Extraction of Cross Language Term Correspondences

Hans Hjelm

Graduate School of Language Technology (GSLT) and Stockholm University
CL Group, Department of Linguistics
SE-106 91 Stockholm, Sweden
hans.hjelm@ling.su.se

Abstract

This paper describes a method for extracting translations of terms across languages, using parallel corpora. The extracted term correspondences are such that they are useful when performing query expansion for cross language information retrieval, or for bilingual lexicon extraction. The method makes use of the mutual information measure and allows for mapping between single word- to multi-word terms and vice versa. The method is scalable (accommodates addition or removal of data) and produces high quality results, while keeping the computational costs low enough for allowing on-the-fly translations in e.g., cross language information retrieval systems. The work was carried out in collaboration with Intrafind Software AG (Munich, Germany).

1. Introduction

One application for a method for extracting term correspondences lies in bilingual lexicon acquisition. Bilingual lexica are used in a number of different settings, including e.g., rule-based machine translation and computer-assisted language learning. Some approaches in cross language information retrieval (CLIR) also rely on the existence of bilingual lexica, for translating the query.

However, translating a query for a CLIR-system word for word does not always produce the desired results – especially not when the query constitutes a multi-word unit. E.g., the English query *heart attack* should not be translated to German *Herz Angriff* (word for word translation), but rather to *Herzinfarkt*. Given that the forming of multi-word units is a productive process in a language, one can not hope to list all such units (just as one can not hope to list all compounds in a compounding language like German). This means that there is a need for producing translations on-the-fly, translations that are relatively cheap computationally, while keeping a reasonably high translational accuracy. The system presented in this article particularly lends itself to solving this type of problems.

2. Method

Our method presupposes the existence of a parallel aligned corpus from the relevant domain (and languages). Before commencing with the extraction, the corpus has to be put through some pre-processing steps.

2.1. Pre-processing

A separate document is created for each alignment unit in the corpus. The content from each language is put in a separate field in the document, to allow for language specific searches (see section 2.2.). Here is an example of how such a document might look (where the alignment has been carried out on a sentence level):

```
<doc>
<de>Durch die Explosion einer Autobombe
ist eine Person ums Leben gekommen.</de>
<en>Someone planted a car bomb and one
person has died.</en>
</doc>
```

These documents are indexed by a full text indexing software. Our system uses the open source full text indexer Lucene[1]. To at least partly deal with problems of sparse data, we use the *LiSa* morphological analyzer (Hjelm and Schwartz, 2006) and add the citation form and part-of-speech for each word to the index. This means that we will be able to retrieve all occurrences of a lemma, regardless of the form in which it appears in the text.

2.2. Extraction

Given a specific source language term (single- or multi-word) for which one wishes to find the equivalent in the target language, the first step consists in placing a source language-specific query over the documents indexed during pre-processing. The query term(s) are also analyzed using the *LiSa* morphological analyzer. We use a threshold to limit the maximum size of the returned document set, for reasons of efficiency. The top-ranking documents in turn define a set of possible single word translations; namely the set of target language words that appear at least once in the document set returned by the query. For each such target language word, we again post a target language-specific query, resulting in a new set of documents. These two sets, along with the size of the intersection between the sets and the total size of the document collection, allow us to calculate a mutual information (MI) value for the source language term and its proposed target language translation. Note that we are not using the *pointwise* mutual information measure, criticized by e.g., Church and Gale (1991), but rather the variant typically used in Information Theory[2]. Ordering the list of target language

[1] http://lucene.apache.org/

[2] We assume that we have two random variables, X and Y, that both can take the values $\{0, 1\}$. X is associated with the source term, Y with the target term. Documents represent random experiments. If the source term appears in a document, X gets the value '1', otherwise it gets the value '0'. Y is treated accordingly.

words by their MI values gives us a ranking of the proposed single word translations.

To account for the case where the translation of a term is a multi-word unit or -term in the target language, we first observe that the constituents of the multi-word unit are likely to be ranked high on the list extracted in the previous step. Next, based on the possible combinations of parts-of-speech for multi-word units in the target language, we form multi-word translation candidates by combining words form the top-*n* candidates from the list of the top-ranked words, giving us a set of bigram translation candidates. These bigrams are used to form so called phrase queries, defining a new set of documents where the bigram under evaluation occurs[3]. This set of documents again allows us to calculate an MI value, this time between the source term and the target bigram translations. The process is iterated, forming candidate n-gram translations from the top (n-1)-gram candidates, until no more candidates are found or we reach a threshold for the maximum length of the multi-word unit[4].

The final step consists in merging the list of words with the list of the n-grams, based on their respective MI values. Since it is possible that a term has more than one translation, we select all candidates with MI values that lie above a certain fraction of the MI value of the top translation candidate[5] and present these as the suggested translations of the source term.

3. Related work

The approach proposed in this paper is related to the one described in (Fung and Church, 1994). They do not make use of sentence aligned texts, but rather split the texts into equal length segments. However, they suggest using a K (dimensionality) equal to the square root of the size of the corpus in words, which would mean around 4.000 segments for our corpus. For our experiments described in section 4., a dimensionality equal to the amount of alignment units (over 660.000) is used, providing a more fine-grained representation. Having smaller alignment units will intuitively increase the translation quality (given that the alignment is carried out correctly), especially for lower frequency terms. Smaller alignment units mean fewer target language terms that the source term co-occurs with, which in turn eliminates a number of incorrect translation candidates.

In (Melamed, 2000), the author describes different approaches for finding translational equivalents among words. Three methods are proposed, all making use of co-occurrence information coupled with e.g., a noise model or statistical smoothing. These methods, though highly effective, are all rather expensive computationally (they use iterative runs to achieve bootstrapping effects) and are restricted to finding 1-to-1 relations between words.

More recently, Tsuji and Kageura (2004) have proposed an extension to Melamed's work, showing some rather impressive results, especially with low-frequency words. They make use of transliteration techniques for performing word alignment, making the method language-pair specific and at least as costly computationally as the methods proposed by Melamed.

4. Experimental setup and results

For our experiments, we use the German and English parts of the Europarl corpus (Koehn, 2002). The texts are sentence aligned and there are over 660.000 alignment units in total.

We randomly selected six groups of words in each language from different frequency ranges; from a frequency of approximately 10.000 occurrences in the most frequent group to only one occurrence in the least frequent. Each group consists of 50 words in the source language. The results of the translations were checked against a German-English online dictionary[6]. If the proposed translation was listed there, it was counted as correct. If not, the result was inspected manually. The manual inspection was necessary for two reasons: first, because some very reasonable translation candidates were missing from the online dictionary (target language) and second, because very few of the low frequency words were listed in the dictionary at all (source language).

All experiments used single words in the source language and initially allowed for multi-word terms in the target language. However, allowing multi-word translations turned out to decrease the accuracy when translating from English to German and we consequently only used this option when translating from German to English. This decrease in accuracy can be explained by German being a compounding language and English not – only in rare cases will there be a one-to-many relationship when translating from an English word to German.

We use one strict and one lenient evaluation scheme. For the strict one, the suggested translation must be complete and no superfluous words are allowed if the translation is a multi-word term. The results of this evaluation method are shown in Figure 1. The lenient method counts a translation as correct if it captures a part of a multi-word term or has one or more superfluous words; see Figure 2 for the results of this evaluation method. The results are measured in "per-

We approximate probabilities by using relative frequencies – no smoothing strategies are used. The following formula is used to calculate the MI value (from (Manning and Schütze, 1999)):

$$\sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

[3]For efficiency reasons, this is only done for bigram candidates with frequencies above a certain threshold. The bigram frequencies can be estimated using information available in the full text index.

[4]We use a threshold of n=5 in our experiments

[5]A fraction threshold of 0.8 was used in our experiments. All parameters were determined empirically during the development of the system and were not tuned for these particular experiments.

[6]http://dict.leo.org/

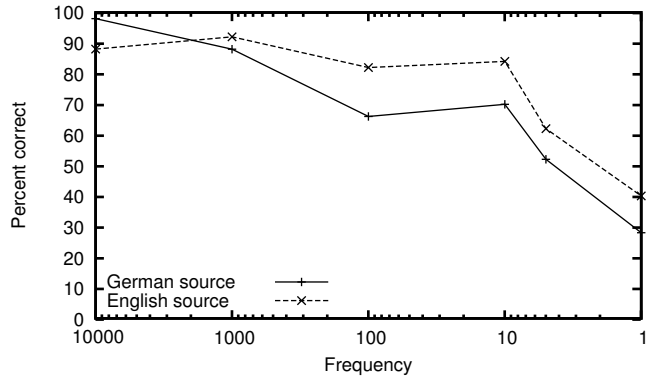

Figure 1: Results of strict evaluation scheme, top translation candidate

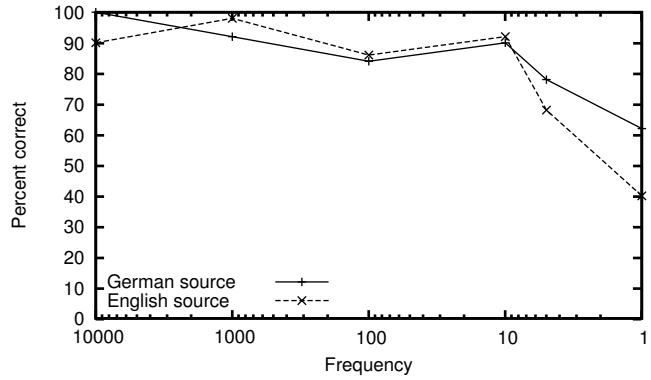

Figure 2: Results of lenient evaluation scheme, top translation candidate

cent correct". We also differentiate between just looking at the correctness of the top scoring translation and looking at all suggested translations (a maximum of three translations were produced for any input). The results of the strict and the lenient evaluation methods when looking at *all* suggested translations are presented in Figure 3 and Figure 4, respectively.

5. Discussion

The results, as displayed in Figures 1– 4, show that the quality of the translations remains relatively stable, going

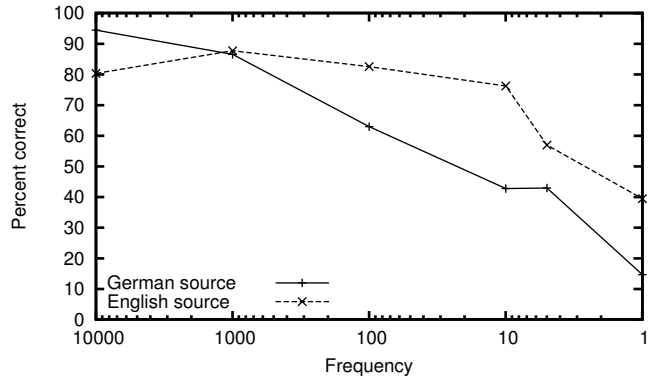

Figure 3: Results of strict evaluation scheme, all suggested translations

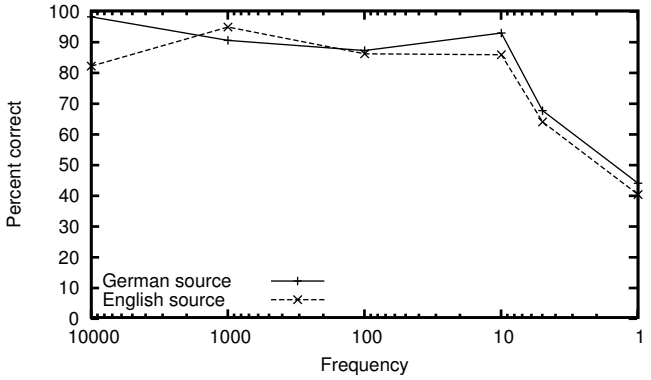

Figure 4: Results of lenient evaluation scheme, all suggested translations

from terms with frequency 10.000 to 10, after which there is a drop in quality. When translating from English to German, the results seem to peak at a term frequency of 1000, though the variations are not statistically significant. Comparing the results between the two directions of translation, translating from English to German generally gives better results than translating from German to English, especially for lower frequency terms. This can be explained by the fact that the percentage of compounds in German gets increasingly higher, the lower in the frequency range we go. Since compounds in German will generally be translated by a multi-word expression in English and finding multi-word translations is harder than identifying 1-to-1 correspondences, this is not unexpected. In fact, if we compare the two directions of translation, looking at the lenient evaluation scheme (where translations containing parts of multi-word terms are also counted as correct, Figures 2 and 4) the difference between the two directions practically vanishes.

For our experiments, we limited the maximum number of translations per input to three. However, as stated in section 2.2., we only suggest more than one translation if the alternative translations have high enough MI values when compared to the top candidate. The average number of suggested translations overall per word was 1.29. The direction from English to German had an average of 1.21 suggestions and the direction from German to English had an average of 1.37. A greater number of suggestions, apart from the cases where more than one translation is correct, might also indicate that the system is "less sure" of its choice. This suspicion is confirmed when we compare the average number of translations suggested for terms with frequency 1000, which resulted in 1.04 suggestions on average, with that for terms with frequency 1, which resulted in 1.92 suggestions on average (assuming that the system gets "less sure" as the frequency of the source language term drops). However, a number of terms with more than one translation were found this way (e.g., *wicked* (en) to *boshaft* and *bösartig* (de)), making the effort of handling more than one translation candidate worthwhile.

One interesting quality of the system, which we were not able to evaluate formally due to time constraints, is its

ability to translate multi-word terms[7]. We hope to be able to perform a full scale evaluation of this aspect of the system in the near future.

Since we have access, through the *LiSa* lemmatizer, to the part-of-speech category of both the source term and the suggested translations, we are using a filter to make sure that only translations with the same part-of-speech as that of the source term are kept. This means that we are assuming that a noun will be translated as a noun, an adjective as an adjective and so forth. Contrary to (Melamed, 2000), we found that this actually improved the accuracy of our system, though we have not yet been able to put any exact numbers on this. This may of course differ, depending on the language pair being used (Melamed used English-French instead of English-German). Note that the LiSa lemmatizer is not a part-of-speech tagger, i.e., its decisions are not context sensitive, rather, the most probable part-of-speech tag is assigned using purely lexical information.

For a system such as this to be of value in a CLIR environment, it is of great importance that its translations are up-to-date, that it is able to handle new terminology as it emerges. Here is one of the main strengths of this system: to update the translations, simply add new parallel text documents to the free text index. Similarly, out-of-date vocabulary can be avoided by removing the outdated documents from the index, or by simply reindexing the text collection without the unwanted documents, if the document collection is not too large. These updates are extremely cheap, computationally, compared to how they are handled by many of the previously suggested methods (see section 3.). Further, adding more data to the system will not make any considerable impact on its speed, since the document retrieval is handled by the full text indexing software (consider the amount of data handled by other full text indexing systems, e.g., search engines such as Yahoo[8]).

For the case when the target language translation will be used for CLIR, it is not necessary for the target language term to be a precise translation of the source language term. The results are also of value if the suggested translation belongs to the same semantic field as the source language term. E.g., one of the suggested translations from our experiments of German *Wohnungskauf* (buying of an apartment) is 'mortgage', which is likely to, if posed as a query, result in a set of relevant documents for someone interested in buying an apartment. This is another area which would be interesting to evaluate formally (such cases were counted as errors in the evalutaions presented in this article).

[7]We have made an informal evaluation on a parallel corpus consisting of the official FIFA soccer rules for English, German, French and Spanish, consisting of approximately 15.000 words per language. In spite of the limited amounts of data and in spite of using automatic methods for the sentence alignment, we were indeed able to establish translations such as "yellow card" (en) to "Gelbe Karte" (de) and "right angle" (en) to "rechtwinklig" (de). A formal evaluation of these results has yet to be undertaken.

[8]http://www.yahoo.com

6. Conclusions

The method, and the system implementing it, presented in this paper, set out to find a "cost-effective" way of extracting term correspondences across languages, using parallel corpora. The results presented in section 4. indicate that the quality of the translations varies with the frequency of the terms, which comes as no big surprise. However, we also note that although the translations of the lower frequency terms are of lower quality, we are still in the 50% range[9], even for hapaxes, and doing a lot better even for only slightly more frequent words. This, coupled with the fact that the system does not make use of iterating steps in the extraction process, makes it suitable for use in applications where both quality (e.g.. bilingual dictionary extraction) and efficiency (e.g., CLIR) are of the essence.

7. Acknowledgements

The author would like to thank Dr. Christoph Goller of Intrafind Software AG for suggesting and refining many of the ideas behind the methods presented in this paper. These methods are also used in Intrafind's *OntologyNet*.

8. References

Kenneth Church and William Gale. 1991. Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research*, pages 40–62.

Pascale Fung and Kenneth Church. 1994. K-Vec: A new approach for aligning parallel texts. pages 1096–1102. COLING.

Hans Hjelm and Christoph Schwartz. 2006. LiSa - morphological analysis for information retrieval. In Stefan Werner, editor, *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, volume 1 of *University of Joensuu electronic publications in linguistics and language technology*. NoDaLiDa, Ling@JoY. In print.

Philipp Koehn. 2002. Europarl: A multilingual corpus for evaluation of machine translation. Draft.

Christopher Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

Keita Tsuji and Kyo Kageura. 2004. Extracting low-frequency translation pairs from japanese-english bilingual corpora. In Sophia Ananadiou and Pierre Zweigenbaum, editors, *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 23–30, Geneva, Switzerland. COLING.

[9]Lenient evaluation method