Heuristic word alignment with parallel phrases

Maria Holmqvist

Department of Computer and Information Science
Linköping University, Sweden
marho@ida.liu.se

Abstract

This paper presents a method for word alignment that uses parallel phrases from manually word aligned sentence pairs to align words in new texts. Experiments on an English–Swedish parallel corpus showed that the heuristic phrase-based method produced word alignments with high precision. Furthermore, alignment recall was improved by generalizing phrases with part-of-speech categories. We also compared the phrase-based method to statistical word alignment and found that a combination of phrase-based and statistical word alignments outperformed pure statistical alignment in terms of Alignment Error Rate (AER).

1. Introduction

This paper presents a new method for identifying corresponding words in parallel texts, a task known as word alignment. The method uses parallel phrases from a set of manually word aligned sentence pairs to find word alignments in new parallel text. Parallel phrases are defined as pairs of aligned sentence segments of arbitrary length which are consistent with the word alignment[1]. Figure 1 shows an example of aligned parallel phrases.

Parallel phrases are also the fundamental building blocks of phrase-based statistical machine translation systems (PBSMT). In phrase-based MT the translation model consists of phrase translations rather than word translations and phrase-based systems have been shown to clearly outperform word-based systems (Koehn et al., 2003). The advantage of phrase-based MT is that phrases capture the translation of words in context and can accommodate local reordering as well as deletions and additions of words. Word alignment should also benefit from using phrase information as phrases will provide context to the word links.

In this paper, word alignment is performed by matching parallel phrases to new sentence pairs and adding the word links suggested by matching phrase pairs. Longer matching phrases are preferred over shorter phrases since the context provided by longer phrases makes them more reliable in suggesting correct word links. In addition, we make the phrases more general by replacing some words with part-of-speech (POS) categories. This generalization of phrases increases the number of matching phrases and improves word alignment recall.

We have also compared word alignment based on parallel phrases to state-of-the-art statistical word alignment with Giza++ (Och and Ney, 2003). In the statistical approach to word alignment, word correspondences are estimated directly from parallel text in an unsupervised manner that does not require any manual annotations. As a rule, statistical word alignment improves with the amount of parallel text. The method we present here produces alignments with higher precision but lower recall than Giza++. For most applications of word alignment, such as machine translation, recall is just as important as precision. We show that by combining the phrase-based word alignments with Giza++ we achieve a better overall alignment.

Although phrases have been used to improve word alignment in different ways, they have not previously been used in the way proposed in this paper. Work on linguistically motivated phrases for word alignment includes Lin and Cherry (2003) and Talbot (2005) who identified phrases and phrase boundaries to put linguistic constraints on statistical word alignment.

The alignment algorithm in described in Section 2. Section 3 presents the result of applying the method to align a Swedish–English parallel corpus, and compares these results to statistical word alignment. Section 4 contains conclusions and directions for further research.

2. Word alignment with parallel phrases

Phrase-based word alignment requires a set of manually word aligned sentence pairs from which we can extract parallel segments, or phrases, of different lengths. The following experiments are based on a set of 1200 manually word aligned sentence pairs from the English–Swedish part of the Europarl corpus (Koehn, 2005). The sentence pairs were POS-tagged with Swedish and English versions of Connexor Machinese Syntax[2] (Tapanainen and Järvinen, 1997) and annotated with word alignments by two bilingual native Swedish speakers. From this data, 1000 sentence pairs were used as the training set and the remaining 200 pairs were kept as a test set for the final evaluation.

2.1. Phrase extraction

A set of parallel phrases was created by extracting all possible phrase pairs with a maximum length of 7 words from the word aligned training data. A parallel phrase consists of a source segment, a target segment and links between words in the segments, as shown in Figure 1. The extracted phrases were consistent with the word alignment so that aligned multi-word units are included as a whole in a phrase. In Figure 1, for example, the words *the port* are

[1] Note that the term **phrase** refers to a number of consecutive words that do not have to constitute a syntactic phrase. This definition of phrase is also used in phrase-based statistical machine translation.

[2] `http://www.connexor.eu`

aligned with *hamnen* and must therefore occur together in the extracted phrases.

```
leave the port   lämna hamnen 0-0 1-1 2-1
leave the port , lämna hamnen 0-0 1-1 2-1
the port         hamnen       0-0 1-0
```

Figure 1: Parallel phrases.

2.2. Phrase matching algorithm

Word alignment of new text is performed by matching source and target segments of parallel phrases to new sentence pairs and applying word links from the phrase to matching words in the new text. Figure 2 illustrates how word alignments are proposed by a matching phrase.

When parallel phrases are matched with new sentence pairs, longer phrase matches are preferred since they provide more context. In addition to the length bias, precision was enhanced with the following constraints: (1) do not apply a phrase if it matches in several positions in a sentence pair, and (2) if two matching phrases overlap by covering the same words in the sentence, only keep the links that both phrases agree on.

The basic algorithm was evaluated on a development test set (devtest) of 107 sentence pairs from the training set. The parallel phrases were extracted from the remaining training data. The results presented in Table 1 show that word alignment with phrases of length 2-7 words resulted in alignments with high precision (92%) on the devtest set but the recall was rather low. Only about 17% of the correct links were found.

Phrase length	Precision	Recall
1-7 words	70.3	41.3
2-7 words	92.2	16.8

Table 1: Word alignment results for phrases of different lengths.

2.3. Phrase generalization

To improve recall, phrases were generalized by substituting words tokens with POS categories. The generalization was applied symmetrically to linked source and target segments, i.e. when a source word was generalized the corresponding target word(s) linked to this word were also generalized. Figure 3 shows parallel phrases generalized with POS from the segment *in the union – i unionen*.

Part-of-speech is a very general substitution of a word. To make the phrases less general, words can also be replaced by POS together with morphological features, or even by the base form of words.

2.4. Phrase selection

A very large set of parallel phrases is created when all phrases are generalized with POS categories and many of these phrases are either too general or too short to produce correct word alignments in new parallel sentences. Therefore, when matching generalized phrases with a new sentence pair we do not only prefer longer phrases over shorter phrases, we also prefer more specific phrases of a certain length over generalized phrases of the same length. Effectively, length is given higher priority than specificity, which is not necessarily the best solution. However, it is not obvious which type of phrase is more reliable – a two word phrase containing only word forms or a five word phrase with just one word form. Word alignment of the devtest set using all generalized phrases at least two words long creates word alignments with higher recall of 55.2% but with precision that drops to 48.4%. Obviously, some phrases tend to produce incorrect word alignments and we want to be able to remove these from our set of phrases while keeping those phrases that will reliably find correct word alignments in new text.

The following sections describe two ways of identifying reliable phrases. In the first experiment we constrain the phrase generalization process using various thresholds. In the second experiment we evaluate the alignment performance of phrases by applying them to sentence pairs from the training data.

2.4.1. Using generalization thresholds

To investigate the effects of phrase generalization, different sets of generalized phrases were evaluated on the devtest set. Each set of phrases were produced using three thresholds that controlled the length of phrases and the amount of generalization:

P minimum phrase length

L minimum length of phrases to generalize

G maximum number of words to generalize in each phrase

Table 2 presents a sample of phrase generalization settings and the resulting precision and recall of the word alignments they produced on devtest data. For example, the set of phrases produced by the threshold values P=2, L=5, G=2, contains all original phrases of length 2-7 and all generalized phrases of at least 5 words containing 1 or 2 POS tokens. This set produced both higher precision (95.42%) and higher recall (17.51%) than the original set of ungeneralized phrases.

Thresholds	Precision	Recall	F-Score
1-3-1	86.8	42.8	57.3
1-4-7	72.0	49.8	58.9
2-5-2	95.4	17.5	29.6
2-5-3	91.5	19.2	31.7
2-2-7	48.8	55.2	51.8

Table 2: Word alignment results on devtest data with different generalization thresholds.

2.4.2. Evaluating phrases on training data

Another way to identify which parallel phrases are likely to produce accurate links on new data is to evaluate their performance on sentences in the training data. Phrases were matched to the sentences in the training set and alignment precision was recorded for each phrase. In addition, we recorded the frequency of each phrase since the reliability

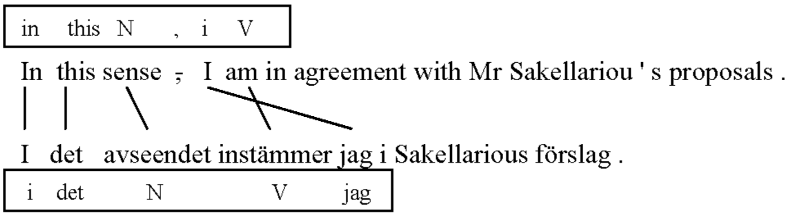
in this N , i V
In this sense , I am in agreement with Mr Sakellariou ' s proposals .
I det avseendet instämmer jag i Sakellarious förslag .
i det N V jag

Figure 2: Matching a parallel phrase to a new sentence pair.

```
in the union   i unionen      0-0 1-1 2-1
P the union    PREP unionen   0-0 1-1 2-1
P DET N        PREP NN        0-0 1-1 2-1
```

Figure 3: Example of generalized parallel phrases.

of the precision estimates increase if the phrase is frequent in the training data.

The generalized set of phrases with thresholds P=1, L=3 and G=1 was chosen for this experiment. All phrases in this set were matched to sentences in the training data and link precision and frequency were recorded for each phrase. Table 3 shows the result of word alignment with different sets of parallel phrases created by varying the thresholds for training data precision (TP) and frequency. The best precision on the devtest set (99.6%) was achieved by applying the set of phrases that occurred at least 5 times and produced at least 95% correct links in training data.

Settings	TP	Frequency	Precision	Recall
1-3-1	0.95	5	99.6	12.2
1-3-1	0.95	3	99.4	15.1
1-3-1	0.95	2	99.1	18.4
1-3-1	0.90	3	98.5	17.9
1-3-1	0.90	2	98.3	20.8
1-3-1	0.85	2	98.1	23.2
1-3-1	0.80	2	97.5	25.3
1-3-1	0.95	1	97.2	24.9
1-3-1	0.90	1	96.5	27.0
1-3-1	0.85	1	96.3	28.9
1-3-1	0.80	1	95.8	30.6

Table 3: Settings with high precision on training data sorted according to link precision on devtest data.

Alignments	Precision	Recall
Phrase combination 1	96.8	31.1
Phrase combination 2	87.3	57.2
Original phrases (1-7 words)	70.3	41.3
Original phrases (2-7 words)	92.2	16.8
Generalized phrases (2-2-7)	48.8	55.2

Table 4: Results of phrase combinations on the devtest set. Results for the original phrases and fully generalized phrases are included in the table for comparison.

2.5. Combining reliable phrases

Based on experiments with generalization thresholds and training data precision we were able to identify reliable subsets of phrases and rank them according to their alignment precision on the devtest set. When aligning a new sentence pair, the links suggested by each of these sets of phrases were combined into a larger set of links, starting with the links produced by the highest ranked phrases and adding links which connect previously unaligned words from the lower ranked link sets in sequence. This way of combining link sets gave high precision whereas the simple method of taking the union of all link sets gave slightly higher recall.

The settings in Table 3 were combined to get a word alignment with high precision (**Combination 1**). Another set of ranked settings with higher recall was combined to get a word alignment with a different precision/recall tradeoff (**Combination 2**). The precision and recall for these two combinations on the devtest set are presented in Table 4.

3. Evaluation

The phrase-based word alignment method was evaluated using the full training set and a test set of 200 sentences. The test set was annotated with word links by two annotators. Their annotations were combined into a gold standard where each link was given a confidence label, either sure (S) or possible (P). The confidence labels reflect the fact that word alignment is a difficult task and while some alignments are very clear others depend on personal judgment or on different alignment strategies. The confidence labels play a role during evaluation so that a missing P link is not punished as severely as a missing S link. Annotators used guidelines similar to Lambert et al. (2005) to distinguish between sure and possible links. Interannotator agreement was calculated as AGR = 2 * I/(A1 + A2) where A1 and A2 is the set of links created by the first and second annotator and I is the intersection of both annotators word links. The agreement was relatively high, with 85.8% labeled agreement and 91.3% unlabeled agreement. The final alignment contained 73% S links and 27% P links (not counting null links)(Holmqvist, 2008).

The results of combination 1 and 2 on the test set is similar to the result of these settings on the devtest set. The evaluation results are shown in Table 5 which also contains the alignment error rate (AER), which is a combined measure of precision and recall that takes into account the distinction between sure and possible links in the gold standard (Och and Ney, 2003). AER is calculated from the set of proposed

Alignments		Precision	Recall	AER
Phrase	combination 1	95.8	28.3	45.7
	combination 2	90.6	41.7	31.5
Giza++	grow-diag 700k	82.3	73.3	15.5
	grow-diag 5k	71.4	62.0	26.6
	intersect 700k	94.8	57.1	16.3
	intersect 5k	93.3	42.8	28.7
Merged	comb. 1 + grow-diag 700k	**84.6**	71.6	**14.2**
	comb. 1 + grow-diag 5k	**75.1**	62.2	**23.9**
	comb. 1 + intersect 700k	93.4	**60.1**	**14.8**
	comb. 1 + intersect 5k	92.2	**48.2**	**24.4**

Table 5: Alignment results on the test set. Improvements over pure statistical word alignment are in boldface.

alignments (A), and the sure (S) and possible (P) links in the gold standard according to the following formula:

$$\text{AER}(A, P, S) = 1 - \frac{|S \cap A| + |P \cap A|}{|S| + |A|} \quad (1)$$

3.1. Comparison with Giza++

The results of phrase-based word alignment were compared to state-of-the-art statistical word alignment with Giza++. Four statistical systems were created, based on different amounts of data and different symmetrization heuristics.

Symmetrization heuristics are used to create symmetric alignments from two statistical alignments in both language directions (Och and Ney, 2003). The simplest symmetrization heuristics take the intersection or the union of links from the two alignments. More complex heuristics start from the intersection of links and then add additional links from the union if they meet particular criteria. The Giza++ results in Table 5 are based on two symmetrization heuristics. The **intersection** results in alignments with high precision but lower recall, similar to what phrase-based alignment produces. The second heuristic is **grow-diag** which is the heuristic that created alignments with the best AER on test data. The systems were also trained on different amounts of parallel training data, either 5 000 or 700 000 sentence pairs. As expected, the Giza++ system trained on the larger corpus size is better than the system trained on the small corpus on all metrics.

All four Giza++ systems outperform phrase-based alignment in terms of AER but the difference is smaller for the systems trained on the small (5K) data set. In terms of precision, however, phrase-based alignment is competitive to statistical alignment.

3.2. Combining phrase-based and statistical alignment

Although phrase-based alignment can not compete with Giza in terms of AER, we see that both methods can be used to create high precision alignments. The question is whether the alignments found using these two methods overlap or not. To investigate this issue, phrase-based alignments from Combination 1 were merged with Giza alignments. The merge was done in different ways for the grow-diag and the intersection Giza alignments. For Giza grow-diag, links from the statistical alignments were added to all words that were unaligned in the phrase-based alignments. For the precision oriented Giza intersection the best results were produced by taking the union of statistical alignments and phrase-based alignments.

The lower part of Table 5 show how the merged set of Giza alignments and high precision alignments from Combination 1 outperformed the AER of pure statistical alignments for both the 5K and the 700K data set. The positive effect of adding phrase-based alignments were larger for the small dataset. Figure 4 presents an example of a successful merge.

4. Conclusions and future work

We have presented a method of heuristic word alignment with parallel phrases. Experiments on English–Swedish parallel text showed that parallel phrases can produce word alignments with high precision and that part-of-speech tags can be used instead of words in the phrases to improve alignment recall. We have also shown how the alignment precision/recall trade-off can be varied to suit different application by selecting appropriate sets of parallel phrases.

Phrase-based word alignment was compared to statistical word alignment with Giza++ and we found that although statistical alignments based on large data sets will outperform phrase-based word alignment, a combination of phrase-based word alignments and statistical alignments will outperform the quality of the statistical alignments alone. This result is important, since it shows that the phrase-based alignments are more than just a subset of the alignments found by Giza++.

The difference between phrase-based and statistical alignments trained on the small data set were not as large as with statistical alignment trained on the large data set. The improvement gained by combining phrase-based and statistical alignments were also larger for the small data set. This result suggests that phrase-based alignment could be a useful addition to statistical alignment when working with smaller parallel corpora.

There are several directions for future research. First of all, there is still room for refinement of the method for finding and combining reliable phrases for word alignment. More features can be used to decide when a phrase is a good match or not, for example relative position in source and target sentences or whether function words or content words match in a sentence. Although two phrases contain the same amount of generalized words, a match of the

<table>
<tr><td colspan="2">. . . (4)through (5)spinelessness (6)or (7)ideology (8), (9)support (10)you . . .
. . . (3)stöder (4)er (5)av (6)slapphet (7)eller (8)av (9)ideologiska (10)skäl . . .</td></tr>
<tr><td>Giza++ grow-diag</td><td>4-6 5-3* 5-4* 5-5* 5-6 6-7 7-6* 7-8* 7-9 7-10</td></tr>
<tr><td>Phrase-based</td><td>4-5 5-6 6-7 9-3</td></tr>
<tr><td>Merged</td><td>4-5 5-6 6-7 7-6* 7-8* 7-9 7-10 9-3</td></tr>
<tr><td>Reference</td><td>4-5 5-6 6-7 7-9 7-10 9-3 10-4</td></tr>
</table>

Figure 4: Example of merging phrase-based word alignments with Giza grow-diag alignments. Incorrect links are marked with *.

phrase *DET A economy - DET A ekonomin* in a sentence pair seems more reliable than matching the function word in *the A N - det A N*.

We showed that a combination of phrase-based and statistical alignment outperformed both approaches. The way we combined the two alignments were rather simplistic and it seems likely that more advanced methods for combination could produce improved results. For example, the phrase-based alignments could be added to the statistical alignments already during the symmetrization process.

It also remains to be investigated whether phrase-based alignment will improve machine translation or other applications of word alignment.

5. References

Maria Holmqvist. 2008. Word alignment by re-using parallel phrases. Licentiate thesis. Department of computer and information science, Linköpings Universitet.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, pages 48–54, Edmonton, Alberta, Canada.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, Phuket, Thailand.

Patrik Lambert, Adrià de Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39:267–285.

Dekang Lin and Colin Cherry. 2003. Word alignment with cohesion constraint. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL 2003)*, pages 49–51, Edmonton, Alberta, Canada.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

David Talbot. 2005. Constrained EM for parallel text alignment. *Natural Language Engineering*, 11(3):263–277.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97)*, pages 64–71, Washington, D.C.