Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité

Amandine Périnet[1], Thierry Hamon[2,3]

[1] LIMICS, INSERM, U1142, LIMICS, F-75006, Paris, France ; Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006, Paris, France ; Université Paris 13, Sorbonne Paris Cité, LIMICS, (UMR_S 1142), F-93430, Villetaneuse, France
amandine.perinet@edu.univ-paris13.fr
[2] LIMSI-CNRS, Orsay – thierry.hamon@limsi.fr
[3] Université Paris 13, Sorbonne Paris Cité, Villetaneuse

Abstract

Approaches based on the distributional analysis have in common to select the contexts of each target word in order, then, to compute a semantic similarity between these words. The definition and selection of contexts as well as the choice of the similarity measure are important parameters when applying this kind of approaches: the choice of these parameters may depend on textual data to analyze (data volume, general or specific language, specialized field, etc.) and on the final application. Our goal is to study the impact of the different values and the definition of those parameters on specialized texts. Working on two corpora in the food domain, we suggest to study the impact of the parameters and the behavior of distributional analysis according to these values. Our experiments show that the Jaccard Index seems to be the measure the more adapted to specialized corpora, when used with a graphical window of two words each side of the target word.

Résumé

Les approches qui reposent sur l'hypothèse distributionnelle ont pour point commun de sélectionner les contextes de chaque mot cible, pour ensuite calculer une similarité entre ces mots. La description et la sélection des contextes mais aussi la mesure de similarité utilisée pour rapprocher des mots cibles sont des paramètres importants lors de la mise en œuvre de ce type d'approche : le choix de ces paramètres peut dépendre des données textuelles à analyser (volume de données, langue générale ou de spécialité, domaine de spécialité, etc.) et l'application finale. Notre objectif est d'étudier l'impact des différentes valeurs et la définition de ces paramètres sur des textes de spécialité. A partir de deux corpus du domaine alimentaire, nous proposons d'analyser l'impact des paramètres et le comportement de l'analyse distributionnelle en fonction des valeurs des paramètres. Ainsi, la mesure de Jaccard semble être la mesure la mieux adaptée aux corpus de spécialité, quand elle est utilisée avec un contexte graphique de deux mots de chaque côté du mot cible.

Mots-clés : Analyse distributionnelle, textes de spécialité, relations sémantiques, paramètres distributionnels.

1. Introduction

Les méthodes d'analyse distributionnelles (AD) se basent sur l'hypothèse harissienne, selon laquelle les mots qui apparaissent dans des contextes similaires ont tendance à être sémantiquement proches (Harris, 1954). Plusieurs paramètres doivent être définis et peuvent varier et nécessiter une adaptation au corpus. Ainsi, le contexte peut être syntaxique (Lin, 1998a ; Curran, 2004) (nom argument d'un verbe), graphique (Wilks et al., 1990 ; Schütze, 1998) (mots dans le voisinage d'un mot cible), ou cross-lingue (Van der Plas et Tiedemann, 2006) (traductions d'un mot dans d'autres langues). Aussi, de nombreuses mesures de pondération et de similarité peuvent être appliquées sur ces contextes pour déterminer la relation sémantique entre deux mots (Panchenko et Morozova, 2012). Tous ces éléments

représentent un nombre important de paramètres à définir avant d'utiliser l'AD dans un contexte applicatif (Weeds et al., 2004). Les méthodes distributionnelles utilisent habituellement des corpus de grande taille constitués de textes en langue générale. Or, les corpus de spécialité sont de plus petite taille et les mots et termes occurrent moins fréquemment. Afin de mettre en œuvre une méthode distributionnelle adaptée aux corpus de spécialité, nous proposons d'étudier le comportement de cinq paramètres distributionnels, sur deux corpus composés de guides alimentaires et de recettes de cuisine : deux paramètres au niveau de la sélection des contextes (taille de la fenêtre graphique et sélection des contextes les plus discriminants) ainsi que trois paramètres autour du calcul de la similarité sémantique (pondération des contextes, mesure de similarité et seuil).

Nous présentons d'abord l'état de l'art, puis la méthode distributionnelle mise en œuvre et les paramètres, et enfin les expériences que nous avons menées. Les résultats obtenus sont ensuite présentés et analysés en termes de macro-précision et MAP.

2. Etat de l'art

La similarité distributionnelle a été beaucoup étudiée et est devenue une méthode établie pour trouver des mots similaires. La mise au point d'une méthode d'analyse distributionnelle pour identifier des similarités entre des mots en fonction de leurs contextes partagés nécessite tout d'abord de définir les mots cibles en relation et le contexte de ces mots, puis de calculer à partir de ces contextes une valeur de similarité entre les mots cibles. Les diverses méthodes diffèrent par la définition du le contexte et du calcul de similarité (Curran et Moens, 2002a).

En ce qui concerne la définition du contexte, la plupart des travaux en langue générale utilise soit des co-occurrences (mots présents autour du mot cible : fenêtres graphiques), soit des informations syntaxiques avec des dépendances de relations grammaticales (Curran et Moens, 2002a ; Brown et al., 1992). Le contexte basé sur des fenêtres graphiques permettrait d'acquérir des relations plus lâches que le contexte syntaxique (Van der Plas et Bouma, 2004 ; Kilgarriff et Yallop, 2000), en établissant des relations entre des mots d'un même domaine, alors que l'analyse syntaxique définirait des contextes plus précis. (Pado et Lapata, 2007) ont d'ailleurs démontré que l'utilisation des dépendances syntaxiques permet de distinguer des classes de relations lexicales. Pour les mesures de similarité, plusieurs travaux se sont intéressés à une présentation et évaluation étendue des différentes mesures de similarité utilisées en langue générale (Weeds, 2003). (Curran et Moens, 2002b) réalisent une expérience d'évaluation à large échelle, dans laquelle ils étudient la performance de plusieurs méthodes communément utilisées. (Van der Plas et Bouma, 2004) présentent une expérience similaire pour le danois, dans laquelle ils testent la plupart des mesures les plus performantes d'après (Curran et Moens, 2002b). Dans leurs expériences, Jaccard et Dice sont les deux mesures qui obtiennent les meilleures performances. Le calcul de similarité est généralement accompagné d'une pondération des contextes distributionnels. La mesure de pondération principalement utilisée en langue générale est l'information mutuelle (Zhitomirsky-Geffet et Dagan, 2009 ; Church et Hanks, 1990 ; Hindle, 1990).

Pour les textes de spécialité en revanche, les travaux en analyse distributionnelle sont moins nombreux. Malgré tout, des tendances se dessinent. Pour la définition du contexte, une fenêtre graphique d'une taille de deux mots autour du mot cible est la taille définie comme idéale par plusieurs auteurs (Rapp, 2003, Généreux et Hamon 2013). Il manque surtout aujourd'hui une étude systématique de la contribution et de la définition des paramètres distributionnels pour les textes de spécialité. C'est dans cette optique que se place notre travail.

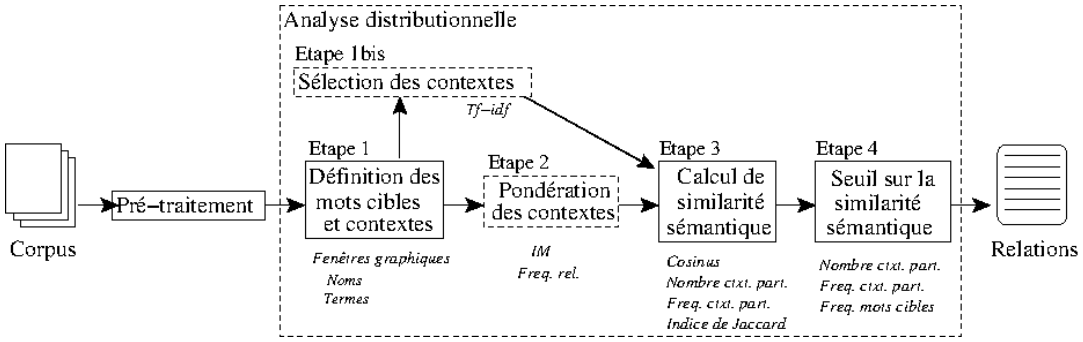

Figure 1. Méthode distributionnelle et paramètres

3. Méthode distributionnelle

Notre méthode s'appuie sur l'hypothèse distributionnelle et suit le schéma de la figure 1. Cette méthode comporte quatre étapes principales que nous avons adaptées aux textes de spécialité : la définition des mots cibles et des contextes (étape 1), la pondération des contextes (étape 2), le calcul de similarité sémantique (étape 3) et le seuil sur la similarité sémantique (étape 4). Nous ajoutons aussi une étape optionnelle de sélection de contextes (étape 1bis) intervenant lors de la définition des contextes. Nous présentons ces étapes.

Étape 1 : Définition des mots cibles et des contextes

Pendant cette étape, les mots cibles et les contextes sont définis. Les termes sont identifiés à l'aide d'un extracteur de termes, à la fois pour les mots cibles et les contextes. De plus, le problème de faibles fréquences auquel nous devons faire face nous amène à sélection les contextes afin de prendre en compte le plus grand nombre de contextes possibles.
Pour définir les mots cibles et contextes, appuyons-nous sur l'exemple suivant :

(1) *Le yaourt contient du calcium.*
Le lait est un composant du fromage.

Mots cibles Les mots cibles sont restreints aux noms et termes (groupes nominaux correspondant à des unités terminologiques) proposés par l'extracteur YaTeA (Aubin&Hamon2006). Nous cherchons par la suite à mettre en relation les mots cibles, du moment qu'ils partagent la même catégorie morphosyntaxique. Dans l'exemple, les mots cibles sont *yaourt*$_{\_NOM}$, *calcium*$_{\_NOM}$, *lait*$_{\_NOM}$, *composant*$_{\_NOM}$ et *fromage*$_{\_NOM}$.

Contextes distributionnels Les contextes sont composés de mots qui co-occurrent au sein d'une fenêtre graphique. Nous avons choisi d'évaluer deux tailles de fenêtres décrites dans la section 4.2. Ces fenêtres sont calculées en prenant en compte uniquement les adjectifs, noms, verbes et termes présents dans les contextes, écartant ainsi les mots non porteurs de sens (déterminants, conjonctions, adverbes, etc.). Par exemple, avec une fenêtre de 5 (4 mots autour du mot cible), le mot cible *lait*$_{\_NOM}$ a pour contextes *composant*$_{\_NOM}$ et *fromage*$_{\_NOM}$.

Étape 1bis : sélection des contextes

Nous ajoutons une étape optionnelle de sélection des contextes. Cette étape nous permet de sélectionner les contextes en fonction de leur caractère discriminant ou non. En effet, parmi les contextes d'un mot, certains sont de meilleurs descripteurs que d'autres (Morlane-Hondère, 2013). Par exemple, avec l'utilisation d'une fenêtre graphique de 5, le fait que le

nom *vanille_NOM* co-occurre avec *avoir_VER* et *exemple_NOM* ne nous donne que peu d'informations sur la nature sémantique de *vanille*, car trop généraux. Ces contextes sont très fréquents : ils apparaissent respectivement 47 297 et 10 950 fois dans le corpus et cooccurrent avec 1 478 et 884 lemmes différents. À l'opposé, les contextes *aromatiser_VER* et *glace_NOM* n'ont qu'une fréquence de 162 et 302 et ne cooccurrent qu'avec un ensemble beaucoup plus réduit de lemmes ; respectivement 98 et 143. Leur cooccurrence avec *vanille* est beaucoup plus significative : ils sont plus caractéristiques du nom *vanille*, et donc beaucoup plus pertinents pour sa description. Nous pourrions utiliser un lexique afin d'écarter ces mots généraux, mais cela ne suffit pas. Aussi, nous proposons de sélectionner automatiquement les mots en contexte les moins discriminants pour un mot cible donné. Généralement, ce rapport d'exclusivité entre un mot et ses contextes se calcule à l'aide de mesures d'association, comme nous le faisons dans l'étape 2, c'est-à-dire une fois que les mots cibles et leurs contextes ont été définis. Nous avons fait le choix d'expérimenter la sélection des contextes les plus discriminants lors de la définition du contexte. Pour cela, nous évaluons l'apport du Tf-Idf en adaptant la méthode standard (Jones, 1972) à nos besoins. Tout d'abord, les documents sont les mots cibles (*w*), et les termes correspondent pour nous aux mots dans le contexte (*w'*). De plus, nous supprimons le log, car celui-ci écrase beaucoup trop les choses. Pour nous différencier du Tf.Idf, nous nommons notre adaptation Cf.Itf.

$$Cf.Itf = f(w,r,w') \times \frac{n(w)}{n(w)|\exists(w,r,w')}$$

où *f(w,r,w')* correspond au nombre d'occurrences des mots *w* et *w'* se trouvant en relation *r*, et *n(w)* est le nombre de mots cibles. Après le calcul du Cf-ITf de chaque contexte d'un mot cible donné, nous sélectionnons les contextes suivant leur Cf-Itf. Pour chaque cible, nous ordonnons les contextes suivant la valeur de Cf-Itf par ordre décroissant, puis nous calculons l'écart moyen entre les valeurs de Cf-Itf des contextes. L'idée est de supprimer les contextes les plus généraux. Il faut donc trouver un bon équilibre, un écart suffisamment important entre deux contextes sans pour autant supprimer trop de contextes. Après observation du comportement des écarts moyens entre deux contextes, nous avons fixé le seuil à l'écart moyen x 1,5. Ainsi, si l'écart entre le contexte courant et le suivant est inférieur à l'écart moyen x 1,5 (défini expérimentalement), nous conservons le contexte. Dès lors où l'on rencontre un écart supérieur à la moyenne x 1,5, nous ignorons le reste des contextes pour ce mot cible.

Étape 2 : Pondération des contextes

Une fois extraits, les contextes sont pondérés. Nous avons décidé d'étudier plusieurs mesures de pondération (*cf.* tableau 1) : l'information mutuelle (Manning et Schütze, 1999) et la fréquence relative : où *p(w,r,w')* est la probabilité de trouver les mots *w* et *w'* en relation de co-occurrence *r*, *p(w,*,*)* la probabilité de trouver le mot cible *w*, et *p(*,r,w')* la probabilité de trouver le contexte *w'*. Nous utilisons la notation en astérisque proposée par (Lin, 1998b).

information Mutuelle (IM)	$poids_{IM} = log(\frac{p(w,r,w')}{p(w,*,*)p(*,r,w')})$
fréquence relative (FreqRel)	$poids_{FreqRel} = \frac{f(w,r,w')}{f(w,*,*)}$

Tableau 1. Mesures de pondération utilisées

Étape 3 : Calcul de similarité sémantique

Après avoir extrait et pondéré les contextes d'apparition de chaque mot cible, l'étape suivante consiste à comparer ces contextes afin de calculer un score de similarité sémantique entre chaque couple de mots cibles, w_m et w_n, partageant la même catégorie morpho-syntaxique. Nous expérimentons les mesures décrites dans le tableau 2.

nombre de contextes partagés (NbCtxt)	$sim_{NbCtxt} = \|(w_m, r, w') \cap (w_n, r, w')\|$
fréquence des contextes partagés (FreqCtxt)	$sim_{FqCtxt} = \sum_{w'} min(\|(w_m, r, w')\|, \|(w_n, r, w')\|)$
mesure de Jaccard (Jacc)	$sim_{Jaccard} = \frac{\|(w_m,*,*)\| \cap \|(w_n,*,*)\|}{\|(w_m,*,*)\| \cup \|(w_n,*,*)\|}$
cosinus (Cos)	$sim_{Cosinus} = \frac{\|(w_m,*,*) \cap (w_n,*,*)\|}{\sqrt{\|(w_m,*,*)\| \times \|(w_n,*,*))\|}}$

Tableau 2. Mesures de similarité utilisées

Étape 4 : Seuils sur les contextes et mots cibles

Lors du calcul de la similarité entre les mots cibles, un très grand nombre de relations est généré. Garder toutes ces relations n'aurait pas de sens : un trop grand ensemble de relations est difficile à exploiter et à analyser a posteriori. L'utilisation des seuils sur les contextes et mots cibles nous permet donc d'éliminer les relations les moins pertinentes. Les deux premiers sont appliqués aux contextes et le dernier aux mots cibles :

• Nombre de contextes partagés : nombre de contextes lemmatisés. Si deux mots partagent *crème*$_{\_NOM}$, *battre*$_{\_VER}$, *poivre*$_{\_NOM}$, *sel*$_{\_NOM}$, le nombre de contextes partagés est 4.

• Fréquences des contextes partagés : nombre d'occurrences des contextes lemmatisés. Dans l'exemple précédent, les fréquences sont toutes égales à 1.

• Fréquence des mots cibles : nombre d'occurrences des mots cibles lemmatisés.

Pour chaque paramètre, un seuil est calculé automatiquement à partir du corpus. Nous avons choisi la moyenne des valeurs prises par chaque paramètre sur l'ensemble du corpus.

4. Expériences

Nous présentons ici d'une part, le matériel utilisé pour les expériences et l'évaluation, d'autre part, les jeux de paramètres que nous expérimentons au sein de notre méthode.

4.1. Corpus

Nous utilisons deux corpus du domaine de l'alimentation en langue française dont le contenu et la taille diffèrent (*cf.* tableau 3). L'intérêt est ici d'évaluer si les paramètres distributionnels se comportent de la même manière quel que soit le corpus de spécialité utilisé.

CORPUS	GUIDES ALIMENTAIRES (GA)	RECETTES
Nombre de textes	21	23 121
Nombre de mots	471 463	3 928 658
Longueur des phrases (moyenne)	19,15 mots	8,97 mots
Contenu	Bonnes Pratiques, conseil médical, règlements, etc.	Recettes de cuisine : Titre + ingrédients + préparation

Tableau 3. Description des corpus

Le premier corpus contient l'ensemble de documents fournis par la compétition scientifique DEFT 2013[1] (3 928 658 mots). Il rassemble des recettes de cuisine. Nous gardons toutes les informations des textes dans notre corpus : le titre, la section ingrédients et le corps de la recette. Le deuxième corpus est une collection de guides alimentaires (GA). Ces textes sont collectés à l'aide de requêtes dans Google telles que « *guide alimentaire* » et « *guide de nutrition* ». Ce corpus est de plus petite taille que le premier, avec 471 463 mots. Les documents sont sélectionnés parmi deux sources principales : le gouvernement et les entités médicales. Il s'agit de textes décrivant les Bonnes Pratiques, des conseils médicaux (destinés aux patients et visant à prévenir les pathologies), les normes, le comportement des professionnels de santé, le bien-être, etc.

Nous avons réalisé un pré-traitement du corpus au sein de la plateforme Ogmios afin d'articuler plusieurs outils de TAL (Hamon et al., 2007). Nous réalisons un étiquetage morpho-syntaxique et une lemmatisation avec *TreeTagger* (Schmid, 1994), et nous utilisons l'extracteur de termes YATEA (Aubin et Hamon, 2006).

MESURES DE SIMILARITE (avec seuils)	SELECTION CF-ITF (avec seuils)	SELECTION CF-ITF (sans seuils)	MESURES DE PONDERATION (avec seuils)
Jaccard	Jaccard + Cf-Itf +seuils	Jaccard + Cf-Itf	Jacc +FreqRel
Cosinus	Cosinus + Cf-Itf +seuils	Cosinus + Cf-Itf	Cos + FreqRel Cos + IM
FreqCtxt	FreqCtxt + Cf-Itf +seuils	FreqCtxt + Cf-Itf	-
NbCtxt	NbCtxt + Cf-Itf +seuils	NbCtxt + Cf-Itf	-

Tableau 4. Paramètres distributionnels évalués pour deux tailles de fenêtre graphique (W5 et W21), pour les noms et les termes

4.2. Paramètres distributionnels

Dans nos jeux d'expériences, nous faisons varier les paramètres distributionnels présentés ci-dessous. Le tableau 4 synthétise ces jeux d'expériences. Nous testons tout d'abord les quatre mesures de similarité avec les seuils sur les valeurs de la mesure. Ensuite, nous évaluons l'apport de la sélection des contextes les plus discriminants à l'aide du Cf-Itf en amont du calcul de similarité, à la fois avec et sans seuil sur la similarité. Et enfin, nous expérimentons les deux pondérations avec les mesures de Jaccard et Cosinus. Ces jeux d'expériences sont réalisés pour deux tailles de fenêtre graphique : une fenêtre de 5 mots (W5) et une fenêtre plus grande de 21 mots (W21) de chaque côté du mot cible.

[1] http://deft.limsi.fr/2013/

		Noms-W21	Termes-W21	Noms-W5	Termes-W5
Recettes	Fréq. des mots cibles	82	5	82	5
	Nb de contextes partagés	13	2	8	1
	Fréq. des ctxt partagés	29	2	15	2
Guides alimentaires	Fréq. des mots cibles	10	2	13	2
	Nb de contextes partagés	7	1	3	1
	Fréq. des ctxt partagés	8	1	4	1

Tableau 5. Valeurs des seuils sur les contextes et mots cibles dans nos expériences

Nous récapitulons les valeurs des seuils sur les contextes partagés et les mots cibles dans le tableau 5 (*cf. Étape 4* pour la définition des seuils).

4.3. Evaluation

Afin d'évaluer la qualité des relations sémantiques acquises, nous comparons nos relations à des références obtenues à partir de trois ressources, à l'aide de deux mesures d'évaluation.

4.3.1. Ressources

Pour l'évaluation, nous utilisons deux ressources existantes et une ressource que nous avons construite à partir de quatre sites Web. Les deux premières ressources correspondent aux 75 222 relations extraites de la partie française d'Agrovoc[2] [AGRO], et aux 1 735 419 relations fournies par la partie française de l'UMLS[3] [UMLS]. Au sein de l'UMLS, nous avons également sélectionné un sous-ensemble de 1 608 relations dédiées au concept Food (type sémantique T168) [UMLS/Food]. La comparaison avec UMLS/Food et Agrovoc se justifie par la disponibilité de ces ressources et la présence de relations entre des termes alimentaires dans les deux ressources. Cependant, dans les recettes de cuisine nous pouvons trouver également d'autres types de relations, telles que les relations entre un terme alimentaire et un terme appartenant à une autre classe sémantique (par exemple, une pathologie). La comparaison avec l'UMLS dans son ensemble est utile afin de détecter d'autres relations que des relations entre ingrédients.

Même si nous n'attendons pas un recouvrement important entre les ressources et le corpus, la comparaison de nos résultats aux relations présentes dans ces ressources nous donne une indication de la contribution de chaque paramètre distributionnel. Etant donné qu'Agrovoc et l'UMLS ne sont pas particulièrement dédiés au domaine alimentaire, nous avons également construit une ressource plus spécifique qui contient 5 058 relations, collectées à partir de quatre sites Web dont le thème est lié à l'alimentaire : une société qui vend une méthode pour perdre du poids[4], le site de Santé Canada[5] (le département du gouvernement canadien responsable de la santé publique nationale), un centre spécialisé dans la chute des cheveux[6], et un site web qui propose des recettes de cuisines[7] (différentes des recettes de notre corpus).

[2] http://aims.fao.org/standards/agrovoc/about

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

[4] http://www.bioweight.com/

[5] http://www.hc-sc.gc.ca/fn-an/securit/addit/diction/index-fra.php

[6] http://www.centre-clauderer.com/acides-bases/femme-2.htm

[7] http://www.cuisine-libre.fr/

Nous avons typé manuellement les relations contenues dans cette ressource, et celle-ci contient des hyperonymes (1 570), des co-hyponymes (2 809), des méronymes (583), quelques variantes morpho-syntaxiques (71) ainsi que des synonymes (25).

Afin de prendre en considération uniquement les relations qui peuvent potentiellement être trouvées, nous avons défini deux références à partir des ressources présentées ci-dessus, une pour chaque corpus. Pour le corpus Recettes, la référence contient 1 551 ([AGRO]), 1 748 ([UMLS]), 871 ([UMLS/Food]) et 2 027 ([WEB]) relations. Pour le corpus GA, la référence contient 2 931 ([AGRO]), 2 812 ([UMLS]), 504 ([UMLS/Food]) et 1 764 ([WEB]) relations. La référence pour le corpus GA contient plus de relations pour Agrovoc et l'UMLS, parce que plus de mots issus de nos corpus sont trouvés dans ces ressources.

4.3.2. Mesures d'évaluation

L'évaluation est réalisée avec la macro-précision et la MAP (Mean Average Precision). Nous calculons la précision pour chaque mot cible : les voisins sémantiques (acquis par notre méthode) trouvés dans la ressource par les voisins sémantiques acquis par notre méthode. Pour chaque mot cible, nous ordonnons les voisins sémantiques obtenus selon leur mesure de similarité, et nous réalisons quatre jeux de voisins : la précision après examen de 1 (P@1), 5 (P@5), 10 (P@10) et 100 (P@100) voisins. Nous évaluons également les résultats avec la MAP (Mean Average Precision) :

$$MAP = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} P(I_i^j)$$

Où $P(I^j_i)$ est la précision non interpolée des voisins sémantiques I^j_i au rang j, N est le nombre de mots cibles, n_i est le nombre de voisins sémantiques I^j_i du mot cible R_i. Il s'agit de la moyenne des valeurs de précision obtenue pour chaque voisin extrait. Nous utilisons le programme standard *trec_eval*[8] pour la calculer.

5. Résultats et discussion

Nous procédons à l'analyse et la discussion des résultats obtenus avec les différents paramètres en fonction de leur macro-précision (tableau 6) et leur MAP (tableau 7). Nous remarquons tout d'abord que les résultats varient suivant la ressource utilisée et le corpus. Aussi, les résultats sont meilleurs pour les noms, avec des valeurs de précision comprises entre 0 et 0,5, que pour les termes dont les précisions varient entre 0 et 0,132. Pour les noms, les meilleurs résultats sont obtenus avec une fenêtre de 21 mots et pour les termes avec une fenêtre de 5 mots. Les résultats obtenus avec la mesure Cosinus sans pondération et avec les deux pondérations (FreqRel et IM) sont semblables. Ainsi, nous analysons ici uniquement les résultats de cette mesure sans pondération. De plus, nous observons que la sélection des contextes avec le Cf-Itf, ne nécessite pas de seuil appliqué sur la similarité, car dans ce cas-là les seuils suppriment beaucoup trop voire toutes les relations. Nous analysons donc uniquement l'utilisation du Cf-Itf sans seuil.

5.1. Noms

Pour les noms, sur le corpus GA, les meilleures précisions sont obtenues avec les mesures NbCtxt (comparaison à [UMLS/Food] : 0,5) et Jaccard pondérée (évaluation avec [WEB] : 0.233), et la meilleure MAP avec la mesure de Jaccard pondérée (avec [UMLS/food] et

[8] http://trec.nist.gov/trec_eval/trec_eval_latest.tar.gz

[WEB] : 1 et 0,490 respectivement). Sur le corpus Recettes, à la fois en termes de MAP et de précision les meilleurs résultats sont obtenus avec la mesure de Jaccard sans pondération (évaluation avec [WEB] : 0,305 et 0,222 respectivement). Cependant, il faut souligner que la valeur 1 obtenue avec la MAP n'est pas significative : elle se justifie par une référence (pour les quatre mesures de similarité) très restreinte contenant seulement deux relations.

		Cosinus	**Jaccard**	**Jacc.Freq**	**FreqCtxt**	**NbCtxt**	**Cos-Cf-Itf**	**Jacc-Cf-Itf**	**FreqCtxt-Cf-Itf**	**NbCtxt-Cf-Itf**
Recettes										
AGRO	N-W21	0,000	0,014	**0,071**	0,057	0,043	0,002	0,002	0,002	0,002
	T-W5	0,000	0,000	**0,022**	0,011	0,000	0,006	0,006	0,006	0,006
UMLS	N-W21	0,086	0,086	0,086	**0,114**	**0,114**	0,011	0,011	0,011	0,011
	T-W5	0,030	0,031	**0,091**	0,061	0,030	0,000	0,000	0,000	0,000
UMLS/Food	N-W21	**0,095**	0,071	0,071	**0,095**	**0,095**	0,023	0,023	0,023	0,023
	T-W5	0,023	0,024	**0,068**	0,045	0,023	0,000	0,000	0,000	0,000
WEB	N-W21	0,111	**0,222**	0,139	0,083	0,111	0,053	0,053	0,053	0,053
	T-W5	0,057	0,080	**0,132**	0,057	0,057	0,000	0,000	0,000	0,000
Guides alimentaires (GA)										
AGRO	N-W21	0,013	**0,038**	**0,038**	0,013	0,000	0,010	0,010	0,010	0,010
	T-W5	0,000	0,008	**0,013**	0,013	**0,013**	0,000	0,000	0,000	0,000
UMLS	N-W21	0,052	0,052	**0,103**	**0,103**	**0,138**	0,008	0,006	0,008	0,008
	T-W5	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
UMLS/Food	N-W21	0,000	0,250	0,312	**0,438**	**0,500**	0,000	0,000	0,000	0,000
	T-W5	0,000	0,000	0,000	**0,071**	**0,071**	0,000	0,000	0,000	0,000
WEB	N-W21	0,067	**0,200**	**0,233**	0,067	0,033	0,033	0,033	0,033	0,033
	T-W5	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Tableau 6. Précision pour le 1er voisin (P@1), pour les noms (N) et les termes (T), avec une fenêtre de 5 mots (W5) et une fenêtre de 21 mots (W21)

Ainsi, pour les noms, la mesure de Jaccard est la plus adaptée, sans pondération quand la taille du corpus est plus grande et pondérée avec la fréquence relative pour les plus petits corpus.

5.2. Termes

Pour les termes, la meilleure précision est obtenue avec la mesure de Jaccard pondérée avec la fréquence relative (0,132 de précision avec la référence [WEB]). Cette constatation est évidente avec le corpus Recettes. En revanche, pour le corpus GA, les résultats obtenus pour les termes sont quasiment tous nuls, à l'exception des mesures FreqCtxt et NbCtxt qui obtiennent 0,071 de précision dans la comparaison avec [UMLS/Food]. Ces faibles résultats sont dus à la taille du corpus et par conséquent aux seuils utilisés sur la mesure de similarité. En effet, pour le corpus GA, le plus petit corpus, nous avons des seuils très bas (*cf.* tableau 5), dont plusieurs sont égal à un, donc inutiles. Les faibles fréquences du corpus ont donc un rôle important non seulement sur le calcul de similarité en lui-même (moins de contextes sont

disponibles), mais aussi sur les valeurs des seuils permettant de sélectionner les relations les plus pertinentes, donc peut-être également sur la manière de calculer ces seuils. Pour ces plus petits corpus, il est nécessaire de définir une autre manière de sélectionner les relations.

		Cosinus	**Jaccard**	**Jacc.Freq**	**FreqCtxt**	**NbCtxt**	**Cos-Cf-Itf**	**Jacc-Cf-Itf**	**Freq-Ctxt-Cf-Itf**	**NbCtxt-Cf-Itf**
Recettes										
AGRO	N-W21	0,087	0,125	**0,225**	0,204	0,167	0,096	0,096	0,096	0,096
	T-W5	0,027	0,011	0,136	0,097	0,067	**0,500**	**0,500**	**0,500**	**0,500**
UMLS	N-W21	0,144	0,127	**0,183**	**0,188**	0,176	0,107	0,107	0,107	0,107
	T-W5	0,078	**0,367**	**0,348**	0,205	0,129	0,015	0,015	0,015	0,015
UMLS/Food	N-W21	**0,166**	0,116	0,164	**0,167**	**0,164**	0,072	0,072	0,072	0,072
	T-W5	0,115	**0,405**	**0,343**	0,180	0,126	0,152	0,152	0,152	0,152
WEB	N-W21	0,208	**0,305**	0,280	0,221	0,217	0,101	0,103	0,101	0,101
	T-W5	0,136	0,281	0,271	0,159	0,134	**0,469**	0,381	**0,469**	**0,469**
Guides alimentaires (GA)										
AGRO	N-W21	0,145	**0,222**	0,233	0,171	0,169	0,218	0,193	0,218	0,218
	T-W5	0,015	0,034	0,063	0,090	0,087	0,117	0,115	**0,194**	**0,194**
UMLS	N-W21	0,222	0,348	**0,405**	0,370	**0,389**	0,238	0,243	0,241	0,241
	T-W5	0,057	0,052	0,059	**0,120**	**0,120**	-	-	-	-
UMLS/Food	N-W21	0,164	0,374	0,392	0,404	0,463	**1,000**	**1,000**	**1,000**	**1,000**
	T-W5	0,014	0,012	0,013	**0,128**	0,130	-	-	-	-
WEB	N-W21	0,201	0,428	**0,490**	0,257	0,278	0,428	0,428	0,428	0,428
	T-W5	0,055	0,049	0,083	0,077	0,075	0,011	0,014	**0,091**	**0,091**

Tableau 7. Valeurs obtenues avec l'évaluation de la MAP), pour les noms (N) et les termes (T), avec une fenêtre de 5 mots (W5) et une fenêtre de 21 mots (W21)

En termes de MAP, les meilleurs résultats sont obtenus avec l'utilisation du Cf-Itf pour la sélection des contextes. Ce constat est valable pour les résultats obtenus avec l'évaluation avec les ressources [AGRO] et [WEB]. Avec [AGRO], une MAP de 0,5 est obtenue avec les quatre mesures de similarité combinées à la sélection par Cf-Itf pour le corpus Recettes, et 0,194 avec FreqCtxt-Cf-itf et NbCtxt-Cf-itf pour le corpus GA. Mais, avec les ressources [UMLS] et [UMLS/Food], pour le corpus Recettes, la mesure de Jaccard obtient les meilleurs résultats avec 0,367 et 0,405 respectivement. Pour le corpus GA, les meilleures mesures sont FreqCtxt et NbCtxt avec 0,120 de MAP pour l'UMLS et 0,130 pour l'UMLS/Food.

De faibles précisions pour le Cf-Itf mais de bons résultats avec la MAP signifient que les meilleurs voisins ne sont pas classés premiers (P@1). La sélection des contextes les plus discriminants a donc un impact sur la mesure de similarité en classant moins bien les voisins.

La faible couverture des ressources utilisées a des conséquences sur l'évaluation : avec le Cf-Itf, très peu des termes en relation sont retrouvés dans les ressources. En effet, nous n'avons pu constituer de références pour l'évaluation de la MAP avec [UMLS] et [WEB]. Cependant, nous avons pu constater que les relations acquises par Cf-Itf contiennent des termes plus rares et plus spécifiques au corpus que ceux acquis sans sélection des contextes.

6. Conclusion

Dans ce travail, nous nous sommes intéressés à l'impact de plusieurs paramètres distributionnels sur des corpus de spécialité, caractérisés par des tailles plus petites que les corpus généraux et par des fréquences des mots et termes plus faibles. Nous avons travaillé sur deux corpus alimentaires en langue française dont la taille et le contenu diffèrent, mais qui restent beaucoup moins volumineux que des corpus de langue générale. Nous nous sommes intéressés aux relations entre noms et entre termes. Nous avons réalisé des jeux d'expériences en faisant varier les tailles des fenêtres graphiques, les mesures de similarité, de pondération, et en expérimentant une méthode de sélection des contextes les plus discriminants à l'aide du Cf-Itf. Nous avons évalué nos résultats avec la macro-précision et la MAP, en nous comparant à des références construites à partir de trois ressources.

Il ressort de ces expériences que pour l'utilisation des mesures de similarité sans sélection préalable des contextes les plus discriminants, les résultats sont meilleurs avec les noms qu'avec les termes. Pour les noms, une fenêtre de 21 et la mesure de Jaccard offre les meilleurs résultats. De plus, si le corpus est de petite taille (pour nous, environ 400 000 mots) il est préférable de pondérer cette mesure avec la fréquence relative. Pour les termes, si le corpus est de taille importante pour un corpus de spécialité (environ 3 millions de mots), la mesure de Jaccard pondérée avec la fréquence relative est le meilleur choix. En revanche, pour les plus petites fréquences, les résultats sont peu probants et des études doivent encore être réalisées. Ainsi, la sélection des contextes les plus discriminants à l'aide du Cf-Itf, avant le calcul de similarité offre de meilleurs résultats pour les termes que pour les noms. Même si avec les termes les valeurs de précision sont très faibles, l'évaluation avec la MAP permet d'identifier que de bonnes relations sont présentes mais mal classées. Une modification de la méthode de sélection des contextes devrait améliorer les résultats. Les premières expériences sur d'autres corpus, notamment en anglais, montrent que la méthodologie proposée est généralisable, mais à condition d'adapter les valeurs des paramètres au corpus traité. Nous avons comme perspective principale l'amélioration de l'extraction de relations pour les termes avec des petits corpus de spécialité. Pour cela, nous envisageons dans un premier temps de renouveler nos expériences avec le corpus médical Menelas (Zweigenbaum, 1994), de taille comparable à notre corpus Guides alimentaires, mais surtout d'analyser plus en détail le comportement de la sélection des contextes les plus discriminants. Nous modifierons ensuite notre calcul actuel du Cf-Itf en fonction de ces analyses.

Références

Aubin S. et Hamon T. (2006). Improving term extraction with terminological resources. *In Advances in Natural Language Processing*, number 4139 in LNAI, pages 380–387. Springer.

Brown P., deSouza P., Mercer R., Della Pietra V., et Lai J. (1992). Class-based n-gram models of natural language. *Comput. Linguist*., 18(4):467–479.

Church K. et Hanks P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29.

Curran J. et Moens M. (2002a). Scaling context space. *Proc. of ACL 2002*, pages 231–238.

Curran J. et Moens M. (2002b). Improvements in automatic thesaurus extraction. *Workshop on Unsupervised lexical acquisition*, volume 9, pages 59–66, Morristown, NJ, USA.

Curran J.-R. (2004). From distributional to semantic similarity. Ph.D. thesis, Institute for Communicating and Collaborative Systems School of Informatics University of Edinburgh.

Généreux M. et Hamon T. (2013). Experiments in synonymy: term extraction and mapping to concepts. *Terminologie et Intelligence artificielle (TIA)*, Paris.

Hamon T., Nazarenko A., Poibeau T., Aubin S. et 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.

Hindle D. (1990). Noun classification from predicate-argument structures. *Proc. of the 28th Annual Meeting of ACL*, pages 268–275, Pittsburgh, Pennsylvania, USA.

Jones K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–20.

Kilgarriff A. et Yallop C. (2000). What's in a thesaurus. *Proc. of LREC 2000*, pp. 1371–1379.

Lin D. (1998a). An information-theoretic definition of similarity. *Proc. of the 15th International Conference on Machine Learning*, pages 296–304. Morgan Kaufmann.

Lin D. (1998b). Automatic retrieval and clustering of similar words. In *Proc. Of ACL*, pages 768.774.

Manning C. et Schütze H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Morlane-Hondère F. (2013). Une approche linguistique de l'évaluation des ressources extraites par analyse distributionnelle automatique. Ph.D. thesis, Université de Toulouse.

Padó S. et Lapata M. (2007). Dependency-based construction of semantic space models. *Comput. Linguist.*, 33(2):161–199.

Panchenko A. et Morozova O. (2012). A study of hybrid similarity measures for semantic relation extraction. *Proc. of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pp. 10–18, Avignon, France.

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

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

Schütze H. (1998). Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123.

Van der Plas L. et Bouma G. (2004). Syntactic contexts for finding semantically related words. Ton van der Wouden, Michaela Poß, Hilke Reckman, et Crit Cremers, editors, *Proc. of CLIN Meeting.*

Van der Plas L. et Tiedemann J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. *Proc. of COLING'06*, pp. 866–873, Stroudsburg, PA, USA.

Weeds J., Weir D. et Mc Carthy D. (2004). Characterising measures of lexical distributional similarity. *Proc. of COLING'2004*, Stroudsburg, PA, USA.

Weeds J. (2003). Measures and Applications of Lexical Distributional Similarity. Ph.D. thesis, Department of Informatics, University of Sussex.

Wilks D., Mcdonald J.E., Plate T. et Slator B.M. (1990). Providing machine tractable dictionary tools. *Journal of Machine Translation*, 2.

Zweigenbaum P. (1994). Menelas : an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, pp. 117-120.