TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Comparaison d'une approche miroir et d'une approche distributionnelle pour l'extraction de mots sémantiquement reliés

Philippe Muller[1, 2] Philippe Langlais[3]
(1) IRIT, Université Paul Sabatier
(2) Alpage, INRIA Paris-Rocquencourt
(3) RALI / DIRO / Université de Montréal
muller@irit.fr, felipe@iro.umontreal.ca

Résumé. Dans (Muller & Langlais, 2010), nous avons comparé une approche distributionnelle et une variante de l'approche miroir proposée par Dyvik (2002) sur une tâche d'extraction de synonymes à partir d'un corpus en français. Nous présentons ici une analyse plus fine des relations extraites automatiquement en nous intéressant cette fois-ci à la langue anglaise pour laquelle de plus amples ressources sont disponibles. Différentes façons d'évaluer notre approche corroborent le fait que l'approche miroir se comporte globalement mieux que l'approche distributionnelle décrite dans (Lin, 1998), une approche de référence dans le domaine.

Abstract. In (Muller & Langlais, 2010), we compared a distributional approach to a variant of the mirror approach described by Dyvik (2002) on a task of synonym extraction. This was conducted on a corpus of the French language. In the present work, we propose a more precise and systematic evaluation of the relations extracted by a mirror and a distributional approaches. This evaluation is conducted on the English language for which widespread resources are available. All the evaluations we conducted in this study concur to the observation that our mirror approach globally outperforms the distributional one described by Lin (1998), which we believe to be a fair reference in the domain.

Mots-clés : Sémantique lexicale, similarité distributionnelle, similarité traductionnelle.

Keywords: Lexical Semantics, distributional similarity, mirror approach.

1 Introduction

Collecter les relations entre les entités lexicales en vue de construire ou de consolider un thésaurus est une activité qui possède une longue tradition en traitement des langues. Les efforts les plus importants ont été dédiés à la recherche de synonymes, ou plus exactement des "quasi-synonymes" (Edmonds & Hirst, 2002), c'est-à-dire des entrées lexicales ayant un sens similaire dans un contexte donné. D'autres relations comme l'antonymie, l'hyperonymie, l'hyponymie, la méronymie ou l'holonymie ont également été étudiées. Certains thésaurus, comme Moby que nous utilisons ici, listent de plus des relations qui sont difficiles à caractériser.

De nombreuses ressources ont été utilisées pour parvenir à acquérir de tels thésaurus. Les dictionnaires électroniques ont tout d'abord été investis, soit pour en extraire des relations sémantiques au niveau lexical (Michiels & Noel, 1982), soit pour définir des mesures de similarité sémantiques entre les entités lexicales (Kozima & Furugori, 1993). L'analyse distributionnelle, qui compare les mots à travers leur contexte d'usage, est également une ressource populaire pour la réalisation d'une mesure de similarité sémantique (Niwa & Nitta, 1994; Lin, 1998).

Plusieurs approches ont montré l'intérêt d'utiliser des corpus dans plusieurs langues et plus particulièrement des corpus parallèles. Dans ces travaux, les entrées lexicales sont dites similaires lorsqu'elles sont alignées avec les mêmes traductions dans une autre langue (van der Plas & Tiedemann, 2006; Wu & Zhou, 2003). Une variante de ce principe proposée par Dyvik (2002) considère comme sémantiquement reliés les mots d'une langue qui sont traduction d'un même mot dans une autre langue ; ces mots sont appelés par l'auteur des *traductions miroir*. Des variantes de cette approche ont été étudiées pour l'acquisition de paraphrases, qui porte sur des associations d'expressions de plusieurs mots : voir par exemple (Bannard & Callison-Burch, 2005) et (Max & Zock, 2008).

Les évaluations des travaux à base de similarité lexicale sont souvent décevantes : différents types de relations lexicales sont typiquement identifiés, qu'il est difficile de distinguer automatiquement. Des travaux comme ceux

de Zhitomirsky-Geffet & Dagan (2009) tentent dans une étape de post-traitement de sélectionner les relations les plus pertinentes qui caractérisent des paires de mots similaires. D'autres, comme Wu & Zhou (2003) tentent de combiner le résultat de différents processus d'extraction de mots reliés (approche distributionnelle, dictionnaires, etc.).

Notre étude s'inscrit dans ce dernier courant. Nous poursuivons l'étude amorcée par Muller & Langlais (2010) où une variante de Dyvik, faisant usage de modèles de traduction statistiques entraînés sur de grands volumes de données, est combinée à une approche distributionnelle. Contrairement à ce travail, nous nous intéressons ici à la langue anglaise pour laquelle des ressources sont disponibles en plus grand nombre. Ceci nous permet de mener une évaluation à l'état de l'art de l'approche miroir, que nous comparons au thésaurus produit par l'approche décrite dans (Lin, 1998) et que Lin tient à la disposition de la communauté. Nous montrons que l'approche miroir se comporte favorablement par rapport à l'approche distributionnelle, et ce, selon différentes évaluations que nous avons menées.

Dans la suite de cet article, nous présentons les ressources mises à profit en section 2 et notre protocole expérimental en section 3. Nous analysons nos résultats en section 4 et discutons les travaux reliés en section 5. Nous concluons cette étude et en dressons les perspectives en section 6.

2 Ressources

Nous avons utilisé deux bases lexicales dans ce travail :
– La base lexicale WordNet [1] que nous interrogeons à travers l'API de NLTK [2]. WordNet encode les relations de synonymie (`gain` / `acquire`), d'antonymie (`gain` / `lose`), d'hyperonymie/hyponymie (`odor` / `stench`) et d'holonymie/méronymie (`wood` / `tree`). Chaque entrée lexicale dans WordNet possède une moyenne de 5 à 6 synonymes et de 8 à 10 termes reliés, toutes relations confondues.
– Le thésaurus Moby [3] est une ressource plus étoffée que WordNet : chaque mot dispose en effet d'environ 80 mots reliés en moyenne. La nature des relations n'est cependant pas annotée.

Afin de comparer les approches miroir et distributionnelle, nous avons sélectionné de manière aléatoire deux ensembles de 1000 mots, un pour les noms et un pour les verbes. Nous appelons ces mots les "cibles" dans la suite. Nous avons imposé arbitrairement un seuil minimal de fréquence sur les mots cibles (> 1000). La fréquence des mots a été calculée à l'aide du corpus libre de droit Wacky [4], qui compte 2 milliards de mots. Les caractéristiques des deux ensembles de cibles ainsi construits sont décrites en table 1.

			nombre d'associations			
Pos	fréquence médiane	référence	moyen	médian	min	max
Noms	3 538	WordNet syns	3,63	2	1	36
Noms	3 538	Moby	73,87	57	3	509
Verbes	11 136	WordNet syns	5,57	4	1	47
Verbes	11 136	Moby	113,23	90	6	499

TABLE 1 – Caractéristiques des deux ensembles de cibles (noms et verbes) : fréquence médiane dans Wacky, nombre moyen de termes associés selon la référence spécifiée, nombre médian, minimum et maximum.

3 Protocole

Nous comparons les termes similaires produits soit par l'approche des miroirs (section 3.1), soit par l'approche distributionnelle (section 3.2). Chaque approche produit un ensemble de termes associés ou *candidats*, classés selon leur degré de similarité. Ces candidats ordonnés sont alors évalués au regard d'une ressource de référence

1. `wordnet.princeton.edu/wordnet`
2. `www.nltk.org`
3. `www.gutenberg.org/dirs/etext02/mthes10.zip`
4. `http://wacky.sslmit.unibo.it/doku.php`

(WordNet ou Moby), soit en gardant les n-meilleurs candidats, soit en gardant ceux dont le score de similarité dépasse un certain seuil (voir les détails plus loin).

À titre d'exemple, la figure 1 montre les candidats proposés par les deux approches pour le mot cible choisi aléatoirement `groundwork`. On observe la grande différence de couverture de WordNet et de Moby.

Candidats Miroir	Candidats Lin	WordNet	Moby
base	**preparation**	base	arrangement
basis	framework	basis	**base**
foundation	timetable	cornerstone	basement
land	rationale	foot	**basis**
ground	impetus	fundament	bed
job	modality	foundation	bedding
field	**foundation**	substructure	bedrock
plan	**prerequisite**	understructure	bottom
force	precondition		briefing
development	blueprint		cornerstone
			... [47 de plus]

FIGURE 1 – Dix premiers candidats proposés par les approches miroirs et distributionnelles pour le mot cible `groundwork`. Les synonymes selon WordNet ainsi qu'un sous ensemble des mots reliés selon Moby sont indiqués. Les candidats soulignés appartiennent à WordNet, tandis que ceux en gras sont présents dans Moby.

3.1 Approche miroir

L'approche miroir est fondée sur l'hypothèse que des mots d'une langue $\mathcal{E}$ qui sont fréquemment alignés avec le même mot dans une autre langue $\mathcal{F}$ sont sémantiquement proches. Dans l'exemple de la figure 2, les mots français `manger` et `consommer` sont tous les deux alignés avec le mot anglais `eat` et sont donc candidats à l'appariement sémantique.

Un	bébé	mange	toutes	les deux heures.
	Babies	eat	every	two hours.
	Canadians	eat	too much	poutine.
Les	Canadiens	consomment	trop de	poutine.

FIGURE 2 – Exemple de traductions miroir.

Notre variante de l'approche miroir repose sur la consultation de deux modèles de traduction statistique p_{e2f} et p_{f2e} qui donnent respectivement la probabilité qu'un mot français soit la traduction d'un mot anglais et la probabilité inverse. Nous calculons la vraisemblance qu'un mot anglais s (pour synonyme) soit relié sémantiquement à un mot anglais w, soit $p(\mathrm{s}|\mathrm{w})$:

$$p(\mathrm{s}|\mathrm{w}) \approx \sum_{\mathrm{f} \in \tau_{e2f}(\mathrm{w})} p^{\delta_1}_{e2f}(\mathrm{f}|\mathrm{w}) \times p^{\delta_2}_{f2e}(\mathrm{s}|\mathrm{f}) \qquad \tau_{e2f}(\mathrm{w}) = \{\mathrm{f} : p_{e2f}(\mathrm{f}|\mathrm{w}) > 0\}$$

Ici $\tau_{e2f}(\mathrm{w})$ désigne l'ensemble des mots français associés par le modèle p_{e2f} au mot anglais w. En pratique, les distributions lexicales utilisées étant bruitées, nous appliquons deux seuils δ_1 et δ_2 (fixés à 0.001 dans cette expérience) qui filtrent les associations peu probables d'un modèle :

$$p^{\delta}_{\bullet}(t|s) = \begin{cases} p_{\bullet}(\mathrm{t}|\mathrm{s}) & \text{si } p_{\bullet}(\mathrm{t}|\mathrm{s}) \geq \delta \\ 0 & \text{sinon} \end{cases}$$

D'autres façons de filtrer les tables de transfert pourraient être déployées. Nous pourrions par exemple utiliser un test de significativité afin de retenir les associations les plus pertinentes. Notre approche au filtrage est certainement perfectible mais présente l'avantage d'être particulièrement simple à mettre en œuvre.

Les modèles lexicaux p_{e2f} et p_{f2e} ont été entraînés sur un bitexte anglais-français de 8,3 millions de paires de phrases extraites des transcriptions des débats parlementaires canadiens (Hansard). Ce bitexte est exploité par le concordancier bilingue `TSRali` [5]. Nous avons lemmatisé les phrases anglaises et françaises du corpus à l'aide de `TreeTager` [6] avant d'entraîner dans les deux directions [7] (anglais→français et français→anglais) des modèles IBM 4 à l'aide de `Giza++` [8] utilisé dans sa configuration par défaut.

Dans l'évaluation qui suit, nous avons considéré les 200 premiers lemmes associés à chaque mot cible par cette approche car c'est le nombre de candidats que produit l'approche distributionnelle que nous avons testée (voir la section suivante).

3.2 Similarité distributionnelle

L'approche distributionnelle que nous utilisons est celle décrite par Lin (1998). Elle représente selon nous une approche de référence dans le domaine. Un thésaurus calculé par l'auteur à l'aide de cette méthode est disponible gratuitement [9].

Pour l'obtenir, Lin a fait usage d'un analyseur grammatical en dépendance afin de comptabiliser les occurrences de triplets `(lemme_de_tête, relation, lemme_dépendant)` où `relation` est une relation (syntaxique) de dépendance. À chaque lemme w est associé un vecteur de compte pour l'ensemble $F(w)$ des traits `(rel, autre_lemme)` où `autre_lemme` est soit un dépendant de w, soit un gouverneur de w.

Par exemple, le verbe `eat` est caractérisé par un ensemble de traits $F(eat)$ contenant `(has_subj,man)`, `(has_obj,fries)`, `(has_obj,pie)`, etc qui correspondent aux contextes syntaxiques de `eat`. Appelons c la fonction de comptage d'occurrence d'un triplet (w, rel, w') et V l'ensemble du vocabulaire, on pose :

$$c(\_, rel, w) = \sum_{w' \in V} c(w', rel, w) \qquad I(w, rel, w') = \log \frac{c(w, rel, w') \times c(\_, rel, \_)}{c(w, rel, \_) \times c(\_, rel, w')}$$

$$c(w, rel, \_) = \sum_{w' \in V} c(w, rel, w')$$

$$c(\_, rel, \_) = \sum_{w' \in V} c(\_, rel, w') \qquad ||w|| = \sum_{(r,w') \in F(w)} I(w, r, w')$$

$I(w, rel, w')$ est alors la spécificité d'une relation (w, rel, w'), définie comme l'information mutuelle entre les éléments du triplet (Lin, 1998). On note $||w||$ la quantité d'information totale associée à w. La similarité entre deux lemmes w_1 et w_2 mesure alors à quel point ils partagent des contextes syntaxiques spécifiques, en utilisant la quantité d'information des contextes qu'ils partagent, normalisée par la quantité d'information totale qu'on peut leur associer séparément.

$$sim(w_1, w_2) = \frac{\sum_{(r,w) \in F(w_1) \cap F(w_2)} [I(w_1, r, w) + I(w_2, r, w)]}{||w_1|| + ||w_2||}$$

D'après (Lin *et al.*, 2003), le corpus utilisé pour obtenir le thésaurus que nous avons utilisé ici serait de 3 milliards de mots, c'est-à-dire plus de 10 fois la taille du corpus que nous avons utilisé pour développer l'approche miroir. Il nous est apparu préférable de prendre ce thésaurus plutôt que de tester notre implémentation de l'approche que nous venons de décrire en particulier parce que nous ne disposons pas d'information sur les réglages des paramètres utilisés par les auteurs pour optimiser leurs sorties. En fait, notre implémentation de l'approche est de moins bonne qualité sur les jeux de tests que nous présentons que ceux obtenus à l'aide du thésaurus compilé par les auteurs. À tout le moins, nous soulignons que la comparaison de l'approche miroir avec l'approche distributionnelle n'est pas biaisée en faveur de l'approche miroir.

Le thésaurus calculé par Lin présente pour chaque mot cible les 200 lemmes les plus proches au sens de cette mesure de similarité.

5. `http://www.tsrali.com/`
6. `www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
7. Les modèles IBM ne sont pas symétriques.
8. `fjoch.com/GIZA++.html`
9. `webdocs.cs.ualberta.ca/~lindek/Downloads/sim.tgz`

4 Expériences

En suivant le protocole décrit plus haut, nous avons évalué la sortie des deux similarités (miroir et distributionnelle) en considérant soit les n-meilleurs candidats de chaque approche, soit en considérant ceux dont le score de similarité dépasse un seuil donné (que nous faisons varier). Nous avons séparé notre jeu de test en deux ensembles de manière à mesurer les différences entre les noms et les verbes : comme le montre la table 1, le nombre de synonymes et autres entités lexicales reliées varie fortement en fonction de la catégorie morpho-syntaxique.

Nous avons considéré lors de l'évaluation les seuls items communs à la référence et au lexique de la ressource utilisée pour le développement d'une mesure de similarité. Par exemple, WordNet contient des synonymes qui ne sont pas présents dans les Hansards que nous avons mis à profit pour développer l'approche miroir : ils sont simplement écartés de notre évaluation.

Les deux approches que nous comparons sont sensibles à la fréquence des mots cibles considérés. Dans les deux approches décrites, tous les sens d'un mot sont regroupés lors des calculs de la similarité et il est vraisemblable que les usages les plus fréquents dominent les autres dans ces calculs. Sachant qu'un mot fréquent a plus de chance d'être polysémique qu'un autre, nous souhaitions prendre en considération dans notre évaluation l'influence de la fréquence des mots cibles étudiés. Nous filtrons à cet effet les candidats dont la fréquence (telle que calculée à l'aide de Wacky) est inférieure à un seuil donné, pour un ensemble de ces valeurs seuils.

Il a été montré (Weeds, 2003) que la plupart des méthodes de similarité lexicale se comportent de façon très différente par rapport à ce critère, sélectionnant selon les réglages plutôt des mots fréquents ou plutôt des mots rares.

Nous avons remarqué la tendance de l'approche miroir à souvent proposer des mots "vides". Cela s'explique par le fait que ces mots sont souvent biens notés par les distributions lexicales que nous utilisons. Ce phénomène a été analysé notamment dans (Moore, 2004). Nous avons arbitrairement éliminé des listes de candidats miroirs les termes apparus dans plus de 25% des listes (ce seuil pourrait être ajusté à l'aide d'un corpus de développement). Ce filtre élimine des noms courants comme `thing` ou `way`, des verbes comme `have`, `be` ou `come`, ainsi que des mots sur-représentés dans les Hansards (ex. : `house`).

Au final, nous avons combiné les listes candidates produites par les deux approches en prenant l'intersection des deux listes. D'autres schémas de combinaison seront étudiés dans des travaux ultérieurs.

Deux aspects nous intéressent plus particulièrement dans cette expérience : la quantité de mots reliés dans la référence que nous sommes capables d'identifier par l'une des approches et la fiabilité avec laquelle ils sont identifiés. En d'autres termes, nous voulons que la tête de liste des candidats soit la meilleure possible au regard d'une liste de référence. Nous évaluons donc les deux approches à l'aide des taux de précision et de rappel [10] que nous mesurons en différents points. Nous résumons ces taux à l'aide des taux MAP (Mean Average Precision) et MRR (Mean Reciprocal Rank) couramment employés en recherche d'information. MAP calcule la précision en chaque point de la liste où un candidat pertinent est identifié ; MRR est calculé comme la moyenne de l'inverse des rangs du premier terme pertinent dans la liste.

Enfin, nous avons également calculé la précision de chaque approche en faisant l'hypothèse d'un "oracle" qui indique le nombre exact de candidats à proposer pour chaque mot cible (il s'agit dans notre cas du nombre de mots reliés dans la référence). Cette mesure est semblable à ce que l'on appelle la R-précision. Par exemple, les 10 candidats de la méthode des miroirs de la figure 1, évalués à l'aide de la référence WordNet, reçoivent une précision de 3/10, un rappel de 3/5 (et pas 3/8 car les mots `understructure`, `substructure` et `fundament` sont absents des Hansards). La R-précision est également de 3/5 car tous les candidats corrects sont proposés à un rang inférieur au nombre de mots reliés dans la référence (5 synonymes). La précision au rang 1 est de 1, alors que la précision au rang 5 est de 3/5. Finalement, le taux MAP est de 0,63 = 6,29/10 = (1/1 + 2/2 + 3/3 + 3/4 + ...+ 3/10) / 10 alors que MRR est de 1 car le premier candidat est correct ; il serait de 1/2 si seulement le second candidat avait été correct, etc.

Il est apparu empiriquement qu'il était préférable de couper une liste candidate à un rang donné que d'essayer de seuiller en fonction d'un score de similarité, et nous détaillons donc uniquement par la suite les résultats avec la première méthode, en faisant varier le rang.

10. Que nous présentons sous forme de pourcentage pour plus de lisibilité.

4.1 WordNet

La table 2 montre les résultats pour les noms évalués selon les synonymes de WordNet. Pour chaque approche, nous indiquons les précisions aux rangs n=1, ..., 100 dans les listes candidates, les taux MAP, MRR, la R-précision, le nombre de synonymes dans la référence ($||ref||$) et le rappel global, pour les 200 premiers candidats de chaque méthode [11]. Nous rapportons également l'influence de différents filtres de fréquence. La ligne f>1000, par exemple, indique que nous retenons des listes candidates et de la référence les seuls mots dont la fréquence (dans Wacky) est supérieure à 1000.

n-meilleur(s)		P1	P5	P10	P20	P100	MAP	MRR	R-prec	$\|\|ref\|\|$	rappel
Miroir	f>1	16,4	5,1	3,8	2,7	1,3	11,9	15,1	16,6	2342	50,0
	f>5000	19,1	5,4	3,8	2,6	1,2	11,3	13,2	17,5	1570	54,8
	f>20000	22,1	5,7	3,9	2,5	1,1	9,8	11,4	22,7	1052	60,6
Lin	f>1	17,4	5,2	3,5	2,5	1,5	11,7	14,3	14,7	2342	35,9
	f>5000	16,5	5,0	3,5	2,5	1,6	9,2	10,8	16,7	1570	36,6
	f>20000	17,5	4,5	3,3	2,5	1,6	7,3	8,4	20,1	1052	36,9
M/L	f>1	25,8	7,5	5,7	4,4	3,8	15,9	17,6	22,0	2342	29,3
	f>5000	27,4	7,4	5,5	4,3	3,8	12,7	13,6	24,6	1570	31,1
	f>20000	26,1	6,4	4,7	3,5	2,6	9,7	10,4	28,9	1052	32,7

TABLE 2 – Résultats pour les noms, micro-moyennés, avec les synonymes de WordNet pour référence.

Comme WordNet répertorie peu de synonymes, les précisions à faible rang (1 et 5) sont les plus pertinentes, ainsi que la R-précision : les autres sont nécessairement très basses. Les autres mesures sont données à des fins de comparaison car elles sont plus pertinentes pour la référence Moby. Ceci étant noté, la table 2 amène plusieurs commentaires [12].

Premièrement, nous observons que la précision de l'approche miroir au rang 1 culmine à 22% alors que le rappel plafonne à un peu plus de 60% : un bien meilleur résultat combiné que l'approche distributionnelle que nous avons testée (moins de 18% de précision au rang 1 et moins de 37% de rappel). Deuxièmement, il apparaît clairement que filtrer les candidats les moins fréquents est beaucoup plus payant pour l'approche miroir [13]. C'est sans doute la conséquence d'un corpus de départ plus petit, pour lequel les occurrences rares de mots peu fréquents entraînent des probabilités d'alignement peu fiables. Troisièmement, nous observons que notre combinaison des deux approches, aussi simpliste soit-elle, s'accompagne d'une augmentation significative de la précision, notamment la R-précision (au détriment cependant du rappel).

Enfin, les résultats sur les verbes sont similaires à ceux présentés ici pour les noms, avec cependant une meilleure précision à rang faible et un compromis sur la fréquence de coupure plus élevé, et ce, même si la précision oracle est globalement la même pour toutes les configurations. Combiner les deux méthodes améliore la précision de manière similaire à ce que nous observons sur les noms, avec une précision oracle qui varie cette fois entre 20% et 27%. Les différences sont toutes significatives sauf cette fois sur P1 à fréquence élevée.

Nous ne montrons pas le détail des scores si on ajoute dans la référence les relations issues de toutes les fonctions lexicales de WordNet, mais on peut noter que les résultats sont très proches entre les deux méthodes sur les noms (R-prec≈13% et P@1=23% quand f>1). En revanche, les traductions miroirs sont légèrement meilleures sur les verbes en R-précision (16% contre 18% pour f>1 et 17% contre 21% pour f>20000), et inférieures en terme de précision au rang 1 (41% contre 37% pour f>1 à 33% contre 34%).

Nous montrerons en section 4.3 que la différence semble se jouer essentiellement sur les relations de synonymie et d'hyperonymie (et hyponymie).

11. Les candidats de Lin étant limités à ce nombre.

12. Sauf précision contraire, les différences entre méthodes discutées plus bas sont tous significatives à p<0.05. Pour toutes les mesures sauf P1 et le rappel global, nous avons utilisé le test de Wilcoxon sur les résultats mot par mot. Dans le cas de P1 qui donne un résultat binaire par cible, nous avons fait un test binomial.

13. Les différences significatives de précision et MAP entre les deux méthodes n'apparaissent que pour les valeurs de fréquence élevée.

4.2 Moby

La table 3 résume les résultats des deux approches pour les noms, en prenant cette fois-ci Moby pour référence. Les relations listées dans ce thésaurus étant du tout venant, nous nous attendions à ce que la référence soit plus proche des sorties produites par une approche distributionnelle. Nous observons que c'est bien le cas sur les noms : la précision de l'approche de Lin est systématiquement supérieure à celle des miroirs, avec presque 10 points de plus au rang 1 ; et ce, même si le rappel est légèrement en faveur de l'approche miroir. Sur les verbes, cependant, les deux approches se comportent de manière comparable. En observant la différence de scores de précision entre

n-meilleur(s)		P1	P5	P10	P20	P100	MAP	MRR	R-prec	$\|ref\|$	rappel
Miroir	f>1	33,7	15,8	13,3	11,0	7,0	18,5	40,1	11,0	60774	18,1
	f>5000	32,7	14,5	12,1	9,8	6,1	18,7	38,1	11,8	43294	21,6
	f>20000	30,3	13,2	10,7	8,6	5,3	18,1	34,9	12,8	28488	26,7
Lin	f>1	44,8	19,9	16,4	13,4	9,5	26,6	46,8	14,7	60774	15,4
	f>5000	40,7	18,5	15,0	12,5	9,3	25,6	41,6	15,0	43294	16,3
	f>20000	39,4	16,1	13,5	11,2	8,4	23,3	35,2	16,8	28488	16,8
M/L	f>1	53,1	25,1	21,4	18,1	35,2	46,6	22,9	25,0	60774	9,4
	f>5000	52,4	23,0	19,3	16,6	13,7	30,7	41,2	23,4	43294	10,9
	f>20000	45,9	19,4	16,5	14,0	11,2	24,6	32,6	21,6	28488	12,5

TABLE 3 – Résultats pour les noms, micro-moyennés, avec les mots reliés de Moby pour référence.

la table 3 et la table 2, il semble que les approches miroir et distributionnelle ramènent bien d'autres entités que des synonymes dans leurs meilleurs candidats. Cela peut sembler un peu surprenant pour l'approche miroir puisque cette approche capitalise à priori sur des relations de traduction. Nous devons analyser cela de façon plus précise afin de savoir si nous sommes en présence de bruit dans les modèles de traduction (ce qui est très probable) ou si Moby contient plus de synonymes que WordNet, ou les deux.

Nous observons également que le rappel de l'approche miroir est plus grand que celui de l'approche distributionnelle, une observation en accord avec notre évaluation sur WordNet, et qui est peut-être due à la différence de performance des deux approches sur les synonymes (qui sont nombreux dans Moby).

Le filtre des entités lexicales offre un rendement mitigé : la précision de la variante f>20000 est légèrement inférieure à celle de la variante f>1 pour les deux approches. Le rappel de l'approche miroir augmente cependant de manière notable et consistante dans les cas où l'on s'intéresse aux mots très fréquents.

4.3 Analyse des erreurs produites par l'approche miroir

Les expériences que nous venons de décrire possèdent quelques limites. La référence WordNet que nous utilisons pour la synonymie possède un nombre relativement restreint de synonymes par mot candidat, ce qui ne permet pas de rendre compte avec précision de la pertinence des autres candidats proposés. Le fait d'utiliser un thésaurus plus vaste comme Moby ne résout que partiellement le problème car la nature des relations encodées dans Moby n'est pas étiquetée et certaines relations présentes dans cette ressource ne correspondent pas à des relations lexicales typiques (ex : `raging` / `abandoned`).

On peut mener une première analyse à l'aide de WordNet afin d'évaluer la présence de termes reliés par d'autres relations que la synonymie. Si l'on regarde les premiers candidats produits pour chaque cible (voir la table 4), on constate que pour les verbes, 19% sont recensés comme hyperonymes et 6% comme hyponymes, les proportions étant de 7% et 4% pour les noms. Les autres fonctions (holonymes, méronymes et antonymes) apparaissent de façon marginale. En regardant les 5 premiers candidats produits par les deux approches, on observe que ceux produits pour les verbes correspondent davantage à des relations présentes dans WordNet. En grande majorité, les candidats identifiés ne correspondent pas à une relation étiquetée dans WordNet. Un peu moins de la moitié des cibles ne reçoit d'ailleurs aucun candidat validé par WordNet (Ø).

Les problèmes de couverture de WordNet se posent malheureusement pour toutes les fonctions lexicales et ceci ne peut être qu'indicatif. Nous avons donc conduit une évaluation manuelle de la sortie produite par l'approche

		top 5							top 1						
		Ø	S	He	Ho	Hl	A	M	Ø	S	He	Ho	Hl	A	M
nom	Miroir	3146	181	175	98	13	5	1	570	64	58	28	5	1	1
	Lin	3078	186	161	123	12	11	2	565	65	49	31	6	4	1
verbe	Miroir	2807	406	428	216	0	7	0	466	95	139	42	0	3	0
	Lin	2882	414	272	212	0	20	0	444	140	106	50	0	5	0

TABLE 4 – Nombre de fonctions lexicales correspondant aux 5 (colonne de gauche) ou 1 (colonne de droite) premiers candidats proposés par chaque approche sans filtre de fréquence, selon WordNet. Ø signifie qu'aucune relation selon WordNet n'est associée à un candidat. Ces occurrences sont comptabilisées pour les cibles traitées par les deux approches, soit 724 noms et 743 verbes. S=synonymes, He=hyperonymes, Ho=hyponymes, Hl=holonymes, A=antonymes, M=méronymes.

miroir en sélectionnant aléatoirement 100 paires de mots `cible` / `candidat` où le candidat est le premier proposé par l'approche miroir, bien qu'il ne soit pas validé comme synonyme par WordNet. Nous avons observé les phénomènes suivants :

– 25% des mots candidats constituent une partie d'une unité composée de plusieurs mots, comme le mot `sea` dans la paire `sea` / `urchin` ;
– 18% des mots candidats non validés par WordNet sont en fait des synonymes selon d'autres thésaurus que nous avons consulté manuellement [14]. C'est par exemple le cas de la paire `torso` / `chest` ;
– 13% des candidats sont en fait des hyperonymes listés dans WordNet ou dans `www.thesaurus.com`, comme la paire `twitch` / `movement` ;
– 6% des paires mettent en relation des mots morphologiquement reliés, comme `accountant` / `accounting`, probablement en raison d'un problème d'étiquetage en partie du discours dans la langue pivot où un mot français comme ici `comptable` peut être aussi bien être un nom qu'un adjectif.

Parmi les erreurs (au sens de WordNet) fréquentes restantes, certaines sont dues à la polysémie d'une traduction pivot, comme par exemple le mot anglais `aplomb` traduit en français par `assurance` qui veut également dire `insurance` en anglais. Ce type de problème est cependant difficile à analyser sans retracer méticuleusement les nombreuses associations utilisées par les modèles de traduction dans notre approche.

D'autres erreurs sont plutôt imputables à des termes peu fréquents dans le corpus des Hansards que nous avons utilisé et que nous aurions dû filtrer au préalable.

Cette analyse suggère que tous les candidats rejetés ne sont pas nécessairement mauvais et qu'il y a donc place à amélioration. La polysémie demeure le problème le plus difficile à résoudre, que ce soit pour notre approche miroir ou pour l'approche distributionnelle.

Nous avons également regardé de manière très informelle les mots cibles pour lesquels WordNet ne propose aucun synonyme alors que l'approche miroir propose des candidats. Dans une proportion non négligeable de cas, les traductions miroir sont pertinentes comme `whopper` / `lie`. Une analyse plus fine est cependant requise pour quantifier plus précisément cette observation.

4.4 Tests de synonymie

Comme évaluation secondaire, plusieurs auteurs utilisent, pour évaluer la pertinence d'une mesure de similarité sémantique, des tests de synonymie semblables à ceux posés dans les examens du TOEFL (Turney, 2008) où la tâche consiste à distinguer parmi quatre candidats, le synonyme d'un mot dans un contexte donné. On peut voir cet exercice comme une version simplifiée d'une tâche de désambiguïsation, où le but est de reconnaître le bon terme dans un ensemble de *distracteurs* (termes à priori sans rapport), au lieu de distinguer les sens d'un même mot. On teste alors la similarité sémantique en prenant celui des candidats qui a le score de similarité le plus élevé avec la cible.

14. Comme par exemple le Roget's 21st Century Thesaurus, `http://www.thesaurus.com`

Les données TOEFL ne sont pas librement disponibles, aussi avons nous utilisé ici un test généré artificiellement à partir des données de WordNet par Freitag *et al.* (2005) [15]. Les auteurs le considèrent comme plus difficile que les équivalents du TOEFL. Ce test est notamment utilisé par Ferret (2010) afin d'ordonner différentes mesures de similarité entre vecteurs de cooccurrences. Le meilleur score qu'il obtient sur ce test est de 71.6% de réponses correctes, ce qui est proche du score de 72% d'exactitude obtenu par (Freitag *et al.*, 2005) à l'aide d'autres méthodes distributionnelles. Les systèmes répondent à toutes les questions.

Les instances de test sont de la forme `house: family obstacle filing surgeon` le premier terme étant la cible, le second un synonyme d'un des sens de la cible selon WordNet, les trois autres sont des distracteurs. Quelques restrictions sont ajoutées pour ne pas rendre le test trop facile : les synonymes dont la forme est proche de la cible (`group`/`grouping`) sont éliminés. Par ailleurs les cibles sont choisies avec une fréquence minimale. Les distracteurs sont choisis complètement au hasard, mais les termes associés sont choisis parmi les synsets de WordNet et privilégient donc les termes polysémiques.

Nous avons appliqué ce test à nos deux approches en choisissant, comme les autres travaux mentionnés, celui des candidats ayant le meilleur rang dans la liste de similarités. Si aucun des candidats du test n'est présent dans les candidats d'une méthode, le système ne répond rien. Nous pouvons donc évaluer l'aptitude d'une mesure de similarité à identifier le bon terme, ce que nous mesurons en terme de précision, rappel et F-score.

La table 5 résume les résultats pour les 200 premiers candidats de chaque méthode. Dans les cas où les candidats sont tous absents de la réponse du système, (Ferret, 2010) renvoie une réponse au hasard, mais cela arrive rarement vue la couverture de son système. Nous avons ici fait le choix de considérer que le système ne répond pas faute de données fiables, car c'est un cas beaucoup plus courant ici. Ceci a pour effet de faire baisser le rappel, et la précision évalue réellement la méthode des miroirs.

Comme nous disposons pour l'approche miroir d'une liste de candidats plus étendue, nous avons aussi évalué cette méthode sans coupure. On constate un nombre important de non-réponses également, cette fois due sans doute à une couverture lexicale limitée de la ressource de départ (les Hansards). Noms et verbes regroupés, cette variante obtient un F-score de 0,73, avec 3908/17285 cibles sans réponse (22%), majoritairement des noms.

À nombres de candidats égaux, on constate donc que la méthode miroir a une précision équivalente à celle de Lin mais un rappel bien supérieur. Sans limite de candidats, elle atteint un F-score comparable aux meilleures méthodes distributionnelles testées dans les travaux susmentionnés.

Noms	F1	P	R	sans réponse
Lin[200]	0,55	0,95	0,38	5885/9887
Miroir[200]	0,63	0,95	0,47	4975/9887
Miroir	0,72	0,87	0,61	2995/9887

Verbes	F1	P	R	sans réponse
Lin[200]	0,55	0,87	0,40	3983/7398
Miroir[200]	0,69	0,89	0,56	2694/7398
Miroir	0,74	0,79	0,70	913/7398

TABLE 5 – Évaluation pour le test de synonymie basé sur WordNet de (Freitag *et al.*, 2005)

Faute de place, nous ne ferons que décrire brièvement une autre façon d'analyser le test effectué, proposée par (Freitag *et al.*, 2005), consistant à mettre les résultats en rapport avec le niveau de polysémie des termes cibles, et qui semblait mettre en évidence que les cibles polysémiques étaient les plus dures à résoudre. Nous n'avons pas constaté ce phénomène ici, la précision reste constante pour les verbes et ne baisse que très légèrement pour les noms, quelle que soit la polysémie des cibles, alors que le rappel augmente, sans doute parce que les miroirs sont plus susceptibles d'avoir une réponse à fournir sur les mots plus fréquents.

5 Travaux reliés

Plusieurs types de travaux peuvent être comparés à la présente étude, ayant des objectifs, des données en entrée et des méthodologies d'évaluation plus ou moins variés. L'extraction de paraphrases partage certains de nos objectifs et ressources, même si elle concerne le rapprochement de termes comportant plus d'une unité lexicale. L'extraction de synonymes, la construction de thésaurus recouvrent aussi nos buts et peuvent être évalués de façon similaire.

15. Ce test et est librement disponible à l'URL `http://www.cs.cmu.edu/~dayne/wbst-nanews.tar.gz`

De façon plus large, les nombreux travaux récents sur la conception et la réalisation de mesures de similarité sémantique peuvent être rapprochés naturellement de la méthode présentée, même si les objectifs sont différents.

L'évaluation de l'acquisition de paraphrases est souvent évaluée par des jugements humains d'acceptabilité des substitutions en contexte, ce qui limite à des petits jeux de test. Barzilay & McKeown (2001) rapportent que 90% des paraphrases extraites par patrons (sur un corpus monolingue de traductions littéraires) sont acceptables, mélangeant synonymes, hyperonymes et termes coordonnés, sans bien sûr pouvoir donner une idée de la couverture d'une telle méthode. En se fondant sur des alignements bilingues et une méthode similaire à la nôtre mais sur plusieurs unités lexicales, Bannard & Callison-Burch (2005) estiment que les meilleurs paraphrases extraites pour chaque cible sont valides dans 75% des cas avec un alignement parfait (48% avec un alignement automatique). De même Lin *et al.* (2003) ou Curran & Moens (2002) évaluent précisément la présence de synonymes dans des listes de similarité dans des petits ensembles de paires de synonymes ou antonymes, ce qui rend difficile une extrapolation sur le genre de données que nous utilisons afin d'atteindre une large couverture.

Plus proche de la méthodologie que nous avons suivie, on trouve des études qui évaluent la classification de paires de mots en synonymes ou non-synonymes. Cela peut être fait directement sur les candidats sélectionnés pour un ensemble de cibles, comme dans l'étude présente, ou sur des ensembles de test rééchantillonnés pour augmenter artificiellement la présence de paires positives et pouvoir appliquer des techniques standards de classification avec une fiabilité raisonnable. Ne pas rééchantillonner est plus réaliste mais donne des scores assez bas, comme nous l'avons constaté : (van der Plas & Tiedemann, 2006) partent de vecteurs d'alignement à la place des vecteurs d'arguments syntaxiques de Lin, en définissant la même similarité et atteignent 12% de F-score par rapport à leur référence ; (Wu & Zhou, 2003) fait de même, ajoutant aussi une distance calculée dans un graphe lexical issu d'un dictionnaire, et apprend des régressions linéaires des différents scores de similarité, tout en restreignant la fréquence des cibles jusqu'à obtenir un maximum de 23% sur les noms et 30% sur les verbes. On peut aussi mentionner (Heylen *et al.*, 2008), qui analysent la répartition des fonctions lexicales dans des listes de similarités de mot en néerlandais. La seconde option, où l'on rééchantillone à l'entraînement et au test, est pertinente seulement si l'on connaît un moyen de présélectionner naturellement les candidats pour atteindre la proportion supposée, ce qui n'est pas le cas pour les études existantes (Hagiwara *et al.*, 2009). Notre étude peut en fait être considérée comme une entrée pour des expériences de ce genre.

L'étude des similarités distributionnelles faite par (Ferret, 2010), qui utilise de la cooccurrence simple, montre des résultats proches de ce que l'on obtient avec les miroirs, plus bas sur WordNet et comparables ou meilleurs sur Moby. Il opère sur un jeu de test beaucoup plus large, sans distinguer les parties de discours, et le jeu de test est découpé différemment par rapport aux fréquences lexicales puisqu'il sélectionne les cibles et les candidats. Sur WordNet il obtient au mieux 11% de R-précision et 17% pour la meilleure P@1 (sur les mots les plus fréquents). Sur Moby, la meilleure R-précision est de 10% et la meilleure P@1 est de 41%, P@5 de 28%. Le rappel est systématiquement inférieur (25% sur WordNet et 10% sur Moby), mais seuls 100 candidats sont gardés par cible. Les résultats sont plus bas que ce que l'on obtient ici avec les données de Lin, et nous pouvons donc supposer que la comparaison que nous faisons est représentative de l'approche distributionnelle dans ce contexte. La combinaison miroir et distribution est par contre supérieure sur tous les scores sauf le rappel.

Avec assez peu de réglage, on voit donc que l'approche des miroirs atteint des résultats comparables ou meilleurs que les similarités d'alignement ou distributionnelle pour isoler des synonymes dans certaines configurations. Les approches distributionnelles peuvent sans doute être améliorées mais l'approche choisie semble représentative. Il faut noter que le calcul qui sous-tend les traductions miroirs est computationnellement bien plus simple que les calculs de similarité entre $n \times n$ vecteurs d'alignement ou de cooccurrences, où n est la taille du vocabulaire.

On peut estimer que l'on peut atteindre un niveau de filtrage des candidats qui rend possible de tenter ensuite la classification des paires restantes. D'après (Hagiwara *et al.*, 2009) [16] , on peut associer (par classification) une fonction lexicale de façon fiable à des paires de mots si la proportion de candidats effectivement à relier par rapport à ceux qui n'ont aucun lien peut atteindre un ratio de 1 pour 6. Une conclusion similaire est tirée par (Piasecki *et al.*, 2008) dans le cas de la détection d'hyperonymie. Nos résultats nous encouragent à penser que l'on peut atteindre une proportion de 1 pour 4 ou 5. L'étude citée ne précise pas la proportion de paires de mots synonymes/non synonymes de départ, mais si on prend les chiffres de (Ferret, 2010), il y a 30000 paires de synonymes sur un vocabulaire de référence de 10000 mots, donc pour environ 50M de paires possibles, soit une proportion de 0,06% de paires de synonymes ou un ratio de 1 pour 1600.

16. (Hagiwara *et al.*, 2009) utilise comme descripteurs des schémas syntaxiques et des vecteurs de cooccurrence.

Une autre façon de juger de la pertinence des mesures de similarité sémantique entre mots dérive des données collectées par (Miller & Charles, 1991) où on demande à des sujets de juger la similarité ou le lien entre des items lexicaux, sur une échelle numérique. C'est une façon intéressante de fournir une évaluation intrinsèque de ces associations, mais le jeu de test ne peut couvrir qu'une part très limitée du vocabulaire (300 mots environ, avec 2 ou 3 associations par mot au plus).

6 Conclusion

Nos expériences confirment la variété des relations lexicales que l'on peut récupérer en appliquant ce que l'on a usage d'appeler des mesures de similarité sémantique. Les deux approches que nous avons étudiées ici semblent corrélées aux ressources de référence que nous avons considérées.

En ce qui concerne les synonymes, nos expériences indiquent que les traductions miroir offrent des candidats plus pertinents que l'approche distributionnelle de Lin (1998). Dans la mesure où les approches miroir ne sont pas aussi prisées que les approches distributionnelles, nous espérons que cette étude contribuera à en montrer l'intérêt pour l'acquisition de relations lexicales. Nous soulignons de plus que l'approche miroir est beaucoup moins coûteuse à développer et à appliquer, pour autant que l'on soit en mesure de trouver des bitextes de taille suffisante mettant en jeu la langue d'intérêt. Patry & Langlais (2011) dressent un portrait des bitextes existants qui indique que de telles ressources sont de plus en plus disponibles pour de nombreuses paires de langues.

L'approche miroir que nous avons mise en place ne tire pas profit du fait que plusieurs bitextes mettant en jeu la langue d'intérêt sont disponibles. C'est par exemple le cas pour la langue française pour laquelle les bitextes des débats parlementaires européens sont disponibles en plus du bitexte que nous avons mis profit ici. L'ajout de telles ressources devrait être en mesure d'augmenter les performances (et la précision en particulier) de notre approche.

La complémentarité des approches testées dans cette étude amène à nous interroger sur la manière optimale de les combiner. La simple intersection que nous avons étudiée ici améliore nettement la précision des listes candidates. Une approche plus originale consisterait à combiner ces approches avec une approche par patron telle que celle de (Barzilay & McKeown, 2001). Le problème de la polysémie discuté en section 4.3 demeure un problème pour toutes les approches dont nous avons discuté, en particulier lorsque deux sens d'une entité lexicale sont fréquents en corpus. Il semble souhaitable d'intégrer l'apport de méthodes qui vise à repérer des groupes de sens équivalents multilingues, comme par exemple dans (Apidianaki, 2008).

Il n'en reste pas moins que notre objectif à moyen terme est de distinguer automatiquement la nature des différentes relations lexicales identifiées. Cette information est pertinente dans bon nombre d'applications (paraphrasage, choix d'une traduction, etc.). Des travaux comme ceux de (Hagiwara *et al.*, 2009) tendent à indiquer qu'il est envisageable d'entraîner de manière supervisée un classificateur à reconnaître certaines fonctions lexicales, pour autant que la proportion de candidats d'une classe particulière soit plus équilibrée que dans les distributions naturelles. Ceci indique qu'il faut être en mesure d'affiner la liste de candidats, ce que notre approche par filtrage ou combinaison réalise.

Remerciements

Nous remercions les relecteurs pour la pertinence de leurs commentaires.

Références

APIDIANAKI M. (2008). Translation-oriented Word Sense Induction Based on Parallel Corpora. In *Actes de LREC Language Resources and Evaluation (LREC)*, p. 3269–3275, Marrakech Maroc.

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, p. 597–604.

BARZILAY R. & MCKEOWN K. R. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*.

CURRAN J. R. & MOENS M. (2002). Improvements in automatic thesaurus extraction. In *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition*, p. 59–66.

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.

EDMONDS P. & HIRST G. (2002). Near-Synonymy and lexical choice. *Computational Linguistics*, **28**(2), 105–144.

FERRET O. (2010). Testing semantic similarity measures for extracting synonyms from a corpus. In *Proceedings of LREC 2010*.

FREITAG D., BLUME M., BYRNES J., CHOW E., KAPADIA S., ROHWER R. & WANG Z. (2005). New experiments in distributional representations of synonymy. In *Proceedings of CoNLL*, p. 25–32.

HAGIWARA M., OGAWA Y. & TOYAMA K. (2009). Supervised synonym acquisition using distributional features and syntactic patterns. *Journal of Natural Language Processing*, **16**(2), 59–83.

HEYLEN K., PEIRSMAN Y., GEERAERTS D. & SPEELMAN D. (2008). Modelling Word Similarity. An Evaluation of Automatic Synonymy Extraction Algorithms. In *Proceedings of LREC 2008*, p. 3243–3249 : ELRA.

KOZIMA H. & FURUGORI T. (1993). Similarity between words computed by spreading activation on an english dictionary. In *Proceedings of the conference of the European chapter of the ACL*, p. 232–239.

LIN D. (1998). Automatic retrieval and clustering of similar words. In *COLING/ACL98*, volume 2, p. 768–774, Montreal.

LIN D., ZHAO S., QIN L. & ZHOU M. (2003). Identifying synonyms among distributionally similar words. In *Proceedings of IJCAI'03*, p. 1492–1493.

MAX A. & ZOCK M. (2008). Looking up phrase rephrasings via a pivot language. In *Coling 2008 : Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, p. 77–85.

MICHIELS A. & NOEL J. (1982). Approaches to thesaurus production. In *Proceedings of Coling'82*.

MILLER G. & CHARLES W. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, **6**(1), 1–28.

MOORE R. C. (2004). Improving IBM word alignment model 1. In *42nd Meeting of the Association for Computational Linguistics (ACL)*, p. 518–525.

MULLER P. & LANGLAIS P. (2010). Comparaison de ressources lexicales pour l'extraction de synonymes. In *Article court au 17e TALN*, Montréal, Canada.

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

PATRY A. & LANGLAIS P. (2011). PARADOCS : l'entremetteur de documents parallèles indépendant de la langue. *TAL*, **51-2**, pp. 41-63.

PIASECKI M., SZPAKOWICZ S., MARCIŃCZUK M. & BRODA B. (2008). Classification-based filtering of semantic relatedness in hypernymy extraction. In A. RANTA & B. NORDSTRÖM, Eds., *GoTAL 2008*, number 5221 in LNAI, p. 393–404 : Springer.

TURNEY P. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*.

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.

WEEDS J. E. (2003). *Measures and Applications of Lexical Distributional Similarity*. PhD thesis, University of Sussex.

WU H. & ZHOU M. (2003). Optimizing synonyms extraction with mono and bilingual resources. In *Proceedings of the Second International Workshop on Paraphrasing*.

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