contextes sont composés de mots qui co-occurent avec le mot cible au sein de la fenêtre graphique. Nous considérons comme contexte les adjectifs, noms, verbes et termes en écartant les mots vides (déterminants, conjonctions, adverbes, etc.). Que ce soit pour les mots cibles ou les contextes, nous considérons leurs formes lemmatisées.

Bien que les contextes soient généralement calculés sur des dépendances syntaxiques, nous avons choisi d'utiliser des contextes graphiques au sein d'une phrase et autour d'un mot cible pour plusieurs raisons :

- les textes de spécialité nécessitent une analyse particulière et nous ne disposons pas d'un analyseur syntaxique adapté.

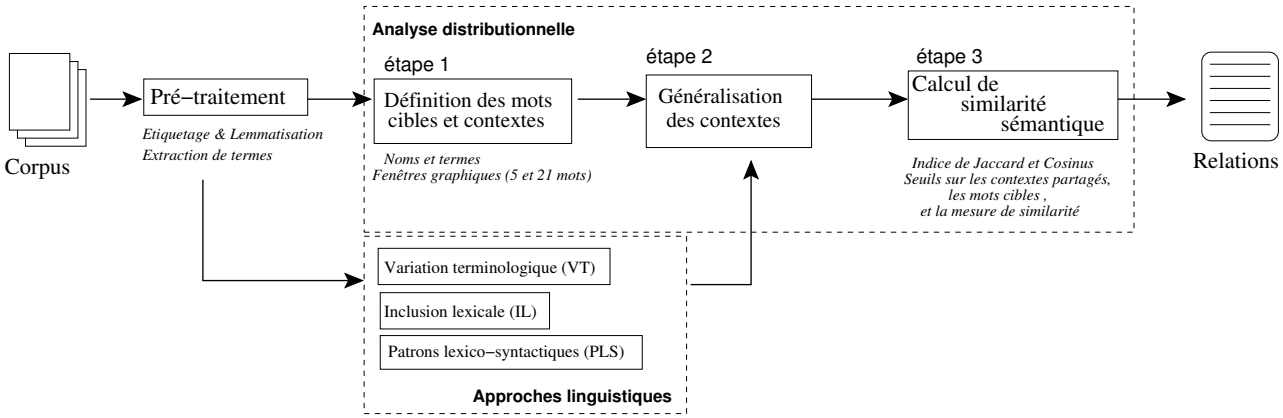

Figure 1: Processus d'analyse distributionnelle

- les fenêtres graphiques étant moins restreintes que l'analyse syntaxique, elles permettent de prendre en compte un plus grand nombre de contextes, ce qui facilite la normalisation.
- l'intégration d'une analyse syntaxique dédiée à la langue générale peut nécessiter une mise en place assez coûteuse sans réellement apporter de plus value lors de l'analyse distributionnelle.

La phase de généralisation des contextes intervient après la définition des contextes distributionnels. Nous décrivons cette étape en détail à la section suivante.
Après avoir extrait et généralisé les contextes d'apparition de chaque mot cible, un score de similarité sémantique est calculé entre chaque couple de mots cibles, en tenant compte des contextes partagés. Nous avons utilisé l'indice de Jaccard, celui-ci étant reconnu pour être adapté aux corpus de spécialité (Grefenstette, 1994) :

$$sim_{JACCARD}(w_m, w_n) = \frac{|ctxt(w_m) \bigcap ctxt(w_n)|}{|ctxt(w_m) \bigcup ctxt(w_n)|}$$

Ainsi, l'indice de Jaccard normalise le nombre de contextes partagés par deux mots cibles w_m et w_n par le nombre total de contextes $ctxt(w)$ de ces mots. Nous utilisons également le cosinus. Cette mesure reflète l'angle entre deux vecteurs représentant chacun un mot cible :

$$sim_{COSINUS}(w_m, w_n) = \frac{|ctxt(w_m) \cap ctxt(w_n)|}{\sqrt{|ctxt(w_m)| \times |ctxt(w_n)|}}$$

Le score de similarité permet ainsi de quantifier dans quelle proportion deux termes sont proches. Il est cependant nécessaire d'appliquer des seuils afin de limiter le nombre de relations proposé et d'écarter les relations potentiellement fausses. Pour cela, il est possible d'appliquer un seuil sur le score de similarité afin de ne retenir que les relations dont la similarité est suffisamment élevée. Nous avons également cherché à densifier la matrice des contextes en appliquant des seuils sur trois paramètres distributionnels : le nombre de contextes partagés, la fréquence des contextes partagés, la fréquence des mots cibles. Pour chaque paramètre, un seuil est calculé automatiquement en fonction du corpus: actuellement ce seuil est déterminé comme étant la moyenne des valeurs prises par chaque paramètre sur l'ensemble du corpus. Lors des expériences présentées à la section 5, nous testerons l'impact de ces seuils sur les résultats.

4.2 Règles de généralisation des contextes distributionnels

Le processus de généralisation des contextes intervient après l'étape de définition de ces contextes. L'objectif est d'une part, de diminuer la diversité des contextes distributionnels, d'autre part d'augmenter le nombre d'occurrences des contextes, c'est-à-dire leur fréquence. Ces contextes résultent de l'application de règles de généralisation. Dans cette perspective, nous avons choisi d'utiliser des informations sémantiques additionnelles, calculées sur le corpus, et fournissant des indices de généralisation.

Ainsi, une fois que les mots cibles et les contextes ont été définis, nous généralisons les contextes avec des relations sémantiques acquises automatiquement sur le corpus de travail à l'aide des méthodes décrites à la section 3.2 : patrons lexico-syntaxiques dédiés à l'hyperonymie, inclusion lexicale, variation terminologique. Les deux premières méthodes proposent en général principalement des relations d'hyperonymie et seront utilisées pour généraliser les contextes. En revanche, la variation terminologique ne propose pas de relations typées sémantiquement. Aussi, étant donné que l'opération d'insertion est la seule utilisée pour acquérir des variantes, nous avons considéré que les relations obtenues étaient des relations d'hyperonymie. Le terme hyperonyme et le terme hyponyme sont identifiés à partir du nombre de mots présents dans chaque terme : le terme le plus court correspond alors à l'hyperonyme (*lésion significative*), et le terme le plus long à l'hyponyme (*lésion coronaire significative*).

Nous disposons alors, pour chaque mot $ctxt_i(w)$ dans le contexte du mot w, de trois ensembles de relations d'hyperonymie, $\mathbb{H}_s(ctxt_i(w)) = \{H_1, \ldots, H_n\}$: $\mathbb{H}_{PLS}$, $\mathbb{H}_{IL}$ et $\mathbb{H}_{VT}$, l'ensemble des hyperonymes pouvant être vide. Nous avons défini deux règles de substitution permettant de généraliser les contextes. Ainsi, pour chaque mot $ctxt_i(w)$ en contexte d'un mot w, nous appliquons l'une des règles suivantes :

1. si $|\mathbb{H}_S(ctxt_i(w))| = 1$, alors $ctxt_i(w) := H_1$

   Si au mot dans le contexte correspond un seul hyperonyme (H_1) acquis par une ou plusieurs méthodes S, le mot est remplacé par cet hyperonyme. Par exemple, si l'inclusion lexicale fournit la relation *restriction / restriction du débit coronaire*, *restriction du débit coronaire* est remplacée par *restriction*.

2. si $|\mathbb{H}_S(ctxt_i(w))| > 1$, $ctxt_i(w) = \text{argmax}_{|H_i|}(\mathbb{H}_S(ctxt_i(w)))$

   Si le contexte correspond à plusieurs hyperonymes acquis par une ou plusieurs méthodes S, nous prenons en compte la fréquence des hyperonymes $|H_1|, \ldots, |H_n|$ dans le corpus, et nous choisissons l'hyperonyme dont la fréquence est la plus élevée dans le corpus.

   Par exemple, si pour le mot *artère coronaire* dans le contexte les patrons lexico-syntaxiques fournissent les hyperonymes suivants : *veine*, *artère*, *vaisseau*, celui qui est le plus fréquent est choisi et utilisé pour remplacer *artère coronaire* dans le contexte.

Quand plusieurs ensembles de relations d'hyperonymie sont disponibles, la phase de généralisation des contextes est réalisée individuellement ou de manière séquentielle : les contextes sont généralisés en utilisant les ensembles de relations les uns à la suite des autres.

5 Expériences et évaluation

5.1 Expériences

Nous avons réalisé plusieurs séries d'expériences sur le corpus Menelas afin d'évaluer l'impact des règles de généralisation proposées. Nous utilisons comme résultats de référence (*baseline*) les résultats obtenus avec l'analyse distributionnelle seule. Nous avons tout d'abord évalué l'importance des seuils sur les paramètres distributionnels (voir section 4.1). Chaque expérience décrite ci-dessous a été réalisée en appliquant ou non les seuils définis. Lorsque les seuils sont utilisés, ceux-ci ont été calculés sur la *baseline*. La table 1 résume les valeurs des seuils utilisées.

Paramètres	Fenêtre de 21 mots	Fenêtre de 5 mots
Score de similarité	Jaccard : $sim > 0,000999$ Cosinus : $sim > 0.9699$	Jaccard : $sim > 0,000999$ Cosinus : $sim > 0.9699$
Nombre de contextes	2	1
Fréquence des contextes	3	2
Fréquence des mots cibles	3	3

Table 1: Définition des valeurs des seuils sur les paramètres distributionnels et sur le score de similarité, en fonction de la taille des fenêtres (de 21 et 5 mots) et des mesures de similarité (Jaccard et Cosinus)

Afin de cerner la contribution de chaque méthode linguistique décrite à la section 3.2, nous avons défini un ensemble d'expériences où la généralisation des contextes est réalisée en utilisant les relations d'hyperonymie proposées par

chaque méthode individuellement. Les règles de généralisation des contextes distributionnels $ctxt_i(w)$ sont alors appliqués en utilisant séparément les ensembles $\mathbb{H}_{PLS}(ctxt_i(w))$ – relations d'hyperonymie acquises à l'aide des patrons lexico-syntaxiques (AD/PLS), $\mathbb{H}_{IL}(ctxt_i(w))$ – relations d'hyperonymie issues de l'inclusion lexicale (AD/IL), et $\mathbb{H}_{VT}(ctxt_i(w))$ – variantes terminologiques (AD/VT).

Puis, de manière séquentielle, nous avons appliqué les règles de généralisation en utilisant les ensembles de relations d'hyperonymie proposées par deux approches linguistiques ($\mathbb{H}_{PLS}(ctxt_i(w))$ puis $\mathbb{H}_{IL}(ctxt_i(w))$ – AD/PLS+IL, $\mathbb{H}_{VT}(ctxt_i(w))$ puis $\mathbb{H}_{PLS}(ctxt_i(w))$ – AD/VT+PLS, etc.). Tous les contextes sont alors généralisés en utilisant les relations proposées par l'un des ensembles (par exemple $\mathbb{H}_{PLS}(ctxt_i(w))$), puis les contextes généralisés ou non sont à nouveau généralisés en utilisant un autre ensemble de relations (par exemple $\mathbb{H}_{IL}(ctxt_i(w))$). De même, nous combinons les trois ensembles de relations (par exemple, $\mathbb{H}_{PLS}(ctxt_i(w))$ puis $\mathbb{H}_{IL}(ctxt_i(w))$ puis $\mathbb{H}_{VT}(ctxt_i(w))$ – AD/PLS+IL+VT). En combinant des sources de relations d'hyperonymie de plusieurs manières, nous souhaitons d'une part, évaluer la complémentarité des approches utilisées pour généraliser les contextes, et d'autre part, étudier l'impact de l'ordre de ces méthodes dans la séquence de généralisation.

Nous avons également considéré toutes les relations d'hyperonymie indépendamment de la méthode utilisée pour les acquérir. On considère alors l'ensemble $H(ctxt_i(w)) = \mathbb{H}_{PLS}(ctxt_i(w)) \cup \mathbb{H}_{IL}(ctxt_i(w)) \cup \mathbb{H}_{VT}(ctxt_i(w))$ – AD/ALL3, pour appliquer les règles de généralisation sur le contexte $ctxt_i(w)$.

L'ensemble des expériences a été réalisé sur deux tailles de fenêtres : 5 mots ($\pm$ 2 mots, centrée sur le mot cible) et 21 mots ($\pm$ 10 mots, centrée sur le mot cible). En effet, la taille des fenêtres a une influence sur le nombre et la qualité mais aussi sur le type des relations obtenues par analyse distributionnelle. En général, une fenêtre de taille restreinte (5 mots) permet de disposer d'un plus grand nombre de contextes pertinents pour un mot cible donné, mais conduit à une dispersion des données plus important qu'avec une fenêtre plus large (Rapp, 2003). De plus, les résultats obtenus avec des fenêtres de taille restreinte sont de meilleure qualité, en particulier pour des relations classiques (synonymie, antonymie, hyperonymie, méronymie, etc.) alors que des fenêtres plus larges sont plus adaptées à l'identification de relations spécifiques au domaine (Sahlgren, 2006; Peirsman *et al.*, 2008).

5.2 Évaluation

La qualité des résultats obtenus lors de nos expériences est évaluée en comparant les relations sémantiques acquises aux 1 735 419 relations fournies par la partie française de l'UMLS[1]. À l'instar de (Curran, 2004) et (Ferret, 2013), nous considérons ici les relations obtenues comme des ensembles de voisins associés à des mots cibles, les voisins étant ordonnés suivant la similarité avec le mot cible.

L'évaluation des résultats est réalisée avec des mesures d'évaluation utilisées habituellement sur les résultats d'une analyse distributionnelle : la macro-précision (Sebastiani, 2002), la moyenne des précisions moyennes (MAP) (Buckley & Voorhees, 2005) et la R-précision.

La macro-précision est la moyenne des précisions $p(w_i)$ obtenues pour chaque mot cible (w_i) et un ensemble de voisins sémantiques I_i^j ($I_i^{j(+)}$ étant un voisin pertinent pour le mot cible considéré, et n_i le nombre de voisins considérés) :

$$p(w_i) = \frac{\sum_{j=1}^{n_i} I_i^{j(+)}}{\sum_{j=1}^{n_i} I_i^j}$$

La macro-précision pour l'ensemble des mots cibles est alors : $P = \sum_{i=1}^{|w_i|} p(w_i)$

Nous considérons quatre sous-ensembles voisins permettant d'obtenir la macro-précision après examen de 1 ($n_i = 1$, P@1), 5 ($n_i = 5$, P@5), 10 ($n_i = 10$, P@10) et 100 voisins ($n_i = 100$, P@100) :

$$P@N = \sum_{i=1}^{|w_i|} p(w_i | n_i = N)$$

La macro-précision permet d'avoir une qualité globale des résultats tout en considérant que tous les mots cibles ont le même poids quel que soit le nombre de voisins, alors que la micro-précision aurait tendance à privilégier les mots cibles

[1] http://www.nlm.nih.gov/research/umls/

comportant beaucoup de voisins, dont une bonne partie ne sont probablement pas pertinents, au détriment de mots cibles ayant peu de voisins. Pour P@1, la macro-précision est équivalente à la micro-précision.

Une alternative consiste à utiliser comme seuil n_i le nombre de voisins corrects attendus pour un mot cible. Pour cela, nous avons utilisé la R-précision (Buckley & Voorhees, 2005).

Pour le calcul de la R-précision, nous comparons nos résultats non plus à l'ensemble des relations de la partie française de l'UMLS, mais à des ensembles de référence constitués à partir de cette ressource. Il s'agit de réduire les relations de référence aux seules relations entre des termes ou des mots présents dans le corpus de travail et dans chaque expérience. Ainsi, nous disposons d'autant de références que d'expériences, avec par exemple entre 24 et 46 relations pour les expériences avec une fenêtre de 21 mots et la mesure de Cosinus.

La moyenne des précisions moyennes (MAP) est obtenue en considérant la précision non interpolée $UAP(I_i^j)$ des voisins sémantiques I_i^j au rang j, n_i est le nombre de voisins sémantiques I_i^j du mot cible w_i. La MAP est alors la moyenne de ces précisions non interpolées :

$$MAP = \frac{1}{|w_i|} \sum_{i=1}^{|w_i|} \frac{1}{n_i} \sum_{j=1}^{n_i} UAP(I_i^j)$$

La MAP est le reflet de la qualité du classement et permet d'évaluer la pertinence de la mesure de similarité utilisée. Ainsi, elle valorise le fait que la méthode ordonne tous les voisins sémantiques corrects proches de la tête de liste. Réciproquement, le fait d'ajouter des voisins sémantiques incorrects en fin de liste (après les voisins corrects) ne pénalise pas la méthode.

6 Résultats et discussion

Dans cette section, nous présentons et discutons les résultats obtenus tout d'abord avec une fenêtre de 5 mots puis avec celle de 21 mots.

	Sans seuil						Avec seuil					
	Rel. acq.	Rel. UMLS	MAP	R-préc	P@1	P@5	Rel. acq.	Rel. UMLS	MAP	R-préc	P@1	P@5
ADSeule	34132	98	0,172	0,057	0,098	0,043	1342	0,496	0,496	0,375	0,375	0,125
AD/VT	25578	84	0,158	0,055	0,068	0,050	1402	0,487	0,487	0,333	0,333	0,133
AD/IL	11658	56	0,164	0,066	0,079	0,047	694	0,443	0,443	0,333	0,333	0,100
AD/PLS	23760	84	0,181	0,100	0,114	0,050	1252	0,570	0,570	0,500	0,500	0,133
AD/VT+IL	12030	56	0,161	0,066	0,079	0,047	696	0,443	0,443	0,333	0,333	0,100
AD/VT+PLS	22176	82	0,185	0,102	0,116	0,051	1188	0,570	0,570	0,500	0,500	0,133
AD/IL+VT	11434	54	0,158	0,069	0,083	0,044	660	0,525	0,525	0,500	0,500	0,100
AD/IL+PLS	10456	52	0,177	0,100	0,114	0,046	610	0,525	0,525	0,500	0,500	0,100
AD/PLS+VT	22176	82	0,185	0,102	0,116	0,051	1188	0,570	0,570	0,500	0,500	0,133
AD/PLS+IL	11280	56	0,166	0,066	0,079	0,047	688	0,443	0,443	0,333	0,333	0,100
AD/VT+IL+PLS	10808	52	0,174	0,100	0,114	0,046	616	0,525	0,525	0,500	0,500	0,100
AD/VT+PLS+IL	11642	56	0,162	0,066	0,079	0,047	694	0,443	0,443	0,333	0,333	0,100
AD/IL+VT+PLS	5608	34	0,212	0,100	0,100	0,067	912	0,229	0,229	0,000	0,500	0,100
AD/IL+PLS+VT	10244	52	0,178	0,100	0,114	0,046	576	0,526	0,526	0,500	0,500	0,100
AD/PLS+VT+IL	6392	42	0,274	0,162	0,177	0,071	1036	0,584	0,585	0,500	0,250	0,133
AD/PLS+IL+VT	6020	40	0,233	0,109	0,125	0,063	970	0,381	0,381	0,250	0,222	0,100
AD/ALL3	11266	56	0,148	0,066	0,079	0,047	694	6	0,4431	0,333	0,333	0,100

Table 2: Résultats obtenus avec la mesure de Jaccard, évalués avec la MAP, R-précision, et précision à 1 et 5 pour une fenêtre de 5 mots – sans et avec seuil sur la mesure de similarité

6.1 Fenêtre graphique restreinte

Pour une fenêtre de 5 mots, nous présentons uniquement les résultats obtenus en utilisant la mesure de Jaccard, sans et avec seuils (tableau 2). La mesure du cosinus donne de très faibles résultats quels que soient les paramètres et ne semble

pas adaptée aux fenêtres de petite taille et lorsque les fréquences sont faibles : la précision (P@1) des résultats obtenus avec le cosinus varie entre 0,02 et 0,06, alors que pour la mesure de Jaccard, nous obtenons des précisions variant entre 0,01 et 0,17. Lorsqu'aucun seuil n'est appliqué sur les paramètres distributionnels, nous constatons que la généralisation à l'aide des relations acquises grâce aux patrons lexico-syntaxiques améliore la qualité des résultats aussi bien en termes de précision ou de R-précision que de MAP. De plus, la variation terminologique tend à dégrader les résultats quand elle est utilisée individuellement alors qu'elle a un impact positif lorsqu'elle est combinée aux relations issues des patrons lexico-syntaxiques. Mais l'apport des relations issues des patrons lexico-syntaxiques est annulé quand on ajoute celles issues de l'inclusion lexicale. Nous pouvons également faire ce constat lorsque l'on considère les relations indépendamment de la méthode utilisée pour les produire. La généralisation des contextes est alors peut-être trop importante pour pouvoir être utile dans l'analyse distributionnelle. Aussi, l'analyse des variations de la MAP permettent également de constater que la généralisation des contextes améliore le classement des relations présentes dans l'UMLS.

A la vue des mesures, la qualité des résultats semble bénéficier de la combinaison des trois sources de relations d'hyperonymie. De même, le nombre de relations retournées est réduit d'au minimum d'un quart avec la généralisation, par rapport à l'ensemble de relations obtenues avec l'AD seule, voire divisé par 6 pour certaines combinaison comme PLS+VT+IL. Cependant, le nombre de relations retrouvées dans l'UMLS est divisé par deux (98 avec l'AD seule, 42 avec la combinaison offrant la meilleure précision) lorsque nous utilisons les relations proposées par plusieurs méthodes.
Si des constats similaires peuvent être réalisés lorsque nous appliquons des seuils sur les paramètres distributionnels, nous remarquons également l'impact positif des relations acquises par inclusion lexicale sur les résultats. Mais ici, les résultats sont probablement peu significatifs. En effet, peu de relations sont retrouvées dans l'UMLS lorsque nous utilisons Jaccard pour mesure la similarité entre les mots, et nous pouvons douter de la significativité statistique des résultats obtenus avec les seuils. Une évaluation manuelle est nécessaire.

6.2 Fenêtre graphique large

Lorsqu'une fenêtre large de 21 mots est utilisée, les observations et les résultats sont différents selon les mesures de similarité utilisées et l'utilisation (cf. tableau 4) ou non (cf. tableau 3) de seuils sur les paramètres distributionnels. Ainsi avec la mesure de Jaccard, la qualité des résultats est améliorée si les contextes sont généralisés avec des relations issues des patrons lexico-syntaxiques et qu'aucun seuil n'est appliqué. La contribution des relations acquises par inclusion lexicale est variable : l'utilisation de ces relations pour généraliser les contextes dégradent les résultats si l'on n'applique pas de seuil sur les paramètres distributionnels, et au contraire permettent d'obtenir les meilleurs résultats lorsque des seuils sont utilisés. L'impact des patrons lexico-syntaxiques seul est beaucoup plus faible avec le cosinus. Mais lorsqu'ils sont pris en compte après l'utilisation des relations acquises par inclusion lexicale, la précision est améliorée et même supérieure à celle obtenue avec Jaccard. De plus, nous pouvons noter que les variantes terminologiques ont un impact nul ou négatif sur la qualité des résultats.

	Rel. acquises		Rel. dans UMLS		MAP		R-précision		P@1		P@5	
	JACC	COS	JACC	COS	JACC	COS	JACC	COS	JACC	COS	JACC	COS
ADSeule	9256	9256	46	46	0,221	0,149	0,142	0,098	0,118	0,088	0,059	0,028
AD/VT	8758	8758	44	44	0,201	0,158	0,120	0,104	0,094	0,094	0,056	0,053
AD/IL	6360	6360	42	42	0,197	0,120	0,075	0,081	0,097	0,065	0,071	0,056
AD/PLS	8418	8418	42	42	0,243	0,165	0,172	0,111	0,133	0,100	0,080	0,026
AD/VT+IL	6312	6312	42	42	0,196	0,120	0,075	0,081	0,097	0,065	0,077	0,060
AD/VT+PLS	7972	7972	42	42	0,244	0,166	0,172	0,111	0,133	0,100	0,080	0,026
AD/IL+VT	6138	6138	40	40	0,175	0,128	0,046	0,086	0,069	0,069	0,069	0,060
AD/IL+PLS	5874	5874	40	40	0,201	0,191	0,046	0,155	0,069	0,138	0,083	0,028
AD/PLS+VT	7972	7972	42	42	0,244	0,166	0,172	0,111	0,133	0,100	0,080	0,041
AD/PLS+IL	6346	6346	42	42	0,220	0,116	0,108	0,065	0,129	0,065	0,084	0,060
AD/VT+IL+PLS	5828	5828	40	40	0,198	0,191	0,046	0,155	0,069	0,138	0,076	0,026
AD/VT+PLS+IL	6310	6310	42	42	0,219	0,115	0,108	0,065	0,129	0,065	0,077	0,041
AD/IL+VT+PLS	5662	5662	40	40	0,202	0,193	0,046	0,155	0,069	0,138	0,083	0,026
AD/IL+PLS+VT	5662	5662	40	40	0,202	0,193	0,046	0,155	0,069	0,138	0,083	0,041
AD/PLS+VT+IL	6310	6310	42	42	0,219	0,115	0,108	0,065	0,129	0,065	0,077	0,041
AD/PLS+IL+VT	6122	6122	40	40	0,199	0,123	0,081	0,069	0,103	0,069	0,083	0,026
AD/ALL3	6306	6306	42	42	0,222	0,120	0,108	0,081	0,129	0,065	0,084	0,026

Table 3: Résultats évalués avec la MAP, R-précision, et précision à 1 et 5 pour une fenêtre de 21 mots – sans seuil sur la similarité sémantique

	Rel. acquises		Rel. dans UMLS		MAP		R-précision		P@1		P@5	
	JACC	COS	JACC	COS	JACC	COS	JACC	COS	JACC	COS	JACC	COS
ADSeule	392	6960	4	40	0,406	0,169	0,250	0,100	0,250	0,100	0,1000	0,0600
AD/VT	418	6576	6	38	0,181	0,180	0,000	0,107	0,000	0,107	0,040	0,064
AD/IL	370	4580	4	34	0,532	0,141	0,000	0,077	0,500	0,077	0,100	0,031
AD/PLS	390	6262	6	36	0,219	0,190	0,500	0,115	0,000	0,115	0,120	0,069
AD/VT+IL	380	4568	4	34	0,531	0,141	0,500	0,077	0,500	0,077	0,100	0,031
AD/VT+PLS	354	5910	6	36	0,220	0,190	0,000	0,115	0,000	0,115	0,120	0,069
AD/IL+VT	352	4414	2	32	0,533	0,152	0,500	0,083	0,500	0,083	0,100	0,033
AD/IL+PLS	334	4216	4	32	0,371	0,228	0,250	0,167	0,250	0,167	0,100	0,050
AD/PLS+VT	354	5910	6	36	0,220	0,190	0,000	0,115	0,000	0,115	0,120	0,069
AD/PLS+IL	404	4514	6	34	0,428	0,134	0,333	0,077	0,333	0,077	0,133	0,031
AD/VT+IL+PLS	338	4208	4	32	0,371	0,228	0,250	0,167	0,250	0,167	0,100	0,050
AD/VT+PLS+IL	404	4514	6	34	0,428	0,133	0,333	0,077	0,333	0,077	0,133	0,031
AD/IL+VT+PLS	316	4056	4	32	0,372	0,230	0,250	0,167	0,250	0,167	0,100	0,050
AD/IL+PLS+VT	314	4056	4	32	0,372	0,230	0,250	0,167	0,250	0,167	0,100	0,050
AD/PLS+VT+IL	404	4514	6	34	0,428	0,133	0,333	0,077	0,333	0,077	0,133	0,031
AD/PLS+IL+VT	378	4352	4	32	0,376	0,144	0,250	0,083	0,250	0,083	0,150	0,033
AD/ALL3	380	4544	6	34	0,430	0,140	0,333	0,077	0,333	0,077	0,133	0,031

Table 4: Résultats évalués avec la MAP, R-précision, et précision à 1 et 5 pour une fenêtre de 21 mots – avec seuils sur la similarité sémantique

En ce qui concerne la précision P@5, bien que les valeurs soient plus faibles que la précision P@1, nous observons que l'inclusion permet d'augmenter les valeurs de précision, en particulier lorsque celle-ci est utilisée après les patrons lexico-syntaxiques. Aussi, en terme de R-précision, les résultats sont plus contrastés qu'avec P@1 : pour les meilleures configurations de généralisation des contextes, les valeurs de R-précision sont identiques ou supérieures. Les relations attendues, sont classées parmi les premières ou à un rang plus élevé qu'avec l'AD seule. En termes de MAP, les valeurs sont assez stables. Les améliorations obtenues suivant les combinaisons montrent qu'il y a plus de relations retrouvées dans l'UMLS parmi les premiers voisins. Mais la variation restant faible entre l'AD seule et l'AD avec des contextes généralisés, l'ordonnancement des voisins n'est pas beaucoup modifié par notre méthode.

Lorsque l'on considère le nombre de relations retournées par l'analyse distributionnelle, nous constatons que quelle que soit la mesure utilisée, la combinaison des trois sources de relations, de manière séquentielle ou globalement, n'a qu'un impact moyen sur l'amélioration de la qualité des résultats. Toutefois, exploiter l'ensemble des relations d'hyperonymie à notre disposition permet de réduire d'un tiers le nombre de relations retournées par l'analyse distributionnelle, tout en conservant le même nombre de relations présentes dans l'UMLS.

Comme lors de l'utilisation d'une fenêtre de taille restreinte, l'application de seuils sur les paramètres distributionnels améliore la qualité des résultats. Mais le faible nombre de relations présentes dans l'UMLS, parmi l'ensemble des relations obtenues avec Jaccard, rend les observations difficiles à interpréter. En revanche, nous pouvons noter que les résultats obtenus avec le cosinus sont meilleurs en appliquant des seuils tout en ne réduisant pas trop le nombre de relations correctes.

7 Conclusion

Nous nous sommes intéressés dans cet article à la réduction de la dispersion des données dans les matrices de vecteurs de contexte utilisés pour mettre en œuvre l'analyse distributionnelle. Pour cela, nous avons proposé une méthode de généralisation des contextes distributionnels s'appuyant sur des relations d'hyperonymie acquises en corpus. Les mots décrivant le contexte distributionnel d'un mot cible sont considérés comme des hyponymes et sont substitués par des hyperonymes identifiés sur le corpus. Nous avons réalisé un certain nombre d'expériences sur un corpus médical en combinant plusieurs paramètres. Les relations d'hyperonymie ont été acquises avec des approches habituellement utilisées sur des textes de spécialité. Bien que l'évaluation des méthodes distributionnelles soit complexe à réaliser, nous avons confronté les résultats aux relations sémantiques proposées par l'UMLS français. Plusieurs mesures d'évaluation ont été utilisées pour évaluer l'impact de la généralisation des contextes sur l'analyse distributionnelle. L'analyse des résultats montre que lorsque la taille des fenêtres graphiques permettant de produire les contextes distributionnels est petite et que l'indice de Jaccard est utilisé comme mesure de similarité, il est préférable d'utiliser les relations proposées par les patrons lexico-syntaxiques pour généraliser les contextes. Il est alors possible d'obtenir un bon compromis permettant d'avoir à

la fois une bonne couverture et une amélioration de la précision. En revanche, lorsque la taille de la fenêtre est large, la généralisation des contextes grâce aux relations issues de l'inclusion lexicale améliore les résultats si le cosinus est utilisé comme mesure de similarité.

Outre une analyse manuelle des relations et de l'impact du processus de généralisation sur les données manipulées, ces résultats ouvrent plusieurs perspectives. Les relations d'hyperonymie que nous avons utilisées ont été exploitées séparément. Or, celles-ci pourraient être considérées comme une ébauche de taxonomie et nous envisageons d'adapter la méthode de généralisation des contextes afin qu'elle prenne en compte ce réseau de relations acquises en corpus. Aussi, l'ensemble des relations acquises en corpus pouvant être bruité, nous envisageons d'utiliser d'autres sources de relations comme celles proposées par des terminologies. Il sera alors possible d'évaluer l'impact de la généralisation et de relations lorsque leur statut terminologique est maîtrisé. Enfin, la réduction de dimensions peut également s'appuyer sur un processus de normalisation des contextes. Il est alors nécessaire de prendre en compte des relations de synonymie ou d'antonymie qui peuvent également être acquises en corpus ou issues de ressources terminologiques existantes.

AUBIN S. & HAMON T. (2006). Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, number 4139 in LNAI, p. 380–387: Springer.

BARONI M., BERNARDINI S., FERRARESI A. & ZANCHETTA E. (2009). The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, **43**(3), 209–226.

BASKAYA O., SERT E., CIRIK V. & YURET D. (2013). Ai-ku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Proceedings of SemEval - 2013*, p. 300–306, Atlanta, Georgia, USA: Association for Computational Linguistics.

BRODA B., PIASECKI M. & SZPAKOWICZ S. (2009). Rank-based transformation in measuring semantic relatedness. In Y. GAO & N. JAPKOWICZ, Eds., *Canadian Conference on AI*, volume 5549, p. 187–190: Springer.

BUCKLEY C. & VOORHEES E. (2005). Retrieval system evaluation. In E. VOORHEES & D. HARMAN, Eds., *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3. MIT Press.

CHATTERJEE N. & MOHAN S. (2008). Discovering word senses from text using random indexing. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'08, p. 299–310, Berlin, Heidelberg: Springer-Verlag.

CURRAN J. R. (2004). *From distributional to semantic similarity*. PhD thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

FERRET O. (2013). Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *TALN 2013*, p. 48–61, Les Sables d'Olonne, France.

FIRTH J. (1957). A synopsis of linguistic theory 1930-1955. *Studies in linguistic analysis*, p. 1–32.

GRABAR N. & ZWEIGENBAUM P. (2004). Lexically-based terminology structuring. *Terminology*, **10**(1), 23–54.

GREFENSTETTE G. (1994). Corpus-derived first, second and third-order word affinities. In *Sixth Euralex International Congress*, p. 279–290.

HAMON T., NAZARENKO A., POIBEAU T., AUBIN S. & DERIVIÈRE J. (2007). A robust linguistic platform for efficient and domain specific web content analysis. In *RIAO 2007*, Pittsburgh, USA.

HARRIS Z. (1954). Distributional structure. *Word*, **10**(23), 146–162.

JACQUEMIN C. (2001). *Spotting and discovering terms through natural language processing*. The MIT Press.

KANERVA P., KRISTOFERSSON J. & HOLST A. (2000). Random indexing of text samples for latent semantic analysis. In L. GLEITMAN & A. JOSH, Eds., *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, volume 1036, Erlbaum, New Jersey.

KARLGREN J. & SAHLGREN M. (2001). From words to understanding. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 294–308: Foundations of Real-World Intelligence.

LANDAUER T. & DUMAIS S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review; Psychological Review*, **104**(2), 211.

LEE D. D. & SEUNG H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**(6755), 788–791.

LUND K. & BURGESS C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers*, **28**, 203–208.

MORIN E. & JACQUEMIN C. (2004). Automatic Acquisition and Expansion of Hypernym Links. *Computers and the Humanities*, **38**(4), 363–396.

PADÓ S. & LAPATA M. (2007). Dependency-based construction of semantic space models. *Comput. Linguist.*, **33**(2), 161–199.

PEIRSMAN Y., KRIS H. & DIRK G. (2008). Size matters. tight and loose context definitions in english word space models. In *ESSLLI Workshop on Distributional Lexical Semantics*, Hamburg, Germany.

POLAJNAR T. & CLARK S. (2014). Improving distributional semantic vectors through context selection and normalisation. In *Proceedings of EACL 2014*. To appear.

RAPP R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *MT Summit'2003*, p. 315–322.

SAHLGREN M. (2006). *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. PhD thesis, Stockholm University, Stockholm, Sweden.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *New Methods in Language Processing*, p. 44–49, Manchester, UK.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, **34**(1), 1–47.

TSATSARONIS G. & PANAGIOTOPOULOU V. (2009). A generalized vector space model for text retrieval based on semantic relatedness. In *EACL 2009*, p. 70–78, Stroudsburg, PA, USA: Association for Computational Linguistics.

TURNEY P. D. & PANTEL P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, **37**, 141–188.

UTSUMI A. (2010). Evaluating the performance of nonnegative matrix factorization for constructing semantic spaces: Comparison to latent semantic analysis. In *Proceedings of SMC*, p. 2893–2900: IEEE.

VAN DER PLAS L. (2008). *Automatic lexico-semantic acquisition for question answering*. Thèse de doctorat, University of Groningen, Groningen.

VOZALIS E. & MARGARITIS K. G. (2003). Analysis of recommender systems' algorithms. In *The 6th Hellenic European Conference on Computer Mathematics & its Applications (HERCMA), Athens, Greece*.

WEEDS J. & WEIR D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Comput. Linguist.*, **31**(4), 439–475.

YURET D. (2012). Fastsubs: An efficient and exact procedure for finding the most likely lexical substitutes based on an n-gram language model. *IEEE Signal Process. Lett.*, **19**(11), 725–728.

ZHITOMIRSKY-GEFFET M. & DAGAN I. (2009). Bootstrapping distributional feature vector quality. *Comput. Linguist.*, **35**(3), 435–461.

ZWEIGENBAUM P. (1994). Menelas: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, **45**.