Explorer le graphe de voisinage pour améliorer les thésaurus distributionnels

Vincent Claveau[1] Ewa Kijak[1] Olivier Ferret[2]
(1) IRISA - CNRS - Univ Rennes 1, Campus de Beaulieu, F-35042 Rennes
(2) CEA, LIST, Laboratoire Vision et Ingénierie des Contenus, F-91191 Gif-sur-Yvette
vincent.claveau@irisa.fr, ewa.kijak@irisa.fr, olivier.ferret@cea.fr

Résumé. Dans cet article, nous abordons le problème de construction et d'amélioration de thésaurus distributionnels. Nous montrons d'une part que les outils de recherche d'information peuvent être directement utilisés pour la construction de ces thésaurus, en offrant des performances comparables à l'état de l'art. Nous nous intéressons d'autre part plus spécifiquement à l'amélioration des thésaurus obtenus, vus comme des graphes de plus proches voisins. En tirant parti de certaines des informations de voisinage contenues dans ces graphes nous proposons plusieurs contributions.
1) Nous montrons comment améliorer globalement les listes de voisins en prenant en compte la réciprocité de la relation de voisinage, c'est-à-dire le fait qu'un mot soit un voisin proche d'un autre et vice-versa.
2) Nous proposons également une méthode permettant d'associer à chaque liste de voisins (i.e. à chaque entrées du thésaurus construit) un score de confiance.
3) Enfin, nous montrons comment utiliser ce score de confiance pour réordonner les listes de voisins les plus proches.
Ces différentes contributions sont validées expérimentalement et offrent des améliorations significatives sur l'état de l'art.

Abstract. In this paper, we address the issue of building and improving a distributional thesaurus. We first show that existing tools from the information retrieval domain can be directly used in order to build a thesaurus with state-of-the-art performance. Secondly, we focus more specifically on improving the obtained thesaurus, seen as a graph of k-nearest neighbors. By exploiting information about the neighborhood contained in this graph, we propose several contributions.
1) We show how the lists of neighbors can be globally improved by examining the reciprocity of the neighboring relation, that is, the fact that a word can be close of another and vice-versa.
2) We also propose a method to associate a confidence score to any lists of nearest neighbors (i.e. any entry of the thesaurus).
3) Last, we demonstrate how these confidence scores can be used to reorder the closest neighbors of a word.
These different contributions are validated through experiments and offer significant improvement over the state-of-the-art.

Mots-clés : thésaurus distributionnel, graphe de k proches voisins, fenêtre de Parzen, algorithme hongrois, T-normes, recherche d'information.

Keywords: distributional thesaurus, k nearest neighbor graph, Parzen window, Hungarian algorithm, T-norms, information retrieval.

1 Introduction

Les thésaurus distributionnels sont utiles à de nombreuses tâches du TAL et leur construction est un problème largement abordé depuis plusieurs années (Grefenstette, 1994). Cela reste néanmoins un champ de recherche très actif, entretenu par la mise à disposition de corpus toujours plus volumineux et de nombreuses applications. Ces thésaurus associent à chacune de leurs entrées une liste de mots qui se veulent proches sémantiquement de l'entrée. Cette notion de proximité est variable selon les travaux (synonymie, autres relations paradigmatiques, relations syntagmatiques (Budanitsky & Hirst, 2006; Adam *et al.*, 2013, pour une discussion)), mais les méthodes utilisées pour la construction automatique de ces thésaurus sont souvent partagées. Pour une grande part, ces méthodes reposent sur l'hypothèse distributionnelle de Firth (1957) : chaque mot est caractérisé par l'ensemble des contextes dans lesquels il apparaît, et la proximité sémantique de deux mots peut être déduite de la proximité de leurs contextes. Cette hypothèse a donc été mise en œuvre de différentes façons, et plusieurs pistes pour en améliorer les résultats ont été suivies (voir section suivante pour un état de l'art).

Les travaux présentés dans cet article s'inscrivent dans ce cadre, et nous proposons plusieurs contributions portant sur la création de ces thésaurus distributionnels et sur leur amélioration. Nous montrons tout d'abord que les modèles de recherche d'information (RI) sont adaptés à la tâche de création de ces thésaurus, offrant des résultats très compétitifs par rapport à l'état de l'art, tout en bénéficiant d'un outillage déjà existant (section 3).

Le cœur de notre travail se situe ensuite sur l'exploitation des relations de voisinage sémantique ainsi mesurées. Les modèles de RI fournissent en effet les listes ordonnées par similarité décroissante de tous les mots avec tous les mots, formant un graphe de plus-proches voisins. Nous proposons de tirer parti de certaines des informations de voisinage contenues dans ce graphe que nous déclinons en trois contributions.
1) Nous montrons comment améliorer globalement les listes de voisins en prenant en compte la réciprocité de la relation de voisinage, c'est-à-dire le fait qu'un mot soit un voisin proche d'un autre et vice-versa (section 4).
2) Nous proposons également une méthode permettant d'associer à chaque liste de voisins (i.e. à chaque entrée du thésaurus construit) un score de confiance (section 5). Cette méthode repose sur l'estimation, à l'aide du graphe de plus proches voisins, des probabilités de trouver un mot donné comme $i$ème voisin d'un autre.
3) Enfin, sur la base de ce travail, nous montrons comment utiliser ce score de confiance et ces probabilités pour réordonner les listes de voisins les plus proches (section 6). Pour ce faire, nous modélisons ce réordonnancement comme un problème d'optimisation de profit, résolu par l'algorithme hongrois (Kuhn & Yaw, 1955).

2 État de l'art

La notion de thésaurus distributionnel est à la fois bien connue et en même temps relativement peu abordée de façon spécifique, sans doute à cause de ses liens étroits avec la notion de similarité sémantique. Beaucoup de travaux portent sur des améliorations concernant les mesures de similarité sémantique de nature distributionnelle, c'est-à-dire directement à la construction du thésaurus. Nous les examinons dans la sous-section suivante. Mais quelques travaux présentés dans la sous-section 2.2 ont aussi cherché, une fois le thésaurus obtenu, à l'améliorer, comme nous nous proposons de le faire en l'examinant comme un graphe de plus proches voisins.

2.1 Construction des thésaurus

Si l'on considère comme point de référence le paradigme défini par Grefenstette (1994), repris à sa suite notamment par Lin (1998) et Curran & Moens (2002), une première voie d'amélioration a porté sur la pondération des éléments constitutifs des contextes distributionnels, simples mots dans le cas de cooccurrents graphiques et paires (mot, relation de dépendance syntaxique) dans le cas de cooccurrents syntaxiques. Dans cette optique, Broda *et al.* (2009) ont ainsi proposé de remplacer les poids associés aux cooccurrents par une fonction tenant compte de leur rang dans ces contextes, ce qui a l'avantage de rendre ce poids moins dépendant d'une fonction de pondération spécifique. Zhitomirsky-Geffet & Dagan (2009) opère cette modification de pondération par le biais d'un mécanisme d'amorçage en faisant l'hypothèse que les premiers voisins d'une entrée sont plus pertinents que les autres et que de ce fait, les cooccurrents qui leur sont le plus fortement associés dans leurs contextes distributionnels sont aussi plus représentatifs de l'entrée du point de vue sémantique. Le poids de ces cooccurrents est alors renforcé pour accroître leur influence lors du réordonnancement des voisins. Yamamoto & Asakura (2010) en est une variante prenant en compte un plus large ensemble de cooccurrents dans les contextes.

Au-delà des changements de pondération, certains travaux se sont attachés au contenu même des contextes distributionnels. De ce point de vue, une première distinction, opérée déjà par Grefenstette (1994) mais explorée plus en détail par Curran & Moens (2002), a été réalisée entre cooccurrents graphiques et syntaxiques, avec un avantage donné à ces derniers. Parallèlement à la nature de l'information contenue dans les contextes, la question de sa forme s'est posée en faisant l'hypothèse que l'information portée par les cooccurrents peut être représentée de façon plus dense par des dimensions sous-jacentes. Cette idée est d'ailleurs renforcée par le constat de Hagiwara *et al.* (2006), par le biais de la sélection de caractéristiques dans un cadre supervisé, que bon nombre de cooccurrents peuvent être filtrés sans altérer significativement l'identification des similarités sémantiques entre mots. Une partie des travaux visant à améliorer l'approche distributionnelle s'est donc focalisée sur l'application de méthodes de réduction de dimensions, depuis l'Analyse Sémantique Latente (Landauer & Dumais, 1997), étendue par Padó & Lapata (2007) aux cooccurrents syntaxiques, jusqu'à la factorisation de matrice non négative (Van de Cruys, 2010) en passant par le Random Indexing (Sahlgren, 2001). Ces méthodes ont cependant donné des résultats limités (Van de Cruys, 2010). Dans ce cadre, l'apprentissage de représentations distribuées

réalisées au moyen de réseaux de neurones (Huang *et al.*, 2012; Mikolov *et al.*, 2013) est également à mentionner, même si ces travaux sortent un peu du cadre distributionnel traditionnel.

2.2 Amélioration des thésaurus

Les travaux que nous considérons ici se concentrent sur des améliorations exploitant plus spécifiquement la structure du thésaurus pour en améliorer la qualité comme nous nous proposons de le faire. Zhitomirsky-Geffet & Dagan (2009) pourrait dans une certaine mesure être rattaché à cette catégorie dans la mesure où sa repondération des éléments de contexte d'un terme dépend de ses voisins sémantiques, donc de la structure du thésaurus.

Deux autres voies ont également été explorées. La première consiste à utiliser un thésaurus initial afin de sélectionner de façon non supervisée un ensemble d'exemples positifs et négatifs de termes sémantiquement similaires ou liés (Ferret, 2012, 2013b). Cet ensemble est utilisé pour entraîner un classifieur permettant ensuite de réordonner les voisins initiaux. Dans le cas de (Ferret, 2012), cette sélection est fondée sur un critère de symétrie de la relation de similarité sémantique : si A est trouvé comme voisin proche de B et B comme voisin proche de A, A et B sont probablement des exemples positifs de mots sémantiquement similaires. On retrouve là la condition de réciprocité que nous explorons dans un autre cadre en section 4. (Ferret, 2013b) s'appuie pour sa part sur l'hypothèse que des constituants similaires, au sens de leur voisinage dans un thésaurus, occupant le même rôle syntaxique dans des mots composés eux-mêmes similaires sont de probables exemples positifs de mots similaires.

La seconde approche, proposée dans (Ferret, 2013a), est plus indirecte. Elle réalise un réordonnancement des voisins sémantiques par le biais d'un processus de détection et de déclassement des voisins les moins susceptibles d'être sémantiquement liés à leur entrée. Cette détection est réalisée en appliquant un modèle discriminant de l'entrée en contexte à un échantillon des occurrences de ses voisins et en jugeant de la proximité entre entrée et voisin sur la base des décisions de ce modèle.

Comme dans ces derniers travaux, nous nous proposons d'améliorer la qualité des thésaurus produits, notamment en réordonnant les listes de voisins. Notre travail repose en partie sur des considérations proches, notamment en ce qui concerne la réciprocité, mais dans une optique différente dans laquelle les listes de plus proches voisins sont directement réordonnées en fonction des voisinages observés dans le thésaurus. En cela, nos travaux peuvent se rapprocher de ceux de (Pedronette *et al.*, 2014), faits dans un tout autre contexte applicatif (recherche d'images), mais reposant également sur l'examen des voisinages observés dans un graphe de k plus proches voisins.

3 Modèle de RI pour la construction de thésaurus distributionnels

3.1 Principes

Comme cela apparaît dans les travaux cités de l'état-de-l'art, le cœur des approches distributionnelles est de calculer des similarités entre représentations textuelles des contextes des mots étudiés. Les méthodes de calcul de similarité utilisées en recherche d'information semblent donc pertinentes pour ce problème. Pour un mot donné, l'ensemble des contextes de ses occurrences est considéré comme un document ; pour un mot w_i, on note ce document $\mathcal{C}_{w_i}$. La proximité entre deux mots est alors mesurée par une fonction de similarité RI sur leur contexte. Cette piste a beaucoup de liens avec les travaux de l'état de l'art mais semble relativement peu explorée en tant que telle, à l'exception de Vechtomova & Robertson (2012) dans le cas particulier de la recherche d'entités nommées similaires. Elle offre pourtant l'avantage d'être très facilement implémentable du fait des nombreux outils de RI disponibles.

Bien entendu, quelques adaptations doivent être faites. Dans les expériences rapportées ci-dessous, le contexte considéré est de deux mots avant et après chaque occurrence. Contrairement à la RI, on souhaite garder les mots outils, mais aussi les positions de ces mots par rapport à l'occurrence du mot examiné. Par exemple, pour l'occurrence de freedom dans l'extrait :
« ... all forms of restrictions on freedom of expression , threats ... »,
les termes d'indexation restrictions-2, on-1, of+1, expression+2 sont ajoutés à la description de freedom (i.e. sont ajoutés à $\mathcal{C}$(freedom)). Pour trouver les voisins distributionnels d'un mot, l'ensemble des contextes collectés pour ce mot sert de requête, qui est alors utilisée pour trouver les mots les plus proches (i.e. dont les contextes sont les plus proches) au sens d'une mesure de similarité RI.

Dans les expériences que nous présentons ci-dessous, nous avons testé quelques unes des mesures les plus classiquement utilisées en RI : la similarité d'Hellinger (Escoffier, 1978; Domengès & Volle, 1979), un TF-IDF/cosinus, et une similarité Okapi-BM-25 (Robertson *et al.*, 1998). Ce dernier modèle peut être vu comme une version plus moderne du TF-IDF, prenant notamment mieux en compte les différences de tailles des documents. Ce point est important puisque dans notre cas, les documents, c'est-à-dire l'ensemble des contextes d'un mot, sont effectivement de tailles très variables du fait du nombre d'occurrences lui-même très variable des différents mots. La similarité Okapi-BM25 entre un mot w_i ($\mathcal{C}(w_i)$ est vu comme une requête), et w_j ($\mathcal{C}(w_j)$ vu comme un document), s'exprime par l'équation 1 dans laquelle les composants correspondent respectivement au poids du mot considéré dans la requête, à son TF et à son IDF dans le document. qtf est le nombre d'occurrence du mot t dans le contexte de la requête $\mathcal{C}(w_i)$, et similairement tf est le nombre d'occurrences dans $\mathcal{C}(w_j)$, dl est la taille des contextes de w_j (nombre de mots dans $\mathcal{C}(w_j)$), dl_{avg} la taille moyenne des contextes, n est le nombre de documents, c'est-à-dire dans notre cas le nombre de mots examinés (nombre d'entrées du thésaurus), $df(t)$ est le nombre de contextes ($\mathcal{C}(\cdot)$) dans lesquels t apparaît, et enfin k_1, k_3 et b sont des constantes, fixées par défaut à $k_1 = 2$, $k_3 = 1000$ et $b = 0.75$.

$$\text{similarité}(w_i, w_j) = \sum_{t \in \mathcal{C}(w_i)} \frac{(k_3+1) * qtf}{k_3 + qtf} * \frac{tf * (k_1+1)}{tf + k_1 * (1 - b + b * \frac{dl(\mathcal{C}(w_j))}{dl_{avg}})} * \log \frac{n - df(t) + 0.5}{df(t) + 0.5} \quad (1)$$

Nous proposons également dans les expériences rapportées ci-dessous une version dite ajustée de la similarité Okapi-BM25, dans laquelle l'influence de la taille du document est renforcée, en prenant $b = 1$, et en mettant l'IDF au carré pour donner plus d'importance aux mots de contexte plus discriminants.

Ces modèles de RI, très classiques, ne sont pas détaillés plus avant ici ; le lecteur intéressé trouvera les notions et détails utiles dans les références citées ou des ouvrages généralistes (Boughanem & Savoy, 2008, par exemple).

3.2 Contexte expérimental

Les données et références utilisées pour nos expériences tout au long de cet article sont celles employées par Ferret (2013a) et mises à notre disposition par l'auteur. Cela nous permet d'avoir un cadre expérimental complètement comparable aux résultats publiés. Pour construire les thésaurus distributionnels, le corpus utilisé est AQUAINT-2, une collection d'articles de presse en anglais d'environ 380 millions de mots. Tous les noms de fréquence > 10 de ce corpus sont considérés, soit 25 000 noms (on note n ce nombre dans la suite) ; ils formeront les entrées des thésaurus. Le corpus est étiqueté en parties-du-discours par TreeTagger, ce qui permet de repérer les noms qui formeront les entrées du thésaurus pour nous comparer aux travaux existants. Ces informations ne sont pas utilisées pour la suite de la construction du thésaurus, ce qui assure une certaine portabilité de la méthode à d'autres langues (Freitag *et al.*, 2005).

Pour évaluer les thésaurus produits, deux références sont utilisées, soit séparément, soit conjointement. Il s'agit d'une part des synonymes de WordNet 3.0 (Miller, 1990), et d'autre part du thésaurus Moby (Ward, 1996). Ces deux ressources ont des caractéristiques assez différentes et complémentaires ; notamment, WordNet indique des liens paradigmatiques assez forts entre les mots (synonymie ou quasi-synonymie) et nous fournit ainsi des listes de 3 voisins en moyenne pour 10 473 noms du corpus, tandis que Moby regroupe des mots partageant des relations syntagmatiques ou paradigmatiques plus larges, et fournit des listes de 50 voisins en moyenne pour 9 216 noms. Les deux ressources combinées donnent une référence de 38 voisins en moyenne pour 12 243 noms. C'est cette combinaison de WordNet et Moby qui sera utilisée comme référence dans toutes les évaluations de cet article.

Cette évaluation intrinsèque porte donc sur une petite moitié des entrées du thésaurus considéré, ce que l'on peut considérer, compte tenu de sa taille, comme un ensemble d'évaluation conséquent par rapport à des jeux de test classiques (e.g. WordSim 353). Ce type d'évaluation intrinsèque est bien entendu limité par la nature des relations présentes dans les ressources de référence. Si cette limite est assez restrictive dans le cas des synonymes de WordNet, elle est beaucoup plus large et pose moins problème dans le cas de Moby, dont les relations sont très diverses.

3.3 Résultats

Pour un nom donné, notre approche par modèles de RI, comme les autres approches de l'état de l'art, ordonne les autres noms par similarité décroissante. La liste obtenue est comparée à la liste de référence, et des mesures classiques d'évaluation sont calculées. Il s'agit de la précision à divers seuils (après 1, 5, 10, 50, 100 voisins, notés P@1, P@5...), la Mean

Référence	Méthode	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
WordNet + Moby	Ferret 2013 *base*	5.6	7.7	22.5	14.1	10.8	5.3	3.8
	Ferret 2013 *best rerank*	6.1	8.4	24.8	15.4	11.7	5.7	3.8
	Hellinger	2.45	2.89	9.73	6.28	5.31	4.12	3.30
	TF-IDF	5.40	7.28	21.73	13.74	9.59	5.17	3.49
	Okapi-BM25	6.72	8.41	24.82	14.65	10.85	5.16	3.66
	Okapi-BM25 ajusté	8.97	10.94	31.05	18.44	13.76	6.46	4.54
WordNet	Ferret 2013 *base*	9.8	8.2	11.7	5.1	3.4	1.1	0.7
	Ferret 2013 *best rerank*	10.7	9.1	12.8	5.6	3.7	1.2	0.7
	Okapi-BM25 ajusté	14.17	12.22	16.97	7.10	4.47	1.41	0.84
Moby	Ferret 2013 *base*	3.2	6.7	24.1	16.4	13.0	6.6	4.8
	Ferret 2013 *best rerank*	3.5	7.2	26.5	17.9	14.0	6.9	4.8
	Okapi-BM25 ajusté	5.69	9.14	32.18	21.37	16.42	8.02	5.69

TABLE 1: Performances des modèles de RI pour la construction des thésaurus distributionnels sur la référence WordNet+Moby

Average Precision (MAP, moyenne des précisions calculées après chaque mot de la référence trouvé), la R-précision (R-prec, précision après R voisins ou R est le nombre de voisins dans la liste de référence pour le nom examiné). Elles sont toutes exprimées en pourcentage dans la suite de l'article.

Le tableau 1 recense les performances des différents modèles de similarités RI. À des fins de comparaison, nous indiquons les résultats obtenus dans les mêmes conditions par Ferret (2013a), avec d'une part une approche standard de l'état de l'art reposant sur des calculs d'information mutuelle sur les contextes, et d'autre part, une version améliorée par apprentissage (cf. section 2). Nous détaillons également certains de ces résultats selon les références WordNet et Moby prises séparément.

Plusieurs éléments méritent d'être notés dans ces premiers résultats. Notons tout d'abord que comme pour les autres travaux de la littérature, nos résultats sont globalement faibles, ce qui atteste de la difficulté de la tâche. Outre ces mesures de précision, le rappel à 100 est par exemple de 21.2 % pour la version ajusté d'Okapi sur la référence WordNet+Moby. On constate par ailleurs que certaines similarités RI sont assez peu performantes, notamment le TF seul ou la similarité d'Hellinger. Cela est peu surprenant puisque ces similarités utilisent des pondérations basiques qui ne permettent pas de mettre en valeur les contextes discriminants des mots. Les similarités incluant une notion d'IDF obtiennent en cela de meilleurs résultats. Les similarités de type Okapi-BM25 offrent de bons résultats ; la version standard d'Okapi obtient des performances similaires à l'état de l'art, et la version ajustée dépasse même largement les deux systèmes de (Ferret, 2013a), notamment en terme de qualité globale (mesurée par la MAP), quelle que soit la référence utilisée. C'est cette dernière version du système qui nous sert de référence pour la suite de cet article.

4 Réciprocité dans le graphe des k-NN

Le calcul de toutes les similarités entre toutes les paires de mots produit un graphe valué de voisinage : chaque mot est lié, avec une certaine force, aux n autres mots. Les résultats obtenus ci-dessus ne tiennent pas compte de cette structure. L'objet des sections suivantes est d'examiner comment tirer parti au mieux des relations de voisinage enfouies dans ce graphe.

Il faut préalablement noter que certaines des mesures de similarités RI que nous avons utilisées, notamment Okapi-BM25, ne sont pas symétriques. La similarité entre un mot w_i, utilisé comme requête, et un autre mot w_j ne donne pas la même valeur que la similarité entre la requête w_j et w_i. Indépendamment de cela, même pour une mesure symétrique, la relation de plus-proche voisin n'est pas non plus symétrique : un mot w_j peut-être dans les k-plus-proches voisins de w_i mais l'inverse peut être faux.

Il semble alors raisonnable de penser que la réciprocité de voisinage entre deux mots (chacun est dans les k plus proches voisins de l'autre) est tout de même un gage de confiance sur la proximité entre ces mots. L'utilisation de cette information pour améliorer les résultats précédents est examinée dans cette section. Dans la suite, on note $\tau_{w_i}(w_j)$ le rang de w_j dans la liste des voisins de w_i, qui varie donc entre 1 et n.

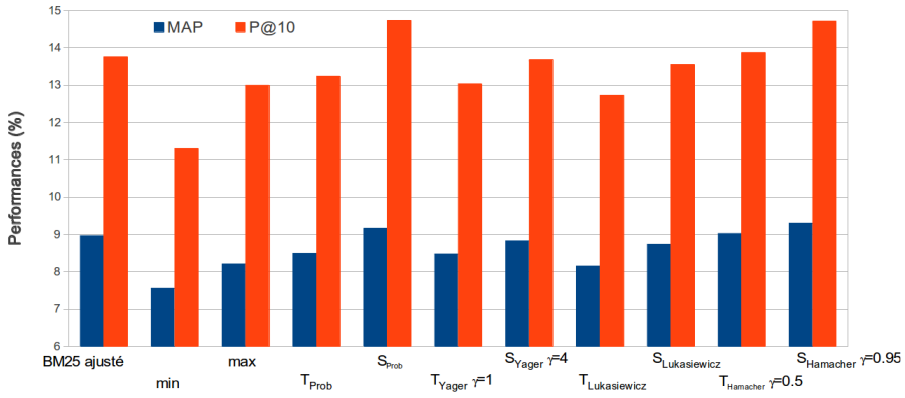


FIGURE 1: Performances de l'agrégation des rangs réciproques sur la référence WordNet+Moby

4.1 Graphe de voisinage distributionnel

La réciprocité de la relation de voisinage distributionnel a déjà été examinée et utilisée dans certains travaux (Ferret, 2013b) en sémantique distributionnelle, ou plus généralement sur des graphes de plus proches voisins (Pedronette *et al.*, 2014). Dans ces derniers travaux, la prise en compte de la réciprocité pour mener à un nouveau score de similarité a été faite simplement. Pour un mot w_i et son voisin w_j, le maximum ou le minimum des rangs ($\tau_{w_i}(w_j)$ et $\tau_{w_j}(w_i)$) est pris comme nouveau rang. Ces deux opérateurs apparaissent comme trop brutaux puisque seul l'un des rangs est pris en considération pour décider du score final, ce qui se transcrit par des performances très dégradées comme nous allons le voir.

Cependant, beaucoup d'autres opérateurs d'agrégation, avec des comportements peut-être plus appropriés à la tâche ont été proposés dans d'autres cadres, notamment en logique floue (Detyniecki, 2000, pour une revue très complète). Ces opérateurs ont une certaine sémantique permettant d'appréhender leur comportement, comme par exemple les T-normes (ET en logique floue) et S-normes (ou T-conormes, OU flou). Dans la suite de cette section, nous testons quelques uns de ces opérateurs sans prétention d'exhaustivité. Ceux-ci étant définis sur le domaine $[0, 1]^2$ et 1 représentant la certitude, ils sont utilisés pour générer un nouveau score de similarité sous la forme :

$$\text{score}_{w_i}(w_j) = \text{Agreg}(1 - \tau_{w_i}(w_j)/n, 1 - \tau_{w_j}(w_i)/n)$$

où Agreg est un opérateur d'agrégation (cf. infra pour le test de quelques fonctions possibles). Les scores obtenus sont alors utilisés pour produire une nouvelle liste de plus proches voisins de w_i (plus le score est élevé, plus la proximité sera avérée). On a bien ainsi la sémantique associée à ces opérateurs ; par exemple, si la fonction d'agrégation est $\max$, on a bien le comportement de OU flou attendu associé à cette S-norme : w_j sera classé très proche de w_i dans la nouvelle liste si w_j était proche de w_i *ou* si w_i était proche de w_j. Pour la T-norme $\min$, il faut que w_j soit proche de w_i *et* que w_i soit proche de w_j.

4.2 Résultats

Pour la fonction d'agrégation Agreg, outre le $\min$ et le $\max$, nous rapportons dans la figure 1 les résultats obtenus avec les T-normes (ou familles de T-normes dépendant d'un paramètre γ) suivantes :

$\text{T}_{\text{Prob}}(x, y) = x * y$

$\text{T}_{\text{Lukasiewicz}}(x, y) = \max(x + y - 1, 0)$

$\text{T}_{\text{Hamacher}}(x, y) = \frac{x*y}{\gamma+(1-\gamma)*(x+y-x*y)}$ with $\gamma \geq 0$

$\text{T}_{\text{Yager}}(x, y) = \max(0, 1 - \sqrt[\gamma]{(1-x)^\gamma + (1-y)^\gamma})$ with $\gamma > 0$

Nous testons aussi les S-normes associées, obtenues par généralisation de la loi de De Morgan : $S(x, y) = 1-T(1-x, 1-y)$. Pour les familles de T-normes dépendant d'un paramètre, nous avons fait varier ce dernier de manière systématique et les résultats rapportés sont ceux maximisant la MAP.

On remarque que ces opérateurs obtiennent des résultats très divers. Ceux qui induisent un seuil (i.e. pour certaines valeurs de $\tau_{w_i}(w_j)$ et $\tau_{w_j}(w_i)$, ces opérateurs renvoient une valeur par défaut générant trop d'ex æquo parmi les voisins, comme le min, max, les normes Lukasiewicz, et d'autres pour certains γ) dégradent la qualité des listes de plus proches voisins.

Référence	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
WordNet + Moby	9.30 (+3.75)	11.06 (+2.03)	30.42 (-2.53)	19.29 (+4.58)	14.71 (+6.92)	7.09 (+9.78)	4.86 (+7.07)
WordNet	15.05 (+6.23)	12.81 (+4.81)	17.55 (+3.41)	7.96 (+12.16)	5.07 (+13.30)	1.63 (+15.69)	0.94 (+12.23)
Moby	5.90 (+3.65)	11.86 (+4.14)	*31.77 (-1.27)*	21.65 (+1.34)	17.0 (+3.53)	8.42 (+5.01)	5.92 (+4.12)

TABLE 2: Performances et gains (%) par agrégation des rangs réciproques sur les références WordNet et Moby séparément avec une agrégation des rangs par S_{Hamacher} $\gamma = 0.95$

Les T-normes, privilégiant les paires de mots proches l'une de l'autre dans les deux sens, sont trop contraignantes. Cela rejoint les conclusions des travaux cités : la condition de réciprocité, appliquée trop strictement, ne permet pas d'améliorer les listes de plus proches voisins sur l'ensemble des mots. En revanche, les S-normes semblent mieux à même de tirer parti du classement. Les améliorations sont dans ce cas modestes en terme de qualité globale (MAP), mais importantes à certains rangs (P@10, P@50).

Enfin, il est important de noter que ces résultats dépendent beaucoup de la ressource utilisée comme référence, comme nous l'illustrons dans le tableau 2 en testant l'agrégation avec S_{Hamacher} $\gamma = 0.95$ sur Moby et WordNet séparément. Pour s'assurer que les différences soient statistiquement significatives, nous effectuons un test de Wilcoxon ($p < 0.05$) (Hull, 1993) ; les résultats non significatifs sont indiqués en italiques. Sur WordNet, basée sur une relation de synonymie assez forte et donc réciproque, les gains obtenus par notre approche sont bien plus importants que sur la référence issue de Moby.

5 Estimation de la confiance d'une liste de voisins distributionnels

Dans la section précédente, le rang de w_i dans la liste des voisins de w_j est utilisé pour améliorer le classement de w_j dans la liste des voisins de w_i. On peut aussi s'intéresser d'une manière plus générale aux positions relatives de w_i et w_j dans toutes les listes de voisins de tous les mots pour en tirer une information peut-être plus complète. Dans un premier temps, nous proposons d'en tirer un critère de confiance associé à chaque liste de plus-proches voisins en se basant uniquement sur des éléments du graphe de voisinage.

5.1 Principe

On fait l'hypothèse suivante : la liste de plus proches voisins d'un mot w est probablement de bonne qualité si la proximité (en terme de rang) entre w et chacun de ses voisins w_i est cohérente avec la proximité observée entre ces mêmes mots (w, w_i) dans les listes de voisins d'autres mots. L'intuition est que des mots supposés proches doivent aussi se retrouver proches des mêmes mots. Par exemple, si w_i est un voisin très proche de w, et que w est un voisin très proche de w_j, on s'attend à ce que w_i soit aussi très proche de w_j. Si les k plus proches voisins de w ont cette qualité, alors on accorde une certaine confiance à cette liste de voisins.

Formellement, nous définissons en terme probabiliste l'indice de confiance de la liste des k plus proches voisins de w par :

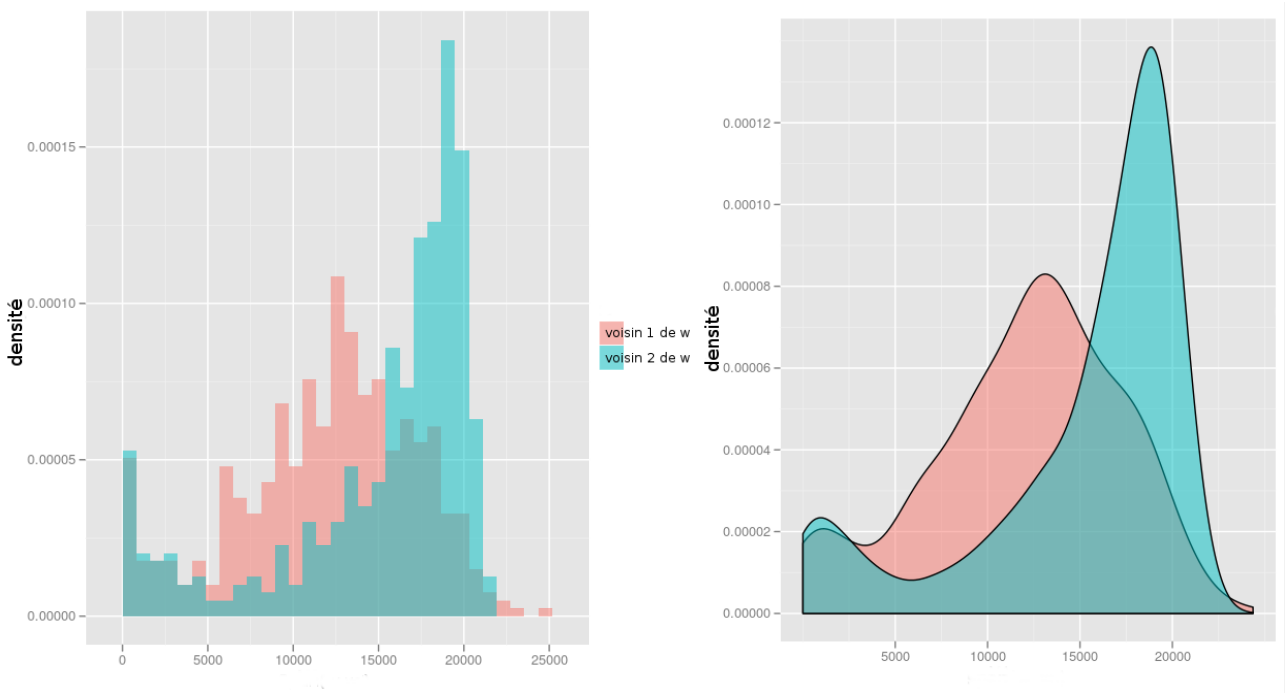
$$Q(w) = \prod_{\{w_i | \tau_w(w_i) \leq k\}} p(\delta(w, w_i) = \tau_w(w_i))$$

avec $p(\delta(w, w_i) = \tau_w(w_i))$ la probabilité que w_i soit le $\tau_w(w_i)$ème voisin de w (i.e. l'écart en terme de nombre de voisins, noté $\delta(w, w_i)$, est de $\tau_w(w_i)$).

Le problème est alors d'estimer pour chaque couple de mots (w, w_i) la distribution de probabilité $p(\delta(w, w_i))$. On utilise pour cela une méthode d'estimation de densité non-paramétrique par fenêtre de Parzen. Nous décrivons comment cette méthode classique (Parzen, 1962; Wasserman, 2005) est appliquée dans notre cas ci-après.

5.2 Estimation par fenêtres de Parzen

Soit x_{ab} la distance (différence entre les rangs) entre deux mots w_a et w_b dans une liste des voisins d'un mot quelconque. On dispose d'un échantillon de n réalisations de x_{ab}, supposées *iid* : (x^1_{ab}, x^2_{ab}, ... , x^n_{ab}), qui sont les distances observées



(a) (b)

FIGURE 2: (a) Deux histogrammes (en bleu et en rouge) des écarts d'un mot et deux de ses voisins. (b) Deux densités de probabilités (en bleu et en rouge) estimées par fenêtres de Parzen, correspondant aux données de la figure 2a.

entre w_a et w_b dans chacune des listes (complètes) des voisins de chaque mot. On les suppose *iid* ; cette hypothèse n'est certainement pas remplie puisque ces distances sont calculées en partie sur des contextes similaires, mais cette simplification permet une formalisation simple et efficace.

En effet, on peut alors estimer la densité de probabilité de x_{ab} avec la technique des fenêtres de Parzen grâce à un estimateur à noyau (équation 2) avec h un paramètre de lissage à fixer, et K un noyau permettant d'estimer la densité localement.

$$\widehat{p_h}(x_{ab}) = \frac{1}{n\,h}\sum_{i=1}^{n} K\left(\frac{x_{ab}-x_{ab}^i}{h}\right) \quad (2) \qquad \text{avec} \quad K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \quad (3) \qquad \text{et} \quad \hat{h} = 0.9\ \min(\hat{\sigma}, \frac{q_3-q_1}{1.34})\ n^{-\frac{1}{5}} \quad (4)$$

Dans notre cas, nous choisissons classiquement un noyau gaussien (équation 3). La probabilité résultante est donc un mélange de gaussiennes centrées réduites sur chaque x^i. Il est montré que le choix du noyau a une influence réduite sur l'estimation. En revanche, ces méthodes sont connues pour être sensibles au choix du paramètre de lissage h, qui contrôle la régularité de l'estimation. Son choix crucial est un problème particulièrement difficile, mais largement abordé dans la littérature. Pour le fixer, nous utilisons la règle empirique de Silverman (Silverman, 1986, page 48, eqn (3.31)). Sous l'hypothèse de normalité de la distribution sous-jacente, cette règle propose une façon simple de calculer le paramètre h optimum lorsque des fonctions gaussiennes sont utilisées pour approximer des données univariées (équation 4 où $\hat{\sigma}$ est l'écart type estimé sur l'échantillon, q_1 et q_3 respectivement les premier et troisième quartiles).

Une fois ces probabilités estimées sur chacun des k plus proches voisins de w, on peut alors calculer le score de confiance $Q(w)$. La complexité de ce calcul pour l'ensemble des listes de voisinage est donc en $\mathcal{O}(k * n^2)$.

5.3 Utilité du score de confiance

L'intérêt attendu du score de confiance est de permettre d'avoir un indice a priori de la qualité d'une liste de voisins pour un mot donné. Un tel score peut ainsi être utile pour de nombreuses applications exploitant les thésaurus produits par notre approche. Une évaluation du score de confiance par le biais de telles applications serait certainement le plus adapté,

mais dépasse le cadre de cet article. Nous utilisons à défaut une évaluation directe vis-à-vis de la MAP : nous mesurons la corrélation entre la MAP et le score de confiance, l'idée étant qu'une entrée avec une liste de voisins de faible qualité correspond à une entrée ayant une MAP faible.

Plusieurs indices de corrélation peuvent être employés. L'indice r de Pearson mesure une corrélation linéaire entre score et MAP. Nous utilisons également les corrélations ρ de Spearman et τ de Kendall qui ne font pas d'hypothèse de linéarité et comparent uniquement l'ordre des mots classés selon la MAP à l'ordre selon le score de confiance. Les résultats de ces trois coefficients sont donnés dans le tableau 3 (1 indique une corrélation parfaite, 0 une absence de corrélation et -1 une corrélation inverse), avec pour chacun la p-valeur du test de significativité associé (une p-valeur faible, par exemple < 0.05, indique un résultat statistiquement significatif). Les scores de confiance sont obtenus avec $k = 20$; d'autres expériences non rapportées ici montrent que ce paramètre, s'il est choisi entre 5 et 100, influence peu les valeurs de corrélation. Ces mesures montrent une corrélation certaine et statistiquement significative entre notre score de confiance et la MAP, mais néanmoins imparfaite et non linéaire. Le score de confiance est tout de même un bon indicateur de qualité comme en témoigne aussi le graphe en figure 3 où est représentée la moyenne des MAP (en ordonné) sur les listes de voisins ayant un score de confiance inférieur à un seuil que nous faisons varier (en abscisse).

Coefficient de corrélation	valeur	significativité statistique
Pearson r	0.16	$p < 10^{-40}$
Kendall τ	0.37	$p < 10^{-64}$
Spearman ρ	0.51	$p < 10^{-64}$

TABLE 3: Mesures de corrélation entre le coefficient de confiance et la MAP, avec leur significativité (p-valeur)

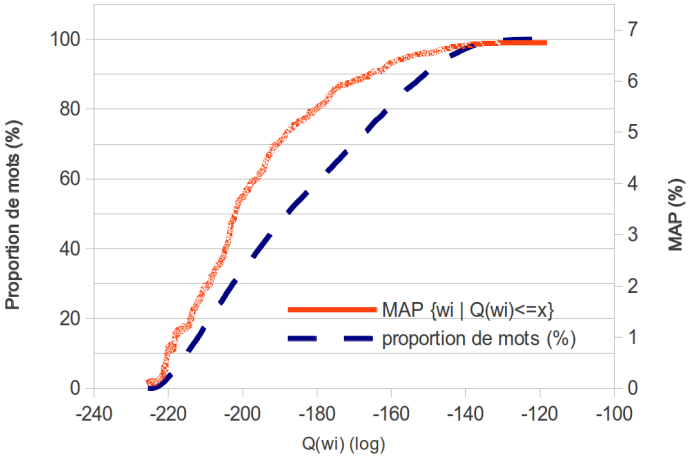


FIGURE 3: MAP des mots dont le score de confiance est inférieur à un certain seuil (donné en abscisse (log)) et proportion cumulée de mots concernés

Le score de confiance peut être utilisé pour améliorer les résultats des techniques d'agrégation vus en section 4. L'idée est simplement d'intégrer le score de confiance dans le score final :
$\text{score}_{w_i}(w_j) = Q(w_j) * \text{Agreg}(1 - \tau_{w_i}(w_j)/n, 1 - \tau_{w_j}(w_i)/n)$
Comme on le voit dans le tableau 4, l'ajout de cette information permet des gains encore plus importants que ceux rapportés dans la section précédente. Comme précédemment, ces gains sont plus sensibles en fin de liste (P@50, P@100). Dans la section suivante, nous tentons d'améliorer également les résultats en début de liste, c'est-à-dire sur les voisins jugés les plus proches, en utilisant différemment les scores de confiance.

Méthode	MAP	R-Prec	P@1	P@5	P@10	P@50	P@100
$S_{\text{Hamacher}}\ \gamma = 0.95$	9.61 (+7.20)	11.59 (+5.85)	*30.86 (-0.53)*	19.52 (+5.83)	14.76 (+7.24)	7.03 (+8.88)	4.93 (+8.67)

TABLE 4: Performances et gains (%) par agrégation des rangs réciproques prenant en compte le score de confiance sur la référence WordNet+Moby

6 Réordonnancement local

La méthode précédente donne un score global à la liste, mais on peut aussi exploiter les probabilités de classements individuelles (les $p(\delta(w_i, w_j))$) calculées selon la méthode des fenêtres de Parzen. Pour un mot donné w, on dispose pour chacun de ses voisins w_j d'un score de confiance individuel lié à son rang actuel ($p(\delta(w, w_j)) = \tau_w(w_j)$), et l'on peut

également calculer les probabilités de voir ce voisin à n'importe quel autre rang τ (probabilité que ce mot soit au rang 1, 2...). Dans cette section, on se propose d'utiliser ces informations plus locales pour améliorer les résultats en réordonnant les k-plus-proches voisins.

6.1 Réordonner par l'algorithme hongrois

Une première approche consisterait à réordonner la liste sur la base de ce critère, des voisins les plus probables aux moins probables. Mais notre critère de qualité associé à chaque mot est imparfait, et un tel réordonnancement dégrade fortement les résultats. On propose donc à la place une méthode permettant de réordonner les k-plus-proches voisins de manière locale (un mot qui n'était pas dans les k-plus-proches ne peut pas y entrer) et contrôlée (un mot ne peut pas s'éloigner trop de son rang initial).

Notre problème s'exprime par la matrice suivante, dite matrice de profit, dans laquelle les lignes correspondent aux mots dans l'ordre du classement actuel (notés w_1 à w_k), et les colonnes correspondent aux nouveaux rangs auxquels assigner ces mots. Étant données les probabilités de chaque mot w_j d'apparaître à un rang τ, l'objectif est de trouver la permutation des k plus proches voisins la plus probable, c'est-à-dire celle qui "profite" le plus.

$$\mathcal{M}_{\text{profit}} = \begin{pmatrix} p(\delta(w,w_1)=1) & \cdots & p(\delta(w,w_1)=k) \\ \vdots & \ddots & \vdots \\ p(\delta(w,w_k)=1) & \cdots & p(\delta(w,w_k)=k) \end{pmatrix} \qquad \mathcal{M}_{\text{pénalité}} = \begin{pmatrix} 1 & \frac{k-1}{k} & \cdots & 0 \\ \frac{k-1}{k} & 1 & \cdots & \frac{1}{k} \\ \vdots & \ddots & \vdots & \\ 0 & \frac{1}{k} & \cdots & 1 \end{pmatrix}$$

Comme nous l'avons souligné, on souhaite par ailleurs limiter les déplacements importants, pour éviter qu'un voisin initialement très proche se retrouve beaucoup plus loin et inversement. On ajoute cette contrainte à la matrice de profit en prenant le produit d'Hadamard (produit de matrices composante par composante, noté $\circ$) avec la matrice de pénalité $\mathcal{M}_{\text{pénalité}}$.

On se trouve alors face à un problème d'optimisation combinatoire, qui peut se résoudre en un temps polynomial en appliquant l'algorithme hongrois (Kuhn & Yaw, 1955, pour une description de l'algorithme) sur $\mathcal{M}_{\text{profit}} \circ \mathcal{M}_{\text{pénalité}}$. Cet algorithme a été initialement proposé pour optimiser l'assignation de travailleurs (dans notre cas les voisins) sur des tâches (dans notre cas, le rang), selon le profit dégagé par chaque travailleur pour chaque tâche (ici, la probabilité que ce voisin soit à ce rang). Il permet de trouver l'assignation optimale étant donnée une matrice de profit. Son résultat nous indique donc un nouveau rang pour chaque mot. L'algorithme converge sur une solution optimale et est de complexité $\mathcal{O}(k^3)$ (pour le réordonnancement des k-plus-proches voisins).

6.2 Résultats

Le tableau 5 présente les performances obtenues par rapport à notre référence Okapi-BM25 ajusté selon les mêmes modalités expérimentales que précédemment. Comme précédemment, on a fixé le voisinage considéré à $k = 20$; les précisions au delà de ce seuil sont donc inchangées et ne sont pas reportées. Nous testons l'efficacité de ce réordonnancement sur l'ensemble des listes de voisins et sur le tiers des listes ayant les scores de qualité les plus faibles. Il en ressort que le réordonnancement sur l'ensemble des listes n'apporte pas de véritable gain ; en revanche, sur les listes dont le score de confiance est faible, le gain est substantiel. Par ailleurs, contrairement aux expériences de la section 4, ces gains portent par construction sur les têtes de listes, qui sont les plus à même d'être utilisées en pratique. La différence entre le traitement de l'ensemble des mots et celui sur le tiers ayant le score de confiance le plus bas s'explique de deux manières. D'une part, les listes ayant les plus forts scores de confiance correspondent en majeure partie aux listes ayant les meilleures MAP, comme attendu (et illustré en figure 3). Celles-ci laissent donc a priori peu de marge à l'amélioration. D'autre part, indé-

Méthode	MAP	R-Prec	P@1	P@5	P@10
tous les mots	9.16 (+2.17)	11.24 (+2.76)	*30.73 (-1.02)*	19.30 (+4.64)	14.37 (+4.44)
tiers avec le plus faible $Q(w_i)$	9.55 (+6.44)	11.81 (+7.99)	31.85 (+2.56)	20.43 (+10.81)	15.46 (+12.37)

TABLE 5: Performances et gains (%) du réordonnancement par l'algorithme hongrois

pendamment de leur MAP, on peut aussi supposer que ces mêmes listes ont déjà un arrangement optimal des probabilités individuelles qui explique le fort score de confiance, le réordonnancement ne concernant alors que peu de voisins.

7 Conclusion et perspectives

Les différentes contributions proposées dans cet article ne se placent pas toutes au même niveau. La construction de thésaurus en utilisant des outils issus de la RI n'est pas une innovation conceptuelle majeure, mais cette approche semble curieusement inexplorée bien qu'elle fournisse des résultats très compétitifs en demandant un minimum de travail de mise-en-œuvre grâce aux outils existants de la RI.

Les différentes propositions exploitant le graphe de voisinage pour améliorer le thésaurus relèvent d'une démarche plus originale où l'ensemble du thésaurus est considéré. Nous y avons en particulier examiné les aspects de réciprocité et de distance, en terme de rang, entre deux mots pour proposer plusieurs contributions. Certaines hypothèses, comme la réciprocité, se défendent aisément pour des relations comme la synonymie, mais restent à valider pour des relations plus complexes. À ce titre, une analyse plus fine par type de relations en s'appuyant sur la typologie de Moby reste à faire. Cependant, les améliorations apportées par l'agrégation sur l'ensemble des voisins ou la technique de réordonnancement à partir des scores de confiance valident globalement notre démarche. Il convient de noter à ce propos que les gains obtenus sont petits en valeur absolue, mais constituent, par rapport à ceux observés dans le domaine, des améliorations significatives. Une analyse contrastive entre Moby et WordNet apporterait également des éléments intéressants et complémentaires à Ferret (2013b).

Les différents aspects de ce travail ouvrent de nombreuses pistes de recherche. Par exemple, beaucoup d'autres fonctions d'agrégation outre celles testées en section 4 existent dans la littérature. Certaines pourraient d'ailleurs offrir la possibilité d'incorporer le score de confiance associé à chaque voisin, comme les intégrales de Choquet ou de Sugeno (Detyniecki, 2000). Plus largement, il serait intéressant d'utiliser itérativement les améliorations des listes de voisins pour mettre à jour les scores de confiance, etc., en s'inspirant par exemple de ce qui est proposé par Pedronette *et al.* (2014). Au delà des thésaurus distributionnels, les méthodes proposées pour calculer des scores de confiance ou réordonner les listes de voisins peuvent s'appliquer à d'autres problèmes où ces graphes de k plus proches voisins sont construits. Notons également que nous n'avons considéré qu'une petite partie de l'information portée par le graphe de voisinage. Nous nous sommes concentrés sur les aspects de réciprocité, mais d'autres travaux, prenant en compte d'autres aspects de ce graphe (la transitivité notamment, ou plus globalement sa topologie), pourraient mener à d'autres améliorations.

Références

ADAM C., FABRE C. & MULLER P. (2013). Évaluer et améliorer une ressource distributionnelle : protocole d'annotation de liens sémantiques en contexte. *TAL*, **54**(1), 71–97.

M. BOUGHANEM & J. SAVOY, Eds. (2008). *Recherche d'information : états des lieux et perspectives*. http ://www.editions-hermes.fr/ : Hermès Science.

BRODA B., PIASECKI M. & SZPAKOWICZ S. (2009). Rank-Based Transformation in Measuring Semantic Relatedness. In *22nd Canadian Conference on Artificial Intelligence*, p. 187–190.

BUDANITSKY A. & HIRST G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, **32**(1), 13–47.

CURRAN J. R. & MOENS M. (2002). Improvements in automatic thesaurus extraction. In *Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, p. 59–66, Philadelphia, USA.

DETYNIECKI M. (2000). *Mathematical aggregation operators and their application to video querying*. PhD thesis, Université de Paris 6.

DOMENGÈS D. & VOLLE M. (1979). Analyse factorielle sphérique : une exploration. *Annales de l'INSEE*, **35**, 3–83.

ESCOFFIER B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de statistique appliquée*, **26**(4), 29–37.

FERRET O. (2012). Combining bootstrapping and feature selection for improving a distributional thesaurus. In *20th European Conference on Artificial Intelligence (ECAI 2012)*, p. 336–341, Montpellier, France.

FERRET O. (2013a). Identifying bad semantic neighbors for improving distributional thesauri. In *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, p. 561–571, Sofia, Bulgaria.

FERRET O. (2013b). Sélection non supervisée de relations sémantiques pour améliorer un thésaurus distributionnel. In *20ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2013)*, p. 48–61, Les Sables d'Olonne, France.

FIRTH J. R. (1957). *Studies in Linguistic Analysis*, chapter A synopsis of linguistic theory 1930-1955, p. 1–32. Blackwell : Oxford.

FREITAG D., BLUME M., BYRNES J., CHOW E., KAPADIA S., ROHWER R. & WANG Z. (2005). New experiments in distributional representations of synonymy. In *Ninth Conference on Computational Natural Language Learning (CoNLL)*, p. 25–32, Ann Arbor, Michigan, USA.

GREFENSTETTE G. (1994). *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers.

HAGIWARA M., OGAWA Y. & TOYAMA K. (2006). Selection of effective contextual information for automatic synonym acquisition. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, p. 353–360, Sydney, Australia.

HUANG E. H., SOCHER R., MANNING C. D. & NG A. Y. (2012). Improving word representations via global context and multiple word prototypes. In *50th Annual Meeting of the Association for Computational Linguistics (ACL'12)*, p. 873–882.

HULL D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93*, Pittsburgh, États-Unis.

KUHN H. W. & YAW B. (1955). The hungarian method for the assignment problem. *Naval Research Logistic Quarterly*, **2**, 83–97.

LANDAUER T. K. & DUMAIS S. T. (1997). A solution to Plato's problem : the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**(2), 211–240.

LIN D. (1998). Automatic retrieval and clustering of similar words. In *17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics (ACL-COLING'98)*, p. 768–774, Montréal, Canada.

MIKOLOV T., YIH W.-T. & ZWEIG G. (2013). Linguistic regularities in continuous space word representations. In *2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (NAACL HLT 2013)*, p. 746–751, Atlanta, Georgia.

MILLER G. A. (1990). WordNet : An On-Line Lexical Database. *International Journal of Lexicography*, **3**(4).

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

PARZEN E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065–1076.

PEDRONETTE D. C. G., PENATTI O. A. & DA S. TORRES R. (2014). Unsupervised manifold learning using reciprocal knn graphs in image re-ranking and rank aggregation tasks. *Image and Vision Computing*, **32**(2), 120 – 130.

ROBERTSON S. E., WALKER S. & HANCOCK-BEAULIEU M. (1998). Okapi at TREC-7 : Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proc. of the 7th Text Retrieval Conference, TREC-7*, p. 199–210.

SAHLGREN M. (2001). Vector-based semantic analysis : Representing word meanings based on random labels. In *ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*, Helsinki, Finland.

SILVERMAN B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. London, Glasgow, Weinheim : Chapman and Hall Boca Raton.

VAN DE CRUYS T. (2010). *Mining for Meaning. The Extraction of Lexico-semantic Knowledge from Text*. PhD thesis, University of Groningen, The Netherlands.

VECHTOMOVA O. & ROBERTSON S. (2012). A domain-independent approach to finding related entities. *Information Processing and Management*, **48**(4).

WARD G. (1996). Moby thesaurus. Moby Project.

WASSERMAN L. (2005). *All of Statistics : A Concise Course in Statistical Inference*. Springer Texts in Statistics.

YAMAMOTO K. & ASAKURA T. (2010). Even unassociated features can improve lexical distributional similarity. In *Second Workshop on NLP Challenges in the Information Explosion Era (NLPIX 2010)*, p. 32–39, Beijing, China.

ZHITOMIRSKY-GEFFET M. & DAGAN I. (2009). Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, **35**(3), 435–461.