# Finding Statistically Motivated Features Influencing Subtree Alignment Performance

**Gideon Kotzé**
University of Groningen
Groningen, The Netherlands
g.j.kotze@rug.nl

## Abstract

In this paper, we present results of an ongoing investigation of a manually aligned parallel treebank and an automatic tree aligner. We establish the features that show a significant correlation with alignment performance. We present those features with the biggest correlation scores and discuss their significance, with mention of future applications of these findings.

## 1 Introduction

A greater emphasis towards syntax-based approaches in machine translation has contributed towards a greater need for the use of subsententially aligned parallel treebanks for training data or example databases (Tinsley et al., 2007a; Vandeghinste and Martens, 2010; Sun et al., 2010). Several methods exist to induce alignments on a phrasal level, for example Wang et al. (2002), Tinsley et al. (2007b), Gildea (2003), Groves et al. (2004), Zhechev and Way (2008) and Tiedemann and Kotzé (2009, 2009b). The latter two papers describe a tree-to-tree based approach to alignment, requiring both sides of the parallel corpus to be syntactically annotated.

We apply this latter implementation to word aligned and parsed parallel sentences to produce links between the nonterminal nodes of phrase-structure parse trees that denote phrasal equivalence. For example, the English noun phrase "yesterday's sitting" is linked to its Dutch equivalent, "de vergadering van gisteren" (NP/NP link).

By evaluating and generating statistics from these links, we hope to find specific features that significantly impact the alignment performance. In this paper, we focus on lexical, structural and link features, all of which may play a statistical role in performance. Additionally, lexical and structural features could be used to help predict an expected score given a syntactically annotated sentence pair, and may help point out more specific linguistic and annotation issues. Our findings may help us to improve future alignment models and may provide us with more insight into the alignment process and the linguistics of the two languages involved.

In section 2, we introduce the software and techniques in our research methodology, and explain how we get our data and statistics. After that, in section 3, we present and discuss our statistical data. Finally, in section 4, we present our conclusion.

## 2 Approach

In (Tiedemann and Kotzé, 2009) and (Tiedemann and Kotzé, 2009b), a discriminative method of automatic tree alignment is presented using a maximum entropy classifier, classifying any given source/target node pair as either linked or unlinked. The software has been developed into a freely available and flexible toolkit called Lingua-Align (Tiedemann, 2010). Features extracted from the training data are used to classify the node pairs of new trees and include structural, lexical, alignment, contextual and history features.

Testing the tree aligner requires a data set consisting of syntactically parsed and translationally equivalent sentence pairs that are also word aligned. We opted for a selection of 140 Dutch/English sentence pairs from the Europarl 3 corpus (Koehn, 2005). The sentences have been aligned with the sentence aligner that is distributed with Europarl. The Dutch sentences were parsed using the Alpino parser (Van Noord, 2006) and the English sentences using the Stanford parser ((Klein and Manning, 2003a), (Klein and Manning, 2003b)). Although the output formats of the parsers differ from each other, it poses no problem as Lingua-Align can process them and is not dependent on any specific tagset.

Since Lingua-Align does not produce its own word alignments, we used the Viterbi alignments of GIZA++ (Och and Ney, 2003). These alignments, as well as the symmetric alignments that are produced by Moses (Koehn et al., 2007) are among the many features used to build the Lingua-Align model. The word alignment model is trained on all sentence aligned text in the Europarl corpus, consisting of 1,080,417 sentence pairs. The resulting word alignments are used when Lingua-Align encounters a new sentence pair to process. To produce our manual training data, we use the Stockholm TreeAligner (Lundborg et al., 2007), which currently requires the Tiger-XML representation format for viewing, to which we converted our trees. A distinction is made between good and fuzzy links, reflecting the level of confidence of the link. This is used by default in Lingua-Align.

We pre-processed the manually produced data set by applying ten-fold cross validation, yielding a balanced F-score of 72.95 when comparing the accuracy of the automatically produced terminal and nonterminal node links with the gold standard.

For every automatically aligned tree pair, we first extract a set of basic statistics. They are:

- based on all links with reference to the gold standard, the alignment precision, recall and balanced F-score
- node counts (terminals and nonterminals)
- link counts (good and fuzzy, terminals and nonterminals)
- sentence lengths and normalized ratios
- tree level/height and normalized ratios
- averages of normalized tree level and sentence length ratios
- average path of terminal nodes to the root node
- standard deviation of these paths

For each tree, we further assign a score based on its parse quality using manual inspection. The scores are on a scale of 1 to 3, where 1 is a good parse, 2 is not so good but reasonable, and 3 is a bad parse.

Ratios were normalized by taking the length of the longer unit into account. For example, if a Dutch tree sentence has a length of 10 tokens and the English tree a length of 12, the sentence length ratio would be 0.83 according to the following formula:

$$1 - ((abs(x - y))/max(x, y)$$

Eventually, we have for each tree pair, and in some cases for each tree, a set of data values that we can investigate for possibly significant correlations with alignment evaluation scores. After extracting these statistics, we can produce distributions of the different variables over the whole set of sentence pairs.

Evaluation scores are based on all links, including those between terminal nodes. Because word alignment links are not produced by Lingua-Align but by GIZA++, the scores also indicate a measure of difference between the word alignments in the training data and those of GIZA++. We would like to study the nonterminal node linking performance of Lingua-Align itself more explicitly by keeping the word alignments fixed. We therefore proceeded to replace the manual word alignment links by those in the GIZA++ output. Naturally, since the word alignment training and testing sets are now similar, this resulted in a significant increase in accuracy, with an F-score of 82.05 when taking all links into account.

Because of the fact that Lingua-Align removes some terminal node links in the output to conform to well-formedness, the word alignment output is still slightly different from the input. However, we now consider training and testing conditions similar enough in order to measure more clearly the performance of the tree aligner itself. Because the evaluation scores take all links into account, we proceeded to calculate the precision, recall and F-scores for the nonterminal node links only. Links between terminal and nonterminal nodes are considered nonterminal node links, since they are produced by Lingua-Align. We obtain a new average F-score of 73.43.

In the next section, we present the distribution of the scores and the most important correlations, with a discussion of our findings.

## 3 Presentation and discussion of statistical data

Figure 1 presents a diagram representing the distribution of F-scores for all sentence pairs as produced by Lingua-Align. We use the F-scores pertaining to precision and recall for all nonterminal node links (including between terminal and nonterminal nodes) involved as a measure, and also present those measures in the diagram. It is clear that alignment accuracy can vary quite extensively, with the line tending towards a logarithmic curve. It is also interesting to note that recall correlates much more with the F-scores than precision does, while it is clear that precision regularly outperforms recall.

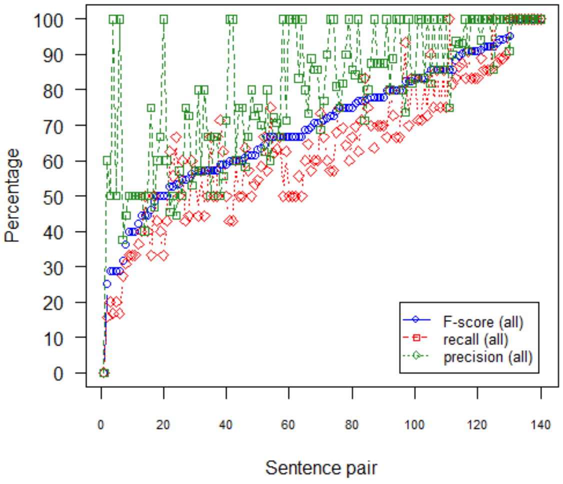

Figure 1: Distribution of F-score, precision and recall of nonterminal link node evaluation scores

We would like to determine which features correlate most strongly with these F-scores. For every feature set, we calculate the Pearson's correlation coefficient when compared to the set of F-scores. We present the correlations with values above 0.5 in figure 2.

| Feature | Correlation coefficient |
|---|---|
| Normalized ratio: Number of linked terminal nodes and all terminal nodes | 0.65 |
| Normalized ratio: Number of linked Dutch nodes and all Dutch nodes | 0.62 |
| Normalized ratio: Number of nonterminal nodes and linked nonterminal nodes | 0.59 |
| Normalized ratio: Number of linked English nodes and all English nodes | 0.57 |
| Normalized ratio: Number of linked Dutch nonterminal nodes and all Dutch nonterminal nodes | 0.56 |
| Normalized ratio: Number of linked English nonterminal nodes and all English nonterminal nodes | 0.54 |
| Normalized ratio: Number of linked terminal nodes and all linked nodes: | 0.51 |

Figure 2: List of strongest correlations

All the top correlations show a clear link between the ratio of linked nodes and F-score. In fact, the top 20 correlations are all link-based, while differences between sentence lengths has only a mild influence at 0.25. The strongest correlation indicates that a sentence pair with relatively many terminal node links is more likely to achieve a good score. One of the features in the tree alignment model specifies calculating a level of link confidence based on the ratio of the number of leaves in the two subtrees. The more leaves that are linked, the more likely the currently considered nonterminal node links are to be linked as well. Since recall is relatively low in comparison with precision, more linked terminal nodes will probably lead to better F-scores.

In general, trees that have relatively more links have generally high scores. This suggests that the alignment model could be improved by lowering the threshold at which to make links, increasing recall.

We also calculated correlations with tree features, such as tree height ratios and average distances to the root node. However, these correlations are mild to low (+-0.25 and lower) and this emphasizes the relative importance of terminal node and link features in comparison with other types of features.

The manual scores given to Dutch and English parse tree quality also show very poor correlations to the F-scores (-0.04 and -0.1 respectively).

## 4 Conclusion and future work

We have presented a statistical study of some features affecting the performance of an automatic tree aligner, given a reasonably good alignment model and reasonably good automatic word alignments. Although the data set is rather small, most of the strongest correlations suggest that more links need to be made, with word alignment links as the most important. We will apply these findings with the hope that accuracy will improve.

It also seems that there is no single dominant linear correlation with any of the extracted features with the presented F-scores. Rather, differences between correlations are gradual, and therefore, many of the features probably have an influence on each other. More sophisticated statistical tests could be employed to clearly outline these dependencies.

Many more features can be extracted. In this study, we have mostly focused on counts and ratios at sentence level, but link-centered features describing the typical contexts of good and bad links may provide more insight.

As always, more data is always better, and using a second data set from a different domain may help strengthen or disprove any findings that resulted from the first data set. Additionally, using different alignment models and even different tree aligners may provide more robustness to any future conclusions that we may draw.

Finally, in the future we hope to gain insight into linguistic issues and be able to apply our findings not only to tree alignment, but also to other domains such as parallel sentence filtering or sentence alignment.

## Acknowledgements

The research presented in this paper was done in the context of the PaCo-MT project, sponsored by the STEVIN programme of the Dutch Language Union.

## References

Daniel Gildea. Loosely Tree-based alignment for Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03)*, pp. 80-87, Sapporo, Japan, 2003.

Declan Groves, Mary Hearne and Andy Way. Robust Sub-Sentential Alignment of Phrase-Structure Trees. In *Proceedings of the 20th International Conference on Computational Linguistics (CoLing 2004)*, pp. 1072-1078, Geneva, Switzerland, 2004.

Dan Klein and Christopher D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. In *Advances in Neural Information Processing Systems 15 (NIPS 2002)*, Cambridge, MA: MIT Press, pp. 3-10.

Dan Klein and Christopher D. Manning. 2003b. Accurate Unlexicalized Parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Philipp Koehn. A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT-Summit*, 2005.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 177-180, Prague, June 2007.

Joakim Lundborg, Torsten Marek, Maël Mettler and Martin Volk. Using the Stockholm TreeAligner. In *Proceedings of the 6th Workshop on Treebanks and Linguistic Theories*, pp. 73-78, Bergen, Norway, 2007.

Franz Josef Och and Hermann Ney. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, volume 29, number 1, pp. 19-51, March 2003.

Jun Sun, Min Zhang and Chew Lim Tan. 2010. Exploring Syntactic Structural Features for Sub-Tree Alignment using Bilingual Tree Kernels. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 306-315, Uppsala, Sweden, 11-16 July 2010.

Jörg Tiedemann. Lingua-Align: An Experimental Toolbox for Automatic Tree-to-Tree Alignment. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta. 2010.

Jörg Tiedemann and Gideon Kotzé. A Discriminative Approach to Tree Alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*. Borovets, Bulgaria. 2009.

Jörg Tiedemann and Gideon Kotzé. Building a Large Machine-Aligned Parallel Treebank. In *Proceedings of Treebanks and Linguistic Theories (TLT-8)*. Milan, Italy. 2009b.

John Tinsley, Mary Hearne, and Andy Way. 2007a. Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. In *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07)*, pages 175-187, Bergen, Norway.

John Tinsley, Ventsislav Zhechev, Mary Hearne and Andy Way. Robust Language Pair-Independent Sub-Tree Alignment. In *Proceedings of Machine Translation Summit XI*, pp. 467-474. Copenhagen, Denmark. 2007b.

Gertjan van Noord. At Last Parsing Is Now Operational. In *Piet Mertens, Cedrick Fairon, Anne Dister, Patrick Watrin (editors): TALN'06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*. pp. 20-42.

Wei Wang, Jin-Xia Huang, Ming Zhou and Chang-Ning Huang. Structure Alignment Using Bilingual Chunking. In *Proceedings of the 19th Conference on Computational Linguistics*, pp. 1-7. Taipei, Taiwan. 2002.

Vincent Vandeghinste and Scott Martens. (2010). Bottom-up transfer in Example-based Machine Translation. In *Proceedings of EAMT 2010*. European Association for Machine Translation. Saint-Raphael.

Ventsislav Zhechev and Andy Way. Automatic Generation of Parallel Treebanks. In *Proceedings of the 22nd International Conference on Computational Linguistics (CoLing)*, pp. 1105-1112, 2008.