Alignment of Portuguese-English syntactic trees using part-of-speech filters

Josué G. Araújo and Helena M. Caseli

Department of Computer Science (DC)
Federal University of São Carlos (UFSCar)
Rod. Washington Luís, km 235 – CP 676
CEP 13565-905, São Carlos, SP, Brazil
{josue_araujo,helenacaseli}@dc.ufscar.br

Abstract. The alignment of syntactic trees is the process of finding the correspondences between internal and leaf nodes of two parsing trees representing parallel sentences in different languages. The resource derived from this process can be used, for instance, in Machine Translation (MT) systems to learn translation rules. The model presented in this paper is based on the Prime Factorization and Alignments algorithm (PFA) [1], which uses prime numbers to align parallel trees. Knowing that the lexical alignment influences the alignment of internal nodes, the experiments described in this paper were designed aiming at improving the accuracy of lexical alignments and, thus, verifying the impact of this improvement on the alignment of internal nodes. To do so we used GIZA++ [2] combined with part-of-speech filters.

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 parallel trees. In this case, the sentences are translations of each other and are represented by syntactic trees generated separately for each language. From a pair of syntactic trees like that, the automatic alignment methods find the correspondences between source and target nodes. The resource derived from the alignment process can be used, for example, to learn translation rules useful in Machine Translation (MT) systems.

Machine translation based on syntactic analysis trees (or just syntactic trees) has been extensively studied in the last years due to the general need of improving the performance of the state-of-the-art phrase-based statistical machine translation (PB-SMT) systems. In many syntax-based approaches, the source and target syntactic trees must be aligned to allow the learning of "translation knowledge". To do so, several tree alignment methods have been proposed and evaluated in the literature [1, 3–7]. Following some ideas of them and also implementing new ones, this paper investigates the impact that the lexical alignment of leaf nodes has in the alignment of internal nodes in two parallel syntactic trees.

We also explore the use of a new source of knowledge in filtering the possible lexical alignments of leaf nodes: the part-of-speech tags.

It is important to say that although the experiments presented in this paper were carried out on a specific language pair (Brazilian Portugese and English), the alignment approach used so far is language independent. By doing so, we aim at advancing in scientific research and also developing work with Brazilian Portuguese MT, still an unexplored language when compared to others such as English. To the best of our knowledge this is the first work focusing on the use of part-of-speech filtering as an attempt to improve the alignment of parallel syntactic trees.

In the next section (section 2) we briefly describe the related work with special attention to [1], based on which our proposal was developed. The proposed syntactic tree alignment method is described in section 3 followed by the experiments designed to evaluate the output alignments (section 4). Finally, we present some conclusions and proposals for future work in section 5.

2 Related work

By studying the related work, we can see that it is common to divide the alignment of syntactic trees in two steps. First, the leaf nodes are aligned by means of a lexical alignment. Second, the internal nodes are aligned based on the alignment of the terminal nodes obtained in the first step.

In addition, many syntactic tree alignment methods follow the well-formedness criteria to limit the possible internal alignments to those that satisfy the rules presented in [3]: (i) a node can only be linked once, that is, only 1 : 1 alignments are possible and (ii) descendants (or ancestors) of a source linked node may only link to descendants (or ancestors) of its target linked counterpart. According to those authors, the well-formedness criteria forbid alignments between constituents that cross each other. Although very interesting and controversial, it is out of the scope of this paper to present a study of the well-formedness criteria and their impact on very different syntactically structured languages or different syntax theories/paradigms.[1] In fact, in the implementation presented in this paper we do not follow the first of these criteria, thus, allowing internal alignments different from 1 : 1.

In the alignment of the internal nodes based on the alignment of leaf nodes, the related works follow various approaches and distinct criteria. [1], for example, performs the syntactic tree alignment in three steps. First, the algorithm assigns unique prime numbers to the leaf nodes and the same prime is assigned to the corresponding aligned words in the parallel sentences. Then, the ascendant nodes receive the product of its child nodes and, finally, the internal nodes with the same value in the parallel trees are aligned. This method is explained in details in section 2.1.

[1] The reader interested in see how the well-formedness criteria behave on more dissimilar syntactic structures can find examples for English and Chinese languages in [1].

Similarly, in [3] the alignment of internal nodes is accomplished using the probability of lexical alignment (leaf nodes) generated by GIZA++ [2]. In this case, the product of the probabilities of lexical alignments (not prime numbers as in [1]) is assigned to parent nodes. This method divides the tree in parts called subtrees in a way similar to Data-Oriented Translation (DOT) [8].

In [4], a similar method also splits the trees into subsets called tree fragments. This approach, as well as [5], uses the best-first strategy to align the dependency structure of the trees. The algorithm automatically aligns fragments of the source tree with the target tree fragments corresponding to the equivalent translation, in a quick and consistent way using composition rules.

The model described in [5] also applies composition rules, for example, a rule that aligns the children of aligned parent nodes which have lexical correspondence. For example, in Figure 1, this rule could be applied to align the child nodes "orangutan" and "orangotango" if they were the only non-aligned children of their aligned parent nodes (NPs in source and target trees). There are also some works like [6] and [7] that use other resources for the alignment of syntactic trees as the prefix analysis, part-of-speech and position of words in the sentence (linear position).

In previous experiments [9] we investigated the Prime Factorization and Alignment (PFA) algorithm proposed in [1], described in the next section (2.1). From the results and conclusions derived from these previous experiments, this paper goes a step further changing the initial method to include new alignment criteria derived from the related work as explained in section 3.

2.1 The Prime Factorization and Alignments algorithm

The algorithm of [1], implemented as described in [9], was used as the baseline method in the experiments described in this paper. As already mentioned, this approach uses prime numbers to align source and target nodes. The PFA (Prime Factorization and Alignment) algorithm, initially assigns prime numbers to terminal nodes (leaf nodes) and spreads them to the rest of the tree from the leaf nodes towards the root by assigning the product of child values to their father. In the next step, it analyzes the nodes in both trees looking for similar values. If a source and a target node with the same value are found, then, they are aligned. By using prime numbers, PFA guarantees that the possible alignments involve the same previous aligned nodes since the product of prime numbers is unique.

The assignment of prime numbers to the terminal (leaf) nodes is performed based on the lexical alignment between the source and target trees. For each pair of aligned terminal nodes is assigned a prime number which is the same for the source and the target node.

Usually, the lexical alignment leaves some terminal nodes unaligned. For these unaligned terminal nodes, PFA assigns the value 1 preventing that a poor quality lexical alignment directly influences the process of propagation of prime numbers to the internal nodes. Also aiming at improving the lexical alignment, the PFA algorithm allows the alignment of a leaf node if it belongs to a subtree where the remaining leaf nodes are aligned. This alignment happens in spite of the order of words in the sentence.

The terminal nodes that have more than one alignment are treated in a special way. To keep the meaning of the translation, if a word is aligned with several words in the parallel tree, PFA assigns a prime number for each terminal node and the product of these numbers is assigned to the common node between them. An example of a multiple alignment between terminal nodes is given in Figure 1 in which the English word "*oldest*" is aligned with the Portuguese words "*mais*" and "*antigo*". Note that the term *oldest* receives the product of values assigned to the terms of *mais* and *antigo*. It is also important to say that the well-formedness criteria are applied in [1].

In the evaluation presented in [1], PFA achieved 81.2% precision and 73.2% recall in the alignment of 30 Chinese–English syntactic trees when manual lexical alignment was used in the leaf nodes. These values decreased to 81.1% precision and 29.1% recall when the automatic alignment of the leaf nodes was used in spite of the manual one. The automatic alignment of leaf nodes, in this case, was obtained by GIZA++[2] [2] when the union of the alignment in both directions (source–target and target–source) was used.

3 The tree alignment method using part-of-speech filters

As already mentioned, the alignment of syntactic trees is the process of finding correspondences between internal and leaf nodes of two parallel trees (syntactic trees representing sentences that are translations of each other). To perform this task in the experiment described in this paper, we applied a model based on the PFA algorithm [1] described in section 2.1.

As in [1], our method assigns a prime number to each pair of terminal nodes aligned by the (manual or automatic) lexical aligner and the value 1 to the unaligned terminal nodes. To those nodes with multiple alignments, our method also assigns the product of prime values corresponding to each aligned node. Different from the PFA algorithm [1] described in section 2.1, the current version of our method does not follow the first well-formedness criterion (only 1 : 1 alignments of internal nodes are allowed). By doing so, the proposed method does not restrict the possible alignments even though this behavior leads to lower precision. This decision was made here since the human expert that aligned the Gold Standard was allowed to perform alignments different form 1 : 1. Future experiments will be carried out to evaluate the impact of applying or not the well-formedness criterion.

Based on the PFA algorithm and aiming at improving its result obtained by using only the lexical alignment performed by GIZA++ [2], the method presented in this paper applies part-of-speech filters to improve the accuracy of lexical alignments. The part-of-speech patterns used as filters were defined based on manual lexical alignments generated by a human expert. For each pair

[2] http://code.google.com/p/giza-pp

of aligned terminal nodes, both part-of-speech tags were extracted and stored in a database indicating a possible pair of part-of-speech tags to be aligned.[3] The proposed tree alignment method takes into account only the lexical alignments generated by GIZA++ that matches one of the possible part-of-speech pairs previously stored in the database.

Thus, the input for our tree alignment method is a pair of parallel syntactic trees lexically aligned by GIZA++ and already filtered based on their part-of-speech tags. As in PFA (see section 2.1), the tree alignment is performed in three steps. First, it is assigned a prime number to each alignment between terminal nodes. Then, the assigned values of terminal nodes are propagated to internal nodes of both trees always considering that the value assigned to a parent node is the product of the values assigned to its child nodes. Finally, the method looks for equal values for source and target internal nodes and if this value is found, the nodes are aligned.

Figure 1 shows a pair of syntactic trees aligned following the described method. Although the alignment between the source and target trees is stored in XML format, it is possible to have a graphical view of them using the TreeAligner[4] tool.

4 Experiments and Results

In the experiments described in this paper we used the corpora (section 4.1) and the evaluation metrics (section 4.2) described in the following sections. The experiments designed to evaluate the proposed method regarding the impact of lexical alignment of leaf nodes in the alignment of internal nodes are reported in section 4.3.

4.1 The test and reference corpora

The corpora used in the experiments described here are composed of Brazilian Portugese and English parallel sentences obtained from articles of the Brazilian scientific magazine *Pesquisa FAPESP*[5]. These sentences were processed by syntactic parsers for Brazilian Portuguese (pt) [10] and English (en) [11] languages. As a result, we built a test corpus containing 108 pairs of pt–en syntactic trees. The English syntactic trees contain 3,273 terminal and 2,743 non-terminal nodes while the Portuguese trees have 3,115 terminal and 1,784 non-terminal nodes. This set of 108 parallel trees is just a small sample of the total set of 16,994 pt–en parallel syntactic trees available for future experiments.

[3] For example, the most frequent alignments of terminal nodes were those involving nouns in Portuguese and English. For the part-of-speech tag “N” (in Portuguese), examples of possible English tags found in our corpus were: “NN” (noun, singular or mass), “NNP” (proper noun, singular), “NNS” (noun, plural), “CD” (cardinal number), “JJR” (adjective, comparative) among others.

[4] http://www.cl.uzh.ch/kitt/treealigner

[5] URL of the online version of the magazine *Pesquisa FAPESP*: http://revistapesquisa.fapesp.br

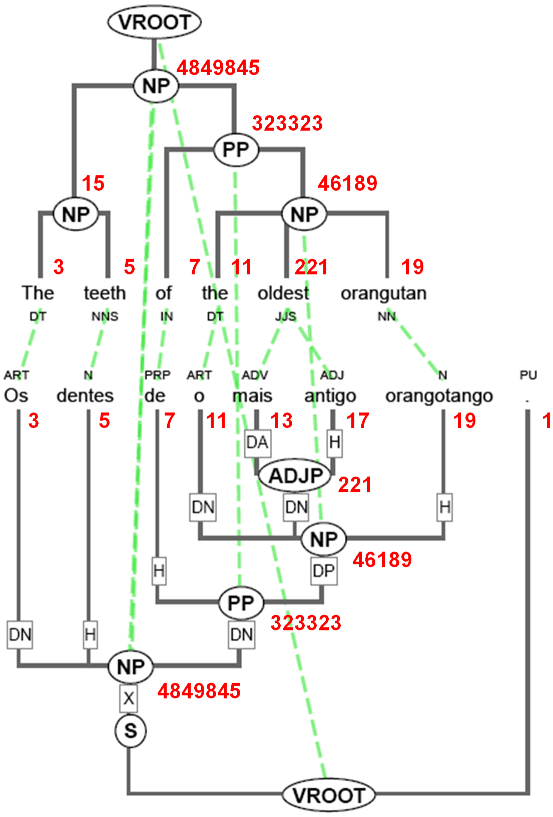

Fig. 1. A pair of parallel syntactic trees aligned by the method proposed in this paper. The prime values are represented by red numbers and the alignment between nodes, by green lines

From this test set we generated a Gold Standard (reference corpus) manually aligned by a human expert. The manual alignment was performed in a two-step process similar to that of the automatic tree alignment methods. In this case, the expert used the TreeAligner tool to graphically view the pairs of syntactic trees originally stored in the TigerXML[6] format. So, from the graphical representation, the human expert aligned the terminal nodes (leaves containing the superficial forms of words) and, then, the non-terminal nodes that represent the syntactic structure of the trees. In this manual alignment process, the expert classified each alignment as very reliable (good) or not (fuzzy) according to the acceptance of such correspondence. The resulting reference corpus has 1,021 internal alignments classified as “good” and only two alignments classified as “fuzzy”.

[6] http://www.ims.uni-stuttgart.de/projekte/TIGER/

4.2 Evaluation metrics

The alignment of syntactic trees was evaluated intrinsically, that is, we evaluated the quality of the alignment itself instead of using the aligned trees in some other application. The metrics used in this evaluation were precision (1), recall (2) and F (3), in which P represents the alignments in the Gold Standard and A, the alignments output by the automatic aligner.

$$\text{Precision} = \frac{|\, P \, \cap \, A \,|}{|\, A \,|} \tag{1}$$

$$\text{Recall} = \frac{|\, P \, \cap \, A \,|}{|\, P \,|} \tag{2}$$

$$\text{F} = 2\, \frac{\text{Precision} \, \times \, \text{Recall}}{\text{Precision} \, + \, \text{Recall}} \tag{3}$$

Precision measures the accuracy of the automatic method, while the recall checks its ability to align the highest possible number of nodes. F, in turn, is the harmonic mean of the two previous values. Values for each of these measures varies between 0 and 1 and the more close to 1, the better is the quality of the alignment. It is important to say that in our experiments only the alignments classified as "good" in the Gold Standard were taken into account to form the P value.

4.3 Experiments

The experiments described in this paper were designed to analyze the impact of the lexical alignment of leaf nodes in the alignment of internal nodes. More specifically, we are interested in measuring the impact of part-of-speech filters applied to lexical alignments following the tree alignment method described in section 3. To allow the comparison with other related methods, we evaluated the alignment of internal nodes regarding the "pure" implementation of PFA (without part-of-speech filters) using the lexical alignment of leaf nodes generated by:

- **Manual Lexical Alignment:** the alignment of the leaf nodes in the Gold Standard manually aligned by the human expert;
- **Automatic Lexical Alignment:** the alignment of the leaf nodes generated by GIZA++ (default configuration).

Table 1 gives the result of the alignment of internal nodes using our implementation of PFA based on the manual and GIZA++'s lexical alignments. Based on the manual alignment, our "pure" PFA method aligned 1,038 internal nodes that were compared with 1,021 aligned internal nodes in the Gold Standard. These values are close to the ones reported in [1] regarding the manual lexical alignment. When GIZA++'s alignment was taken into account, three lexical alignments were considered: pt–en alignment, en–pt alignment and the union of both directions. The values obtained here are also close to those reported in [1].

Table 1. Alignment of **internal nodes** produced by our implementation of PFA based on manual and GIZA++'s lexical alignments [9]

lexical alignment	precision	recall	F
manual	82.6%	84.0%	83.3%
GIZA++ pt–en	73.4%	27.6%	40.1%
GIZA++ en–pt	68.4%	22.5%	33.8%
GIZA++ union	72.6%	22.0%	33.7%

From the values on Table 1 we can reach the same conclusion of [1]: the quality of lexical alignment has impact on the alignment of internal nodes in parallel syntactic trees. In our experiments, this impact is greater in recall (that decreased from 84% in manual to 22–27.6% in automatic lexical alignment) than in precision (which decreased from 82.6% in manual to 68–73.4 % in automatic lexical alignment). It is important to say that the bigger impact in recall is directly related to how the alignment method described in section 3 works. More specifically, in this method a very high weight is given to the lexical alignments, since the unaligned leaf nodes are ignored (by assigning the value 1) and the wrong lexical alignments are propagated (through the multiplication of prime numbers) to generate the internal node alignment.

Trying to decrease the impact of wrong lexical alignments in the internal node alignment, mainly on recall, the proposed method improve the quality of automatic lexical alignment decreasing the number of lexical nodes incorrectly aligned by applying part-of-speech filters as mentioned in section 3.

Table 2 brings the values of precision, recall and F for the alignment of internal nodes using our proposed method. Comparing the values in Tables 1 and 2, we can say that although the precision of internal node alignment has decreased, the recall increased what reflected in the F measure. It is important to mention that the decision to prioritize a better precision or a better recall is related to the application that will use the aligned trees. In the case of learning translation rules, for example, high recall can be more important than high precision.

Table 2. Alignment of **internal nodes** generated by our method (PFA based on the part-of-speech filtered lexical alignments generated by GIZA++)

lexical alignment	precision	recall	F
Proposed method pt–en	60.6%	36.7%	45.7%
Proposed method en–pt	50.9%	31.9%	39.2%
Proposed method union	62.2%	30.5%	40.9%

Analyzing the values on Tables 1 and 2, it is possible to conclude that a lexical alignment with high precision leads to a better recall on the internal node alignment. To make this conclusion clearer, Table 3 shows the quality of automatic lexical alignment using the part-of-speech as a restriction criterion to eliminate erroneous alignments performed by GIZA++ (our proposal).

Table 3. Quality of **lexical alignment** generated by GIZA++ and our method (GIZA++ and part-of-speech filters)

	GIZA++			Our method		
	precision	recall	*F*	precision	recall	*F*
pt-en	40.8	39.9	40.3	62.7	39.9	48.8
en-pt	40.0	36.8	38.3	62.7	36.8	46.4
union	33.0	42.8	37.2	54.1	42.8	47.8

The values on Table 3 were calculated using the same metrics of section 4.2. In this case the terminal nodes aligned automatically were compared with the aligned terminal nodes of the Gold Standard. It is interesting to notice that the recall values remain the same in both cases. This happens since we reduced the number of incorrectly aligned nodes and maintained the number of correctly aligned nodes. By doing so, the precision increased and no impact was noticed in recall since the denominator of the equation (2) (number of aligned nodes in the Gold Standard) remained constant.

About the values of lexical alignment evaluation on Table 3 we can say that our method produced less alignments than the version based only on GIZA++'s basic configuration, but the not generated alignments were those with high probability to be wrong. In pt–en automatic lexical alignment, the amount of aligned terminal nodes reduced from 3,067 output by GIZA++ to 1,998 output by our method, while in en–pt alignment the reduction was from 2,887 to 1,844. In the union of both alignment directions, the amount of aligned terminal nodes decreased from 4,071 output by GIZA++ to 2,483 output by our method.

So, based on the results of our experiments on Tables 1 and 2, we can conclude that: (i) a lexical alignment with high precision leads to a better recall on the internal node alignment (an increase of almost 10% in recall and about 5% in *F*) but (ii) this improvement is still far from the values reached when manual lexical alignment is used (83.3% X 45.7% in *F*, for example).

5 Conclusions and Future Work

In this paper we proposed a way to improve internal node alignment quality by improving the automatic lexical alignment. The experiments were carried out using an implementation of PFA [1] and a test/reference corpora composed of 108 pairs of Brazilian Portuguese and English syntactic trees. From these experiments, we conclude that our proposed strategy of using part-of-speech filters on lexical alignment can increase the precision of lexical alignment performed by GIZA++ and, hence increasing the recall of internal node alignment.

A strong feature of PFA that will be better explored in future experiments is that it seems to be language-independent since the results for Chinese–English [1] and Portuguese–English (this work) were very similar. However, it is not totally independent of the structural representation used in more complex languages such as [6]. Specifically in our work we are interested in dealing with constituent syntactic trees since it is the main approach in literature and the one that fits better to the resources and tools for the language pair under study.

As future work we intend to perform the following changes to the current version of our tree alignment method: (i) to follow the all well-formedness criteria, (ii) to take into account other properties of terminal nodes in addition to their part-of-speech aiming at improving even more the lexical alignments and (iii) to consider new alignment criteria such as the combination metrics of probabilities proposed in [3].

Acknowledgments

We thank the financial support of Brazilian agencies CAPES and FAPESP.

References

1. Lavie, A., Parlikar, A., Ambati, V.: 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, Morristown, NJ, USA, Association for Computational Linguistics (2008) 87–95
2. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1) (2003) 19–51
3. Tinsley, J., Zhechev, V., Hearne, M., Way, A.: Robust language pair-independent sub-tree alignment. In: Proceedings of the MT Summit XI, Copenhagen, Denmark (2007) 467–474
4. Groves, D., Hearne, M., Way, A.: Robust Sub-Sentential Alignment of Phrase-Structure Trees. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING) 2004, Morristown, NJ, USA (2004) 1072–1078
5. Menezes, A., Richardson, S.: 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, Toulouse, France (2001) 39–46
6. Marecek, D., Zabokrtsky, Z., Novak, V.: Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In: Proceedings of XII EAMT conference, Hamburg, Germany (2008)
7. Tiedemann, J., Kotzé, G.: Building a large machine-aligned parallel treebank. In: Proceedings of the 8th International Workshop on Treebanks and Linguistic Theories (TLT'08), Milão, Italy (2009) 197–208
8. Poutsma, A.: Data-Oriented Translation. In: Proceedings of the 18th conference on Computational linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2000) 635–641
9. Araújo, J.G., Caseli, H.M.: Alinhamento de árvores sintáticas português-inglês. In: Proceedings of the V Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence (WTDIA-2010). (2010) 1–10
10. Bick, E.: The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. In: PhD thesis - Aarhus University, Aarhus, Denmark (2000)
11. Collins, M.: Headdriven statistical models for natural language parsing. In: PhD thesis - University of Pennsylvania, verificar (1999)