High-precision Word Alignment with Parallel Phrases

Maria Holmqvist, Lars Ahrenberg, and Sara Stymne
Linköping University
Linköping, Sweden
{marho,lah,sarst@ida.liu.se}

Abstract

This paper presents a method for word alignment which re-uses parallel phrases from manually word aligned texts. We show the method's potential on two corpora and discuss which phrasal features are important for word alignment quality.

1 Introduction

Phrase-based models have been shown to be superior to word-based models for machine translation (Koehn, 2003). An advantage of phrases over word-based models, both for translation and word alignment, is that they capture non-monotone alignments at the token level, including deletions, additions and inversions. When the monolingual data is tagged, e.g., by lemmas, parts-of-speech and/or morphological categories, there is also the possibility to generalize phrases by replacing words by categories.

In this study we investigate phrase-based algorithms for high-precision word alignment. Beside machine translation, a specific application of such algorithms is for the extension of a small parallel treebank, where word alignments have been validated, to a larger one. For a human validator, adding links usually takes less time than correcting erroneous ones.

2 Word alignment with phrases

The idea behind phrase-based word alignment is that parallel segments from manually word aligned sentences can be used to align words in new sentences. First, parallel phrases of different lengths are extracted from a set of manually word aligned sentences. Each parallel phrase contains a source phrase, a target phrase and an internal word alignment as shown in Table 1. If a parallel phrase matches words in a new sentence pair, the word alignments within the phrase can be applied to the new sentence.

Manually word aligned corpora tend to be small in size and although extracted phrases produce correct links in new sentence pairs, only a small percentage of the words in the new corpus will be covered. By generalizing words in the extracted phrases with category information, e.g., part-of-speech, extracted phrases will cover more words.

Source	Target	Links
in the union	i unionen	0-0 1-1 2-1
P the union	P unionen	0-0 1-1 2-1
P DET N	P N	0-0 1-1 2-1
in this N , i V	i det N V jag	0-0 1-1 2-2 5-3 4-4

Table 1: Parallel phrases

2.1 Generalized phrases

To investigate the potential of phrase-based word alignment, all phrases of length 2-7 words were extracted from 900 sentences of a manually word aligned corpus of English-Swedish database software manuals.

When matching phrases to sentences in a test set, links proposed by longer phrases were preferred over links from shorter phrases. The resulting word alignments had a precision of 98% and a recall of 48% when evaluated against the manual word alignment of the test set. To improve the coverage of the extracted phrases, new generalized phrases were created by exchanging a number of corresponding source and target words in each phrase with part-of-speech information. The amount of generalization was controlled by two thresholds:

1. **L** The minimum length of phrases that should be generalized.

2. **M** The maximum number of words to generalize in each phrase.

During matching, longer phrases were preferred over shorter and in addition, more specific phrases were preferred over general phrases. Using all possible generalizations of phrases (L=2 and M=7), recall increased considerably (80%), but at the expense of precision (75%). The greatest improvement in recall that maintained high precision was found by generalizing phrases at least 3 words long and generalizing at most 1 word in each phrase with part-of-speech information. This resulted in a precision of 98% and a recall of 60%, which is an increase of 12 percentage points.

Phrases	Precision	Recall
Word	0.98	0.48
Word + PoS, L=2, M=7	0.75	0.80
Word + PoS, L=3, M=1	0.98	0.60

Table 2: Alignment of software manuals.

The results in Table 2 show that phrase-based alignment will indeed produce high precision alignments for the software manuals. However, these texts contain a rather limited vocabulary and simple grammatical constructions compared to open domain text such as the EU parliament proceedings (Europarl) which contains longer sentences, a larger vocabulary and more grammatical variation.

2.2 Heuristic alignment

The alignment method described above resulted in very low precision scores using the same amount of parallel text from the Europarl corpus. On this corpus we have performed experiments with a heuristic method that builds a matrix of all possible word alignments in a sentence pair where rows represent source words and columns target words. Scores are added to each cell in the matrix based on a combination of "clues" such as lexicon probability or string similarity (Tiedemann, 2005).

We use the information in parallel phrases to assign a score to each link proposed by our phrases. Following the intuition from the previous experiment, we consider longer phrases and phrases containing word forms to be more reliable than shorter phrases and phrases that mostly contain parts-of-speech. Other features that might be relevant are the phrase's precision on training data (TP), and the difference in position of source and target words in a link (POS). These features were combined into a score for each proposed link. After adding scores to the matrix, the best link for each word was selected. A score threshold t was used to remove links with very small scores.

Features	Precision	Recall
t=4,specific	0.83	0.52
t=4,specific,length	0.80	0.55
t=4,specific,length,TP	0.84	0.55
t=4,specific,length,TP,POS=abs	0.88	0.56
t=4,specific,length,TP,POS=rel	0.88	0.57
t=10,specific,TP,POS=rel	0.97	0.27

Table 3: Contribution of phrase features.

Table 3 shows how each new feature contributes to the word aligment quality. The length feature improves recall at the expense of precision because it improves the score of links proposed by longer generalized phrases. There is plenty of room for optimization of the weight of each feature as well as of the features themselves. For example, comparing an absolute and a relative measure of word distance showed that the relative measure improved both precision and recall.

3 Conclusion

We have presented phrase-based word alignment and investigated its potential on two different corpora. We have shown that phrase features can be used to find correct word alignments. To investigate the full potential of the heuristic alignment method we need to optimize the phrase features and their weights, e.g., using *minimum error rate training* (Och, 2003). There are also alternative algorithms for making use of the information contained in parallel phrases that remains to be tried. Another line of research is to do a qualitative comparison of the links we produce and those of Giza++ (Och and Ney, 2003).

References

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical Phrase-based Machine Translation In *Proceedings of HLT-NAACL '03*, pp. 48–54, Edmonton, Canada.

F. J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL '03*, pp. 160–167, Sapporo, Japan.

F. J. Och and H. Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

J. Tiedemann. 2005. Optimisation of Word Alignment Clues *Natural Language Engineering*, 11(3):279–293.