TALN 2010, Montréal, 19–23 juillet 2010

Comparaison de ressources lexicales pour l'extraction de synonymes

Philippe {Muller[(1)], Langlais[(2)]}
(1) IRIT, Université de Toulouse & Alpage, INRIA
(2) DIRO, Université de Montréal

1 Introduction

La construction automatique de thesaurus par collecte d'un ensemble de relations entre unités lexicales (synonymie, antonymie, méronymie, etc) est un objectif relativement ancien en traitement automatique des langues. Il est parfois étendu par l'ajout d'associations thématiques, aux contours variables. Les tentatives initiales étaient fondées soit sur des dictionnaires numériques, soit sur des analyses distributionnelles rapprochant des termes ayant des contextes de cooccurrence similaires, soit les deux (Niwa & Nitta, 1994), prenant éventuellement en compte les fonctions syntaxiques (Lin, 1998). De nombreuses instances de ces idées ont été proposées plus récemment, que ce soit en exploitant les structures de dictionnaires ou des données distributionnelles. Quand ils sont évalués, ces efforts se révèlent plutôt décevants : sauf à faire des restrictions importantes, les similarités ainsi définies rapportent un mélange hétérogène de fonctions lexicales et de termes sémantiquement apparentés sans que les contours de cette parenté soient évidents à délimiter. La synonymie est sans doute la fonction lexicale la plus testée parmi ces tentatives, car on dispose de références qui permettent de l'évaluer, jusqu'à un certain point. Le bilan à tirer de ces travaux est que les données utilisées pour extraire des relations lexicales n'ont sans doute pas livré tout le potentiel que les auteurs leur attribuent et que l'on connaît encore mal la place réelle occupée par les termes en relations de synonymie parmi les associations calculées.

Il a été aussi proposé, de façon plus marginale, d'utiliser des corpus parallèles multilingues pour retrouver une similarité entre termes (Dyvik, 2002) en se fondant sur l'hypothèse que des termes proches doivent partager largement leurs traductions. Dans cette optique, on considère les termes associés par des traductions en "miroir" : les traductions d'un terme d'une langue source 1 dans une langue cible 2 sont mis en rapport avec leur traduction dans la langue 1. L'hypothèse est que les termes obtenus après cet aller-retour 1-2-1 sont des bons candidats à la synonymie avec le terme de départ.

Notre but ici est d'évaluer la présence de la synonymie dans des données d'alignement et des données distributionnelles pour le français, en prenant en compte de façon conjointe les phénomènes de fréquence des termes dans l'une ou l'autre de ces ressources. L'une des originalités de cette étude par rapport aux travaux mentionnés est d'utiliser et d'évaluer un score d'association en miroir entre termes (Dyvik, 2002), quand les efforts d'évaluation existants se sont concentrés uniquement sur des similarités de vecteurs d'alignements.

La suite de l'article présente les ressources utilisées (section 2), les expérimentations menées (section 3) et les premières analyses que nous en tirons (section 4). Nous revenons enfin à une discussion des approches comparables à la nôtre en section 5.

2 Protocole

Nous tentons de comparer l'aptitude d'une approche distributionnelle et d'une approche miroir à identifier des synonymes. Nous avons pour cela sélectionné un nombre arbitraire de substantifs et de verbes (4000 de chaque environ) vérifiant des seuils de fréquence minimale afin d'éviter les termes trop spécifiques. Les fréquences dans cette étude ont été calculées à partir de la version francophone de Wikipédia[1], soit environ 200 millions de mots. Les substantifs étant plus nombreux dans notre référence, nous les avons échantillonnés avec deux seuils différents de fréquence minimale, fixés arbitrairement à 100 et 1000.

Les deux approches (distributionnelle et miroir) produisent pour chacun de ces termes "cibles" une liste de termes candidats synonymes, ordonnés selon un score indiquant leur degré d'association (voir la figure 1). Ces listes sont évaluées par leur aptitude à identifier les synonymes d'un lexique de référence. Nous calculons donc les taux de précision/rappel/f-mesure, soit des n meilleurs candidats dans la liste (en faisant varier n), soit en gardant tous les candidats dont le score dépasse un certain seuil (en faisant varier le seuil).

avoir (0,046), **consommer** (0,039), être (0,032), *nourrir* (0,020), **gruger** (0,007), aller (0,007), faire (0,006), **prendre** (0,005), **dévorer** (0,005), **absorber** (0,005), **dîner** (0,005), déposer (0,005), **déjeuner** (0,004), **alimenter** (0,003), servir (0,003), **ronger** (0,003), **avaler** (0,003), **engloutir** (0,003), ...

FIG. 1 – Verbes candidats à la synonymie produits par l'approche miroir pour le verbe *manger*. Ceux en gras appartiennent à `DicoSyn`, celui en italique y apparaît sous forme pronominale (se nourrir).

Notre lexique de référence est l'union des dictionnaires de synonymes que constitue la ressource électronique `DicoSyn`, initiée par Ploux & Victorri (1998) à partir de sept dictionnaires classiques[2]. Ces dictionnaires sont assez hétérogènes ; ainsi les trois plus gros éléments de cet ensemble qui ont une taille comparable (Du Chazaud, Robert, Larousse) ont un taux de recouvrement moyen assez faible (entre 0,42 et 0,55 de f-score si on les compare deux à deux). La ressource suffit cependant à un objectif de comparaison de différentes configurations.

La fréquence moyenne des verbes (calculée sur Wikipedia), des noms moyennement fréquents et fréquents est respectivement de 2500, 4300 et 9700, alors le que le nombre moyen de synonymes dans la référence est de 26, 12 et 15, respectivement, avec un maximum d'environ 180, hors stop-words (*abri* pour les noms, *battre* pour les verbes).

Comme évaluation alternative, on peut aussi envisager des tests comme ceux du TOEFL où la tâche consiste à distinguer, parmi plusieurs candidats, un synonyme d'un mot donné dans une phrase exemple. La synonymie peut être aussi testée au coup par coup par des lexicographes (Falk *et al.*, 2009).

3 Approches

Nous avons utilisé deux types de ressources pour l'expérimentation : un corpus parallèle pour calculer les associations d'alignement, et des données distributionnelles collectées à partir de Wikipedia.

Ressources distributionnelles : La base de "voisins" distributionnels utilisée a été générée à partir du corpus de l'ensemble des articles de la version francophone de Wikipédia. Elle a été constituée en suivant l'approche de Lin (1998) : un analyseur syntaxique fournit des comptes de triplets syntaxiques (gouverneur,

[1]Les données distributionnelles, ainsi que tout ce qui est extrait de Wikipedia, nous ont été fournies par le laboratoire CLLE-ERSS, où elles ont été collectées par Frank Sajous.

[2]La ressource a été de plus catégorisée en verbes/noms/adjectifs par l'équipe CLLE-ERSS.

relation, dépendant) à un module d'analyse développé par Bourigault (2002) porté sur Wikipedia par Frank Sajous. Ce module calcule la similarité de Lin entre des couples de "prédicats" (gouverneur, relation), ou bien entre des couples de dépendants syntaxiques ("arguments"). Si on considère un prédicat X et un argument Y, Lin introduit une notion de quantité d'information associée à X ou Y, noté $q_{info}(X)$, qui est le logarithme du nombre d'arguments différents du prédicat X rapporté au nombre total d'arguments possibles. Un score intermédiaire d'information mutuelle entre deux prédicats est alors donné par la somme des quantités d'informations des arguments qu'ils partagent et le score d'association de Lin est défini en normalisant par rapport aux sommes de q_{info} de tous les arguments des deux prédicats. Certains seuils étant appliqués aux différentes étapes du traitement, chaque item lexical possède un nombre de voisins variables, qui peut être nul. Le vocabulaire couvert par les voisins de nos termes cibles représente 25% des verbes de la référence (hors locutions), et 18-19% des noms selon la restriction fréquentielle. L'ensemble des paires cibles-candidats couvre respectivement 24, 25 et 29% des paires recensées par `DicoSyn` pour chaque catégorie de cible. Il est évident que les choix du corpus, de la mesure de similarité choisie et de l'analyseur syntaxique ont une influence en bout de chaîne et seront évalués à terme.

Ressources bilingues : Nous avons mis à profit une des bases de traductions de l'application `TSRali` (Bourdaillet *et al.*, 2010). Cette base contient 8.3 millions de paires de phrases des débats parlementaires canadiens alignés au niveau des phrases. Ces textes ont ensuite été lemmatisés à l'aide de `TreeTager`[3]. La boîte à outil `Giza++`[4] a enfin été utilisée afin d'entraîner un modèle de traduction statistique. Les distributions lexicales (IBM4) p_{s2t} et p_{t2s} des modèles obtenus en changeant la langue considérée comme source au moment de l'entraînement[5] ont ensuite été utilisées de manière à réaliser une approche miroir. Formellement, nous calculons pour chaque mot de test w :

$$p(s|w) \approx \sum_{t \in \tau_{s2t}(w)} p_{s2t}(t|w) \times p_{t2s}(s|t)$$

pour tout mot de la langue source s atteignable depuis w en utilisant la langue cible comme pivot ($\tau_{s2t}(w)$ désigne l'ensemble des mots associés à w dans le modèle p_{s2t}). En pratique, deux seuils (source et cible) contrôlent le bruit présent dans les distributions lexicales.

Cette ressource concerne 93 458 (resp. 103 770) formes anglaise (resp. françaises). L'ensemble des termes "proposés" par la ressource représente 70% des verbes mentionnés dans la référence `DicoSyn`(hors locutions), 40% des noms (F>100) et 44% des noms (F>1000). Tous les verbes proposés sont présents dans `DicoSyn`, alors qu'environ 20% des noms n'apparaissent pas dans cette référence. L'ensemble des paires cibles-candidats couvre respectivement 50, 40 et 43% des paires recensées par `DicoSyn` pour chaque catégorie de cible. Au vu de ces statistiques, nous avons restreint les évaluations au vocabulaire couvert en commun par les ressources et la référence.

4 Expérimentations et résultats

Sur le principe des expériences présentées précédemment (étude des n meilleurs candidats, ou passant un seuil variable), nous avons étudié l'influence des paramètres suivants : catégorie syntaxique (noms, verbes) classe de fréquence des termes cibles (pour les noms, deux classes de fréquence "forte" et "moyenne à forte"), fréquence minimale des termes candidats (en faisant varier ce seuil de 0 à 5000 par palliers

[3] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[4] `http://fjoch.com/GIZA++.html`

[5] Les modèles IBM ne sont pas symétriques.

exponentiels) et influence d'un filtre de mots tabous sur les noms ("stop words") dans le cas des données d'alignement[6]. Pour les verbes nous avons enlevé les 15 items les plus fréquents dans tous les cas. Dans le cas des n meilleurs candidats, nous avons ajouté un cas où un oracle donne à la méthode le nombre de synonymes de la référence pour choisir le n adapté à chaque terme. Enfin nous avons étudié la combinaison des ressources en testant l'intersection des candidats proposés par les deux méthodes.

La figure 2 illustre pour une fréquence de filtre donnée l'évolution du F-score des candidats proposés par la méthode des voisins et des miroirs, en faisant varier le nombre de termes candidats retenus (gauche), ou le seuil de score (droite). Dans ce dernier cas, même si les scores d'association des termes ne sont pas comparables, on voit nettement la dominance de la méthode miroir en observant les maxima des deux courbes dans ce cas précis.

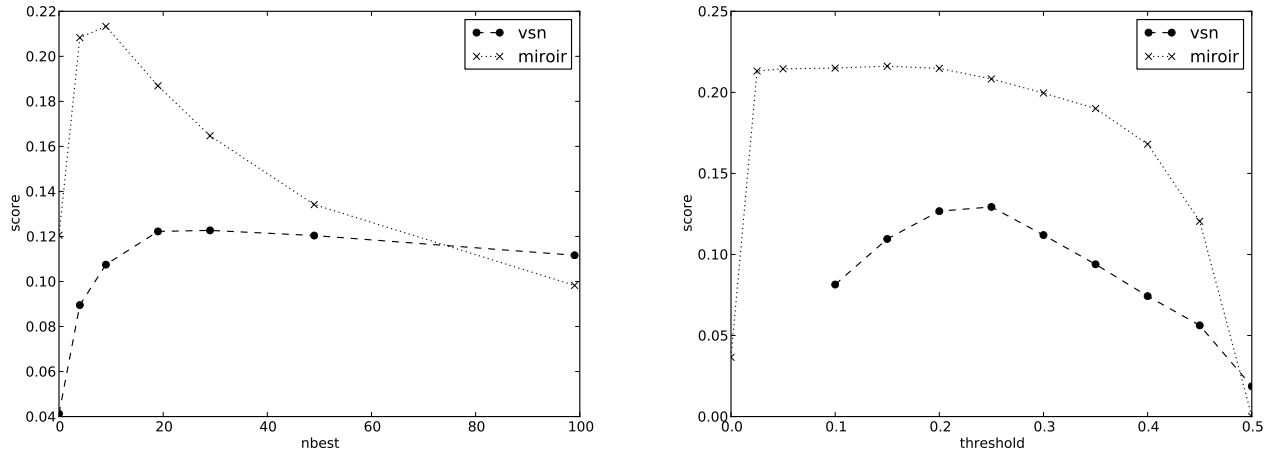

FIG. 2 – Evolution du F-score moyen pour les verbes en fonction des n candidats gardés (gauche) et d'un seuil minimal d'association (droite). Un filtre de fréquence>1000 des candidats a été appliqué.

La table 1 détaille les résultats en fonction des n premiers candidats considérés, en fixant quelques valeurs des paramètres. Les scores sont meilleurs quand on filtre les candidats en fréquence, les scores évoluant de façon régulière logarithmiquement. Nous ne montrons ici que les valeurs mimale et maximale des paramètres de fréquence des candidats. On peut constater que la méthode du miroir est supérieure de façon à peu près uniforme à celle utilisant les voisins, surtout pour les candidats de fréquence plus élevée. L'élimination des mots tabous donne un gain de 2 ou 3% supplémentaires, en enlevant une source de bruit. Le résultat est similaire sur les types de cibles, noms moyennement ou très fréquents et verbes. Il faut noter que beaucoup des termes cibles n'ont pas de voisins (environ 60%), même sans filtrage. Un certain nombre de constatations peuvent être faites, dont nous ne pouvons présenter le détail par manque de place, mais que nous synthétisons ici. Pour les résultats séparés en précision et rappel, on retrouve la même supériorité de la méthode miroir, qui diminue quand n augmente (les rappels finissent par se confondre). Les meilleurs résultats sont sur les noms, avec à chaque fois 1 ou 2% de différence selon les configurations. Les scores considérés par seuil présentent des évolutions comparables, et ne font pas apparaître de valeur clé qui permettrait de définir une valeur générale de filtrage. Nous avons donc montré seulement ceux avec les n meilleurs candidats, plus faciles à interpréter. On peut noter aussi que combiner les deux ressources améliore encore un peu la fiabilité des termes proposés. Les scores en gras sont les maximums d'une ligne,

[6]Les termes présents dans trop de listes candidates sont éliminés, comme *avoir* ou *aller* dans l'exemple de la figure 1.

n		1	5	10	20	30	50	100	oracle
(N) voisins	freq=1	0,034	0,079	0,097	**0,106**	0,105	0,097	0,083	0,089
	freq=5000	0,081	0,145	**0,153**	0,145	0,133	0,117	0,092	0,123
(N) miroir	freq=1	0,059	0,130	0,148	**0,149**	0,139	0,122	0,093	0,139
	freq=5000	0,123	**0,201**	0,193	0,163	0,141	0,111	0,076	0,179
(N) combinaison m/vsn	freq=1	0,067	0,143	0,166	**0,172**	0,169	0,151	0,125	0,138
	freq=5000	0,146	0,231	**0,232**	0,214	0,193	0,163	0,127	0,186
(N) stop	freq=1	0,073	0,151	0,170	**0,171**	0,161	0,140	0,110	0,136
+ combinaison m/vsn	freq=5000	0,170	**0,248**	0,235	0,199	0,173	0,140	0,104	0,193
(V) voisins	freq=1	0,030	0,066	0,081	0,098	0,103	**0,106**	0,102	0,087
	freq=5000	0,047	0,103	0,123	**0,132**	0,130	0,126	0,117	0,097
(V) miroir	freq=1	0,063	0,143	0,168	**0,171**	0,162	0,142	0,111	0,161
	freq=5000	0,115	**0,211**	0,209	0,176	0,151	0,119	0,083	0,186
(V) combinaison m/vs	freq=1	0,060	0,154	0,188	**0,205**	0,205	0,193	0,163	0,079
	freq=5000	0,124	0,250	**0,273**	0,262	0,245	0,211	0,163	0,127

TAB. 1 – F-score moyen par mot, en gardant n candidats pour les verbes, et les noms cibles de fréquence>1000, et en faisant varier la fréquence minimale des candidats. L'oracle correspond à n=nombre de synonymes de la référence, pour chaque mot.

et les scores grisés sont les maximums d'une colonne pour chaque catégorie (N/V).

On observe que les scores sont assez bas. La meilleure méthode qui combine l'approche distributionnelle et l'approche miroir en éliminant les candidats à la synonymie dont la fréquence (dans Wikipédia) est inférieure à 5000 obtient un f-score de 0,273. Ces scores doivent donc plutôt être considérés comme une indication de la pertinence des ressources pour une tâche ultérieure de classification de paires synonymes. Notre objectif à terme est de replonger les termes dans leurs contextes d'emploi pour affiner cette première approche.

Finalement, nous observons que les scores augmentent systématiquement avec la fréquence minimale des termes candidats, de façon logarithmique, dans une plage de 10% environ. Ce phénomène est à peu près similaire entre les deux méthodes (voisins/miroir). Nous omettons le cas verbe+liste de stops, peu différent des autres.

5 Discussion

Nos différentes expérimentations montrent qu'aucune des deux approches décrites ne permet d'identifier seule des relations synonymiques avec fiabilité. Bien que de nombreux facteurs puissent être responsables de ce constat, nous observons cependant la supériorité de l'approche miroir. Les données d'alignement bilingue ont été les moins exploitées pour la recherche de synonymes. On peut citer quelques précurseurs expérimentaux, tels que (van der Plas & Tiedemann, 2006), qui comparent deux items avec une mesure de similarité entre leurs "vecteurs d'alignement" (la fréquence d'alignement d'un mot avec les autres mots du lexique) dans différentes langues, et (Wu & Zhou, 2003) qui ont tenté de combiner linéairement des classifieurs regroupant tous les types de ressources mentionnés : similarité d'alignement, de distribution syntaxique, et de proximité dans un graphe de dictionnaire. Les premiers rapportent des f-scores maximaux de 12%. Les seconds combinent plusieurs classifieurs intégrant des données distributionnelles et des

dictionnaires ; les tests portent sur des classes de termes de fréquence variable, la meilleure combinaison donnant 23% sur les noms et 30% sur les verbes. Les données d'alignement seules donnent respectivement 22 et 26%.

Les performances de notre approche miroir sont donc comparables, même si l'approche que nous décrivons est bien plus simple puisque les travaux sus-mentionnés doivent calculer entièrement la matrice $N \times N$ de similarité des alignements, où N est la taille du lexique, et où chaque terme est codé par un vecteur de tous les termes que l'on peut aligner avec lui.

À terme, notre objectif est d'évaluer la faisabilité de l'apprentissage d'une relation lexicale à partir de ces données, avec comme horizon la collecte d'un corpus complémentaire où les termes seraient évaluables en contexte. La complémentarité de ces ressources est aussi une question ouverte. Le principal obstacle théorique à ce genre d'approche est lié à la polysémie des termes, qu'elles ne peuvent guère distinguer, et à la variabilité en fréquence des emplois différents. C'est pourquoi nous avons aussi analysé le rôle de la fréquence dans les résultats. D'une part on peut supposer que les mots peu fréquents ne fourniront pas des données fiables, et d'autre part qu'il sera plus difficile de discriminer les différentes fonctions des mots très fréquents.

Références

BOURDAILLET J., HUET S., LANGLAIS P. & LAPALME G. (2010). TransSearch : from a bilingual concordancer to a translation finder. *Mach. Transl.*, p. 35 pages. To appear.

BOURIGAULT D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9ieme conférence sur le Traitement Automatique de la Langue Naturelle*, p. 75–84, Nancy.

DYVIK H. (2002). Translations as semantic mirrors : From parallel corpus to wordnet. In *The Theory and Use of English Language Corpora, ICAME 2002*. http ://www.hf.uib.no/i/LiLi/SLF/Dyvik/ICAMEpaper.pdf.

FALK I., GARDENT C., JACQUEY E. & VENANT F. (2009). Sens, synonymes et définitions. In *Conférence sur le Traitement Automatique du Langage Naturel - TALN'2009*, Senlis France.

LIN D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL '98*, volume 2, p. 768–774, Montreal.

NIWA Y. & NITTA Y. (1994). Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of Coling 1994*.

PLOUX S. & VICTORRI B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement automatique des langues*, **39**(1), 161–182.

VAN DER PLAS L. & TIEDEMANN J. (2006). Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, p. 866–873, Sydney, Australia : Association for Computational Linguistics.

WU H. & ZHOU M. (2003). Optimizing synonyms extraction with mono and bilingual resources. In *Proceedings of the Second International Workshop on Paraphrasing*, Sapporo, Japan : Association for Computational Linguistics.