Swedish Graduate School of Language Technology (GSLT):
Term Paper in Lexical Semantics. Spring 2006.

Using a Third Language to Improve Extraction of Bilingual Term Correspondences

Hans Hjelm

GSLT and CL Group, Department of Linguistics
Stockholm University, S-106 91 Stockholm, Sweden
hans.hjelm@ling.su.se

Abstract

This paper describes a method for improving the quality of the extraction of bilingual term correspondencies. The basic intuition is that using a parallel text consisting of three (or more) languages, the third language will provide additional information when extracting term correspondencies between the original source and target languages. First experiments show encouraging results, with overall accuracy increasing about two percent, varying slightly with the direction of translation and method of evalutation. Further extensions of the method are suggested, which are hoped to be effective in cases where the current method is unable to improve the preformance.

1 Introduction

The basic, bilingual, task is to, given a list of terms (single- or multi-word) or words in the source language, produce their most likely correspondences in the target language, based on information derived from a parallel text. For a given word in the source language, it is not always possible for a system to come up with a single translation to that word. This can have several different causes:

- The source word is polysemic or homonymic and the different senses of the word give rise to different translations in the target language.
- The target language has more than one word with more or less the same meaning, used interchangeably as translations of the source words (i.e., the target language words are synonyms).
- The two languages have different ways of "carving up" the semantic field. E.g., English *go* corresponds to both German *gehen* (go by foot) and *fahren* (go by some means of transportation) [God01].

In all cases mentioned above, a system extracting correspondences would be correct in suggesting more than one translation of the source language word. However, this also means that the system has to be able to differentiate the cases where more than one translation is correct from other cases where the apparent need for more than one translation is purely accidental (depending on the method used for extracting the correspondences, this situation can arise in different ways). This paper proposes a method of using a third language as a kind of sanity check for all alternative suggestions, in order to weed out the unwanted ones.

2 Background

This section first presents an overview over what has been done towards solving the basic problem of extracting term correspondences. The basic system, used as a backbone in the system and method presented in this paper, is described. We then proceed to present previous efforts of incorporating a third language into traditionally bilingual alignment tasks.

2.1 Automatic extraction of term correspondences

The basic, bilingual, method for extracting term correspondences used in this paper is described in detail in [Hje06]. The method, developed in collaboration with Intrafind AG[1], presupposes a parallel sentence aligned text. The texts are preprocessed by a system for morphological analysis called LiSa [HS06], which provides information about lemmas and part-of-speech. The distributions of the words

[1] http://www.intrafind.de

in the alignment units of the respective languages are used to calculate a correlation measure between any two given words or phrases. The mutual information measure is used for this purpose, though one should point out that this is not the *pointwise* mutual information measure, critisized by e.g. Church and Gale [CG91], but rather the measure typically used in Information Theory (from [MS99]):

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

The strengths of this system are its efficiency coupled with a high level of accuracy and its ability to handle many-to-many relations (e.g., where a three word phrase in the source language is translated with a four word phrase in the target language).

Most notable among the other approaches to solve this problem is probably the one described by Melamed in [Mel00]. Correlation measures for the distribution of words are coupled with a noise model and statistical smoothing, giving some impressive results. On the downside, these approaches use iterative runs in the extraction process, making them computationally expensive. Also, a one-to-one relationship between source and target words is assumed.

Continuing in Melamed's line of work, Tsuji and Kageura [TK04] have attained even higher accuracy rates, especially for low frequency words. They use transliteration techniques to perform word alignment, building on top of Melamed's methods. This of course means that the computational cost is at least as high as for Melamed's approaches.

2.2 Using a third language in bilingual alignment tasks

Though there have been some articles written on exploiting a third (or "additional") language in some areas of NLP (e.g., Yarowsky et al. [YNW01] use aligned texts to induce monolingual text analysis tools, such as part-of-speech taggers), not many articles have been written about using a third language in automatic text alignment. One of the select few is Simard's article on automatic sentence alignment, using a third language to improve the results [Sim99]. He is able to show an increased alignment accuracy for the languages tested and generally concludes that:

> ...the more languages, the merrier!

meaning that the addition of further languages would continue to increase the accuracy rate.

Where Simard could make use of the fact that there are no *crossing alignments* in sentence alignment (meaning that segment *k* of one language always corresponds to segment *k* of another) this principle clearly does not hold for word alignment, which makes it a harder problem to solve. Borin [Bor00] makes use of a third language in an automatic bilingual word alignment task. He refers to the process as *pivot alignment* and, like Simard, sees a clear beneficial effect of using the extra, *pivot*, language.

Where Borin's experiments deal with word alignment on a *token* level (this being what is commonly understood by word alignment), the experiments carried out in this paper could be seen as word alignment on a *type* level. That is, we are not interested in the translation of a particular *occurrence* of a word or a phrase, but rather in finding the most probable translation(s) when considering the parallel text as a whole.

3 Method

Using the system described in section 2.1 as a starting point, we modified that system in two major ways to achieve the effect we were looking for. The first deals with filtering out incorrect suggestions and the second involves reordering the results.

3.1 Filtering out incorrect translations

Assume we want to translate `wordA1` from `langA` to `langB`. At our disposal we have a parallel, sentence aligned text in the languages `langA`, `langB` and `langC`. The first step is then to produce translation models for each language pair, using the method mentioned in section 2.1. The models produced by this system are bi-directional, which means that, for three languages we would produce three different models. We then translate `wordA1` into `langB`, using the translation model `langA-langB`. Assume further that the translation model has three suggestions for translations of `wordA1` into `langB`: `wordB1`, `wordB2` and `wordB3`, of which only the last two are correct. The next step is to translate these three suggestions into `langC`, which we, for sake of this example, suppose would give rise to the following results, or "chains" or "paths" of translations:

- `wordA1 - wordB1 - wordC1`
- `wordA1 - wordB1 - wordC2`
- `wordA1 - wordB2 - wordC3`
- `wordA1 - wordB3 - wordC2`
- `wordA1 - wordB3 - wordC3`
- `wordA1 - wordB3 - wordC4`

Next, we translate `wordC1` - `wordC4` back into `langA`. Ideally, we would get back to the word we started with, `wordA1`. Our method now says that, if there is a path, leading over a translation suggestion in `langB` leading back to `wordA1`, this word is a correct translation; if not, it is incorrect. However, if no

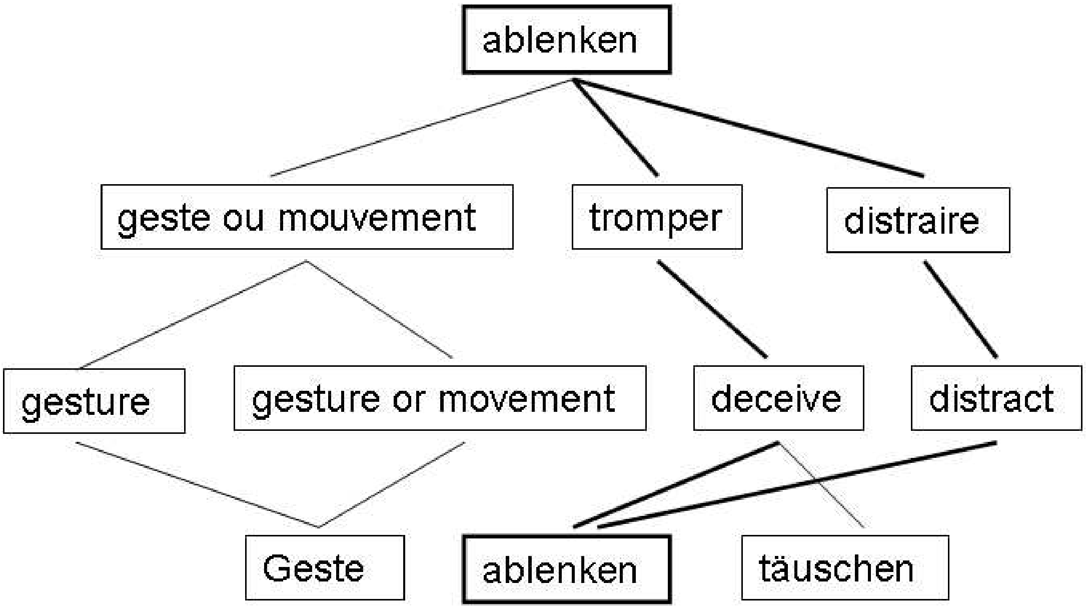

Figure 1: Results of going via English when translating from German to French

alternative leads back to `wordA1`, we accept all alternatives – we have gained no information. Continuing with our example, the paths marked in bold in the following are considered correct, meaning that they lead back to `wordA1`. Since no path leads over `wordB1`, it is filtered out as a possible translation of `wordA1`:

- `wordA1 - wordB1 - wordC1 - wordA2`
- `wordA1 - wordB1 - wordC2 - wordA2`
- **`wordA1 - wordB2 - wordC3 - wordA1`**
- `wordA1 - wordB2 - wordC3 - wordA3`
- `wordA1 - wordB3 - wordC2 - wordA2`
- **`wordA1 - wordB3 - wordC3 - wordA1`**
- `wordA1 - wordB3 - wordC3 - wordA3`
- **`wordA1 - wordB3 - wordC4 - wordA1`**

A real world example from an experiment translating from German to French, via English, is shown in figure 1. The thick line through the figure indicates a correct path, others are filtered out. Obvously, there are some mistranslations along the way, but these do not influence the end outcome.

3.2 Reordering the results

Any two pairs of words or phrases are given a correlation value by the original system, a value that ranges between 0 – 1. When more than one translation is suggested for a particular input word, the suggestions are ordered (in descending order) by their level of correlation with the source word. This means that the most highly correlated word will be suggested first, the second most second and so on. For these experiments, we instead used the average level of correlation between all three word pairs in the chain:

$$\frac{I(wordA; wordB) + I(wordB; wordC) + I(wordC; wordA)}{3}$$

When, for a word in `langB`, there is more than one path leading back to `langA` (like in the example in the previous section: `wordB3` can reach `wordA1` over `wordC3` or over `wordC4`) the highest scoring path is used and the other path is simply ignored.

4 Experimental setup and Results

Through the two measures described in section 3, we hoped to achieve two things:

- Lower the number of suggested translations without lowering the accuracy of the system.
- Increase the number of correct suggestions appearing at the top of the list of translation suggestions.

We conducted an experiment to evaluate these issues. For the experiment we used FIFA's *Laws of the Game 2005*[2] in the languages German, French and English. Using a method for term extraction, based again on mutual information, we extracted 254 German words and phrases[3] that we translated into English and 241 English words and phrases that we translated into German. We performed the translation once with the original system and once with the system as described in section 3, using French as our "pivoting" language. We deliberately chose to work with these relatively small corpora (around 15,000 words per language) to present the systems with a more challenging task. (In [Hje06] we used corpora containing about 18,000,000 words per language, which is a more commonly used order of size in this area of research).

Results are measured in "percent correct". We use two ways of measuring correctness: one strict and one lenient. For the strict part, we only consider those translations correct which are complete (no words are missing from multi-word expressions) and with no superfluous words added. For the lenient approach, also translations capturing only part of a multi-word expression are considered correct. We also differentiate between measuring the accuracy of all suggested translations and measuring the accuracy of the first (best) translation suggestion. We thus get four evaluation measures: one strict for the first candidate, one lenient for the

[2] International soccer rules, available through http://www.fifa.com

[3] the phrases were restricted to a maximum length of two words

first candidate, one strict for all candidates and one lenient for all candidates. The results are presented in tables 1 and 2.

German-English	strict 1st	lenient 1st	strict all	lenient all
regular	58.7%	72.8%	42.3%	62.0%
third lang.	60.6%	74.8%	45.0%	64.6%

Table 1

English-German	strict 1st	lenient 1st	strict all	lenient all
regular	55.6%	75.1%	44.4%	65.7%
third lang.	55.6%	76.3%	45.9%	68.5%

Table 2

However, in a large number of these cases, our method really has no chance on improving the results – namely the cases where only one translation is suggested by the system (it would not make sense to filter out the single translation available). Further, in cases where more than one translation is suggested, but none of the suggestions are correct (not even partly), there is also no real room for improvement. We therefore singled out the cases where there was a theoretical possibility of improvement (or deterioration). This was the case for 88 (of 254) of the German and 63 (of 241) of the English words and phrases. The results, when looking strictly at that subset, are presented in tables 3 and 4.

German-English	strict 1st	lenient 1st	strict all	lenient all
regular	51.1%	76.1%	38.3%	72.4%
third lang.	56.8%	81.8%	42.2%	77.6%

Table 3

English-German	strict 1st	lenient 1st	strict all	lenient all
regular	50.8%	81.0%	42.1%	76.6%
third lang.	50.8%	85.7%	43.2%	82.4%

Table 4

5 Discussion and future work

Looking especially at tables 3 and 4, we see a clear tendency that the results are improving, with the one exception of the strict evaluation of the top ranking suggestion, translating from English to German. It is not entirely clear why this particular evaluation measure does not show any improvements, we would have to repeat the experiments on different data to see if this is coincidental or an actual tendency. For all other measurements, we see improvements ranging from 1.1%

to 5.8%. The results overall are worse than what you typically find in this line of research, however, this is due to the small sizes of the test corpora (more than a factor 1,000 smaller than was used in [Hje06]). Considering this, the lenient evaluation of the top scoring translation being over 80% correct for both directions of translation should be seen as encouraging.

According to Sager [Sag94], the notion of equivalence is central to the field of translation. Equivalence relations, in turn, have the properties of being reflexive (any word is a translation of itself), symmetrical (if A is a translation of B, then B is a translation of A) and transitive (if A is a translation of B, and B of C, then A is also a translation of C) [BJ89]. The method presented in this paper makes use of both the transitive and symmetrical properties, when assuming that the translations will be preserved when transferred across languages and back again to the original language. Judging from the positive results presented in section 4, we seem to be at least partly justified in making these assumptions. Of course, in pratice, the notion of equivalence should be modified to a notion of relative equivalence.

In future experiments, it would be interesting to investigate whether the nature of the third, or *pivoting*, language used influences the quality of the results. Borin [Bor00] reports that Spanish works better as a pivoting language than Polish, when aligning Swedish and English. This he attributes to the closer "genetic" relation between the three languages involved when using Spanish rather than Polish. It is plausible that the same would hold in our case.

In some cases, where none of the suggested translations "survive" the process described in section 3 (meaning that no translation is translated back to the source language word), the system is unable to take any action. One way to take advantage of this otherwise unfortunate state would be to use it as a sort of alarm clock: if no suggestions survive the process, perhaps something has gone wrong (i.e., all suggestions are false). For these cases, it would be interesting to see if the results would improve if the order of the translations described in section 3 were reversed. One would then for these cases start by translating `wordA` from `langA` to `langC` and then continue from `langC` to `langB`, to see if one could somehow circumvent the shaky connection between `langA` and `langB` for this particular input.

6 Conclusions

In this project we have developed a method for increasing the quality of word or term correspondences extracted from a parallel text, developed a system that implements the method and evaluated the quality of this system. The main idea of the method was to use a third language as a kind of quality control, to be able to filter out faulty suggestions. The evaluations performed, though limited in size, indicate that the method fulfills its purpose, producing improvements in the range between 1.1% and 5.8% for all but one out of eight evaluation measures, the eighth remaining unchanged. We have also presented some ways of improving the system in future experiments and some alternative approaches that we would like to see evaluated.

References

[BJ89] George Boolos and Richard Jeffrey. *Computability and Logic*. Cambridge University Press, Cambridge, 3 edition, 1989.

[Bor00] Lars Borin. You'll take the high road and i'll tahe the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th International Conference on Computational Linguistics*, volume 1, pages 97–103. COLING, 2000.

[CG91] Kenneth Church and William Gale. 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, 1991.

[God01] Cliff Goddard. Universal units in the lexicon. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, and Wolfgang Raible, editors, *Language Typology and Language Universals*, volume 2, pages 1178–1190. Walter de Gruyter, Berlin, New York, 2001.

[Hje06] Hans Hjelm. Extraction of cross language term correspondences. In *Proceedings of LREC 2006*. LREC, 2006. To appear.

[HS06] Hans Hjelm and Christoph Schwartz. 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, 2006. In print.

[Mel00] Dan Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000.

[MS99] Christopher Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA, 1999.

[Sag94] Juan Sager. *Language Engineering and Translation Consequences of automation*. John Benjamins Publishing Company, Amsterdam, 1994.

[Sim99] Michel Simard. Text-translation alignment: Three languages are better than two. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 2–11, 1999.

[TK04] Keita Tsuji and Kyo Kageura. 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, 2004. COLING.

[YNW01] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of HLT 2001, First International Conference on Human Language Technology Research*, 2001.