In *Proceedings of the 18th International Conference on Computational Linguistics. COLING 2000. Vol. 1*, Saarbrücken: Universität des Saarlandes. 2000. 97–103.

You'll Take the High Road and I'll Take the Low Road: Using a Third Language to Improve Bilingual Word Alignment

Lars Borin
Department of Linguistics, Uppsala University
Box 527
SE-751 20 Uppsala, Sweden
Lars.Borin@ling.uu.se

Abstract

While language-independent *sentence* alignment programs typically achieve a recall in the 90 percent range, the same cannot be said about *word* alignment systems, where normal recall figures tend to fall somewhere between 20 and 40 percent, in the language-independent case. As words (and phrases) for various reasons are more interesting to align than sentences, we need methods to increase word alignment recall, preferably without sacrificing precision. This paper reports on a series of experiments with *pivot alignment*, which is the use of one or more additional languages to improve bilingual word aligment. The conclusion is that in a multilingual parallel corpus, pivot alignment is a safe way to increase word alignment recall without lowering the precision.

1 Introduction

For about a decade and a half now, researchers in Natural language processing (NLP) and general and applied linguistics have been working with parallel corpora, i.e., in the prototypical case corpora consisting of original texts in some *source language* (SL) together with their translations into one or more *target languages* (TL). In general linguistics, they are used—in the same fashion as monolingual corpora—as handy sources of authentic language. In computational linguistics and language engineering, various methods for (semi-)automatic extraction from such corpora of, among others, translation equivalents, have been explored.

2 Why is word alignment more interesting and why is it difficult?

Alignment—the explicit linking of items in the SL and TL texts judged to correspond to each other—is a prerequisite for the extraction of translation equivalents from parallel corpora, and the granularity of the alignment naturally determines what kind of translation units you can get out of these resources. With sentence alignment, you get data which can be used in, e.g., translation memories. If you want to build bi- or multilingual lexica for machine translation systems (or for people), however, you want to be able to align parallel texts on the word (and phrase) level. This is because, in the last two decades, NLP grammars have become increasingly lexicalized, and grammars for machine translation—as opposed to translation memories, or example-based machine translation, neither of which uses a grammar in any interesting sense of the word—form no exception in this regard. The entries of the lexicon, which is the major repository of linguistic knowledge in a lexicalized grammar, are mainly made up of units on the linguistic levels of words and phrases.

The problem here is that sentence alignment is a fairly well-understood problem, but word alignment is much less so. This means that while language-independent sentence alignment programs typically achieve a recall in the 90 percent range, the same cannot be said about word alignment systems, where normal recall figures tend to fall somewhere between 20 and 40 percent, in the language-independent case. Thus, we need methods to increase word alignment recall, preferably without sacrificing precision.[1]

There are many conceivable reasons for word alignment being less 'effective' than sentence alignment. Different language structures ensure that words comparatively more seldom stand in a one-to-one relationship between the languages in a parallel text, because, e.g.,

- SL function words may correspond to TL grammatical structural features, i.e. morphology or syntax, or even to nothing at all, if the TL happens not to express the feature in question. At the same time, function words tend to display a high type frequency, both because of high functional load (i.e., they are needed all over the place) and because they tend to be uninflected (i.e. each function word is typically represented by one text word type, while content words tend to appear in several inflectional variants). This of course means that function words will account for a relatively large share of the differences in recall figures between sentence and word alignment;

- orthographic conventions may disagree on where word divisions should be written, as when compounds are written as several words in English, but as single words in German or Swedish, the extreme case being that some orthographies get along entirely without word divisions;

- word alignment must by necessity (because word orders differ between languages) work with word *types* rather than with word *tokens*, while sentence alignment *always* works with sentence tokens,[2] i.e., it relies on linear order. This means that polysemy (one type in the SL corresponding to several types in the TL), homonymy (several types in the SL corresponding to one type in the TL), and combinations of polysemy and homonymy will disrupt the correspondence even between structurally similar languages;

Thus, the circumstance that linear order cannot be used to constrain word alignment —beyond the restriction that putative word alignments must appear in one and the same sentence alignment unit—together with the other factors just mentioned, conspire to make word alignment a much harder problem than sentence alignment in the language-independent case.[3]

3 Improving word alignment by combining knowledge sources

The project in which the research reported here has been carried out, the ETAP project (see section 8, below), is a parallel translation corpus project, the aim of which is to create an annotated—understood as part-of-speech (POS) tagged and aligned—multilingual translation corpus, which will be used as the basis for the development of methods and tools for the automatic extraction of translation equivalents.

Lately, we have been concentrating on finding good ways to improve word alignment. The word alignment system we currently use (which was developed in a sister project in our department, the PLUG project; see Sågvall Hein (to appear)) works iteratively with many kinds of information sources, and it seems that this is a good way to proceed. Distributional parallelism, coocurrence, string

[1] Alignment *recall* is here understood as the number of units aligned by the alignment program divided by the total number of correct alignments (established by independent means, normally by human annotation). *Precision* is the number of correct alignments (again established by independent means) divided by the number of units aligned by the alignment program (i.e., the numerator in the recall calculation). We will not in this paper go into a discussion of null alignments (source language units having no correspondence in the target language expression) or partial alignments (part, but not all, of a phrase aligned), as we believe that the results we present here are not dependent on a particular treatment of these—admittedly troublesome—phenomena.

[2] In parallel corpus alignment, that is, but *not* e.g. in searching in translation memories.

[3] We must stress that we are talking about the *language-independent* case here. For any particular language pair, language-specific linguistic (and possibly other) information can be used to improve both sentence and word alignment, although the former will probably still stay ahead of the latter in terms of performance.

similarity (both between and within languages), and part of speech are some of the information sources used, and also (heuristically based) stemming to increase type frequencies for the distributional measures (see, e.g. Tiedemann (to appear a), Tiedemann (to appear b); Melamed (1995), Melamed (1998)). In our work in the ETAP project we are looking for additional such information sources, and so far we have concentrated our efforts on exploring linguistically rich information, such as word similarity (Borin, 1998) and the combination of word alignment and POS tagging (Borin, to appear a).

There must certainly exist other sources of information, in addition to those mentioned above, that can be used to improve word alignment. This paper discusses one particular such source, namely the use of a third language in the alignment process. Apart from an earlier presentation by the present author (Borin, to appear b), I have not seen any mention in the literature of the possibility of using a third language in this way for improving word alignment. Simard (1999) describes how the use of a third language can be brought to bear upon the simpler problem of *sentence* alignment, but he does not consider the harder problem of word alignment. Perhaps it has not being thought of for the simple reason 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.

4 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;
2. Check whether there exist aligned words on the indirect 'alignment path'[4] 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 aligments.[5]

Hypothesis (1) seems plausible, since the word alignment system used (Tiedemann (to appear a), Tiedemann (to appear b)) actually already utilizes several kinds of information to align the words in the two texts. In particular, it uses distributional information, cooccurrence statistics, iterative size reduction, 'naive' stemming, and string similarity to select and rank word alignment candidates (but *not* linear order; cf. also section 3 above). 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,[6] while synonymy or polysemy (i.e., distributional differences; see above) will

[4]It is this metaphor of the alignments going by different 'paths' or 'roads' to the same goal which has inspired me to borrow the first part of the title of this paper from the chorus of the song "Loch Lomond".

[5]Incidentally, the indirect path could be extended with more languages, e.g. Swedish→ Polish→ English→ Spanish, etc., but we have not investigated this possibility, although we explore the possibility of using several additional languages in parallel, below.

[6]This is perhaps intuitively the most likely situation in this particular case, since Polish and Serbian-Bosnian-Croatian are fairly closely related Slavic languages that share many easily recognizable cognates, while both are much more remotely related to Swedish

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: First pivot alignment experiment results (null links in standard not counted) [From Borin (to appear b)]

prevent the first link to be made on the direct path SE→SBC.

5 An experiment with pivot alignment

In recent work (Borin, to appear b), we reported on a small preliminary experiment to test the feasibility of the method. We proceeded as follows:

1. The ETAP IVT1 corpus was used for the experiment. This is a five-language parallel translation corpus 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. (to appear)) was used to produce evaluation standards (“gold 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. 500 words were sampled randomly from the 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 varied 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 evaluation function in the alignment system was used to calculate recall and precision for each 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 null links mentioned in (2) above were largely due to the sampling procedure choosing many function words, which often (also in this case) are troublesome in the context of finding good translation equivalents, since they may not correspond to words in the TL (see section 2 above).

The results of the preliminary experiment are shown in Table 1.

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

languages aligned (standard)	*correct links*	*not correct*	*accumulated correct*	*recall*	*precision*
se-pl (501)	112	11		24.55%	91.06%
+ se-en-pl	2		+2	24.95%	91.20%
+ se-es-pl	2		+2	24.95%	91.20%
+ se-sbc-pl	6		+5	25.75%	91.47%
			+9	26.35%	91.67%
se-es (501)	167	13		35.93%	92.78%
+ se-en-es	9	1	+9	37.92	92.63%
+ se-pl-es	6		+3	37.13%	93.01%
		1	+12	38.52%	92.75%
se-en (501)	139	12		30.14%	92.05%
+ se-es-en	7	1	+7	31.74%	91.82%
+ se-pl-en	2		+1	30.54%	92.16%
		1	+8	31.94%	91.87%
se-sbc (501)	137	8		28.94%	94.48%
+ se-pl-sbc	2		+2	29.34%	94.56%

Table 2: New pivot alignment experiment results (null links in standard not counted; correct and partly correct links counted together)

alignment (through EN and SBC, respectively), while the result was insignificant for SE→SBC and perhaps even detrimental in the case of SE→EN.

We saw these results as suggestive, rather than conclusive. It certainly seemed that the closer genetic relatedness of the two Slavic languages worked to our advantage, but we concluded that we needed to do more experiments, both with more language combinations and with a modified sampling procedure. In particular, we wanted to get rid of the problematic function words (see above).

Since the recall is fairly low to start with, even a few correct additional alignments mean a great deal for the overall performance of the word alignment system. Thus, we thought that this approach would be worth pursuing further.

6 A new experiment with pivot alignment

To confirm these results, we redesigned slightly and extended our experimental procedure, in the following way. A new sampling of the same corpus was performed, but this time we first constructed a stop word list consisting of the 50 most frequent word types in the Swedish part of the IVT1 corpus, as a language-independent way of approximating the set of function words in the language. Thus, we had a new sample, with more content words, to compare with the previous one, the hypothesis being that a larger percentage of content words would be able to contribute more links in the pivot alignment process.

We also added some new language combinations, so that we now would be able to whether there is a difference in using Spanish as a pivot in aligning Swedish and English, as opposed to using Polish. We also investigated what the result would be of using more than one additional language in parallel.

The new pivot alignment paths investigated (in addition to the ones investigated in the first experiment) are represented by the following 'language triads':

- SE→EN→PL
- SE→ES→PL
- SE→PL→EN
- SE→PL→ES

The hypothesis was that the new setup would make the possible effect of close genetic relatedness more discernible, which indeed seems to be the case (see below).

The results of the new experiment are shown in Table 2. We see that

- initial (non-pivot alignment) recall has gone up quite a bit, presumably because function words have been avoided in the standard;
- initial alignment precision still remains at the same high level as before;
- all but two of the alignments added by pivot alignment are correct, i.e. recall is raised without a decrease in precision;
- different pivot languages add different alignments, i.e. there seems to be a cumulative positive effect from adding more languages;
- the degree of relatedness of the languages in a triad seems to play a role for how well pivot alignment will work for the particular triad.

7 Discussion and conclusions

With the new experimental setup, we confirmed the results from the earlier experiment, i.e., recall increases, but precision does not suffer. This tendency is even more marked in the new series of experiments. In addition, there seems to be a clear division along genetic lines; Polish is the best pivot language for Swedish–Serbian-Bosnian-Croatian alignment, and vice versa, while Spanish works better together with English. Another subcorpus in the ETAP project contains a Finnish part, and we aim at investigating the effects of using this non-Indo-European language (all the languages are Indo-European in the two experiments described here) as one of the languages in a similar experiment

It seemed that the choice of content words (or rather: lower-frequency words) over function words did lead to a better result, but this should be further investigated.

We also see that the more languages we add, the better the results become, i.e., different additional languages complement each other. In general, there was little overlap in the contributions that each language added to the final result.

It should be mentioned at this point, that the sampling and annotation procedure used did not allow us to check up on incorrect alignments which may have propagated through the pivot language. The sampling procedure would have to be redesigned for this to be possible,[7] which we plan to do in the future.

For the same reason, we do not have all the data needed to calculate the significance of the results. Thus, the results will have to remain suggestive for the time being, although the suggestion is strong that pivot alignment works the way it was hypothesized to work.

In summary, the results are encouraging, in that the links added through pivot alignment were largely correct links, i.e. pivot alignment could be expected to make a positive and safe contribution—i.e. increasing recall without lowering precision—in a word alignment system as one of many independent knowledge sources.

8 Acknowledgements

The research reported here was carried out within the ETAP project (Borin, to appear c), 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 for generously letting us use the Uplug system and the PLUG link annotator.

[7] To do this, you would sample sentences instead of sampling words randomly throughout the corpus, which is the way it is done at present. Actually, the sampling and annotation software was devised for strictly bilingual word alignment evaluation, and not for the purpose which it has been pressed into serving here.

References

Lars Borin. 1998. Linguistics isn't always the answer: Word comparison in computational linguistics. In *The 11th Nordic Conference on Computational Linguistics. NODALIDA '98 Proceedings*, pages 140–151. Center for Sprogteknologi and Dept. of General and Applied Linguistics, University of Copenhagen.

Lars Borin. to appear a. Alignment and tagging. In *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Lars Borin. to appear b. Pivot alignment. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (Nodalida99)*.

Lars Borin. to appear c. The ETAP project — a presentation and status report. ETAP research report etap-rr-01, Dept. of Linguistics, Uppsala University.

I. Dan Melamed. 1995. Automatic evaluation and uniform filter cascades for inducing N-best translation lexicons. In *Proceedings of the Third Workshop on Very Large Corpora*.

I. Dan Melamed. 1998. Word-to-word models of translational equivalence. Technical Report IRCS Technical Report #98–08, Department of Computer and Information Science, University of Pennsylvania.

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

Magnus Merkel, Mikael Andersson, and Lars Ahrenberg. to appear . The PLUG Link Annotator – interactive construction of data from parallel corpora. In *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Anna Sågvall Hein. to appear. The PLUG project: Parallel corpora in Linköping, Uppsala, Göteborg: Aims and achievements. In *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Michel Simard. 1999. 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.

Jörg Tiedemann. to appear a. Uplug – a modular corpus tool for parallel corpora. In *Parallel Corpora, Parallel Worlds*. Dept. of Linguistics, Uppsala University.

Jörg Tiedemann. to appear b. Word alignment step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (Nodalida99)*.