Comparing distributional and mirror translation similarities for extracting synonyms

Philippe {Muller[(1)], Langlais[(2)]}

(1) IRIT, Univ. Toulouse & Alpage, INRIA (2) DIRO, Univ. Montréal

Abstract. Automated thesaurus construction by collecting relations between lexical items (synonyms, antonyms, etc) has a long tradition in natural language processing. This has been done by exploiting dictionary structures or distributional context regularities (coocurrence, syntactic associations, or translation equivalents), in order to define measures of lexical similarity or relatedness. Dyvik had proposed to use aligned multilingual corpora and defines similar terms as terms that often share their translations. We evaluate the usefulness of this similarity for the extraction of synonyms, compared to the more widespread distributional approach.

1 Introduction

Automated thesaurus construction by collecting relations between lexical items has a long tradition in natural language processing. Most effort has been directed at finding synonyms, or rather "quasi-synonyms", [1], lexical items that have similar meanings in some contexts. Other lexical relations such as antonymy, hypernymy, hyponymy, meronymy, holonymy are also considered, and some thesauri also consider semantically associated items with less easily definable properties (e.g. the Moby thesaurus).

From the beginning, a lot of different resources have been used towards that goal. Machine readable dictionaries appeared first and generated a lot of effort aiming at the extraction of semantic information, including lexical relations, [2] or were used to define a semantic similarity between lexical items [3]. Also popular was distributional analysis, comparing words via their common contexts of use, or syntactic dependencies [4, 5] in order to define another kind of semantic similarity. These approaches went on using more readily available resources in more languages [6]. More recently, a similar approach has gained popularity using bitexts in parallell corpora. Lexical items are considered similar when they are often aligned with the same translations in another language, instead of being associated to the same context words in one language [7, 8]. A variation on this principle, proposed by [9], is to consider translation "mirrors": words that are translations of the same words in a parallell corpus, as they are supposed to be semantically related. Although this idea has not been evaluated for synonyms extraction, it is the basis of some paraphrase extraction work, i.e. finding equivalent phrases of varying lengths in one language, see for instance [10].

Evaluations of this line of work vary but are often disappointing. Lexical similarities usually bring together heterogeneous lexical associations and semantically related terms, that are not easy to sort out. Synonymy is probably the easiest function to check as references are available in many languages, even though they may be incomplete (e.g. WordNet for English) and synonym extraction is supposed to complement the existing thesauri.
If these approaches have the semantic potential most authors assume, there is still a lot to be done to harness that potential. One path is to select the most relevant associations output by the aforementioned approaches (dictionary-based, distribution-based, or translation-based), as in the work of [11], hopefully making possible a classification of lexical pairs into the various targeted lexical relations. Another is to combine these resources and possibly other sources of information; see for instance [8].
We make a step in this latter direction here, by testing Dyvik's idea on lexical relation extraction. Translation mirrors have not been precisely evaluated in such a framework, and the way it can be combined with distributional information has not been investigated yet. We also pay particular attention to the frequency of the words under consideration, as polysemy and frequency variations of semantic variants seem to play an important role in some existing evaluations. Indeed, we show that mirror translations fare better overall than a reference distributional approach in the preselection of synonym candidate pairs, both on nouns and verbs, according to the different evaluations we performed.
The remainder of this paper is organised as follows: we present in section 2 the resources we considered and our experimental protocol in section 3. We analyze our results in section 4. We relate our results to comparable approaches addressing the same issue in section 5 and finally conclude our work in section 6.

2 Resources and input

We considered two reference databases in this work:

- the WordNet lexical database,[1] provided through the NLTK package API.[2] WordNet provides a reference for the following lexical relations: synonyms, antonyms, hypernyms, hyponyms, holonyms, meronyms. Each lemma present in WordNet has on average 5-6 synonyms, or 8-10 related terms if all lexical relations are taken together.
- the Moby thesaurus[3] which provides not only synonyms but more loosely related terms. This resource is much richer and less strict than WordNet, as each target has an average of about 80 related terms.

To estimate the frequencies of the words considered, we used data provided by the freely available Wacky corpus. [4]

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

In order to compare similarities induced by distributional and mirror approaches, we have selected at random two sets of 1000 lexical items, a set of nouns and a set of verbs, that we will call "targets". We imposed an arbitrary minimal frequency threshold on the targets (> 1000). The statistics of the two references, with respect to the test sets of targets considered, are shown in table 1.

Table 1. Reference characteristics with the two target sets considered: median frequency in the Wacky corpus, mean number of associated terms, median, minimum and maximum number; (NB: Moby mixes verbs and nouns, so we considered terms having a noun form or a verb form in each case).

			number of associations			
Pos	Median frequency	reference	mean	med	min	max
Nouns	3,538	WordNet syns	3.6	2.0	1	36
Nouns	3,538	Moby	73.8	57.0	3	509
Verbs	11,136	WordNet syns	5.6	4.0	1	47
Verbs	11,136	Moby	113.2	90.0	6	499

3 Protocol

We consider similar terms derived either by a translation mirror approach (section 3.1) or a syntactic distributional approach (section 3.2). Each approach provides a set of associated terms, or "candidates", ranked according to the similarity considered. These ranked candidates are then evaluated with respect to a reference for different lexical relations, either keeping n-best candidates or candidates above a given threshold. Details of the evaluation are presented below. As an example, table 2 shows candidates proposed by the translation mirrors (see section 3.1) for the randomly chosen target term `groundwork`. Note the huge difference in coverage of WordNet and Moby.

3.1 Translation mirrors

The translation mirror approach is based on the assumption that words in a language $\mathcal{E}$ that are often aligned in a parallel corpus with the same word in another language $\mathcal{F}$ are semantically related. For instance, the french words `manger` and `consommer` are often both aligned with, and probable translations of, the english word `eat`.
For the translation mirror based approach, we used a French-English bitext of 8.3 millions pairs of phrases in translation relation coming from the Canadian Hansards (transcripts of parliamentary debates). This bitext is used by the bilingual concordancer `TSRali`[5] and was kindly made available to us by the maintainers of the application. We lemmatized both French and English sentences

[5] http://www.tsrali.com/

Table 2. First ten candidate associations proposed by our translation mirror approach for the target term `groundwork` and synonyms according to WordNet as well as a sample of related terms according to Moby. Underlined candidates belong to the WordNet reference, while those in bold are present in Moby; both are also reported in the reference they belong to. Words marked with * are absent from the Hansards.

Candidates	WordNet	Moby
base	base	arrangement
basis	basis	**base**
foundation	cornerstone	basement
land	foot	**basis**
ground	fundament*	bed
job	foundation	bedding
field	substructure*	bedrock
plan	understructure*	bottom
force		briefing
development		cornerstone
		... [47 more]

using `TreeTager`.[6] Then, we trained in both directions[7] (English-to-French and French-to-English) statistical translation models, running the `Giza++` toolkit in its standard setting.[8] Our translation mirror approach makes use of the lexical distribution of the two models,[9] p_{e2f} and p_{f2e}, we obtained this way (see tab 3 for an example). More specifically, we compute the likelihood that the word s is related to the target word w as:

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

where $\tau_{e2f}(\mathrm{w})$ stands for the set of French words associated to w by the model p_{e2f}. In practice, two thresholds, δ_1 and δ_2, control the noise of the lexical distributions:

$$p_\bullet^\delta(t|s) = \begin{cases} p_\bullet(\mathrm{t}|\mathrm{s}) & \text{if } p_\bullet(\mathrm{t}|\mathrm{s}) \geq \delta \\ 0 & \text{otherwise} \end{cases}$$

In the evaluations below we considered only the first 200 lemmas for each target, in order to compare it with the available distributional candidates presented in the following section.

3.2 Distributional Similarity

The distributional similarity we used is taken straight from the work of [5], as we believe it represents well this kind of approach. Also, a thesaurus computed

[6] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[7] IBM models are not symmetrical.
[8] http://code.google.com/p/giza-pp/
[9] We used IBM models 4.

Table 3. The 10 most likely associations to the words `consommer` and `eat` according to the lexical distributions $p_{f2e}(\bullet|\text{consommer})$ and $p_{e2f}(\bullet|\text{eat})$ respectively.

$p_{f2e}(\bullet|\text{consommer}) \triangleright$ consume *(0.22)* use *(0.18)* be *(0.1)* eat *(0.092)* consumption *(0.048)* consuming *(0.037)* take*(0.023)* drink *(0.019)* burn *(0.012)* consumer *(0.011)* ...
$p_{e2f}(\bullet|\text{eat}) \triangleright$ manger *(0.39)* consommer *(0.08)* se *(0.036)* de *(0.031)* nourrir *(0.028)* avoir *(0.027)* du *(0.023)* alimentation *(0.017)* gruger *(0.016)* qui *(0.014)* ...

by Lin is freely available,[10] which eases reproducibility. Lin used a dependency-based syntactic parser to count occurrences of `(head lemma,relation,dep. lemma)`, where `relation` is a syntactic dependency relation. Each lemma is thus associated with counts for a set F of features `(rel,other lemma)`, either as head of a relation with another lemma or as dependent. For instance, the verb `eat` has the features `(has-subj,man)`, `(has-obj,fries)`, `(has-obj,pie)`, etc. Let c be the function giving the number of occurrences of a triplet (w, rel, w') and let V be the vocabulary :

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

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

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

I is the specificity of a relation (w, rel, w'), defined as the mutual information between the triplet elements [5]. Let's note $\|w\|$ the total information quantity associated to w. Finally, similarity between two lemmas w_1 and w_2 measures the extent to which they share specific syntactic contexts, using the information quantity of their shared contexts, normalised by the sum of their total information quantities.

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

The available thesaurus lists the closest 200 lemmas for each word in a given vocabulary.

4 Experiments and results

Following the protocol introduced above, we evaluated the outputs of lexical similarities based on the n-best candidates, varying n, or based on varying similarity

[10] http://webdocs.cs.ualberta.ca/∼lindek/Downloads/sim.tgz

thresholds, both for the distribution-based approach and the mirror approach. We have two different test sets to evaluate differences between nouns and verbs. As shown in table 1, syntactic categories differ in the number of synonyms or other lexical related items they possess, and it is likely that they impact as well the approaches we investigated; see [12] on the role of frequency in that perspective. We considered for evaluation only the items that were common to the reference and the lexicon covered by the resources used. For instances some synonyms from WordNet have no occurrences in the Hansard or in Lin's database and this can be seen as a preprocessing filter of rare items.

Moreover, both approaches we compare are sensitive to the typical frequencies of the targets considered. In both cases, all senses of a word are conflated in the computation and it is likely that more frequent usages dominate less frequent ones. We wanted to evaluate the role played by this factor and we took this into account in our evaluations by adding a varying frequency threshold on the candidates considered. For a set of values c_i, we filtered out candidates with frequencies less than c_i in a reference corpus (the Wacky corpus, mentioned above).[11]

Additionally, we took out a list of the most common items in the candidates of the target sets. We arbitrarily removed those terms that appear in more than 25% of the candidate lists (this threshold could be tuned on a development set in further experimentations). This includes very common nouns (e.g. `thing`, `way`, etc) and verbs (e.g. `have`, `be`, `come`), as well as terms that are over-represented in the Hansard corpus (e.g. `house`), since alignment errors induce some noise for very frequent items. Finally, we combined the candidate lists produced by the two approaches by filtering candidates for one approach that are not present in the other's candidate list.

We are interested in two aspects of the evaluation: how much of the reference is covered by our approaches, and how reliable they are, that is, we want the top of the candidate list to be as precise as possible with respect to an ideal reference. In order to do so, we evaluate our approaches according to precision and recall[12] at different points in the n-best list or at different threshold values. We also compute typical information retrieval measures to estimate the relevance of the ranking: mean average precision (MAP), mean reciprocal rank (MRR). MAP computes the precision at each point where a relevant term appears in a list of candidates; MRR is the average of the inverses of the ranks of the first relevant term in a list. Last, we looked at the precision of each method assuming an "oracle" gives them the right number of candidates to consider for each target, a measure called R-precision in the information retrieval literature.

So for instance, the 10 candidates of table 2 evaluated against the WordNet reference would receive a precision of 3/10 and a recall of 3/5, (and not 3/8, because `understructure`, `substructure` and `fundament` are absent from the Hansard). R-precision would also be 3/5, since all correct candidates are found at ranks less than the reference size (5 synonyms). Precision at rank 1 would be

[11] The thresholds were chosen to correspond to different ranges of lexical items.

[12] For the sake of readability, we report precision and recall as percentage.

1 while precision at rank 5 would be 3/5. The MAP would be 0.63 = 6.29/10 = (1/1 + 2/2 + 3/3 + 3/4 + ... + 3/10) / 10 and MMR would be 1 in this case because the first candidate is correct. It would be 1/2 if only the second were correct, etc.
Our experiments led to the observation that it is better to cut the n-best list at a given rank than to try to find a good similarity threshold, and we thus only detail results for the first method.

4.1 WordNet

Table 4 shows the results for nouns with respect to synonyms in WordNet. For each approach we report precision at ranks n=1, ..., 100 in the candidate list, MAP, MRR, the R-precision, the number of considered synonym pairs from the reference ($\|ref\|$), with respect to which the overall recall is computed. We also report the influence of different frequency filters. A line with f>5000 means we consider only candidates and reference items with a frequency above 5000 in Wacky.
As the WordNet reference has few synonyms, one should focus on precisions at low ranks (1 and 5) as well as the oracle, R-precision: all others are bound to be quite low. The other cutoffs make more sense for the evaluation with respect to Moby, and are here for comparison. This being noted, table 4 calls for several comments. First of all, we observe that the precision of the mirror approach at rank 1 culminates at 22% while overall recall tops at 60%, a much better score than the distributional approach we tested. Second, it is noticeable that filtering out less frequent candidates benefits the mirror approach much more than the distributional one. It might be a consequence of having a smaller corpus to start with, in which rarer words have less reliable alignment probabilities.
Third, we observe that combining the candidates of both approaches yields a significant boost in precision at the cost of recall. This is encouraging since we tested very simple combination scenarios: the one reported here consists in intersecting both lists of candidates.

Table 4. Results (percentages) for nouns, micro-averaged, with respect to synonyms in WordNet.

n-best		P1	P5	P10	P20	P100	MAP	MRR	R-prec	$\|ref\|$	recall
Mirror	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

Last, the results on verbs are quite similar to those for nouns, with a better precision at low ranks, and at higher frequency cutoffs, even though the oracle evaluation is roughly the same for all configurations. Again, filtering one method with the other yields better results, with oracle precision between 20% and 27%, similarly to what is observed on nouns.

Also, it is noteworthy that Lin's approach is close to the mirror approach on nouns: R-prec≈13% and P1=23% for both when f>1; the latter fares better on verbs: R-prec=16 for Lin, 18 for mirror translation, and then the gap widens with higher f-values to reach 17 vs. 21 for f>20000; P1 goes from 41 to 33 for Lin, and from 37 to 34 for mirror translation.

4.2 Moby

Table 5 shows the results for nouns with respect to the related terms in the Moby thesaurus. We expected that this reference would be closer to what is recovered by a distributional similarity, and that is indeed the case for nouns: Lin's precision is superior across the board, even by 10 points at rank 1. However, both methods are comparable on verbs. One notable fact is that both similarities capture a whole lot more than just synonymy so the scores are much higher than on WordNet, and this can be considered somewhat of a surprise for the mirror translations, since this method should capitalise on translation relations only.

Also, in almost all cases, the overall recall is higher with the translation mirror method, an observation consistent with our experiments on WordNet. Filtering out low frequency words has mixed effects: precision is slightly less for f>20000 than f>1 but the corresponding recall of high frequency related terms is higher. The combinations of the two methods consistently improve precision (again to the detriment of recall). As a conclusion, related terms do appear in mirror translations, even if they seem to do so with lower similarity scores than synonyms, and we have to investigate more precisely what is happening (translations ap-

Table 5. Results (percentages) for nouns, micro-averaged, with respect to related terms in Moby.

n-best		P1	P5	P10	P20	P100	MAP	MRR	R-prec	$\|ref\|$	recall
Mirror	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	15.2	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

proximations or errors or a better coverage of synonymy in the Moby thesaurus than in WordNet).

4.3 Error analysis

The kind of evaluation we presented above has a few shortcomings. The main reference we used for synonymy does not have a large number of synonyms per entry, and if one of our objectives is to extend existing resources, we cannot estimate the interest of the items we find that are absent from that reference. Using a larger thesaurus such as Moby only partially solves the problem, since there is no distinction between lexical relations, and some related terms do not correspond to any classical lexical function. In order to evaluate more precisely our output, but on a much smaller scale, we have looked at a sample of items that are absent from the reference, to measure the amount of actual errors. To do this, we took a number of terms which are the first candidates proposed by the mirror approach for a target, but are absent from WordNet. We found a number of different phenomena, on a sample of 100 cases:

- 25% of words that are part of a multi-word expression which were probably aligned to the same target translation, such as `sea/urchin`;
- 18% of words that are actually synonyms, according to other thesauri we could check manually,[13] such as `torso/chest`;
- 13% hypernyms, listed in WordNet or in www.thesaurus.com, e.g. `twitch/movement`.
- 6% morphologically related items such as `accountant/accounting`, probably because of a pos-tag ambiguity in the pivot language, here the French word `comptable`, which can be a noun or an adjective.

Among the remaining errors that are probably common, some are due to a polysemy of a pivot translation (e.g.: English word `aplomb` translated into French as `assurance` which can also mean `insurance` in English). This is hard to quantify exactly in the sample without looking in detail at all related aligned word pairs. On the remaining various errors, some bear on rare occurrences in the input corpus, that we should have filtered out beforehand. All in all, we can see there is room for easy improvement. Only polysemy is a hard problem to address, and this is so for any kind of approach relying on distributional data.
In addition to that, we are currently looking at items that were not considered in the evaluation because there was no synonym for them in Wordnet, but for which there are mirror translations (such as `whopper/lie`). Although we cannot yet quantify the presence of truly related lexical items, the few examples we looked at seem to reflect the analysis above.

5 Related work

There are several lines of work that are comparable to what we presented here, with a variety of objectives, evaluation methodologies and input data. Para-

[13] Such as http://www.thesaurus.com.

phrase extraction shares some of our objectives and some of the resources we considered. Synonym extraction and thesaurus building also overlap our goals and evaluation methods. Also, work striving to design and compare semantic similarities is the closest in nature, if not in the objectives.

Paraphrase acquisition is usually evaluated on the acceptability of substitutions in context, and only small-scale human judgments of the results give an indication of the lexical functions captured: [13] reports that 90% of their pattern-based extracted paraphrases are valid, mixing synonyms, hypernyms and coordinate terms, but with no indication of coverage. Similarly, [14] or [15] precisely evaluate the presence of synonyms on similarity lists on a small subsets of synonym/antonym pairs, which makes it hard to extrapolate on the kind of data we used, where we aim at a much larger coverage.

Closer to our methodology, several studies evaluate the classification of a set of word pairs as synonyms or not; either directly on the candidates selected for each target, as we do here, or on resampled word pairs test sets that make the task accessible to common learning techniques. The former method (non-resampled, which is also ours) is more realistic and of course gives rather low scores: [7] use alignment vectors on a set of language pairs, and syntactic argument vectors, and similarity is defined in a comparable way between the vectors; [8] also use a syntactic distributional similarity and a distance in a dictionary-based lexical network. The first study only looks at the first three candidates in Dutch, with respect to a synonym reference (Euro WordNet) and considers only nouns. Scores P1 range from 17.7 to 22.5% on alignment candidates, with distributional similarity at 8%, and the combination at 19.9%. The authors have an updated experiment in [16], still on Dutch nouns, and reach 30% for P1, but do not explain the differences in their setup. The second study applies linear regressions on all similarity scores, with different target frequencies and similarity thresholds, and reaches a maximum f-score of 23% on nouns and 30% on verbs on one of its many settings. The reference here was the union of WordNet and the Roget's, which places it somewhere between WordNet and Moby with respect to coverage.

A different setting is resampled evaluation, where a classifier is trained and tested on a set of word pairs with an priori ratio of synonyms and non-synonyms. It is only relevant if a good preselection method allows one to reach the assumed proportions of synonyms in the training and test sets [17]. Our results could actually be considered as an input to such methods.

Taken alone, distributional similarities in [18] show results that are comparable to ours or better on Moby, but slightly lower on WordNet. His test set is larger, and split differently with respect to word frequencies. His results are lower than what we obtain here with Lin's own data (as we noticed also about [7]), so we can assume that our comparison is representative with respect to the distributional approach and is a fair comparison.

Mirror translations thus reach comparable or better results than distributional similarity and alignment similarities for synonyms in English, and we have shown that the different methods can be usefully combined in a simple way. Besides,

mirror translations are simpler to compute than the best similarities between $n \times n$ alignment or cooccurrent vectors, where n is the size of the vocabulary.

As a secondary evaluation, authors often use TOEFL synonymy tests [19, 6] where the task is to distinguish the synonym of a given word in a given context, among four candidate items. This is a sort of easier word disambiguation test where the task is to separate a word from unrelated distractors, instead of distinguishing between word senses. We are planning to test the mirror translations against such available benchmarks in a near future. Another way to evaluate the relevance of similarity measures between words is derived from the data collected by [20] where humans are asked to judge the similarity or relatedness of items on a scale. This is an interesting way of providing an intrinsic evaluation of these associations, but it covers only a very limited part of the vocabulary (about 300 words, with only a couple of associations for each).

6 Conclusion

Our different experiments confirm the variety of lexical associations one can find for word paired with so-called semantic similarity measures. While the mirror and the distributional approaches we considered in this work both seem correlated to the references considered, our objective is to be able to pinpoint more precise lexical functions, as they are needed for different tasks (paraphrase substitution, translation lexical choice, etc). With respect to synonyms, our experiments indicate that mirror translations provide a better filter than syntactic distribution similarity. While alignment data have been less studied as a source of similarity than syntactic distributions, we hope we succeeded in showing that they are worth the investigation. We also note that finding mirrors is computationally simpler than finding the better similarities between alignment or distributional vectors, the latter method being the closest in spirit to our approach.

Our longer-term objective is to reproduce synonymy word pair supervised classification; any similarity alone scores quite low as a synonymy descriptor, but experiments, such as [17], show it is doable to reliably label word pairs with lexical functions if the proportion of candidates is more balanced than the very low natural proportion, and this means designing a filter as we do here.

The complementarity of resources considered here is still an open question, although we show that intersecting similarities as simply as we did here is already providing some gain in precision. A more interesting path is probably to combine this with pattern-based approaches, either as another filter or to help selecting productive patterns to start with. The main problem for word similarity measures based on any kind of distribution regularity remains to deal with polysemy, especially when different senses have very different frequency use. Lastly, we plan to investigate the use of multiple language pairs to improve the precision of the predictions of the mirror approach.

References

1. Edmonds, P., Hirst, G.: Near-Synonymy and lexical choice. Computational Linguistics **28**(2) (2002) 105–144
2. Michiels, A., Noel, J.: Approaches to thesaurus production. In: Proceedings of Coling'82. (1982)
3. H.Kozima, Furugori, T.: Similarity between words computed by spreading activation on an english dictionary. In: Proceedings of the conference of the European chapter of the ACL. (1993) 232–239
4. Niwa, Y., Nitta, Y.: Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In: Proceedings of Coling 1994. (1994)
5. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of Coling 1998. Volume 2., Montreal (1998) 768–774
6. Freitag, D., Blume, M., Byrnes, J., Chow, E., Kapadia, S., Rohwer, R., Wang, Z.: New experiments in distributional representations of synonymy. In: Proceedings of CoNLL. (2005) 25–32
7. van der Plas, L., Tiedemann, J.: Finding synonyms using automatic word alignment and measures of distributional similarity. In: Proceedings of the COLING/ACL Poster Sessions. (2006) 866–873
8. Wu, H., Zhou, M.: Optimizing synonyms extraction with mono and bilingual resources. In: Proceedings of the Second International Workshop on Paraphrasing, Sapporo, Japan, Association for Computational Linguistics (2003)
9. Dyvik, H.: Translations as semantic mirrors: From parallel corpus to wordnet. In: The Theory and Use of English Language Corpora, ICAME 2002. (2002)
10. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. (2005)
11. Zhitomirsky-Geffet, M., Dagan, I.: Bootstrapping distributional feature vector quality. Computational Linguistics **35**(3) (2009) 435–461
12. Weeds, J.E.: Measures and Applications of Lexical Distributional Similarity. PhD thesis, University of Sussex (2003)
13. Barzilay, R., McKeown, K.R.: Extracting paraphrases from a parallel corpus. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. (2001)
14. Lin, D., Zhao, S., Qin, L., Zhou, M.: Identifying synonyms among distributionally similar words. In: Proceedings of IJCAI'03. (2003) 1492–1493
15. Curran, J.R., Moens, M.: Improvements in automatic thesaurus extraction. In: Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition. (2002) 59–66
16. Lonneke, P., Tiedemann, J., Manguin, J.: Automatic acquisition of synonyms for French using parallel corpora. In: Proceedings of the 4th International Workshop on Distributed Agent-Based Retrieval Tools. (2010)
17. Hagiwara, M., Ogawa, Y., Toyama, K.: Supervised synonym acquisition using distributional features and syntactic patterns. Journal of Natural Language Processing **16**(2) (2009) 59–83
18. Ferret, O.: Testing semantic similarity measures for extracting synonyms from a corpus. In: Proceeding of LREC. (2010)
19. Turney, P.: A uniform approach to analogies, synonyms, antonyms, and associations. In: Proceedings of Coling 2008. (2008) 905–912
20. Miller, G., Charles, W.: Contextual correlates of semantic similarity. Language and Cognitive Processes **6**(1) (1991) 1–28