Improving syntactic tree alignment through rule-based error correction

Gideon Kotzé

University of Groningen

Abstract. Automatic alignment of parallel treebanks often display regular errors that can be corrected by improving the alignment model. However, if the aligner is statistical, often much more training data is needed to properly address these errors. In some cases, a rule-based approach to error correction may provide a quick and convenient solution. We present an approach that highlights problematic phenomena which enables us to pinpoint systematic error patterns for which we can devise rules for correction. Finally, we investigate the application of two manually constructed rules on a large parallel treebank.

1 Introduction

Parallel treebanks are generally considered a useful resource in fields such as machine translation. However, machine translation systems generally require very large corpora for training. As a result, the manual construction of such a parallel treebank is costly. To address this issue, various automatic alignment techniques have been proposed (such as Gildea [2], Groves et al [3], Zhechev and Way [13], Tiedemann and Kotzé [12]). Our study focuses on improving the output of a tree to tree alignment tool that automatically aligns phrase structure trees. In this paper, we present one way to find systematic errors in a test set for which we can construct correcting rules. We show how the application of just two rules significantly increases alignment coverage.

In (Gildea [2]) the alignment method involves changing the tree structure and is applied as a step during machine translation. Groves et al ([3]) applies some rules based on category and POS labels and tree structure to assign links as part of a best-first algorithm in the context of the Data-Oriented Translation (DOT) paradigm (Poutsma [10]; Hearne and Way [4]). Zhechev and Way ([13]) calculates scores based on existing word alignments to assign links between non-terminal nodes. This is integrated as one of the features in the tool Lingua-Align (Tiedemann [11]), which is also our tool of choice for the research presented here. Lingua-Align is a statistical aligner that aligns trees by training and applying a maximum entropy model based on user-defined features and parameters. In this paper, we present a method to facilitate finding systematic errors and investigate whether we can utilize a rule-based approach to improve alignment by correcting these errors.

2 Data and setup

For our parallel corpus, we use a sentence aligned and parsed version of Europarl 3 (Koehn [5]), with Dutch as the source and English as the target language. It is also word aligned by GIZA++ (Och and Ney [9]), using the implementation of Moses (Koehn [6]). The parse trees are converted to the phrase-structure format of Tiger-XML, so that they can be viewed and edited by the Stockholm TreeAligner (Lundborg et al. [8]), our tool of choice. For alignment above the word level, we implement, as mentioned, the tree-to-tree alignment tool Lingua-Align.

We use a set of 140 manually aligned sentences to train our Lingua-Align alignment model. After doing ten-fold cross validation, we obtain an average F-score of 72.95. Note that this also reflects differences between the manual word alignments and those produced by GIZA++, which we obtain during the automatic alignment. Next, we proceed to tree align the full Europarl corpus.

3 Error finding and rule creation

For better error analysis, we decided to make tables containing counts of features and examples of all non-matching links, when compared between an alignment output and its equivalent gold standard. Since we applied ten-fold cross validation, we have for every sentence pair a set of output links and a set of manual links. We distinguish between false positives, where a link is present in the output but not in the gold standard, and false negatives, where a link is present in the gold standard but not in the output. For this study, we focus on the counts and examples of specific combinations of category labels in the case of non-matching links. For example, we found that all 15 cases of PP/NP combinations (Dutch to English) are mismatches, of which 1 is a precision mismatch and 14 are recall mismatches. Table 1 is a small extract of the abovementioned table, with a list of mismatch examples, including both false positives and false negatives, corresponding to a category label combination. These examples include sentence IDs and matching word phrases, so that they are easily referred to using the Stockholm TreeAligner.

Table 1. Examples of mismatches by category label combination

Source cat.	Target cat.	Mismatch examples
pp	NP	15_16(ook_tijdens...—15_523(those_made_at...
pp	VP	114_18(van_het_vergaderrooster)—114_516(determine_its...
inf	S	19_14(het_werken_vanaf...—19_524(quite_simply...

A study of these cases reveals a few systematic errors that can be corrected using a rule-based approach. NPs do not generally link well to PPs, so all these

cases have a fuzzy link in the training data (more or less equivalent, but not exactly). Because of that, and the fact that this phenomenon is not so well represented in the training data, the aligner does not link these nodes. Therefore, as a statistical feature, the category label combination NP/PP plays a negative role.

Since we suggested in a previous study (Kotzé [7]) that word alignments as a feature have a significant impact on eventual tree alignment accuracy, this naturally leads us to the following question: What if there are cases where a link should be made based on existing word alignments regardless of category label combination? Of particular interest is the question whether we can describe, at least in some cases, the conditions for a link in clearly specified rules, as a complement to the already existing statistical approach.

We notice that there are cases where no link is made between two nonterminal nodes even though all of their linked terminal node children are only linked to each other. Since this is a regular occurrence and it seems that with a majority of these cases, the nodes should be linked, with the result also conforming to general principles of well-formedness (see, for example, Zhechev and Way [13]), we decided to formulate and apply an automatic linking rule. Briefly stated, the rule specifies that a link is made between two unlinked nonterminal nodes if they both have only terminal nodes as children, if at least one terminal node pair is linked and if all links involving the terminal node children only occur in the two subtrees involved. We will call these nonterminal nodes "first-level" nonterminals. Figure 1 displays the visual output of a corrected alignment. The word "hun" is correctly linked to its equivalent "their", but the word "koopkracht" is incorrectly linked to "purchasing", whereas it should really be linked to both "purchasing" and "power". However, the two phrases are exactly equivalent as a whole. This indicates a possible advantage of such a general rule in that it does not require all word alignments to be exactly correct, as long as they are all shared between the aligned subtrees.

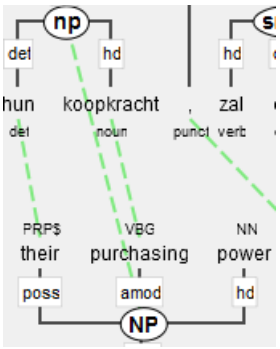

Figure 1: An example of an added link: the two NPs in the picture are linked after application of the rule

Next, we devise an additional rule that is a superset of the first rule, which states that children do not necessarily need to be terminal nodes. Figure 2 is

an example of where such an alignment is necessary. In this case, the words are perfectly aligned to each other, but note that the NPs involved here would not have been aligned if only the first rule was considered, since the Dutch NP does not only have terminal nodes as children.

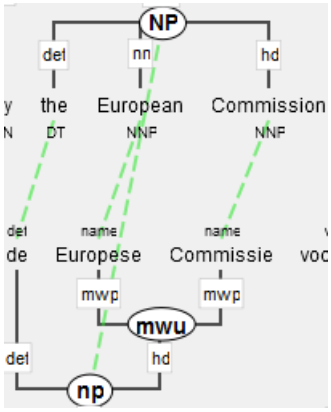

Figure 2: Example of an output of a relaxed form of the first rule: the two NPs in the picture are linked after application of the rule

4 Application and evaluation of created rules

After having defined our experimental rules, we apply them to the fully aligned Europarl corpus. Links are naively added in a greedy fashion while iterating through a list of nodes. The number of existing links is 30,118,058. If we only apply the first rule, a total of 1,090,818 links are added, which is a 3.6% increase in coverage. If we consider the links under each nonterminal node to determine whether any of the two rules can be applied, a total of 194,571 links are added for the second (relaxed) rule, which is a 0.65% increase in coverage, but only 573,236 links now for the first rule. This appears to be because the application of the second rule can link a node that has nonterminal children with a first-level node, ruling out the possibility of the latter node to be linked to another first-level node by application of the first rule.

First, we evaluate the results of the first rule application, where only links between first-level nodes were added. We proceeded to extract a random sample of 150 changed sentence pairs, containing 247 added links. 189 of these links were deemed correct, with 35 unsure or of fuzzy quality and 23 wrong. Taking only the correct and incorrect categories into account, we achieve an accuracy of 89.15% on the manually analyzed set. Even though this is an extremely small sample, so far this confirms our initial suspicion that non-terminal nodes with exclusively linked terminals should generally be linked.

To evaluate the results of the simultaneous application of rules, we extracted a random set of 120 links added by the first rule and 120 links added by the second rule, and proceeded to score them as well. Table 2 summarizes the result.

Table 2. Results of manual evaluation of combined rule application

Rule	Correct	Incorrect	Fuzzy/unsure
1 (strict)	90	10	20
2 (relaxed)	46	44	30

We have not yet tested the performance of the second rule only, or how it affects the outcome if one of them is applied only after the other one was applied for the whole sentence. However, it is clear that the first rule greatly outperforms the second rule. We have found many cases where the second rule did match the correct general segments of the trees, but that on either the source or target side, the span was too great. For example, when an NP on the source side should have been aligned with an NP on the target side, it was instead aligned with its parent, a VP. From this it is clear that the second rule is too general, and should rather be applied only in the presence of additional constraints. A bottom-up approach to adding links seems therefore very reasonable.

5 Conclusion and future work

We have showed a relatively simple way to systematically improve a tree aligner by pinpointing regular error patterns and devising rules to correct them. Moreover, by adding corrected versions of this systematic error output to our training data, we can expand the size of our training data in a meaningful way. Training and testing on the new sentence pairs may also reveal other systematic errors that can be corrected using our error correction system.

At the moment, the finding of candidates for rule correction and the application of the rules are hard coded in a Perl script. However, we hope to develop this into a modular tool that can accept rule conditions and implementations in a formally defined way. We intend to include rules handling word alignment links as well, and will also look at deletion of erroneous links. We might also incorporate features that are based on tree structure rather than only existing alignment links, such as those implemented by Groves et al [3]. Although the tool described in the latter uses more refined rules that also involves using word alignments as a feature, it might also benefit from our approach although it depends largely on the quality of word alignment. The latter approach is in some ways also more language dependent, whereas we would like our rules to be as general as possible. For Zhechev and Way's ([13]) tool, we expect a similar effect on performance as on Lingua-Align, since it also largely depends on word alignments. A parallel treebank with high recall alignments would not benefit as much from our proposed rules. However, our intention was merely to investigate the feasibility of a rule-based approach as a complement to statistical tree alignment, and as the results are promising, we can now proceed, after developing our module, to implement a machine learning tool to learn rules automatically, pos-

sibly integrating it with Lingua-Align as well. The transformation-based learning algorithm (as proposed by Brill [1]) for iterative rule-based error correction that complements an existing system seems particularly suited to our goals.

Acknowledgements

This research was done in the context of the PaCo-MT project (STE07007), sponsored by the STEVIN programme of the Dutch Language Union.

References

1. Eric Brill. Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. *Computational Linguistics*, 21(4):543-565, 1995.
2. 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.
3. 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.
4. Mary Hearne and Andy Way. 2003. Seeing the Wood for the Trees: Data-Oriented Translation. *MT Summit IX*. New Orleans, LO., pp.165172.
5. Philipp Koehn. A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT-Summit*, 2005.
6. 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.
7. Gideon Kotzé. (2011) Finding statistically motivated features influencing subtree alignment performance. In *Proceedings of the 18th Nordic Conference of Computational Linguistics*. Riga, Latvia.
8. 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.
9. 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.
10. Arjen Poutsma. 2000. Data-Oriented Translation. In *18th COLING*, Saarbrucken, Germany, pp.635641.
11. 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.
12. 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.
13. 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.