Synonym Acquisition across Domains and Languages

Lonneke van der Plas, Jörg Tiedemann, and Jean-Luc Manguin

Abstract. We describe an approach to the automatic extraction of synonyms that is easy to port across domains and across languages. The approach relies on automatic word alignments in parallel texts and uses distributional methods to compute the semantic similarity of words based on these word alignments. As a result the system outputs ranked lists of candidate synonyms for a given word. We apply the method to French, a language for which an extensive electronic synonym dictionary is available, that serves to evaluate the method. We compare the performance with a system that uses syntactic contexts to acquire synonyms automatically. We show that the alignment-based method outperforms the syntactic method by a large margin. In addition, we show that we can adapt to the domain of colloquial language use by replacing the parallel corpus with one that contains a lot of conversational speech: a corpus of movie subtitles. Furthermore, we apply the method to another language, Dutch, with similar performances.

1 Introduction

Support for semantics has been mentioned as one of the goals of next generation information retrieval tools. Synonymy is a type of lexico-semantic knowledge that helps to overcome the so-called terminological gap for tasks such as information retrieval, information extraction, and question answering. Imagine a French student

Lonneke van der Plas
University of Geneva, Switzerland
e-mail: lonneke.vanderplas@unige.ch

Jörg Tiedeman
Uppsala University, Sweden
e-mail: jorg.tiedemann@lingfil.uu.se

Jean-Luc Manguin
CNRS/University of Caen, France
e-mail: jean-luc.manguin@unicaen.fr

V. Pallotta, A. Soro, and E. Vargiu (Eds.): Advances in DART, SCI 361, pp. 41–57.
springerlink.com

searching for a job types *cherche boulot* into a search engine. The student might be ignorant to the fact that the word *boulot* is a colloquial term for synonyms such as *travail, poste*, while these latter terms (*travail, poste*) are used in the majority of job announcements. Hence, simple word matching will not retrieve the information needed by the student. Resources that group French synonyms could help this user in fulfilling his/her information need.

Synonym dictionaries are a common source of semantic information that could be used to deal with the problem described above. However, the drawback of dictionaries is that they are static and based on common knowledge, and therefore not personalised. Personalisation and context awareness have been mentioned as goals to improve information retrieval tools in addition to support for semantics. Providing domain-dependent lexical information is a first step towards context-aware search engines. Search engines need to know when to relate a word like *bank* with the establishment for the custody of money (in the financial domain, for example) and when to relate it to the shore of a river. There are domain-specific dictionaries available, but the number of domains covered is limited.

Automatic methods for synonym acquisition are more flexible and therefore more easily adjustable to emerging needs. For example, work on the acquisition of synonyms using distributional models has shown that syntactic contexts can be applied to any large corpus of text that is analysed syntactically to acquire semantically related words [12, 15] and the method has been applied to corpora from different domains [21] with reasonable success. However, one of the prerequisites for this method is a large parsed corpus or at least a syntactic parser for the target language. For English there are many parsers available but for the majority of languages such tools do not exist. The syntax-based method for the acquisition of synonyms cannot be applied to those languages. A personalised search tool would at least want to serve the user in his/her own language. Therefore, we need automatic methods for synonym acquisition that are easily portable across different languages and that rely as little as possible on language-specific pre-processing.

In this paper we will present a method that is particularly well-suited to be ported across different languages and across different domains. Moreover, the method outperforms the syntax-based approach described above for the task of synonym acquisition.

2 The Distributional Hypothesis

Before we move to describing the methodology we need to explain the hypothesis that underlies our work, the distributional hypothesis. It states that semantically related words are distributed similarly over contexts [11]. In other words, you can grasp the meaning of a word by looking at its contexts.

Context can be defined in many ways. Previous work has been mainly concerned with the syntactic contexts a word is found in. For example, the verbs that are in a subject relation with a particular noun form a part of its context. These contexts can be used to determine the semantic relatedness of words. For instance, words that

occur in a object relation with the verb *to drink* have something in common: they are liquid.

With the advent of multilingual parallel corpora, yet another type of context has been born, the multilingual, translational context. Moreover, tools from the machine translation community such as automatic word alignment tools, that we will discuss in more detail in the next section, have opened the way to rather precise multilingual contexts on the word level. We can extract a list of probable translations for a word in the languages included in the parallel corpus from the output of word-alignment tools. In addition, we can compute the number of times the tool has found a given translation pair. The translational context of a word is composed of the set of translations it gets in other languages. For example, the translational context of *cat* is *kat* in Dutch and *chat* in French. How do we get from translational contexts to synonymy? The idea is that words that share a large number of translations are similar. For example both *autumn* and *fall* get the translation *herfst* in Dutch, *Herbst* in German, and *automne* in French. This indicates that *autumn* and *fall* are synonyms.

Bilingual dictionaries are another source of translations of words. Although dictionary information is very precise and less noisy than the translational contexts automatically acquired from multilingual parallel corpora, there are several advantages associated with automatically extracted translational contexts. Dictionaries are not always publicly available for all languages. Dictionaries are static and often incomplete resources, and they do not provide frequency information. For the acquisition of translational contexts, any multilingual parallel corpus can be used. It is thus possible to focus on a special domain. Furthermore, the automatic alignment provides us with frequency information for every translation pair, useful for handling ambiguity.

3 Translational Context

We rely on automatic word alignment in parallel corpora to find the most probable translation pairs.

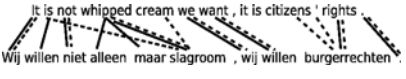

Fig. 1 Example of bidirectional word alignments of two parallel sentences

Figure 1 illustrates the automatic word alignment between a Dutch and an English phrase as a result of using the IBM alignment models [4] implemented in the open-source tool GIZA++ [19]. The alignment of two texts is bidirectional. The Dutch text is aligned to the English text and vice versa (dotted lines versus continuous lines). The alignment models produced are asymmetric. Several heuristics exist to combine directional word alignments. The intersection heuristic, for example, only accepts translation pairs that are found in both directions.

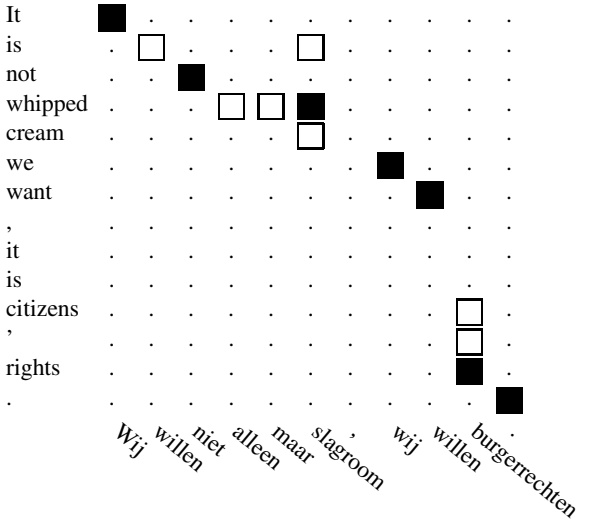

Fig. 2 A word alignment matrix

In Figure 2 we see the intersection of links illustrated by filled boxes. Additional alignment points from the union of links are shown as empty boxes.

Translational contexts are used to find distributionally similar words. We give some examples of translational co-occurrence vectors in Table 1. Every cell in the vector refers to a particular translational co-occurrence type. For example, *chat* 'cat' gets the translation *Katze* in German. The value of these cells indicate the number of times the co-occurrence type under consideration is found in the corpus.

Each co-occurrence type has a cell frequency. Likewise each head term has a row frequency. The row frequency of a certain head term is the sum of all its cell frequencies. In our example the row frequency for the term *chat* 'cat' is 60. Cut-offs for cell and row frequency can be applied to discard certain infrequent co-occurrence types or head terms respectively.

Table 1 Translational co-occurrence vector for *poste* 'job' *boulot* 'job', and *chat* 'cat' based on four languages

	Arbeit-DE	baan-NL	lavoro-IT	job-EN	cat-EN	Katze-DE
poste	17	26	8	13	0	0
boulot	6	12	7	10	0	0
chat	0	0	0	0	26	34

For comparison, an example of a syntax-based co-occurrence vector is given in Table 2.

Table 2 Syntactic co-occurrence vector for *chat*

	avoir_obj 'have_obj'	nourrir_obj 'feed_obj'	noir_adj 'black_adj'	pliant_adj 'folding_adj'
kat 'cat'	50	10	5	1

3.1 Measures for Computing Similarity

The more similar the vectors are, the more distributionally similar the head terms are. We need a way to compare the vectors for any two head terms to be able to express the similarity between them by means of a score. Various measures can be used to compute the distributional similarity between terms. We will explain in section 4 what measures we have chosen in the current experiments.

Furthermore, it has been shown that distributional methods benefit from using feature weights. For example in syntax-based approaches selectionally weak [27] or *light* verbs such as *hebben* 'to have' are given a lower weight than a verb such as *uitpersen* 'squeeze' that occurs less frequently. We have used weights for the translational context to counter balance the alignment errors that often occur with frequent words.

3.2 Related Work

Multilingual parallel corpora have been used for tasks related to word sense disambiguation such as target word selection [9] and separation of senses [28, 10, 13].

Automatic acquisition of paraphrases using multilingual corpora is discussed in [30, 2, 5], of which only the last two are based on automatic word alignment.

[2] use a method that is rooted in phrase-based statistical machine translation. Translation probabilities provide a ranking of candidate paraphrases. These are refined by taking contextual information into account in the form of a language model. The Europarl corpus [14] is used. A precision of 55.3% is reached when using context information. A precision score of 55% is attained when using multilingual data. Manual alignment improves the performance by 26%. In a more recent publication, [5] improved this method by using syntactic constraints and multiple languages in parallel.

Improving the syntax-based approach for synonym identification using bilingual dictionaries and parallel corpora has been discussed in [16], [34], [23], and [25].

[16] try to tackle the problem of identifying synonyms in lists of nearest neighbours in two ways: Firstly, they look at the overlap in translations of semantically similar words in multiple bilingual dictionaries. Secondly, they design specific patterns designed to filter out antonyms. They evaluate on a set of 80 synonyms and 80 antonyms from a thesaurus that are also found among the top-50 distributionally similar words of each other. The pattern-based method results in a precision of 86.4 and a recall of 95.0. The method using bilingual dictionaries gets a higher precision score (93.9). However, recall is much lower: 39.2.

[34] report an experiment on synonym extraction using bilingual resources (an English-Chinese dictionary and corpus) as well as monolingual resources (an English dictionary and corpus). Their monolingual corpus-based approach is very similar to our monolingual corpus-based approach. The bilingual approach is different from ours in several aspects. Firstly, they do not take the corpus as the starting point to retrieve word alignments. They use the bilingual dictionary to retrieve multiple translations for each target word. The corpus is only employed to assign probabilities to the translations found in the dictionary. The authors praise the method for being able to find synonyms that are not in the corpus as long as they are found in the dictionary. However, the drawback is that the synonyms are limited to the coverage of the dictionary. The aim of automatic methods in general is precisely to overcome the limited coverage of such resources. A second difference with our system is the use of a bilingual parallel corpus whereas we use a multilingual corpus containing 11 languages in total. The authors show that the bilingual method outperforms the monolingual methods both in recall and precision. However, a combination of different methods leads to the best performance. A precision of 27.1 on middle-frequency nouns is attained.

In [23] the distributional alignment-based method is introduced. A comparison is made between the alignment-based method and the syntax-based method for Dutch synonym acquisition. The alignment-based method outperforms the syntax-based method. The setup in section 8 of this article does not follow [23], but experiments undertaken in [22], because the evaluation framework in [22] is more carefully designed. The latter includes a subsubsection (Section 4.5.4) on comparing different corpora for the acquisition of synonyms for Dutch.

In [25] the syntax-based and alignment-based method for French synonym acquisition on the general domain are compared. The results in Section 6 are taken from this study.

4 Materials and Methods

In the following subsections we describe the data collection we used and the similarity measure and weighting function we chose.

4.1 Data Collection

We need a parallel corpus of reasonable size with French either as source or as target language. Furthermore, we would like to experiment with various languages aligned to French. The freely available Europarl corpus [14] includes 11 languages in parallel, it is sentence aligned [32], and it is of reasonable size. Each language contains around 1 million sentences, which corresponds roughly to 28 million words. Thus, for acquiring French synonyms we have 10 language pairs with French as source language: Danish (DA), German (DE), Greek (EL), English (EN), Spanish (ES), Finnish (FI), Dutch (NL), Italian (IT), Portuguese (PT), and Swedish (SV). We applied a lemmatiser [29] to the French part of the language pairs in order to 1) reduce

data sparseness, and 2) to facilitate our evaluation based on comparing our results to existing synonym databases.

Context vectors are populated with the links to words in other languages extracted from automatic word alignments. We applied GIZA++ and the intersection heuristic as explained in section 3. From the word-aligned corpora we extracted translational co-occurrence types, pairs of source and target words in a particular language with their alignment frequency attached. Each aligned target word is a feature in the (translational) context of the source word under consideration. We removed word type links that include non-alphabetic characters to focus our investigations on real words and we transformed all characters to lower case.

In Section 3, we explained that each headword has a corresponding row frequency, and each cooccurrence type has a cell frequency. Cut-offs for cell and row frequency can be applied to discard certain infrequent and often less reliable co-occurrence types or head terms respectively. We applied a cell and row frequency cutoff of 4 and 10 respectively, because these settings performed well in previous work [25].

Note that we rely entirely on automatic processing of our data. Thus, the results from automatic tagging, lemmatisation and word alignment include errors.

4.2 Comparing Vectors

To estimate the similarity of words by means of their associated translational co-occurrence vectors we need a similarity measure. We explained in 3.1 that some attributes contain more information than other attributes. We want to account for that using a weighting function, that will modify the cell values.

We have limited our experiments to using Dice†, a variant of Dice as our similarity measure and and Pointwise mutual information (MI, [6]) as weight. Dice†[1] is defined as:

$$Dice\dagger = \frac{2\sum min(weight(W1,*_r,*_{w'}),weight(W2,*_r,*_{w'}))}{\sum weight(W1,*_r,*_{w'}) + weight(W2,*_r,*_{w'})}$$

We describe the functions using an extension of the notation used by [15], adapted by [7]. Co-occurrence data is described as tuples: ⟨word, language, word′⟩, for example, ⟨chat, EN, cat⟩.

Asterisks indicate a set of values ranging over all existing values of that component of the relation tuple. For example, $(w,*,*)$ denotes for a given word w all translational contexts it has been found in any language. For the example of *chat*, this would denote all values for all translational contexts the word is found in: *Katze_DE*:17, *chat_FR*:26 etc. There is a placeholder for the weighting function: *weight*.

Pointwise mutual information (MI) measures the amount of information one variable contains about the other. MI is computed as follows:

[1] Note that Dice † gives the same ranking as the well-known Jaccard measure, i.e. there is a monotonic transformation between their scores. Dice † is easier to compute and therefore the preferred measure [8].

$$MI = log \frac{P(w,r,w')}{P(w,*,*)P(*,r,w')}$$

Here, $P(w,r,w')$ is the probability of seeing *chat* aligned to *the* in a French-English parallel corpus, and $P(w,*,*)P(*,r,w')$ is the product of the probability of seeing *chat* aligned to any word in the corpus and the probability of seeing *the* aligned to any word in the corpus.

5 Evaluation

There are several evaluation methods available to assess lexico-semantic data. [7] distinguishes several. We decided to compare against a gold standard, because there is a large French synonym dictionary available. We evaluated our results on the *Dictionnaire Electronique des Synonymes* (DES, [26]), which is based on a compilation of seven French synonym dictionaries. It contains 49,149 nodes connected by 200,606 edges that connect synonymous words.

We compare our results to the syntax-based method for French by [3]. They present an explorative study of using distributional similarity to extract synonyms for French. They use two corpora: a 200 million-word corpus of newspaper text from *Le Monde*, and a 30 million-word corpus consisting of 515 twentieth century novels. Several syntactic relations are extracted.

The test set was chosen by looking at the pairs of candidate synonyms resulting from the syntax-based method that receive a score not lower than 0.16. This resulted in a list of approximately 1000 nouns. Of this list 950 can be found in the data of the alignment-based method. This list of 950 word constitutes the test set.

6 Results and Discussion

The results can be seen in Figure 3 and Figure 4. The x-axis indicates the threshold we set for the similarity score. For each candidate synonym the system calculates a similarity score. The dot at 0.20 in Figure 3 shows the average precision of all candidate synonyms for all testwords that have a similarity score of 0.20 or higher. Precision and recall are calculated as well as the coverage of the system at varying thresholds. Coverage indicates for how many of the testwords the system finds a synonym at the given similarity threshold.

Coverage of both systems decreases when the threshold for the similarity score is augmented. That is expected since not many words have candidate synonyms with a high similarity score. The alignment-based method never reaches 100% coverage. However, it should be noted that the test set was chosen in a way that favours the syntax-based method. The test set is composed of pairs of candidate synonyms resulting from the syntax-based method that are above the threshold 0.16. Thus, the coverage of the syntax-based method is 100% at 0.16. The coverage of the alignment-based method is approximately 70% for that threshold. However, the

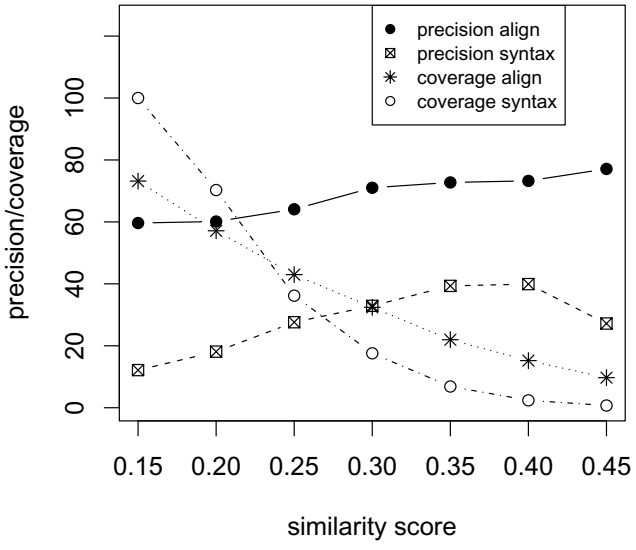

Fig. 3 Precision and coverage for the two methods at several thresholds of similarity score

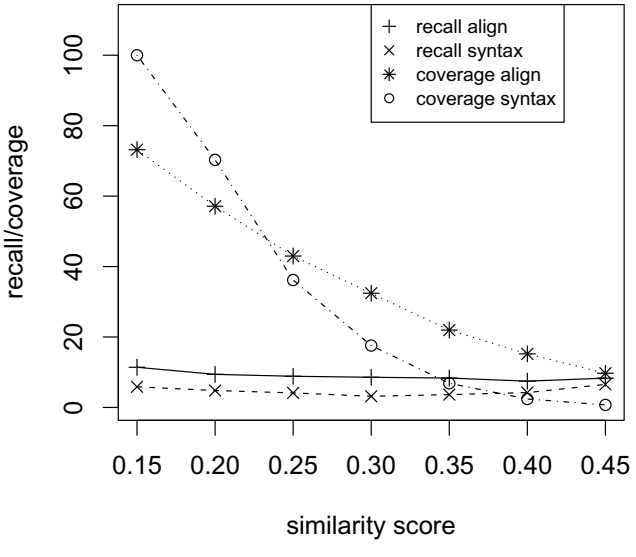

Fig. 4 Recall and coverage for the two methods at several thresholds of similarity score

coverage of the syntax-based method decreases more rapidly as the thresholds are raised. At threshold 0.45 the coverage of the syntax-based method is close to zero.

If we compare the precision of the candidate synonyms for both systems at the same level of coverage (50%) we see that the syntax-based method has a precision score of 25%, whereas the alignment-based method produces candidate synonyms

with a precision of 60% to 65%. The precision of the alignment-based method ranges between a little under 60% to a little under 80% at threshold 0.45. The precision of the syntax-based method ranges between 10% at threshold 0.16 and a little under 40% for threshold 0.4. It is striking that the precision drops at the end of the line, when the threshold is set to 0.45. The candidate synonym with the highest scores are not the best. However, it should be noted that due to limited coverage (close to 0) the numbers at this threshold are unreliable.

With respect to recall, it can be concluded that there is a smaller difference between the two methods and the scores are less satisfactory in general. It should be noted that the dictionaries often include synonyms from colloquial language use. We do not expect to find these synonyms in the Europarl corpus. We will see in the next section that we find more synonyms from colloquial language use when we change the parallel corpus we collect our data from.

A closer inspection of the candidate synonyms resulting from the alignment-based method shows that many of the candidate synonyms judged incorrect are in fact valuable additions, such as *sinistre* 'disaster' for *accident* 'accident'.

Many errors stem from the fact that the alignment-based method does not take multiword units into account. For the French data this typically results in many related adjectives and adverbs being selected as candidate synonyms. For example, *majoritaire* 'majority (adj)' is returned as a synonym for *majorité* 'majority (noun)', stemming from the multiword unit *parti majoritaire*. Also *majoritairement* and *largement* are among the candidate synonyms. Words that would be translated in English as *for the most part*. These translations that are composed of multiple words cause problems for the alignment method and hence for the synonyms extracted. In [24] we propose a adaptation of the alignment-based method that uses standard phrase extraction techniques commonly used in statistical machine translation to handle multi-word terms.

Furthermore, we have to stress that we rely entirely on automatic processing of our data. [2] show that when using manual alignment the percentage of correct paraphrases significantly rises from 48.9% to 74.9%.

7 Porting to a Different Domain

We mentioned in Section 5 that the synonym sets acquired from the proceedings of European Parliament do not contain many colloquial terms, whereas the French synonym dictionary does. Recall the example in the introduction of this article, the student was looking for a job and typed the colloquial term *boulot* in the interface of the search engine.

To be able to find synonyms typical for every day language use, we need to give the system access to corpora of every day language use. The conversations in a large variety of movies are probably closer to the every day language use of an average French student than the proceeding of the European parliament. Instead of extracting translation pairs from the parallel corpus Europarl, we used a multilingual parallel corpus of movie subtitles, the OpenSubs corpus[31].

This method allows us to acquire synonyms that are specific to any particular domain, for example, in [24] we extended the alignment-based method to find medical term variants.

7.1 Materials and Methods

The OpenSubs corpus contains about 21 million aligned sentence fragments in 29 languages. We used all language pairs that include French, 23 language pairs in total. Still the corpus is much smaller than the Europarl corpus. The number of translation pairs (hapaxes excluded) we extract from this corpus is more than 50 times smaller than the number of translation pairs extracted from Europarl.

The domain is different from the domain of the Europarl corpus. There is a world of difference between the working day of a member of the European Parliament and the adventures of Nemo. Moreover, movie subtitles consist mainly of transcribed speech. In principle this is the same for the Europarl corpus. However these proceedings are edited and far less spontaneous than the speech data from the movies.

Another difference is the amount of pre-processing we applied. For the Europarl corpus we had access to lemma information for the words because, as we explained, the corpus was lemmatised. For the OpenSubs corpus we did not have access to lemma information and thus used the words instead. These experiments are therefore a good testbed for the usefulness of the method when there is limited data available and limited resources, as is the case for many resource-poor languages. We again applied a cell and row frequency cutoff of four and ten.

7.2 Evaluation of the Synonyms from the OpenSubs Corpus

Because the OpenSubs corpus has not been lemmatised, the resulting synonyms have all kinds of inflections. The synonym dictionary we use for the evaluation contains lemmas. This means that if we evaluate the synonyms on the dictionary it would not recognise the inflected wordforms (plurals and feminine forms) and count them as incorrect. We therefore converted every plural form to its singular form, using the Morphalou database from the : *Centre National de Ressources Textuelles et Lexicales*[2] and we converted feminine forms to their masculine counterpart. We then removed the doubles from the lists of candidate synonyms and ran the evaluations using the synonym dictionary.

Note that despite this pre-processing, the acquisition of synonyms from the non-lemmatised corpus of subtitles is a more difficult task than the acquisition of synonyms from the lemmatised Europarl corpus. Lemmatisation does not only make evaluation on the dictionary easier, it also reduces data sparseness. For a small corpus such as the OpenSubs corpus data sparseness is more problematic than for larger corpora. On the other hand this setting illustrates the performance of the alignment-based method in a setting that might be realistic for many languages: only a small amount of data and no lemmatiser is available. In the end, our aim is to use as little

[2] The resource is available from `http://www.cnrtl.fr/lexiques/morphalou/`

language-specific pre-processing as possible to assure a wide applicability of the method.

7.3 Results for the OpenSubs Corpus

The main reason for applying the method to a corpus from another domain is to find synonyms that are typical for that domain. From the examples given in Table 3 we get the impression that the synonyms stemming from the OpenSubs corpus contain more slang and colloquial language use, such as *nana* 'babe' as a candidate synonym for *fille* 'girl'. At the same time we see that the proceedings of the European parliament constitute a specific domain as well. In the context of the European parliament a friend is like a comrade and an ally, synonyms very specific to the the particular domain, whereas in movies friends are pals and buddies. Similarly, the adjective *malade* for which the default translation would be 'ill', gets synonyms from the OpenSubs corpus like *fou* 'mad' and *dingue* 'crazy'

Table 3 Examples of candidate synonyms at the top-3 ranks for two corpora

Testword	Corpus			
ami	Europarl	amitié	camarade	allié
'friend'		'friendship'	'comrade'	'ally'
	Subtitles	copain	pote	amie
		'pal'	'buddy'	'girlfriend'
fille	Europarl	fillette	enfant	filial
'girl'		'small girl'	'child'	'relative to a daughter'
	Subtitles	fillette	nana	fiiie
		'small girl'	'babe'	(mistake in optical character recognition)
malade	Europarl	patient	maladie	souffrant
'ill'		'patient'	'illness'	'suffering'
	Subtitles	souffrant	fou	dingue
		'suffering'	'mad'	'crazy'

We also calculated precision, recall, and coverage for the synonyms acquired from the OpenSubs corpus as can be seen in Figure 5. The scores are not as good as when using the Europarl corpus. We expect that using the subtitle corpus leads to poorer performance. There are at least three reasons for this. The first reason is that the corpus is smaller. The second reason is that the subtitle corpus is more noisy than the Europarl corpus, for example due to mistakes in optical character recognition, as we see in Table 3. The last is the absence of lemmatisation. Still, despite the small amount of data, the noise and the absence of language-specific pre-processing the system is able to find synonyms for more than 20% of the words in the testset with a precision of around 40% at the lowest threshold, which is promising. It means that the method can be applied to any language and any domain for which there exists a relatively small, possibly noisy, parallel corpus, without the need of language-specific pre-processing. This is good news from the point of view of sense induction

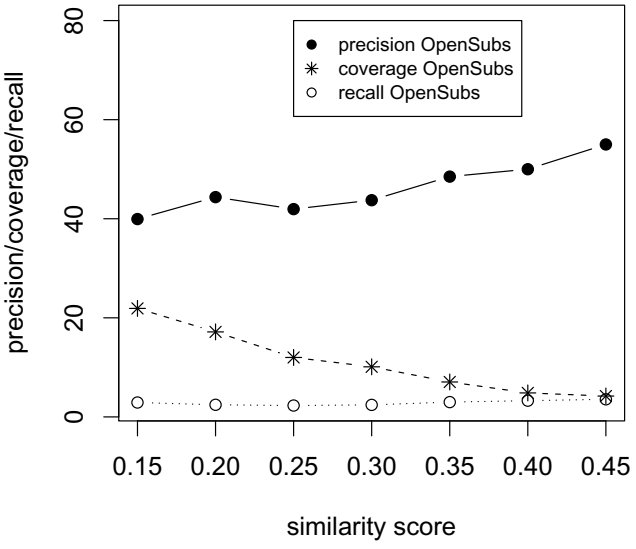

Fig. 5 Precision, recall and coverage for the synonyms stemming from the OpenSubs corpus

as well. The several senses a word has are often limited by a specific domain. We can deduce from the synonyms the word *malade* gets in Table 3 that in the context of the proceedings of the European Parliament the word gets the sense ill, whereas in the context of movie subtitles another sense of the word *malade*, namely the sense crazy, is apparent as well.

8 Porting to a Different Language

We explained in the introduction that the method easily ports to different languages. As an example we include results of applying the alignment-based method to Dutch taken from [21].

8.1 Materials and Methods

Instead of collecting translations for all French words, we collected translations for all Dutch words in the same parallel corpus used to acquire the French synonyms, the Europarl corpus. We post-processed the alignment results in various ways for Dutch. We applied a simple lemmatiser to the Dutch part of the bilingual translational co-occurrence types. For this we used two resources: CELEX, a linguistically annotated dictionary of English, Dutch, and German [1], and the Dutch snowball stemmer implementing a suffix-stripping algorithm based on the Porter stemmer. We removed word type links that include non-alphabetic characters and we transformed all characters to lower case.

We restricted our study to Dutch nouns. Hence, we extracted translational co-occurrence types for all words tagged as nouns in CELEX. We also included words that are not found in CELEX[3].

In the Dutch experiments we also compared the alignment-based method to a syntax-based method. The data for the syntax-based method comprises 500 million words of Dutch newspaper text: the Twente Nieuws Corpus (TwNC, [20]) that is parsed automatically using the Alpino parser [18]. The result of parsing a sentence is a dependency graph according to the guidelines of the Corpus of Spoken Dutch [17].

From these dependency graphs, we extracted tuples consisting of the (non-pronominal) head of an NP (either a common noun or a proper name), the dependency relation, and either (1) the head of the dependency relation (for the object, subject, and apposition relation), (2) the head plus a preposition (for NPs occurring inside PPs which are prepositional complements), (3) the head of the dependent (for the adjective and apposition relation) or (4) the head of the other elements of a coordination (for the coordination relation).

For these experiments the well-known Cosine measure was used instead of the Dice† measure. The overall performances of the more standard syntax-based method were highest when using Cosine, so comparative evaluations in [21] were done using Cosine. For the same reason cell and row cutoffs were set to two.

Evaluations for Dutch were done on a large test set of 3000 nouns selected from Dutch EuroWordNet. The test set is split up in equal amounts (1000) of high-frequency (HF), middle-frequency (MF) and low-frequency (LF) words. This was done to be able to study the effect of frequency on the performance of the system. We used the synsets in Dutch EuroWordnet [33] for the evaluation of the proposed synonyms.

8.2 Results on Synonym Extraction for Dutch

In Table 4 we see the comparative results on Dutch. The percentage of synonyms for the top-k candidate synonyms is given for the two methods and the three testsets. We see that 31.71% of the candidate synonyms produced by the alignment-based method as the most probable candidate synonym for the words in the high-frequency testset are indeed synonyms, whereas the syntax-based method only finds 21.31%.

Table 4 Percentage of synonyms over the k candidates for the alignment-based and syntax-based method for the three frequency bands

Method	HF		MF		LF	
	k=1	k=5	k=1	k=5	k=1	k=5
Alignment-based	31.71	19.16	29.26	16.20	28.00	16.22
Syntax-based	21.31	10.55	22.97	10.11	19.21	11.63

[3] Discarding these words would result in losing too much information. We assumed that many of them will be productive noun constructions.

The performance of the syntax-based method decreases rapidly when we go down the list of candidate synonyms, i.e. at higher values of k. For the high-frequency test set this is most apparent: From k=1 to k=5 the syntax-based method precision score is halved. At k=5 the alignment-based method receives still 2/3rd of the score at k=1. The syntax-based method retrieves about 2/3rd of the synonyms the alignment-based method retrieves for the high-frequency test set.

Although the results are less clear-cut than the results for french, they show a similar pattern. In spite of data sparseness, it is clear from Table 4 that the alignment-based method is better at finding synonyms than the syntax-based method.

Differences in performance between the experiments on Dutch and on French can be partly explained by the difference in the goldstandards used in the evaluation. The French dictionary contains 49,149 nodes connected by 200,606 synonym edges. EWN contains a total of 56,283 entries. However, the degree of synonymy is higher. For EWN the degree of synonymy is expressed by the ratio of senses per synset: 1.59. The ratio of edges per entry in the case of the DES is 4. A high degree of synonymy favours high precision scores.

In an evaluation with human judgements [23] showed that in 37% of the cases the majority of the subjects judged the synonyms proposed by the system to be correct even though they were not found to be synonyms in Dutch EuroWordnet. The results from human judgements lead us to believe that the method performs better than the scores in Table 4 indicate. Over and above, this indicates that we are able to extract automatically synonyms that are not yet covered by available resources.

Of course the differences in nature of the two languages also play their roles. Dutch uses single-word compounding, contrary to the majority of languages that use multiple words to describe a concept. For example, in English, compounds are mostly composed of two words orthographically, e.g *table cloth* and *hard disk* versus *database*. The Dutch word *slagroom* 'whipped cream' is a single word. These single-word compounds introduce errors in the word alignments, where *slagroom* is attached to either *whipped* or *cream*. French behaves like English in this respect.

9 Conclusions

We have shown that the alignment-based method outperforms the traditional syntax-based method for the task of automatic synonym acquisition by a very large margin on the task of French synonym acquisition. The precision is more than twice as high for the alignment-based method and it manages to find valuable additions not present in the large synonym dictionary on which it was evaluated. In addition, we showed that the method can be easily ported across languages and domains. We showed that we can retrieve synonyms of a different nature by using a corpus of movie subtitles instead of a corpus that consists of proceedings from the European parliament. The method can be easily adapted to a specific domain. Moreover, the method works reasonably well, even when using small, noisy corpora, and no language-specific pre-processing. This opens the way to the acquisition of synonyms for specialised domains and resource-poor languages. The method can be easily

applied to other languages and results in similar performances. It compares favourably to the syntax-based methods for the acquisition of Dutch synonyms as is the case for French synonym acquisition.

Acknowledgements

Part of this work has received funding from the EU FP7 programme (FP7/2007-2013) under grant agreement nr 216594 (CLASSIC project: www.classic-project.org). It is based on research carried out in the project *Question Answering using Dependency Relations*, which is part of the research program for *Interactive Multimedia Information eXtraction*, IMIX, financed by NWO, the Dutch Organisation for Scientific Research.

References

1. Baayen, R., Piepenbrock, R., van Rijn, H.: The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, Philadelphia (1993)
2. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the annual Meeting of the Association for Computational Linguistics, ACL (2005)
3. Bourigault, D., Galy, E.: Analyse distributionnelle de corpus de langue générale et synonymie. In: Lorient, Actes des Journées de la Linguistique de Corpus, JLC (2005)
4. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19(2), 263–296 (1993)
5. Callison-Burch, C.: Syntactic constraints on paraphrases extracted from parallel corpora. In: Proceedings of EMNLP (2008)
6. Church, K.W., Hanks, P.: Word association norms, mutual information and lexicography. In: Proceedings of the Annual Conference of the Association of Computational Linguistics, ACL (1989)
7. Curran, J.: From distributional to semantic similarity. Ph.D. thesis, University of Edinburgh (2003)
8. Curran, J.R., Moens, M.: Improvements in automatic thesaurus extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 222–229 (2002)
9. Dagan, I., Itai, A., Schwall, U.: Two languages are more informative than one. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL (1991)
10. Dyvik, H.: Translations as semantic mirrors. In: Proceedings of Workshop Multilinguality in the Lexicon II (ECAI) (1998)
11. Harris, Z.S.: Mathematical structures of language. Wiley, Chichester (1968)
12. Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the Annual Meeting of the Association of Computational Linguistics, ACL (1990)
13. Ide, N., Erjavec, T., Tufis, D.: Sense discrimination with parallel corpora. In: Proceedings of the ACL Workshop on Sense Disambiguation: Recent Successes and Future Directions (2002)

14. Koehn, P.: Europarl: A multilingual corpus for evaluation of machine translation (2003)
15. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of COLING/ACL (1998)
16. Lin, D., Zhao, S., Qin, L., Zhou, M.: Identifying synonyms among distributionally similar words. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2003)
17. Moortgat, M., Schuurman, I., van der Wouden, T.: CGN syntactische annotatie, Internal Project Report Corpus Gesproken Nederlands (2000), http://lands.let.kun.nl/cgn
18. van Noord, G.: At last parsing is now operational. In: Actes de la 13eme Conference sur le Traitement Automatique des Langues Naturelles (2006)
19. Och, F.: GIZA++: Training of statistical translation models (2003), http://www.isi.edu/~och/GIZA++.html
20. Ordelman, R.: Twente nieuws corpus (TwNC). Parlevink Language Techonology Group. University of Twente (2002)
21. van der Plas, L.: Automatic lexico-semantic acquisition for question answering. Groningen dissertations in linguistics (2008)
22. van der Plas, L.: Automatic lexico-semantic acquisition for question answering. Ph.D. thesis, University of Groningen (2008)
23. van der Plas, L., Tiedemann, J.: Finding synonyms using automatic word alignment and measures of distributional similarity. In: Proceedings of COLING/ACL (2006)
24. van der PLas, L., Tiedemann, J.: Finding medical term variations using parallel corpora and distributional similarity. In: Proceedings of the Coling Workshop on Ontologies and Lexical Resources (2010)
25. van der Plas, L., Tiedemann, J., Manguin, J.L.: Extraction de synonymes à partir d'un corpus multilingue aligné. Actes des 5èmes Journées de Linguistique de Corpus à Lorient (2008)
26. Ploux, S., Manguin, J.: Dictionnaire électronique des synonymes français (1998, released 2007)
27. Resnik, P.: Selection and information, Unpublished doctoral thesis, University of Pennsylvania (1993)
28. Resnik, P., Yarowsky, D.: A perspective on word sense disambiguation methods and their evaluation. In: Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? (1997)
29. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, pp. 44–49 (1994), http://www.ims.uni-stuttgart.de/~schmid/
30. Shimota, M., Sumita, E.: Automatic paraphrasing based on parallel corpus for normalization. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC (2002)
31. Tiedemann, J.: News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In: Nicolov, N., Bontcheva, K., Angelova, G., Mitkov, R. (eds.) Recent Advances in Natural Language Processing, vol. V, pp. 237–248. John Benjamins, Amsterdam (2009)
32. Tiedemann, J., Nygaard, L.: The OPUS corpus - parallel & free. In: Proceedings of the International Conference on Language Resources and Evaluation, LREC (2004)
33. Vossen, P.: EuroWordNet A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers, Dordrecht (1998)
34. Wu, H., Zhou, M.: Optimizing synonym extraction using monolingual and bilingual resources. In: Proceedings of the International Workshop on Paraphrasing: Paraphrase Acquisition and Applications, IWP (2003)