G.J. Kotzé

RULE-INDUCED ERROR CORRECTION OF ALIGNED PARALLEL TREEBANKS[1]

1. Abstract

Automatic sub-tree 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 can provide a quick and convenient solution. We present an approach that highlights problematic phenomena which enables us to pinpoint regular error patterns for which we can devise rules for correction.

2. Introduction

A parallel treebank is generally defined as a set of translationally equivalent and syntactically annotated sentence pairs. Often, they are aligned on word and/or phrase level. The fast and automatic construction of very large parallel treebanks is considered very useful, especially in the field of machine translation (MT) for acting as training data[2]. In this paper, we focus on the application of a tree-to-tree aligner

[1] The research presented in this paper was done in the context of the PaCo-MT project (STE07007), sponsored by the STEVIN programme of the Dutch Language Union.

[2] *Tinsley J., Hearne M., Way A.* Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation // Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07). Bergen, Norway, 2007. P. 175–187; *Vandeghinste V., Martens S.* Bottom-up transfer in Example-based Machine Translation // Proceedings of EAMT 2010. European Association for Machine Translation. Saint-Raphael; *Jun Sun, Min Zhang, Chew Lim Tan*. Exploring Syntactic Structural Features for Sub-Tree Alignment using Bilingual Tree Kernels // Proceedings of the 48th Annual

called Lingua-Align[1] which requires both sides of the sentence pair to be syntactically annotated. We propose a method of highlighting systematic alignment errors. Using this information, we build a set of rules that we apply to correct some errors in a large parallel treebank. We show how the application of just two sets of rules significantly increases alignment coverage while a manual inspection of the output shows promise.

2. Data and setup

Lingua-Align, our tree aligner of choice, is a statistical tree aligner requiring training data to build alignment models. In the current setup, word alignment is done separately by other software. The tree aligner then uses a combination of existing word alignments and user-defined features to build a maximum entropy model which predicts an alignment probability for every nonterminal node pair considered.

Our data set of choice is the Dutch/English versions of the Europarl3 corpus[2]. The sentences are aligned using the aligner distributed with the corpus, which amounts to a total of 1,080,417 sentence pairs.

To create word alignments, we utilize GIZA++[3] and Moses[4]. We

---

Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 11–16 July 2010. P. 306–315.

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

[2] *Koehn P.* A Parallel Corpus for Statistical Machine Translation// Proceedings of MT-Summit, 2005.

[3] *Och F. J., Ney H.* A Systematic Comparison of Various Statistical Alignment Models // Computational Linguistics. Vol. 29, Nr. 1. March 2003. P. 19–51.

[4] *Koehn P., Hoang H., Birch A., Callison-Burch Ch., Federico M., Bertoldi N., Cowan B, Shen W., Moran Ch., Zens R., Dyer Ch., Bojar O., Constantin A., Herbst E.* Moses: Open Source Toolkit for Statistical Machine

use a combination of heuristics in an attempt to strike a good balance between high precision and high recall.

We select a random set of 140 sentence pairs which already contain word alignments, and draw links between the nonterminal nodes using the Stockholm TreeAligner[1]. This produces an XML file that can be used by Lingua-Align to train an alignment model.

Specifying various parameters and features such as relative node positions and number of word alignments in the subtrees, we train a model on the whole set using ten-fold cross validation, yielding an average balanced F-score of 73.43.

We proceed to use this model to align the whole Europarl corpus.

3. Error finding and rule creation

We now have for every sentence pair a gold standard consisting of manually created links and a set consisting of automatically created links. In both sets, the word alignments are fixed. We distinguish between precision mismatches, where a link is present in the output but not in the gold standard, and recall mismatches, 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 set of statistics, with a list of mismatch examples corresponding to a category label combination. These examples include sentence IDs and matching word phrases, so that they are easily found when viewing the alignments with the Stockholm

---

Translation // Proceedings of the ACL 2007 Demo and Poster Sessions. Prague. June 2007. P. 177–180.

[1] *Lundborg J., Marek T., Mettler M., Volk M.* Using the Stockholm TreeAligner// Proceedings of the 6th Workshop on Treebanks and Linguistic Theories. Bergen, Norway, 2007. P. 73–78.

TreeAligner. Note that we conventionally chose Dutch as the source and English as the target language.

Table 1. Examples of link mismatches

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(...
inf	S	19_14(het_werken_vanaf...-19_524(quite...

A study of those cases reveal a few systematic errors that can be corrected using a rule-based approach. We notice that in general, recall is much lower than precision (68.94 versus 78.63), and we therefore opted for the correction of systematic recall mismatches.

Often, a pair of nonterminal nodes should generally be linked since they govern terminal nodes that are exclusively linked between the two subtrees. In our first experiment, we apply a rule consisting of the following steps:

1. For every unlinked source nonterminal node:
2. If all the children are terminal nodes:
3. Get the target side nodes to which these terminal nodes link.
4. If these nodes share the same unlinked parent and the parent's children all link to children of the node in (1):
5. Link this parent with the current source nonterminal node.

Figure displays an example of such an added link.

Example of an added link (np/NP)

The already aligned Europarl corpus consists of 4,721,670 nonterminal links, of which this rule alone added 573,236 links, or 12.14% of the former. To also take nonterminals with exclusively linked terminals that do not only have terminals as children into account, we adapt our previous rule in the following way:

1. For every unlinked source tree nonterminal node:
2. If there are any linked terminal nodes in the subtree:
3. Get the set of target side nodes to which these terminal nodes link.
4. Get the lowest common parent that they share.
5. Get all the leafs of this parent's subtree.
6. If these leafs link to source side terminals that have the current nonterminal source node as a shared parent:
7. Link this target side parent node, if it is unlinked, with the current source side node.

This added another 194,571 links (4.12% of the number of existing nonterminal links in the original version). To gauge the accuracy, we extracted a random set of 120 links added by the first rule

and 120 links added by the second rule, and gave each link a score of good, bad or fuzzy/unsure. The results are displayed in table 2.

Table 2. Results of manual evaluation

	Good	**Bad**	**Fuzzy/unsure**
Rule 1	90	10	20
Rule 2	46	44	30

Clearly, the first rule outperforms the second rule by far. Even though this is an extremely small sample, the results are promising and we can proceed to test the effect of rule 1 on a machine translation system. Many of the results of rule 2 are syntactic mismatches rather than outright wrong alignments. For example, a VP is often misaligned with an NP which forms part of the VP that must be aligned instead. It should be possible to refine the rule to take into account the relative heights of the nodes or the relative lengths of the phrases in question.

4. Conclusion and future work

We have shown a relatively simple way to systematically improve a statistical 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 its size 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.

In the future, we hope to utilize machine learning to automatically learn rules for error correction. We may also look at correcting tree structures. Finally, our work may provide more insight into general shortcomings of current statistical tree alignment and how to overcome them.