Finding Medical Term Variations using Parallel Corpora and Distributional Similarity

Lonneke van der Plas
Department of Linguistics
University of Geneva
lonneke.vanderplas@unige.ch

Jörg Tiedemann
Department of Linguistics and Philology
Uppsala University
jorg.tiedemann@lingfil.uu.se

Abstract

We describe a method for the identification of medical term variations using parallel corpora and measures of distributional similarity. Our approach is based on automatic word alignment and standard phrase extraction techniques commonly used in statistical machine translation. Combined with pattern-based filters we obtain encouraging results compared to related approaches using similar data-driven techniques.

1 Introduction

Ontologies provide a way to formally represent knowledge, for example for a specific domain. Ontology building has received a lot of attention in the medical domain. This interest is reflected in the existence of numerous medical ontologies, such as the Unified Medical Language System (UMLS) (McCray and Hole, 1990) with its metathesaurus, semantic network, and specialist lexicon. Although the UMLS includes information for languages other than English, the coverage for other languages is generally smaller.

In this paper we describe an approach to acquire lexical information for the Dutch medical domain automatically. In the medical domain variations in terminology often include multi-word terms such as *aangeboren afwijking* 'birth defect' for *congenitale aandoening* 'congenital disorder'. These multiple ways to refer to the same concept using distinct (multi-word) terms are examples of synonymy[1] but are often referred to as term variations. These term variations could be used to enhance existing medical ontologies for the Dutch language.

Our technique builds on the distributional hypothesis, the idea that semantically related words are distributed similarly over contexts (Harris, 1968). This is in line with the Firthian saying that, 'You shall know a word by the company it keeps.' (Firth, 1957). 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 (Lin, 1998; Curran, 2003). For example, the verbs that are in a subject relation with a particular noun form a part of its context. In accordance with the Firthian tradition 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. Other work has been concerned with the bag-of-word context, where the context of a word are the words that are found in its proximity (Wilks et al., 1993; Schütze, 1992).

Yet another context, that is much less studied, is the translational context. The translational context of a word is the set of translations it gets in other languages. For example, the translational context of *cat* is *kat* in Dutch and *chat* in French. This requires a rather broad understanding of the term context. 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.

[1] Spelling variants are a type of term variations that are not included in the definition of synonymy.

Proceedings of the 6th Workshop on Ontologies and Lexical Resources (Ontolex 2010), pages 28–37,
Beijing, August 2010

A straightforward place to start looking for translational context is in bilingual dictionaries. However, these are not always publicly available for all languages. More importantly, dictionaries are static and therefore often incomplete resources. We have chosen to automatically acquire word translations in multiple languages from text. Text in this case should be understood as multilingual parallel text. Automatic alignment gives us the translations of a word in multiple languages. The so-called *alignment-based distributional methods* described in Van der Plas (2008) apply the translational context for the discovery of single word synonyms for the general domain. Any multilingual parallel corpus can be used for this purpose. It is thus possible to focus on a special domain, such as the medical domain we are considering in this paper. The automatic alignment provides us also with domain-specific frequency information for every translation pair, which is helpful in case words are ambiguous.

Aligned parallel corpora have often been used in the field of word sense discovery, the task of discriminating the different senses words have. The idea behind it is that a word that receives different translations might be polysemous. For example, a word such as *wood* receives the translation *woud* and *hout* in Dutch, the former referring to an area with many trees and the latter referring to the solid material derived from trees. Whereas this type of work is all built upon the divergence of translational context, i.e. one word in the source language is translated by many different words in the target language, we are interested in the convergence of translations, i.e. two words in the source language receiving the same translation in the target language. Of course these two phenomena are not independent. The alleged conversion of the target language might well be a hidden diversion of the source language. Since the English word might be polysemous, the fact that *woud* and *hout* in Dutch are both translated in English by *wood* does not mean that *woud* and *hout* in Dutch are synonyms. However, the use of multiple languages overshadows the noise resulting from polysemy (van der Plas, 2008).

Van der Plas (2008) shows that the way the context is defined influences the type of lexico-semantic knowledge that is discovered. After gold standard evaluations and manual inspection the author concludes that when using translational contexts more tight semantic relations such as synonymy are found whereas the conventional syntax-based approaches retrieve hypernyms, co-hyponyms, and antonyms of the target word. The performance on synonym acquisition when using translational contexts is almost twice as good as when using syntactic contexts, while the amount of data used is much smaller. Van der Plas (2008) ascribed the fact that the syntax-based method behaves in this way to the fact that loosely related words, such as *wine* and *beer*, are often found in the same syntactic contexts. The alignment-based method suffers less from this indiscriminant acceptance because words are typically translated by words with the same meaning. The word *wine* is typically not translated with a word for *beverage* nor with a word for *beer*, and neither is *good* translated with the equivalence of *bad*.

In this paper we are concerned with medical term variations that are in fact (multi-word) synonyms. We will use the translational context to compute similarity between terms. The translational context is not only very suitable to find tight relations between words, the transition from single-word synonyms to multi-word term variations is also straightforward due to advances in phrase-based machine translation. We will use word alignment techniques in combination with phrase extraction techniques from statistical machine translation to extract phrases and their translations from a medical parallel corpus. We combine this approach with Part-of-Speech (PoS) patterns from the term extraction literature to extract candidate terms from the phrase tables. Using similarity measures used in distributional methods we finally compute ranked lists of term variations.

We already noted that these term variations could be used to enhance existing ontologies for the Dutch language. On top of that we believe that the multi-lingual method that uses translations of multi-word terms in several languages could be used to expand resources built for English with translations in other languages (semi-) automatically. This last point falls outside the scope of this paper.

In the following section we will describe the alignment-based approaches to distributional similarity. In section 3 we will describe the methodology we followed in this paper in detail. We describe our evaluation in section 4 and discuss the results in section 5. Section 6 concludes this paper.

2 Alignment-based methods

In this section we explain the alignment-based approaches to distributional similarity. We will give some examples of translational context and we will explain how measures serve to determine the similarity of these contexts. We end this section with a discussion of related work.

2.1 Translational context

The translational context of a word or a multi-word term is the set of translations it gets in other languages. For the acquisition of translations for the Dutch medical terms we rely on automatic word alignment in parallel corpora.

Figure 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 (Brown et al., 1993) implemented in the open-source tool GIZA++ (Och, 2003). The alignment of two texts is bi-directional. 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 which is usually called "symmetrization". In order to cover multi-word terms standard phrase extraction techniques can be used to move from word alignment to linked phrases (see section 3.2 for more details).

2.2 Measures for computing similarity

Translational co-occurrence vectors are used to find distributionally similar words. For ease of reading, we give an example of a single-word term *kat* in Table 1. In our current setting the terms can be both single- or multi-word terms such as *werkzame stof* 'active ingredient'. Every cell in the vector refers to a particular translational co-occurrence type. For example, *kat* '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 *kat* 'cat' is 65. Cut-offs for cell and row frequency can be applied to discard certain infrequent co-occurrence types or head terms respectively.

	DE Katze	FR chat	IT gatto	EN cat	total
kat	17	26	8	13	64

Table 1: Translational co-occurrence vector for *kat* ('cat') based on four languages

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 methods can be used to compute the distributional similarity between terms. We will explain in section 3 what measures we have chosen in the current experiments.

2.3 Related work

Multilingual parallel corpora have mostly been used for tasks related to word sense disambiguation such as separation of senses (Resnik and Yarowsky, 1997; Dyvik, 1998; Ide et al., 2002).

However, taking sense separation as a basis, Dyvik (2002) derives relations such as synonymy and hyponymy by applying the method of semantic mirrors. The paper illustrates how the method works. First, different senses are identified on the basis of manual word translations in sentence-aligned Norwegian-English data (2,6 million words in total). Second, senses are grouped in semantic fields. Third, features are

assigned on the basis of inheritance. Lastly, semantic relations such synonymy and hyponymy are detected based on intersection and inclusion among feature sets .

Improving the syntax-based approach for synonym identification using bilingual dictionaries has been discussed in Lin et al. (2003) and Wu and Zhou (2003). In the latter parallel corpora are also applied as a reference to assign translation likelihoods to candidates derived from the dictionary. Both of them are limited to single-word terms.

Some researchers employ multilingual corpora for the automatic acquisition of paraphrases (Shimota and Sumita, 2002; Bannard and Callison-Burch, 2005; Callison-Burch, 2008). The last two are based on automatic word alignment as is our approach.

Bannard and Callison-Burch (2005) use a method that is also 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 (Koehn, 2005) is used. It has about 30 million words per language. 46 English phrases are selected as a test set for manual evaluation by two judges. When using automatic alignment, the precision reached without using contextual refinement is 48.9%. A precision of 55.3% is reached when using context information. Manual alignment improves the performance by 26%. A precision score of 55% is attained when using multilingual data.

In a more recent publication Callison-Burch (2008) improved this method by using syntactic constraints and multiple languages in parallel. We have implemented a combination of Bannard and Callison-Burch (2005) and Callison-Burch (2008), in which we use PoS filters instead of syntactic constraints to compare our results with. More details can be found in the Section 5.

Apart from methods that use parallel corpora mono-lingual pattern-based methods have been used to find term variations. Fahmi (2009) acquired term variation for the medical domain using a two-step model. As a first step an initial list of synonyms are extracted using a method adapted from DIPRE (Brin, 99). During this step syntactic patterns guide the extraction of candidate terms in the same way as they will guide the extraction in this paper. This first step results in a list of candidate synonyms that are further filtered following a method described in Lin et al. (2003), which uses Web pages as an external source to measure the synonym compatibility hits of each pair. The precision and recall scores presented in Fahmi (2009) are high. We will give results for this method on our test set in Section 5 and refer to it as the pattern- and web-based approach.

3 Materials and methods

In the following subsections we describe the setup for our experiments.

3.1 Data collection

Measures of distributional similarity usually require large amounts of data. For the alignment method we need a parallel corpus of reasonable size with Dutch either as source or as target language coming from the domain we are interested in. Furthermore, we would like to experiment with various languages aligned to Dutch.

The freely available EMEA corpus (Tiedemann, 2009) includes 22 languages in parallel with a reasonable size of about 12-14 million tokens per language. The entire corpus is aligned at the sentence level for all possible combinations of languages. Thus, for acquiring Dutch synonyms we have 21 language pairs with Dutch as the source language. Each language pair includes about 1.1 million sentence pairs. Note that there is a lot of repetition in EMEA and the number of unique sentences (sentence fragments) is much smaller: around 350,000 sentence pairs per language pair with about 6-7 million tokens per language.

3.2 Word alignment and phrase extraction

For sentence alignment we applied *hunalign* (Varga et al., 2005) with the 'realign' function that induces lexical features from the bitext to be combined with length based features. Word alignment has been performed using GIZA++ (Och, 2003). We used standard settings defined in the Moses toolkit (Koehn et al., 2007) to generate Viterbi word alignments of IBM model 4 for sentences

not longer than 80 tokens. In order to improve the statistical alignment we used lowercased tokens and lemmas in case we had them available (produced by the *Tree-Tagger* (Schmid, 1994) and the Alpino parser (van Noord, 2006)).

We used the *grow* heuristics to combine the asymmetric word alignments which starts with the intersection of the two Viterbi alignments and adds block-neighboring points to it in a second step. In this way we obtain high precision links with some many-to-many alignments. Finally we used the phrase extraction tool from Moses to extract phrase correspondences. Phrases in statistical machine translation are defined as sequences of consecutive words and phrase extraction refers to the exhaustive extraction of all possible phrase pairs that are consistent with the underlying word alignment. Consistency in this case means that words in a legal phrase are only aligned to words in the corresponding phrase and not to any other word outside of that phrase. The extraction mechanism can be restricted by setting a maximum phrase length which is seven in the default settings of Moses. However, we set the maximum phrase length to four, because we do not expect many terms in the medical domain to be longer than 4 words.

As explained above, word alignment is carried out on lowercased and possibly lemmatised versions of the corpus. However, for phrase extraction, we used surface wordforms and extracted them along with the part-of-speech (PoS) tags for Dutch taken from the corresponding Alpino parse trees. This allows us to lowercase all words except the words that have been tagged as *name*. Furthermore, the inclusion of PoS tags enabled us to filter the resulting phrase table according to typical patterns of multi-word terms. We also removed phrases that consist of only non-alphabetical characters. Note that we rely entirely on automatic processing of our data. Thus, the results from automatic tagging, lemmatisation and word alignment include errors. Bannard and Callison-Burch (2005) show that when using manual alignment the percentage of correct paraphrases significantly rises from 48.9% to 74.9%.

3.3 Selecting candidate terms

As we explained above we can select those phrases that are more likely to be good terms by using a regular expression over PoS tags. We apply a pattern using adjectives (A), nouns (NN), names (NM) and prepositions (P) as its components based on Justeson and Katz. (1995) which was adapted to Dutch by Fahmi (2009):

```
((A|NN|NM)+|((A|NN|NM)*
(NN|NM P)?)(A|NN|NM)*))NN+
```

To explain this regular expression in words, a candidate term is either a sequence of adjectives and/or nouns and/or names, ending in a noun or name or it consists of two such strings, separated by a single preposition.

After applying the filters and removing all hapaxes we are left with 9.76 M co-occurrences of a Dutch (multi-word) term and a foreign translation.

3.4 Comparing vectors

To compare the vectors of the terms we need a similarity measures. We have chosen to describe the functions used in this paper using an extension of the notation used by Lin (1998), adapted by Curran (2003). Co-occurrence data is described as tuples: ⟨word, language, word′⟩, for example, ⟨kat, 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 in any language. For the example of *kat* in, this would denote all values for all translational contexts the word is found in: *Katze_DE*:17, *chat_FR*:26 etc. Everything is defined in terms of co-occurrence data with non-zero frequencies. The set of attributes or features for a given corpus is defined as:

$$(w, *, *) \equiv \{(r, w') | \exists (w, r, w')\}$$

Each pair yields a frequency value, and the sequence of values is a vector indexed by $r{:}w'$ values, rather than natural numbers. A subscripted asterisk indicates that the variables are bound together:

$$\sum (w_m, *_r, *_{w'}) \times (w_n, *_r, *_{w'})$$

The above refers to a dot product of the vectors for term w_m and term w_n summing over all the $r{:}w'$ pairs that these two terms have in common. For example we could compare the vectors for *kat* and some other term by applying the dot product to all bound variables.

We have limited our experiments to using Cosine[2]. We chose this measure, since it performed best in experiments reported in Van der Plas (2008). Cosine is a geometrical measure. It returns the cosine of the angle between the vectors of the words and is calculated as the dot product of the vectors:

$$Cosine = \frac{\sum (W1, *_r, *_{w'}) \times (W2, *_r, *_{w'})}{\sqrt{\sum (W1, *, *)^2 \times \sum (W2, *, *)^2}}$$

If the two words have the same distribution the angle between the vectors is zero.

3.5 Post-processing

A well-known problem of phrase-based methods to paraphrase or term variation acquisition is the fact that a large proportion of the term variations or paraphrases proposed by the system are super- or sub-strings of the original term (Callison-Burch, 2008). To remedy this problem we removed all term variations that are either super- or sub-strings of the original term from the lists of candidate term variations output by the system.

4 Evaluation

There are several evaluation methods available to assess lexico-semantic data. Curran (2003) distinguishes two types of evaluation: direct evaluation and indirect evaluation. Direct evaluation methods compare the semantic relations given by the system against human performance or expertise. Indirect approaches evaluate the system by measuring its performance on a specific task.

Since we are not aware of a task in which we could test the term variations for the Dutch medical domain and ad-hoc human judgments are time consuming and expensive, we decided to compare against a gold standard. Thereby denying the common knowledge that the drawback of using gold standard evaluations is the fact that gold standards often prove to be incomplete. In previous work on synonym acquisition for the general domain, Van der Plas and Tiedemann (2006) used the synsets in Dutch EuroWordnet (Vossen, 1998) for the evaluation of the proposed synonyms. In an evaluation with human judgments, Van der Plas and Tiedemann (2006) 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. For evaluating medical term variations in Dutch there are not many gold standards available. Moreover, the gold standards that are available are even less complete than for the general domain.

4.1 Gold standard

We have chosen to evaluate the nearest neighbours of the alignment-based method on the term variations from the Elseviers medical encyclopedia which is intended for the general audience containing 379K words. The encyclopedia was made available to us by Spectrum B.V.[3].

The test set is comprised of 848 medical terms from *aambeeld* 'incus' to *zwezerik* 'thymus' and their term variations. About 258 of these entries contain multiword terms. For most of the terms the list from Elseviers medical encyclopedia gives only one term variation, 146 terms have two term variations and only one term has three variations. For each of these medical terms in the test set the system generates a ranked list of term variations that will be evaluated against the term variations in the gold standard.

[2]Feature weights have been used in previous work for syntax-based methods to account for the fact that co-occurrences have different information values. Selectionally weak (Resnik, 1993) or *light* verbs such as *hebben* 'to have' have a lower information value than a verb such as *uitpersen* 'squeeze' that occurs less frequently. Although weights that promote features with a higher information value work very well for syntax-based methods, Van der Plas (2008) showed that weighting only helps to get better synonyms for very infrequent nouns when applied in alignment-based approaches. In the current setting we do not consider very infrequent terms so we did not use any weighting.

[3]http://www.kiesbeter.nl/medischeinformatie/

5 Results and Discussion

Before we present our results and give a detailed error analysis we would like to remind the reader of the two methods we compare our results with and give some more detail on the implementation of the second method.

5.1 Two methods for comparison

The first method is the pattern- and web-based approach described in Fahmi (2009). Note that we did not re-implement the method, so we were not able to run the method on the same corpus we are using in our experiments. The corpus used in Fahmi (2009) is a medical corpus developed in Tilburg University (http://ilk.uvt.nl/rolaquad). It consists of texts from a medical encyclopedia and a medical handbook and contains 57,004 sentences. The system outputs a ranked list of term variation pairs. We selected the top-100 pairs that are output by the system and evaluated these on the test set described in Subsection 4.1. The method is composed of two main steps. In the first step candidate terms are extracted from the corpus using a PoS filter, that is similar to the PoS filter we applied. In the second step pairs of candidate term variations are re-ranked on the basis of information from the Web. Phrasal patterns such as $XorY$ are used to get synonym compatibility hits as opposed to $XandY$ that points to non-synonymous terms.

The second method we compare with is the phrase-based translation method first introduced by Bannard and Callison-Burch (2005). Statistical word alignment can be used to measure the relation between source language items. Here, one makes use of the estimated translation likelihoods of phrases ($p(f|e)$ and $p(e|f)$) that are used to build translation models in standard phrase-based statistical machine translation systems (Koehn et al., 2007). Bannard and Callison-Burch (2005) define the problem of paraphrasing as the following search problem:

$$\hat{e_2} = argmax_{e_2:e_2 \neq e_1} p(e_2|e_1) \quad \text{where}$$

$$p(e_2|e_1) \approx \sum_f p(f|e_1)p(e_2|f)$$

Certainly, for paraphrasing we are not only interested in $\hat{e_2}$ but for the top-ranked paraphrase candidates but this essentially does not change the algorithm. In their paper, Bannard and Callison-Burch (2005) also show that systematic errors (usually originating from bad word alignments) can be reduced by summing over several language pairs.

$$\hat{e_2} \approx argmax_{e_2:e_2 \neq e_1} \sum_C \sum_{f_C} p(f_C|e_1)p(e_2|f_C)$$

This is the approach that we also adapted for our comparison. The only difference in our implementation is that we applied a PoS-filter to extract candidate terms as explained in section 3.3. In some sense this is a sort of syntactic constraint introduced in Callison-Burch (2008). Furthermore, we set the maximum phrase length to 4 and applied the same post-processing as described in Subsection 3.5 to obtain comparable results.

5.2 Results

Table 2 shows the results for our method compared with the method adapted from Bannard and Callison-Burch (2005) and the method by Fahmi (2009). Precision and recall are given at several values of k. At k=1, only the top-1 term variations the system proposes are taken into account. At k=3 the top-3 candidate term variations are included in the calculations.

The last column shows the coverage of the system. A coverage of 40% means that for 40% of the 850 terms in the test set one or more term variations are found. Recall is measured for the terms covered by the system.

From Table 2 we can read that the method we propose is able to get about 30% of the term variations right, when only the top-1 candidates are considered. It is able to retrieve roughly a quarter of the term variations provided in the gold standard[4]. If we increase k precision goes down and recall goes up. This is expected, because the system proposes a ranked list of candidate term variations so at higher values of k the quality is lower, but more terms from the gold standard are found.

[4]Note that a recall of 100% is not possible, because some terms have several term variations.

Method	k=1		k=2		k=3		Coverage
	P	R	P	R	P	R	
Phrase-based Distr. Sim	28.9	22.8	21.8	32.7	17.3	37.2	40.0
Bannard&Callison-Burch (2005)	18.4	15.3	16.9	27.3	13.7	32.3	48.1
Fahmi (2009)	38.2	35.1	37.1	35.1	37.1	35.1	4.0
Phrase-based Distr. Sim (hapaxes)	25.4	20.9	20.4	32.1	16.1	36.8	47.8

Table 2: Percent precision and recall at several values of k and percent coverage for the method proposed in this paper (plus a version including hapaxes), the method adapted from Bannard and Callison-Burch (2005) and the output of the system proposed by Fahmi (2009)

In comparison, the scores resulting from our adapted implementation of Bannard and Callison-Burch (2005) are lower. They do however, manage to find more terms from the test set covering around 48% of the words in the gold standard. This is due to the cut-off that we use when creating the co-occurrence vector to remove unreliable data points. In our approach we discarded hapaxes, whereas for the Bannard and Callison-Burch approach the entire phrase table is used. We therefore ran our system once again without this cut-off. As expected, the coverage went up in that setting – actually to 48% as well.[5] However, we can see that the precision and recall remained higher, than the scores we got with the implementation following Bannard and Callison-Burch (2005). Hence, our vector-based approach seems to outperform the direct use of probabilities from phrase-based MT.

Finally, we also compare our results with the data set extracted using the pattern- and web-based approach from Fahmi (2009). The precision and recall figures of that data set are the highest in our comparison. However, since the coverage of this method is very low (which is not surprising since a smaller corpus is used to get these results) the precision and recall are calculated on the basis of a very small number of examples (35 to be precise). The results are therefore not very reliable. The precision and recall figures presented in Fahmi (2009), however, point in the same direction. To get an idea of the actual coverage of this method we would need to apply this extraction technique to the EMEA corpus. This is especially difficult due to the heavy use of web queries which makes it problematic to apply this method to large data sets.

[5]The small difference in coverage is due to some mistakes in tokenisation for our method.

5.3 Error analysis

The most important finding we did, when closely inspecting the output of the system is that many of the term variations proposed by the system are not found in the gold standard, but are in fact correct. Here, we give some examples below:

arts, dokter (‘doctor’)
ademnood, ademhalingsnood (‘respiratory distress’)
aangezichtsverlamming, gelaatsparalyse (‘facial paralysis’)
alvleesklierkanker, pancreaskanker (‘cancer of the pancreas’)

The scores given in Table 2 are therefore pessimistic and a manual evaluation with domain specialist would certainly give us more realistic and probably much higher scores. We also found some spelling variants which are usually not covered by the gold standard. Look, for instance, at the following examples:

astma, asthma (‘asthma’)
atherosclerose, Artherosclerosis (‘atherosclerosis’)
autonoom zenuwstelsel, autonome zenuwstelsel (‘autonomic nervous system’)

Some mistakes could have been avoided using stemming or proper lemmatisation (plurals that are counted as wrong):

abortus, zwangerschapsafbrekingen (‘abortion’)
adenoom, adenomen (‘adenoma’)
indigestie, spijsverteringsstoornissen (‘indigestion’)

After removing the previous cases from the data, some of the remaining mistakes are related to the problem we mentioned in section 3.5: Phrase-

based methods to paraphrase or term variation acquisition have the tendency to propose term variations that are super- or sub-strings of the original term. We were able to filter out these super- or sub-strings, but not in cases where a candidate term is a term variation of a super- or sub-string of the original term. Consider, for example the term *bloeddrukverlaging* 'blood pressure decrease' and the candidate *afname* 'decrease', where *afname* is a synonym for *verlaging*.

6 Conclusions

In this article we have shown that translational context together with measures of distributional similarity can be used to extract medical term variations from aligned parallel corpora. Automatic word alignment and phrase extraction techniques from statistical machine translation can be applied to collect translational variations across various languages which are then used to identify semantically related words and phrases. In this study, we additionally apply pattern-based filters using part-of-speech labels to focus on particular patterns of single and multi-word terms. Our method outperforms another alignment-based approach measured on a gold standard taken from a medical encyclopedia when applied to the same data set and using the same PoS filter. Precision and recall are still quite poor according to the automatic evaluation. However, manual inspection suggests that many candidates are simply misjudged because of the low coverage of the gold standard data. We are currently setting up a manual evaluation. Altogether our approach provides a promising strategy for the extraction of term variations using straightforward and fully automatic techniques. We believe that our results could be useful for a range of applications and resources and that the approach in general is robust and flexible enough to be applied to various languages and domains.

Acknowledgements

The research leading to these results has received funding from the EU FP7 programme (FP7/2007-2013) under grant agreement nr 216594 (CLASSIC project: www.classic-project.org).

References

Bannard, C. and C. Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the annual Meeting of the Association for Computational Linguistics (ACL)*.

Brin, S. 99. Extracting patterns and relations from the World Wide Web. In *WebDB '98: Selected papers from the International Workshop on The World Wide Web and Databases*.

Brown, P.F., S.A. Della Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–296.

Callison-Burch, C. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of EMNLP*.

Curran, J.R. 2003. *From Distributional to Semantic Similarity*. Ph.D. thesis, University of Edinburgh.

Dyvik, H. 1998. Translations as semantic mirrors. In *Proceedings of Workshop Multilinguality in the Lexicon II (ECAI)*.

Dyvik, H. 2002. Translations as semantic mirrors: from parallel corpus to wordnet. *Language and Computers, Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*, 16:311–326.

Fahmi, I. 2009. *Automatic Term and Relation Extraction for Medical Question Answering System*. Ph.D. thesis, University of Groningen.

Firth, J.R. 1957. A synopsis of linguistic theory 1930-1955. *Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32.

Harris, Z.S. 1968. *Mathematical structures of language*. Wiley.

Ide, N., T. Erjavec, and D. Tufis. 2002. Sense discrimination with parallel corpora. In *Proceedings of the ACL Workshop on Sense Disambiguation: Recent Successes and Future Directions.*

Justeson, J. and S. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1:9–27.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M.Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A.Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*, pages 79–86, Phuket, Thailand.

Lin, D., S. Zhao, L. Qin, and M. Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Lin, D. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING/ACL*.

McCray, A. and W. Hole. 1990. The scope and structure of the first version of the umls semantic network. In *Symposium on Computer Applications in Primary Care (SCAMC-90)*, IEEE Computer Society, pages 126–130, , Washington DC, IEEE Computer Society. 126-130.

Och, F.J. 2003. GIZA++: Training of statistical translation models. Available from `http://www.isi.edu/~och/GIZA++.html`.

Resnik, P. and D. Yarowsky. 1997. 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?*

Resnik, P. 1993. Selection and information. Unpublished doctoral thesis, University of Pennsylvania.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, September. http://www.ims.uni-stuttgart.de/~schmid/.

Schütze, H. 1992. Dimensions of meaning. In *Proceedings of the ACM/IEEE conference on Supercomputing*.

Shimota, M. and E. Sumita. 2002. Automatic paraphrasing based on parallel corpus for normalization. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*.

Tiedemann, Jörg. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248, Borovets, Bulgaria. John Benjamins, Amsterdam/Philadelphia.

van der Plas, L. and J. Tiedemann. 2006. Finding synonyms using automatic word alignment and measures of distributional similarity. In *Proceedings of COLING/ACL*.

van der Plas. 2008. *Automatic lexico-semantic acquisition for question answering*. Groningen dissertations in linguistics.

van Noord, G. 2006. At last parsing is now operational. In *Actes de la 13eme Conference sur le Traitement Automatique des Langues Naturelles*.

Varga, D., L. Nmeth, P. Halcsy, A. Kornai, V. Trn, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of RANLP 2005*, pages 590–596.

Vossen, P. 1998. EuroWordNet a multilingual database with lexical semantic networks.

Wilks, Y., D. Fass, Ch. M. Guo, J. E. McDonald, and B. M. Slator T. Plate. 1993. Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154.

Wu, H. and M. Zhou. 2003. Optimizing synonym extraction using monolingual and bilingual resources. In *Proceedings of the International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP)*.