Combining Models for the Alignment of Parallel Syntactic Trees

Josue G. Araújo[1], **Helena M. Caseli**[1]

[1]Department of Computer Science
Federal University of São Carlos (UFSCar)
Rod. Washington Luís, km 235 – CP 676
CEP 13565-905, São Carlos, SP, Brazi

{josue_araujo,helenacaseli}@dc.ufscar.br

Abstract. *The alignment of syntactic trees is the task of aligning the internal and leaf nodes of two sentences in different languages structured as trees. The output of the alignment can be used, for instance, as knowledge resource for learning translation rules (for rule-based machine translation systems) or models (for statistical machine translation systems). This paper presents some experiments carried out based on two syntactic tree alignment algorithms presented in [Lavie et al. 2008] and [Tinsley et al. 2007]. Aiming at improving the performance of internal nodes alignment, some approaches for combining the output of these two algorithms were evaluated in Brazilian Portuguese and English parallel trees.*

1. Introduction

The alignment of syntactic trees is the task of aligning the internal and leaf nodes of two sentences in different languages structured as trees. More specifically, the parallel sentences are represented by syntactic trees generated separately for the source and target languages. From this pair of syntactic trees, automatic methods determine the correspondences between source and target nodes (internal and leaf ones). The alignment produced by the automatic methods can be very useful for Machine Translation (MT).

This paper, therefore, proposes the combination of two syntactic tree alignment methods —[Lavie et al. 2008] (a bottom-up approach) and [Tinsley et al. 2007] (a top-down approach) — aiming at improving their performance evaluated on Brazilian Portuguese (pt) and English (en) pair of languages. Moreover, some lexical alignment filters are proposed to filter out the misaligned leaf nodes. The investigated hypotheses are: (i) it is possible to combine the baseline alignment methods and their features and also (ii) a good lexical alignment of leaf nodes can improve the quality of internal nodes alignment.

2. Related Work

According to related work, it is possible to note that the alignment of syntactic trees is divided in two steps. First, the lexical alignment is applied to align the leaf nodes, then, the other (internal) nodes are aligned. Furthermore, there is a wellformedness criterion for creating internal alignments which states that an ascendant node in the source tree may only be aligned with an ascendant node in the target tree, regarding the previously aligned node. The same is true for descendant nodes: a descendant node in the source tree can only be aligned with a descendant node in the target tree.

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 169–173,
Cuiabá, MT, Brazil, October 24–26, 2011. ©2011 Sociedade Brasileira de Computação

After the alignment of leaf nodes, the internal nodes are aligned following various approaches and distinct criteria. For instance, the method presented in [Lavie et al. 2008] assigns a prime number to each pair of aligned leaf nodes in source and target trees based on the lexical alignment. This alignment is propagated to the highest nodes in a way that the ascendant nodes receive the product of their children, and the internal nodes of both trees with the same resultant value are aligned.

Similarly, in [Tinsley et al. 2007] the alignment of internal nodes is accomplished using the alignment probabilities of leaf nodes generated by GIZA++ [Och and Ney 2003]. In this case, the product of the probabilities of lexical alignment (not prime numbers as [Lavie et al. 2008]) is assigned to parent nodes. In [Menezes and Richardson 2001] and [Groves et al. 2004], the proposed methods automatically align fragments of the source tree with the equivalent target tree fragment quickly and consistently using a best-first approach and composition rules. Some other methods, such as [Marecek et al. 2008] and [Tiedemann and Kotzé 2009], use different resources for the alignment of syntactic trees as: prefix analysis, part-of-speech and organization of words in the sentence (linear position).

For the experiments presented in this paper, the baseline models were implemented based on [Lavie et al. 2008] and [Tinsley et al. 2007] mainly because they do not require rich resources such as [Marecek et al. 2008] neither use manually created composition rules as [Menezes and Richardson 2001] and [Groves et al. 2004].

3. Models for Aligning Parallel Syntactic Trees

3.1. Model 1 – Based on [Lavie et al. 2008]

Following an idea similar to that described in [Lavie et al. 2008], our implementation (model 1) assigns prime numbers to each pair of aligned terminal nodes[1]. For those non-aligned terminal nodes, model 1 assigns the value 1 and for those nodes with multiple alignments, it assigns the product of the prime numbers of each alignment.

Then, in a second step, the values are propagated to the internal nodes (a bottom-up approach): the value assigned to a parent node is the product of the values assigned to its child nodes. Finally, the value of each node in the source tree is compared to the values of each target node and the source and target nodes with the same value are aligned.

3.2. Model 2 – Based on [Tinsley et al. 2007]

As proposed in [Tinsley et al. 2007], the model 2 uses the probability generated by GIZA++ [Och and Ney 2003]. For each node in the source tree, model 2 calculates the alignment probability regarding each target node in the parallel tree. These values are organized in a matrix and, in each interaction, the pair of nodes with the highest score is aligned. The restriction of aligning each node only once has also been followed. So, different from model 1, model 2 only generates 1-to-1 alignments.

3.3. Models 3-5

Aiming at improving the performance of baseline alignment models, three extended models were implemented as shown in Figure 1. Note that the input of all extended models is the output of model 1 and model 2.

[1]The original model presented in [Lavie et al. 2008] uses prime factorization.

Model 3 is the union between model 1 and model 2, and was implemented in an attempt to improve the recall of the parallel syntactic tree alignment process by joining the output of both baseline models. Model 4, in turn, implements the intersection between the alignment generated by models 1 and 2. By doing so model 4 tries to improve the precision of the parallel syntactic tree alignment process and it is composed by all the pairs of parallel nodes aligned by both models and not only one of them as in model 3.

Finally, model 5 is the merge of models 1 and 2. In this case, the merge is the application of model 2 to filter out ambiguous alignments generated by model 1: when a node has more than one alignment (remember that model 1 is able to generate 1-to-many alignments), model 5 outputs only the one aligned by model 2.

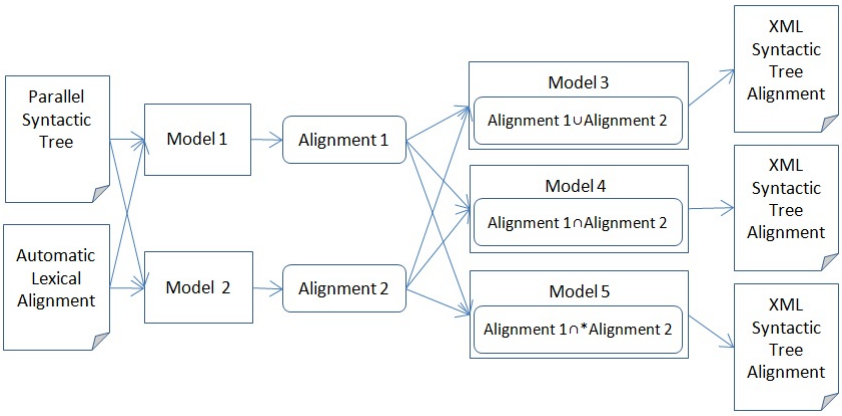

Figure 1. Diagram of the five syntactic tree alignment models

3.4. The Lexical Alignment Filters

The lexical alignment was generated by the union of GIZA++'s output in source-target and target-source alignment directions. This union-alignment was evaluated regarding the aligned terminal nodes in the gold standard and achieved 74.63% precision, 93.42% recall and 82.97% F-measure. In order to improve the lexical alignment automatically generated by GIZA++, two new features were defined to filter the alignments based on their *part-of-speech* or *neighborhood*.

In the *part-of-speech filter*, the labels of each leaf node of the source and the target syntactic trees are compared. If they belong to different groups of labels extracted from the lexical alignment of the gold standard, the alignment between them is filtered out. The *neighborhood filter*, in turn, allows only alignments between source and target nodes that occur in similar positions in source and target trees, respectively. By doing so, we try to solve some ambiguities in the lexical alignment filtering out the less probable ones.

4. Experiments and Results

The experiments were carried out in a Brazilian Portuguese (pt) and English (en) parallel corpus containing 108 pairs of syntactic trees. These trees were generated by syntactic parser for pt [Bick 2000] and en [Collins 1999], separately. These trees represent sentences derived from articles of the Pesquisa FAPESP[2] Brazilian magazine. From this test corpus, a gold standard manually aligned by a human expert was produced to serve as reference in the automatic comparison.

[2]http://revistapesquisa.fapesp.br.

Table 1 shows the precision, recall and F scores for the five models. The evaluation was performed considering the 1-to-many alignments (1:n) of model 1 (lines 1, 3-4) and also restricting this model to provide only the 1-to-1 alignments (1:1, lines 5-8). In the left part of this table we can see the results using the GIZA++'s union of source-target and target-source lexical alignments; and on the right, GIZA++'s output filtered by part-of-speech and neighborhood filters.

Table 1. Precision, recall and F scores for models 1-5

	GIZA++			GIZA++ & Filters		
	Precision (%)	Recall (%)	F (%)	Precision (%)	Recall (%)	F (%)
Model 1 (1:n)	94.09	82.63	87.99	91.64	86.97	89.24
Model 2 (1:1)	91.47	76.96	83.59	92.81	76.77	84.03
Model 3 (1:n)	91.10	**91.88**	**91.49**	90.91	**93.48**	**92.18**
Model 4 (1:n)	**95.22**	67.71	79.14	**93.94**	70.25	80.39
Model 1 (1:1)	96.84	66.67	78.97	95.59	69.59	80.54
Model 3 (1:1)	91.81	**87.91**	**89.82**	92.74	**89.24**	**90.96**
Model 4 (1:1)	**97.36**	55.71	70.87	**96.34**	57.13	71.73
Model 5 (1:1)	94.59	72.62	82.16	93.05	75.83	83.56

From this table it is possible to notice that, as expected, model 4 (intersection) improved the precision of the baseline models (1 and 2) while model 3 (union) improved their recall. These results confirm our first hypothesis since we see that it is possible to combine the baseline alignment methods and improve their performance.

Regarding our second hypothesis, applying the filters on GIZA++'s output lead to a better lexical alignment precision but a worser recall.[3] The better precision in lexical alignments improved the recall of internal nodes alignment since the 93.48% of recall in model 3 (1:n) was the best recall achieved in our experiments. However, the same improvement was not achieved for the precision of internal nodes alignment.

5. Conclusions and Future Work

In this paper, we evaluated some models to automatically align the internal nodes of two parallel syntactic trees. The best precision (97.36%) was obtained by the intersection (model 4) while the union (model 3) achieved the best recall (93.48%) and F (92.18%) scores. Model 5 was not the best in any measure, but it improved the precision of model 1 mitigating the decline in recall of model 4.

The next steps in this research are: (i) to apply the best models in the whole corpus of 16,994 pairs of Brazilian Portuguese-English parallel syntactic trees, (ii) to extract translation rules from the aligned parallel trees and (iii) to apply the extracted rules in a machine translation system.

Acknowledgements

We gratefully acknowledge support for this research from CAPES, FAPESP (#2010/07517-0) and PIADRD/UFSCar.

[3]The application of both filters (part-of-speech and neighborhood) lead to an improvement of 10 points (from 74.63% to 84.91%) in precision and a decaying of 2 points in recall (from 93.42% to 91.91%) in the lexical alignment of GIZA++ (the union of both directions).

References

Bick, E. (2000). The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. In *PhD thesis - Aarhus University*, Aarhus, Denmark.

Collins, M. (1999). Headdriven statistical models for natural language parsing. In *PhD thesis - University of Pennsylvania*.

Groves, D., Hearne, M., and Way, A. (2004). Robust Sub-Sentential Alignment of Phrase-Structure Trees. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 1072–1078.

Lavie, A., Parlikar, A., and Ambati, V. (2008). Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In *SSST '08: Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation*, pages 87–95, Columbus, Ohio. Association for Computational Linguistics.

Marecek, D., Zabokrtsky, Z., and Novak, V. (2008). Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In *Proceedings of XII EAMT conference*, Hamburg, Germany.

Menezes, A. and Richardson, S. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at ACL-2001*, pages 39–46, Toulouse, France.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Tiedemann, J. and Kotzé, G. (2009). Building a large machine-aligned parallel treebank. In *Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT'08)*, pages 197–208, Milan, Italy.

Tinsley, J., Zhechev, V., Hearne, M., and Way, A. (2007). Robust language pair-independent sub-tree alignment. In *Proceedings of the MT Summit XI*, pages 467–474, Copenhagen, Denmark.