To appear in *Proceedings of NoDaLiDa'99*,
Trondheim, 9–10 December, 1999.

Pivot alignment

Lars Borin
Department of Linguistics
Uppsala University
Lars.Borin@ling.uu.se

Abstract

Word alignment of parallel texts is typically carried out using many kinds of knowledge, or information sources, in concert, i.e., it is profitably viewed as a kind of cooperative process, where e.g. distribution, string similarity, cooccurrence statistics, and other information sources are used together. We investigate a novel such information source in this paper, namely the use of a third language as a 'pivot' to increase alignment recall, hence the name *pivot alignment*. The results of the preliminary experiments reported here indicate that pivot alignment increases word alignment recall, without sacrificing precision. We conclude that the method is well worth exploring further, by examining more languages and language combinations.

1 Introduction

Parallel texts aligned on the word level have a number of potential uses. Given suitable browsing and search tools, linguists can use aligned parallel corpora in the same way that they already use monolingual corpora, i.e. as a rich source of authentic language data, in this case data on translation equivalence (see, e.g., Olsson & Borin forthcoming). Bilingual lexicography, translator training, and foreign language instruction all stand to benefit from the use of such corpora. In computational linguistics, the application which springs to mind first is the automatic or semi-automatic extraction of translation equivalents for machine translation systems from word-aligned parallel texts, but there are also possible applications in the fields of computer-assisted language learning and cross-lingual information retrieval.

The ETAP project is a parallel translation corpus project funded by the Bank of Sweden Tercentenary Foundation. The aim of this project is to create an annotated and word-aligned multilingual translation corpus, which will be used as the basis for the development of methods and tools for automatic extraction of translation equivalents on the word and phrase levels (see Borin forthcoming a).

2 Word alignment as a cooperative process

Sentence alignment is a fairly well-understood problem, with state-of-the-art sentence alignment algorithms routinely achieving accuracies close to a hundred percent,[1] even without the use of language-specific information. The best word alignment systems, on the other hand, typically achieve a recall in the 25 to 45 percent range in the language-independent case (but with high precision, typically over ninety percent).[2]

In common with many other nontrivial linguistic tasks, the decisions of which words to link up with each other, i.e. the ability to make correct word alignments, seem to draw on many different knowledge sources simultaneously. The word alignment system that we use in the ETAP project, the UWA (Uppsala Word Aligner; see Tiedemann this volume), uses several kinds of information in an iterative word alignment process, where a text-specific translation dictionary is accumulated, and aligned units are removed after each step. The following kinds of information are used to align words (this is an extremely simplified account of how the UWA works; see Tiedemann this volume for details).

- single-word 'sentences', which may be the result of previous removals of words from multi-word sentences[3]
- identical and highly similar words
- distributionally similar words

Additionaly, language-internal cooccurrence statistics are used to find multi-word units ('phrases') in both languages, which can then be aligned in the same way as single words, while lowercasing and stemming reduce the number of types, thus increasing the average type frequency, making statistical methods more effective.

Thus, we see that the UWA uses many kinds of knowledge to achieve its objective. In the same spirit, we have explored the possibilities of combining word alignment and part-of-speech (POS) tagging (Borin forthcoming c), as well as combining different POS taggers using linguistically motivated rules, so that the combination achieves greater accuracy than the best individual tagger (Borin forthcoming d).

All this have led us to a view of word (and phrase and sentence) alignment, and also POS tagging, as a *cooperative process*, where many independent 'experts', using various kinds of information sources, access and modify the same, increasingly richer linguistic representation, performing POS tagging, alignment, and possibly other kinds of linguistic analysis and annotation as well, utilizing the relevant information that other experts have left there.[4]

We already know that distributional parallelism, language-internal and cross-language coocurrence, string similarity (also both between and within languages), and part of speech are useful information sources for word alignment (Tiedemann 1998, this volume, forthcoming; Melamed 1995, 1998; Borin forthcoming c).

The view of word alignment as being achieved by the use of many (mutually independent) kinds of knowledge in concert naturally makes one look for additional such kinds of knowledge, information sources that could be used to further improve word alignment. This paper discusses one such source which to the best of my knowledge has not been considered earlier, namely the use of a third language in the alignment process. Perhaps the reason that it has not been considered earlier is that it is possible only with *multi*lingual parallel corpora, and—for obvious reasons—not with *bi*lingual corpora, which has been the kind of parallel corpus that has received most attention from researchers in the field.

3 Pivot alignment

Since the third language acts as, as it were, a pivot for the alignment of the two other languages, we refer to the method as *pivot alignment*, and it works as follows, with three languages, e.g. Swedish (SE), Polish (PL) and Serbian-Bosnian-Croatian (SBC), where the aim is to align Swedish with the other two languages on the word level.

1. Perform the pairwise alignments SE→PL, SE→SBC, PL→SBC, and SBC→PL;[5]

2. Check whether there exist aligned words on the indirect 'alignment path' SE→SBC→PL, which are not on the direct path SE→PL. If there are, add them to the SE→PL alignments;

3. Do the same for the indirect path SE→PL→SBC and the direct path SE→SBC.

In order for this procedure to work, we must believe that

1. there will be differences in the SE→PL and SE→SBC alignments, and

2. that these differences will 'survive' the PL→SBC and SBC→PL alignments.[6]

In other words, the indirect alignment path must *add* information to the one yielded by the direct path. If we can conceive of some plausible reason for this to happen, we may believe in the first hypothesis. One such good reason could be the fact that, as mentioned earlier, the word alignment system used, UWA, utilizes several kinds of information to align the words in the two texts. Thus it is fully conceivable, e.g., that distributional information will provide one of the links and word similarity the other in a three-language path, such as SE→PL→SBC, while synonymy or polysemy (i.e., distributional differences) prevents the first link to be made on the direct path SE→SBC. Intuitively, this is perhaps the most likely situation in this particular example, since Polish and Serbian-Bosnian-Croatian are fairly closely related Slavic languages which share many easily recognizable cognates, while both are much more remotely related to Swedish.

4 An experiment with pivot alignment

To test these hypotheses, we performed a small experiment with pivot alignment, as follows.

1. The ETAP IVT1 corpus was used for the experiment. This is a five-language parallel translation corpus consisting of text from the Swedish newspaper for immigrants (*Invandrartidningen*; the English version is called *News and Views*). Swedish is the source language, and the other four languages are English (EN), Polish, Serbian-Bosnian-Croatian and Spanish (ES). The IVT1 corpus has roughly 100,000 words of text in each language;

2. The PLUG Link Annotator (Merkel 1999; Merkel *et al*. forthcoming), was used to produce evaluation standards for the following alignment directions: SE→PL, SE→SBC, PL→SBC, SBC→PL in one group, and SE→EN, SE→ES, EN→ES, ES→EN in the other. A total of 500 words were sampled randomly from the full Swedish source text, and the standards with Swedish as the source were made manually by me from this sample. The target units of these standards were then used as the basis for the manual establishment (again by me) of the various target language alignment evaluation standards. Because of null links, misaligned or differently aligned sentences, etc., the size of the evaluation standards varies from 366 to 500 words;

3. In addition to the already word aligned SE→{EN,ES,PL,SBC}, we aligned the other language pairs necessary for the experiment;

4. The word link evaluation tool of the Uplug system, a parallel corpus toolbox of which the UWA is one component (see Tiedemann forthcoming), was used to calculate recall and precision for each language pair (i.e., SE→{EN,ES,PL,SBC}) word alignment. In addition to this, we manually extracted the additional links, if any, that would be found on the indirect path through the third language.

The results of the experiment are shown in Table 1.

languages aligned	*found links*	*links in standard*	*recall*	*correct (C)*	*partly corr. (PC)*	*not corr.*	*precision, correct*	*precision C + PC*
se-sbc	82	429	19.11%	57	17	8	69.51%	90.24%
+ se-pl-sbc	1			1				
=	83		19.35%	58	17	8	69.88%	90.36%
se-pl	57	370	15.41%	37	14	6	64.91%	89.47%
+ se-sbc-pl	4			4				
=	61		16.49%	41	14	6	67.21%	90.16%
se-es	87	454	19.16%	65	14	8	74.71%	90.80%
+ se-en-es	8			7		1		
=	95		20.92%	72	14	9	75.79%	90.53%
se-en	95	442	21.49%	70	14	11	73.68%	88.42%
+ se-es-en	4			2		2		
=	99		22.40%	72	14	13	72.73%	86.87%

Table 1: Pivot alignment experiment results (null links in standard not counted)

The "partly correct" alignments are those where part(s) of a multi-word-unit, but not all of it, have been correctly aligned. As you can see in Table 1, the potentially thorny issue did not arise of how to count a partially correct link added by pivot alignment.

We see that only a few units survived the trip through two languages, but out of those that did, most contributed positively to the total result. SE→ES and SE→PL were the alignments which benefitted most from pivot alignment (through EN and SBC, respectively), while the result was insignificant for SE→SBC and perhaps even slightly detrimental in the case of SE→EN.

5 Discussion

The material examined is fairly small, and it would be fair to say that the results presented above are best treated as suggestive, rather than conclusive. I think we may be said to have made a case for the usefulness of pivot alignment, as it tended to increase the overall recall, without lowering precision. In other words, the links added by pivot alignment tend to be good links.

Several ways suggest themselves in which the research presented here could be extended to see whether the case holds upon closer scrutiny.

In the results, there are differences between languages, even in this small material, but not exactly those that we had expected. Recall that we speculated (at least implicitly) that using a language closely related to the target language as a pivot would be more effective than using a combination of relatively more remote languages. Thus, we would have predicted that SE→PL and SE→SBC would come out on top in Table 1, which obviously was not the case. This could be due to chance, but also to some other factor. There is also the circumstance that English is actually very close to the Romance languages (of which Spanish is one) in its vocabulary, so that we may in fact have been comparing two quite similar cases.

To investigate this, we intend to perform the same kind of experiment with the other possible pivot languages in the IVT1 corpus, still using Swedish as the source, e.g. aligning Swedish and Polish, using Spanish as pivot. In this way, genetic factors should be more clearly discernible. We will also include at least Finnish in future experiments, as a representative of another language family (all the languages in the experiment were Indo-European languages; Finnish is the only non-Indo-European language included in the IVT corpus at present).

The planned experiments where the same language pair will be aligned with different pivot languages will make it possible to investigate whether pivot alignment is 'cumulative', i.e., whether

1. each pivot language contributes positively to the alignment, and
2. different pivot languages contribute different additional alignments.

In this case, we would have, not only pivot alignment in general as an additional 'expert', but each new language in a multilingual parallel corpus could then, potentially, make the annotation of all the other languages in the corpus richer.

The Plug Link Annotator is a very useful tool, without which the experiments described here could not have been carried out. It was originally developed with another goal in mind, however, that of evaluating word alignment systems. Hence, it is not surprising that we found, in the course of our work, that the PLA could be made even more useful for our purposes. Two modifications in particular would facilitate further experiments with pivot alignment, one more trivial and one more fundamental:

1. The sampling procedure should be modified to exclude function words. They tend to have a high text frequency, and thus make up a sizeable part of any random sample. Most of the null links in the experiment reported here resulted from function words, the typical case here being that of personal pronouns in Swedish,

where the equivalent information is normally expressed by person marking on the verb in Polish, Serbian-Bosnian-Croatian and Spanish, and only rarely by a separate pronoun.

2. At the moment, at most one word is sampled in each sentence alignment unit. For our purposes, it would be better if sentences were sampled, instead of words, and that the annotator be allowed to link as many words as desired in the sentence alignment unit of the sampled sentence. This would allow us to follow up on the misaligned source language units, which at present cannot be tracked through the pivot language, because the 'sample' for the pivot language is made up of the correct target words only. As the UWA aligns words only within sentence alignment units, working with sentences instead of words as sampling units would hopefully make it possible to follow up also on incorrect alignments.

A simple approximation in the first case would be to exclude high-frequency items from the sampling, if it is deemed desirable to avoid introducing language-specific information. This is a comparatively simple measure to take, and certainly one that we will take in the next round of experiments with pivot alignment.

The second problem requires for its solution a major redesign of the Plug Link Annotator, which is something that might be worth undertaking in case further experiments confirm the preliminary conclusions reached here.

It would seem that pivot alignment is suited mainly for parallel *translation corpora*, and not for the kind of corpora sometimes called simply parallel corpora, sometimes *comparable corpora*, i.e., corpora, where 'the same kind' (comparable with regard to topic, style, etc.) of text material has been collected in several languages, and mainly statistical (distributional) methods are used to locate equivalent items in the different language versions. It is possible that (a kind of) pivot alignment could be used also with comparable multilingual corpora, and this is certainly an idea worth pursuing.

6 Conclusion

In conclusion, we may say that the results of the experiments presented here are encouraging, although not conclusive. It turned out that the links added by pivot alignment were largely correct links, i.e. pivot alignment could be expected to make a positive contribution in a word alignment system using many independent information sources.

We saw that the sampling procedure and annotating program used could be optimized for this kind of investigation. The results also pointed to natural ways of extending the work reported here, by the investigation of

- more language combinations and more pivot languages, including non-Indo-European ones
- the effect of using two or more pivot languages in parallel
- the possibilities of using (a procedure similar to) pivot alignment also on comparable (parallel non-translation) corpora

Notes

[0]The research reported here was carried out within the ETAP (*Etablering och annotering av parallellkorpus för igenkänning av översättningsekvivalenter*; in English: "Creating and annotating a parallel corpus for the recognition of translation equivalents") project, supported by the Bank of Sweden Tercentenary Foundation as part of the research programme *Translation and Interpreting—a Meeting between Languages and Cultures*. See http://www.translation.su.se/. Leif-Jöran Olsson, who is responsible for systems development in the ETAP project, wrote most of the software which made the experiment reported here possible. I wish to thank the members of the PLUG project (see Ahrenberg *et al.* 1998; Sågvall Hein forthcoming) for generously letting us use the Uplug system, including the Uppsala Word Aligner, and the PLUG Link Annotator.

[1]Although there are still some unresolved issues even in sentence alignment (see McEnery & Oakes 1996; Borin forthcoming b). Our empirical experience shows that its accuracy is dependent upon many factors, such as text type, the quality of the translation, the tokenization algorithm used, etc.

[2]By the *recall* of a word alignment system, we here mean the number of (total or partial) alignments (or links) returned, divided by the number of alignments established in the text pair by a human annotator (i.e., we work with a manually established evaluation standard; see below, and also Merkel 1999; Merkel *et al.* forthcoming), while *precision* is the number of correct alignments returned divided by the total number of returned links. Thus, if the human annotator has established a standard containing 200 links in a text corpus, and the word alignment system returns links for 80 of the *source language words* in the standard, its recall is 40% (80/200). If 74 of those 80 links are correct (according to the standard), the precision becomes 92.5% (74/80). In this paper, we disregard the question of how to count null links—source language words in the standard which explicitly should remain unlinked—when calculating recall and precision, not because it is unimportant, but because we cannot see that it bears directly upon the issues discussed here.

[3]The UWA presupposes a sentence-aligned input corpus, and performs word alignments only within the existing sentence alignment units (thus, if the sentence alignment is wrong, for some reason, the word aligner will not be able to correct it).

[4]Our picture of what the ideal word alignment system would look like has much in common with the "blackboard model", which was once popular in Artificial Intelligence (see, e.g., Patterson 1990).

[5]It may seem strange that we make *both* the PL→SBC *and* the SBC→PL alignments. Intuitively, one would think that the direction would not matter, i.e., that these two alignments would result in the same set of word links. However, we have not checked whether the alignment system used (the Uppsala Word Aligner; see Tiedemann this volume) actually works in this way (this could be the topic of an interesting investigation in its own right). Thus, in order not to introduce a possibly confounding extra variable in the experiment, we decided to treat the alignment as directional (guilty until proven innocent, as it were), and to use both alignments.

[6]Incidentally, the indirect path could be extended with more languages, e.g. Swedish→Polish→English→Spanish, etc., but we have not investigated this possibility.

References

Ahrenberg, L., Merkel, M., Mühlenbock, K., Ridings, D., Sågvall Hein, A. & Tiedemann, J. 1998. Automatic Processing of Parallel Corpora. A Swedish Perspective. Linköping: Electronic University Press.

Borin, L. 1998. Linguistics Isn't Always the Answer: Word Comparison in Computational Linguistics. *NODALIDA '98 Proceedings*. Center for Sprogteknologi & Dept. of General and Applied Linguistics, University of Copenhagen. 140–151.

Borin, L. forthcoming a. The ETAP Project – a Presentation and Status Report. ETAP Technical Report etap-rr-01. Dept. of Linguistics, Uppsala University.

Borin, L. forthcoming b. …and Never the Twain Shall Meet? *Parallel Corpora, Parallel Worlds*, ed. by Lars Borin. Dept. of Linguistics, Uppsala University.

Borin, L. forthcoming c. Alignment and Tagging. *Parallel Corpora, Parallel Worlds*, ed. by Lars Borin. Dept. of Linguistics, Uppsala University.

Borin, L. forthcoming d. Something Borrowed, Something Blue: Rule-Based Combination of POS Taggers. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, Athens, Greece, 31 May–2 June, 2000.

Melamed, I. D. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-Best Translation Lexicons. *Proceedings of the Third Workshop on Very Large Corpora*. Boston, Massachusetts.

Melamed, I. D. 1998. Word-to-Word Models of Translational Equivalence. IRCS Technical Report #98–08. Dept. of Computer and Information Science, University of Pennsylvania.

Merkel, M. 1999. *Understanding and Enhancing Translation by Parallel Text Processing*. Dept. of Computer and Information Science, Linköping University.

Merkel, M., Andersson, M. & Ahrenberg, L. forthcoming. The PLUG Link Annotator – Interactive Construction of Data from Parallel Corpora. *Parallel Corpora, Parallel Worlds*, ed. by Lars Borin. Dept. of Linguistics, Uppsala University.

McEnery, T. & Oakes, M. 1996. Sentence and Word Alignment in the CRATER Project. *Using Corpora for Language Research. Studies in Honour of Geoffrey Leech*, ed. by Jenny Thomas and Mick Short. London: Longman. 220–230.

Olsson, L.-J. & Borin, L. forthcoming. A Web-Based Tool for Exploring Translation Equivalents on Word and Sentence Level in Multilingual Parallel Corpora. Paper to be presented at the VAKKI Symposium, 12–13 February, 2000, Vaasa University, Finland.

Patterson, D. W. 1990. *Introduction to Artificial Intelligence and Expert Systems*. Englewood Cliffs, N.J.: Prentice-Hall.

Sågvall Hein, A. forthcoming. The PLUG Project: Parallel Corpora in Linköping, Uppsala, Göteborg: Aims and Achievements. *Parallel Corpora, Parallel Worlds*, ed. by Lars Borin. Dept. of Linguistics, Uppsala University.

Tiedemann, J. 1998. Extraction of Translation Equivalents from Parallel Corpora. *NODALIDA '98 Proceedings*. Center for Sprogteknologi & Dept. of General and Applied Linguistics, University of Copenhagen. 120–128.

Tiedemann, J. this volume. Word Alignment Step by Step.

Tiedemann, J. forthcoming. Uplug – a Modular Corpus Tool for Parallel Corpora. *Parallel Corpora, Parallel Worlds*, ed. by Lars Borin. Dept. of Linguistics, Uppsala University.