Using Tectogrammatical Alignment in Phrase-Based Machine Translation

David Mareček

Charles University, Institute of Formal and Applied Linguistics, Prague, Czech Republic.

Abstract. In this paper, we describe an experiment whose goal is to improve the quality of machine translation. Phrase-based machine translation, which is the state-of-the-art in the field of statistical machine translation, learns its phrase tables from large parallel corpora, which have to be aligned on the word level. The most common word-alignment tool is GIZA++. It is very universal and language independent. In this text, we introduce a different approach – the tectogrammatical alignment. It works on content (autosemantic) words only, but on these words it widely outperforms GIZA++. The GIZA++ word-alignment can be therefore improved using tectogrammatical alignment and if we use this improved alignment for training phrase-based automatic translators, the translation quality also slightly increases.

Introduction

Machine translation is a standard and increasingly popular NLP task. There exists a lot of approaches, from rule-based ones through many hybrid ones to purely statistical approaches. They differ also in the level of transfer. The Moses toolkit [Koehn et al., 2007] represents the phrase-based machine translation and operates only on the word level. On the other side, TectoMT system [Žabokrtský et al., 2008] first analyzes the source sentence up to the deep syntactic tree and then it makes the transfer step and generates back the target sentence.

In this work, we describe an improvement of the word-alignment step. This step is very useful for many types of machine translation systems, because translation dictionaries and phrase tables can be very easily automatically generated from the alignment.

In several works, it was shown that changing the quality of alignment does not influence the quality of translations much. Lopez and Resnik [2006] made an experiment concerned with degrading the word alignment. They applied GIZA++ on smaller chunks of the parallel corpus instead of one run over the whole corpus. They found that although the alignment error rate considerably decreased, the translation quality did not change so much. Fraser and Marcu [2007] presented an empirical study of alignment evaluation metrics. They conclude that the standard metric AER [Och and Ney, 2003] is not a good metric for word alignment which is supposed to be used for training statistical machine translators, because it prefers sparser alignment.

Our effort is to improve GIZA++ word alignment using the alignment of tectogrammatical trees. In tectogrammatics (as introduced in Functional Generative Description by Sgall [1967], and implemented in the Prague Dependency Treebank [Hajič et al., 2006]), each sentence is represented by a tectogrammatical tree (t-tree for short). T-tree is a rooted dependency deep-syntactic tree. Example of an English t-tree is shown in Figure 1. Unlike in the surface syntax, only content (autosemantic) words have their own nodes in the t-trees. Function words such as prepositions, articles, auxiliary verbs, subordinating conjunctions, articles, and modal verbs are represented differently: for instance, there is no node representing auxiliary verb *has* in the t-tree example, but one of the functions it conveys is reflected by attribute *tense* attached to the autosemantic verb's node (*be*). Other attributes describe several cognitive, syntactic and morphological categories. The presence of an attribute in a node is determined by the node type.

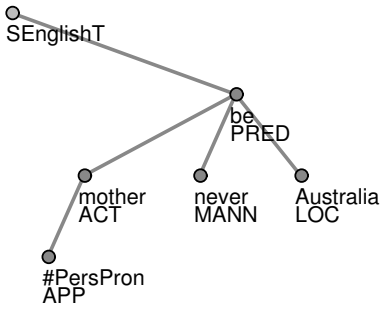

Figure 1. Tectogrammatical representation of the English sentence "My mother has never been to Australia". The visualization is simplified: only t-lemma and functor attributes are depicted with the nodes.

Differences between word and tectogrammatical alignment

There is an example of parallel sentence aligned on the word level in Figure 2. The solid arrows represent the obvious alignment pairs, whereas the correspondence expressed by the dashed arrows is not straightforward. For example, there is only one negation word *No* in the English sentence while in the Czech one, there is the negation in both *Žádné* and *nebylo*. The

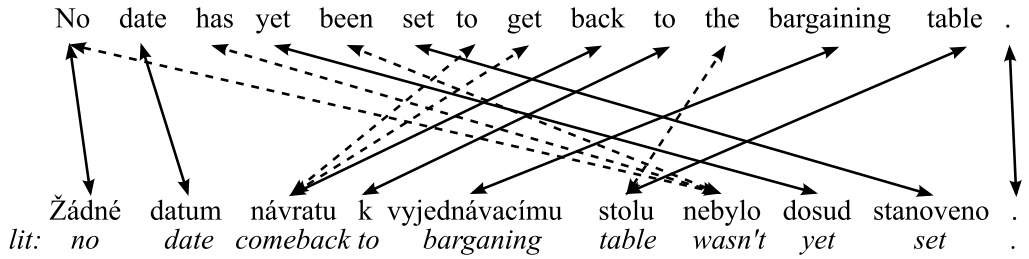

Figure 2. Example of word alignment. Connections, which are only "possible", are dashed.

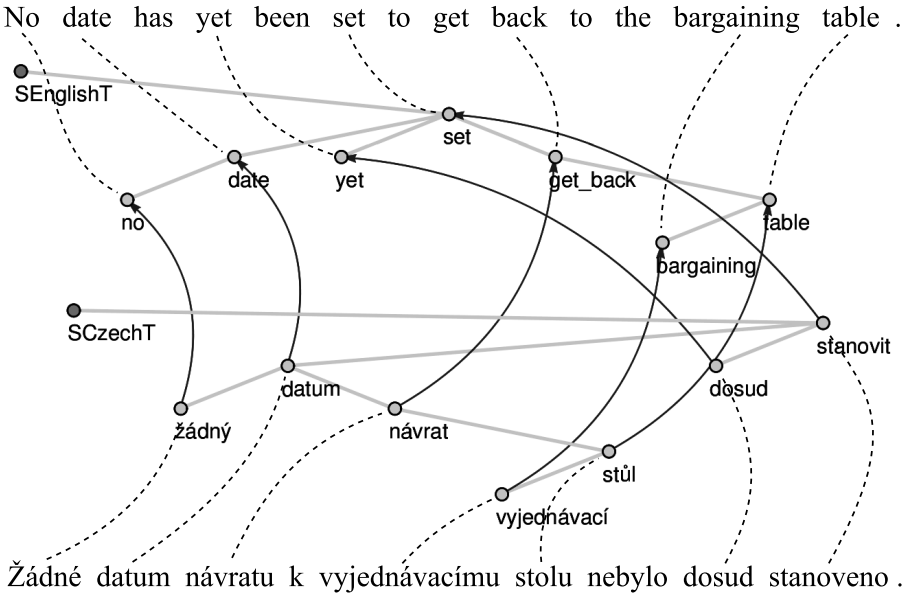

Figure 3. Tectogrammatical alignment and its transfer to the surface.

word *nebylo* can be translated into English as *wasn't*, but if the word *dosud* follows, the only possibility is the present perfect tense form *has been*. The word *dosud* has thus a relationship with the present perfect tense and should be linked besides *yet* also with *has* and *been*. This example illustrates the fact that the word-alignment of Czech-English sentence pairs is rather complex.

On the tectogrammatical, layer the Czech and English sentence trees are more similar compared to the similarity of their surface shapes. Alignment on t-layer for the same sentence pair as in Figure 2 is shown in Figure 3. We can see that the alignment pairs made in t-trees are exactly those that were aligned as evident (solid arrows) on the surface.

GIZA++ Word Alignment

GIZA++ [Och and Ney, 2003] is a standard alignment tool which implements all IBM models [Brown et al., 1993] and Hidden Markov Models. It is based on unsupervised learning so that it is very universal and language independent. GIZA++ takes a set of tokenized parallel sentences on the input and returns a set of connections between corresponding words.

For morphologically richer languages like Czech it is advisable to lemmatize the sentences before running GIZA++. Bojar and Prokopová [2006] showed that lemmatization of the input text reduces the Czech vocabulary size to a half. Thus the vocabulary sizes of Czech and English become comparable. The data are thus not so sparse, which helps alignment error rate by about 10% absolute.

The GIZA++ output alignment is asymmetric. For each word from the source sentence, at most one counterpart in the target sentence is found. There can be only many-to-one mappings but no one-to-many mapping. To symmetrize the output, we run GIZA++ twice in both directions (source-to-target and target-to-source) and then merge these two outputs together using any of the following symmetrization methods.

The apparent symmetrization is *union* and *intersection* of the acquired two alignments. Other symmetrization methods, *grow*, *grow-diag*, *grow-diag-final*, and *grow-diag-final-and*, described in [Och and Ney, 2003], are somewhere between. They include all the connections form *intersection* alignment and add some connections from the *union*. The following inclusions hold for them:

$$\textit{intersection} \subset \textit{grow} \subset \textit{grow-diag} \subset \textit{grow-diag-final-and} \subset \textit{grow-diag-final} \subset \textit{union}$$

T-aligner

This section briefly describes the tectogrammatical alignment tool T-aligner [Mareček, 2008]. The algorithm is simple, for each node in the tectogrammatical tree the most probable counterpart in the other tree is found. The scores of the potential connections are obtained using an implementation of the discriminative reranker described by Collins [2002] (basically a modification of averaged perceptron).

A score of each connection is computed from values of features. We use features reflecting similarities of tectogrammatical lemmas, equality of their semantic part-of-speech, similarity in relative linear position of the nodes within the sentences, and similarities of their child and parent t-nodes. Several features take into account whether GIZA++ aligned the examined pair on the surface or not; some features carry information from our probabilistic translation dictionary.

The output from T-aligner is also asymmetric, since only one counterpart is found for each node. Therefore, similarly to GIZA++, we have to run it twice in both directions and then we can make the *intersection* or the *union* alignment. Figure 3 depicts a simple method to transfer the tectogrammatical alignment to the word layer. Each tectogrammatical node has an attribute which points to the word which it obtained its lexical meaning from.

Alignment Evaluation and Combined Alignment

For the purposes of evaluating alignment systems, we compiled a manually annotated data set [Mareček, 2008]. It consists of 2,500 pairs of sentences from several different domains: newspaper articles, E-books, short stories and also EU law. Each sentence pair was manually aligned on the word level independently by two annotators. We used the same annotation schema as Bojar and Prokopová [2006] and their data are also included in our set. The two obtained annotations were then merged in order to get only one "golden" alignment, that includes only *sure* and *possible* connections as it is in our example in Figure 2.

Evaluation Metrics

We use the following metrics for evaluating the alignment quality: precision, recall, and alignment error rate (AER) described by Och and Ney [2003]. Precision and recall are defined by the formulas:

$$Prec = \frac{|(P \cup S) \cap A|}{|A|}, \qquad Rec = \frac{|S \cap A|}{|S|},$$

where S is the set of *sure* links, P is the set of the *possible* links and A is the set of links suggested by the evaluated automatic aligner. Obviously, asserting connections that were neither sure nor possible causes lower precision, whereas omitting sure connections causes lower recall. AER combines both views and is computed using the formula

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

Evaluation Results

Table 1 presents the evaluation of GIZA++ against people annotations. The evaluation is done for all the words and also for the content words only (for those which have its own node in the tectogrammatical tree), so that we could compare the results with the T-aligner.

Table 1. Evaluation results for GIZA++ on all the words and on the content words only.

Symmetrization	All words			Content words only		
	precision	recall	AER	precision	recall	AER
intersection	95.8	79.0	**13.2**	97.8	82.2	**10.6**
grow-diag-final	71.5	92.0	20.3	78.5	93.6	14.7
union	68.5	93.2	22.1	74.3	94.7	17.1

We can see that while the *intersection* symmetrization is too sparse (precision is much higher than recall), the *grow-diag-final* and *union* symmetrizations have the opposite problem. Also the problem with AER [Fraser and Marcu, 2007] is confirmed. The lowest AER is achieved by *intersection* alignment which is however too sparse for machine translation training. We also documented that the alignment of content words is less problematic than the alignment of other (mostly functional) words.

Table 2. T-aligner evaluation.

Alignment tool	All words			Content words only		
	precision	recall	AER	precision	recall	AER
GIZA++ (intersection)	95.8	79.0	13.2	97.8	82.2	10.6
T-aligner	–	–	–	96.0	89.7	**7.3**
Combined alignment	94.3	84.6	**10.7**	–	–	–

The T-aligner evaluation is compared with the results of GIZA++ (*intersection* symmetrization) in Table 2. Our T-aligner outperforms GIZA++ on content words by 3.3% absolute. The problem is that no functional words are aligned by T-aligner.

Therefore we introduce a "combined alignment", which uses both of the two presented approaches. We take the GIZA++ alignment and substitute the connections between the content words with the connections acquired by the T-aligner. This new word alignment has 2.5% lower AER compared to the word alignment produced by GIZA++.

Application of Combined Alignment in Machine Translation

Besides the intristic word alignment evaluation using AER, it is desirable to test the usefulness of the alignment in machine translation. For this purposes we have a chosen commonly used phrase-based statistical machine translation toolkit Moses [Koehn et al., 2007]. Moses is trained, tuned, and evaluated separately for each examined word alignment, so that the translation quality could be compared. We translate English sentences into Czech and use the data set form WMT08[1]. The data consists of about 80,000 training parallel sentences (commentaries from Project Syndicate parallel corpus) and other 1,000 and 2,000 sentence pairs for tuning and evaluation respectively. In all the experiments we tuned the parameters using MERT.

We examine both GIZA++ alignment and combined alignment. Both the alignments are symmetrized using all the six presented symmetrization methods. The results are in Table 3. The quality of translation is measured using two different metrics – BLEU score and SemPOS [Kos and Bojar, 2009], which has a better correlation with the human judgements.

Table 3. BLEU and SemPOS scores for the GIZA++ and combined word alignment using various symmetrization methods.

	BLEU		**SemPOS**	
Symmetrization method	**GIZA++ alignment**	**Combined alignment**	**GIZA++ alignment**	**Combined alignment**
intersection	12.37	12.46	44.34	44.86
grow	12.53	12.60		
grow-diag	12.80	12.82		
grow-diag-final-and	12.93	**13.00**	45.52	**46.20**
grow-diag-final	12.91	12.64		
union	12.96	12.64	45.99	45.40

For most of the symmetrizations the translation quality is slightly higher when combined alignment is used instead of GIZA++. The best translation was achieved with *grow-diag-final-and* symmetrization an it holds both for BLEU score and for SemPOS metric. However, the differences in scores are very low (only 0.07 BLEU points).

Conclusions

We have presented a method for improving standard GIZA++ word alignment. The tectogrammatical alignment, which outperforms GIZA++ on content words, can be used for correcting links between content words. This new word alignment has much lower error-rate than GIZA++ (comparing to the human annotations) and also if the phrase-based machine translation is trained on this new alignment, the translation quality slightly increases.

Acknowledgments. The research reported on in this paper has been supported by the Charles University Grant Agency under Contract GAUK 99409 and the Czech Grant Agency under Contract GA201/09/H057.

[1] http://www.statmt.org/wmt08/

References

Bojar, O. and Prokopová, M., Czech-English Word Alignment, in *Proceedings of LREC'06*, pp. 1236–1239, ELRA, 2006.

Brown, P. F., Pietra, S. D., Pietra, V. J. D., and Mercer, R. L., The Mathematic of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, *19*, 263–311, 1993.

Collins, M., Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, in *Proceedings of EMNLP*, vol. 10, pp. 1–8, 2002.

Fraser, A. and Marcu, D., Measuring word alignment quality for statistical machine translation, *Computational Linguistics*, *33*, 293–303, 2007.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., and Mikulová, M., Prague Dependency Treebank 2.0, LDC, Philadelphia, 2006.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E., Moses: Open Source Toolkit for Statistical Machine Translation, in *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 177–180, Association for Computational Linguistics, Prague, Czech Republic, 2007.

Kos, K. and Bojar, O., Evaluation of Machine Translation Metrics for Czech as the Target Language, *Prague Bulletin of Mathematical Linguistics*, *92*, 2009.

Lopez, A. and Resnik, P., Word-based alignment, phrase-based translation: Whats the link?, in *Proceedings of AMTA 2006*, p. 9099, 2006.

Mareček, D., *Automatic Alignment of Tectogrammatical Trees from Czech-English Parallel Corpus*, Master's thesis, Charles University in Prague, 2008.

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

Sgall, P., *Generativní popis jazyka a česká deklinace*, Academia, Prague, Czech Republic, 1967.

Žabokrtský, Z., Ptáček, J., and Pajas, P., TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer, in *Proceedings of the 3rd Workshop on Statistical MT, ACL*, 2008.