Incorporating Linguistic Information to Statistical Word-Level Alignment[*]

Eduardo Cendejas, Grettel Barceló, Alexander Gelbukh, and Grigori Sidorov

Center for Computing Research,
National Polytechnic Institute,
Mexico City, Mexico
{ecendejasa07,gbarceloa07}@sagitario.cic.ipn.mx
http://www.gelbukh.com,
http://cic.ipn.mx/~sidorov

Abstract. Parallel texts are enriched by alignment algorithms, thus establishing a relationship between the structures of the implied languages. Depending on the alignment level, the enrichment can be performed on paragraphs, sentences or words, of the expressed content in the source language and its translation. There are two main approaches to perform word-level alignment: statistical or linguistic. Due to the dissimilar grammar rules the languages have, the statistical algorithms usually give lower precision. That is why the development of this type of algorithms is generally aimed at a specific language pair using linguistic techniques. A hybrid alignment system based on the combination of the two traditional approaches is presented in this paper. It provides user-friendly configuration and is adaptable to the computational environment. The system uses linguistic resources and procedures such as identification of cognates, morphological information, syntactic trees, dictionaries, and semantic domains. We show that the system outperforms existing algorithms.

Keywords: Parallel texts, word alignment, linguistic information, dictionary, cognates, semantic domains, morphological information.

1 Introduction

Given a bilingual or multi-lingual corpus, i.e., a set of texts expressing the same meaning in various languages, the text alignment task establishes a correspondence between structures, e.g., words, of the texts in the two languages. For example, given the two texts: English *John loves Mary* and French *Jean aime Marie*, word alignment task consists in establishing the correspondences *John* ↔ *Jean*, *loves* ↔ *aime*, *Mary* ↔ *Marie*.

Text alignment is useful in various areas of natural language processing, such as automatic or computer-aided translation, cross-lingual information retrieval

[*] Work done under partial support of Mexican Government (SNI, CONACYT grant 83270 and 50206-H, and SIP-IPN grant 20090772 and 20091587).

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 387–394, 2009.

and database querying, computational lexicography, contrastive linguistics, terminology, and word sense disambiguation, to mention only a few.

In recent years, many text alignment techniques have been developed [1], [2]. Most of them follow two main approaches: linguistic and statistical or probabilistic [3], [4]. Linguistic approaches use linguistic information, which implies its reliance on availability of linguistic resources for the two languages. Probabilistic approaches involve difficult-to-implement probabilistic generative models. Statistical approaches are simpler since they are based on frequencies of occurrences of words, though they imply high computational cost and usually give lower precision.

In this paper, we present a hybrid alignment algorithm based on a combination of traditional approaches. Its flexibility allows adapting it to the computational environment and to available linguistic resources.

2 Related Work

Ideally, the units (words, sentences, paragraphs) of the two texts ought to be in direct one to one correspondence. However, the alignment task is complicated by many effects that break such an ideal model. One effect is that sometimes the correspondence is not $1 \leftrightarrow 1$ (one word from the source text corresponding to one word in its translation) but $1 \leftrightarrow M$, $M \leftrightarrow 1$, $M \leftrightarrow M$, $1 \leftrightarrow \emptyset$ and $\emptyset \leftrightarrow 1$, where M stands for many words and $\emptyset$ stands for none (empty string). Another effect, specific mainly for the word level, is that the words in the two texts can follow in different order, e.g., English *a difficult problem* vs. Spanish *un problema difícil*.

Most of the alignment systems are oriented on low-inflective languages, for this reason they use wordforms as the basic unit. In the case of highly inflective languages this leads to high data sparseness, rendering statistic translation nearly impossible. For instance, Spanish is a rather highly inflective language, especially in its verbal system, where the complex conjugation produces many wordforms from the same verbal root [5].

It is possible to construct alignment methods based on generative models [6]. Although the standard models can, theoretically, be trained without supervision, in practice several parameters should be optimized using labeled or tagged data. What is more, it is difficult to add characteristics to the standard generative models [7].

Other systems are based on linguistic resources [8], [9]. The use of linguistic resources can present yet another problem for word alignment task. There are two cases as to the use of resources: limited or unlimited [10]. We believe that the more resources are available to the system the better the alignment accuracy. This leads us to the idea of a hybrid combined method for word-level alignment.

This approach is not new. In [11], for instance, a hybrid system is presented, in which the outputs of different existing alignment systems are combined. However, in that approach, interpreting the outputs of the systems is necessary and the user has to define the confidence threshold for each system. De Gispert et al.

proposed in [12] a method to incorporate linguistic knowledge in statistical phrase-based word alignment, but the linguistic information is only used to takes final decisions on unaligned tokens. In [13], parse trees and a few phrase reordering heuristics were incorporated after using the alignment lexicon generated by a statistical word aligner. Another systems have just included morphosyntactic knowledge, as in [14].

3 Alignment Algorithm

The proposed alignment algorithm combines statistical and linguistic approaches. Due to the simplicity of statistical algorithms, approaches of this kind are a starting point for the alignment in our system. Nevertheless, the morphological and syntactical differences between the languages cause multiple errors in such alignment. It is for this reason that, at a later stage, linguistic-based processing is carried out that reinforces or weakens the alignment hypotheses previously obtained with the statistical methods.

3.1 Statistic Processing

There are many well-known statistical alignment methods. Some of them intent to align texts written in very different characters sets, such as English vs. Chinese. The approaches of this paradigm are classified as associative or estimation-based. K-Vec [15] and IBM Models 1 and 2 [16] are examples of associative statistical methods.

The statistical stage of the proposed system relies on three different techniques: (1) Modified K-Vec algorithm, boolean, (2) Modified K-Vec algorithm, with frequencies and (3) IBM Model 2.

Our modified K-Vec algorithm is slightly changed as compared to the original K-Vec presented by Fung & Church [15]. K-Vec algorithm starts with segmentation of the input texts: the texts are divided into small parts and each of the parts is processed independently. The original K-Vec algorithm allows the text to be divided into small pieces or segments. Our modification allows the pieces to be paragraphs, sentences, or a specific number of words (a window). These very convenient division options streamline the statistical process, since its use largely depends on the size of the text segments.

The next step consists in generating a list of words with an associated vector. This vector contains the occurrences of the word in each one of the segments resulting from the division of the text. In the first technique, (modified K-Vec), only boolean values are used to indicate the presence (1) or absence (0) of the word. In the second technique, the frequency of occurrences (i.e., the number of times that the word occurs in a segment) is recorded.

The list of words founded in the text is also used to optimize the later linguistic processing and can also contain the frequency of occurrences of each word in the complete text.

After the list has been completed, the vector corresponding to each word in the source language is compared to all the vectors obtained in the translation,

$$\boxed{\begin{array}{ll}
V(\underline{\textit{este}}) & = \{0, \underline{1}, 0, \underline{1}, 0, 0, \underline{1}, 0, 0, 0\} \\
 & \\
V(\underline{\textit{this}}) & = \{0, \underline{1}, 0, \underline{1}, 0, 0, \underline{1}, 0, 0, 0\} \\
V(\textit{is}) & = \{1, 0, 0, 1, 0, 1, 1, 0, 1, 0\} \\
V(\textit{my}) & = \{0, 0, 0, 0, 0, 0, 1, 1, 0, 0\} \\
V(\underline{\textit{car}}) & = \{0, \underline{1}, 0, \underline{1}, 0, 0, \underline{1}, 1, 0, 1\}
\end{array}}$$

Fig. 1. Comparison among vectors

with the purpose of finding those words that match as to their occurrences in each segment. For example, in Fig. 1, the occurrences of the vector corresponding to *este* coincide with those of the words *this* and *car*, with occur in the segments 2, 4 and 7.

Using the correspondences between the vectors of both languages, a contingency table is built to represent information on each pair of related words. Then, the similarity of the pair is calculated for each table. The similarity of words is determined by means of an association test. Our system incorporates the following similarity measures: – Pointwise Mutual Information (PMI), – T-score, – Log-likelihood ratio, and – Dice coefficient.

After all association values have been calculated, the word with the greatest level of association is selected and the other candidates are discarded. In this way, a dictionary is created from the translation words that better correspond to each source word. If the algorithm is used in a bidirectional way, then the same process is carried out interchanging the source and target languages [17] and the best averages of both results are obtained to acquire the best word pairs.

If the algorithm does not use linguistic information, then after this stage a file of final alignments is created, indentifying each word and its position in both texts.

3.2 Linguistic Processing

The methods developed following the linguistic approach make use of diverse resources, such as bilingual dictionaries [16], lexicons with morphological information [9] and syntactic trees [8]. In addition to these, in our algorithm we incorporate the use of semantic domains.

Dictionaries allow for extraction of lexical information. In this way, the word from the source text is considered along with all its possible translations in the target text. These data can then be employed in the calculation or adjustment of the probabilities of the correspondences obtained in the statistical phase.

Similarly, the morphological and syntactical information are knowledge sources useful for increasing or decreasing the certainty of each alignment hypothesis. Using morphological information, it is possible to compare lemmas and verify grammatical categories [9]. On the other hand, knowing the syntax of the sentences allows the identification of its parts (subject, predicate, etc.) and facilitates comparisons with its counterparts in the target language.

Finally, semantic domains provide a natural way to establish semantic relations between the meanings of a word. Roughly speaking, the method consists of rejecting those translations that lay in different domains from the original word, and giving greater weight to those that lay in the same domains. We used WordNet Domains [18] to extract the domain labels. This is a rarely used concept, utilized to locate or train the aligner in a specific domain [19].

In addition to the above linguistic resources, we use a heuristic of cognates. Shared sequences of characters are looked in both texts, for example: English *organization* and Spanish *organización*. In this way it is easier to align words that totally or partially coincide (for example, proper nouns). The minimum percentage of coinciding letters in the two words to consider them as cognates is a user-defined parameter of the system. False cognates are taken into account by using a predefined list of known false cognates.

Unlike most of the alignment models, where training is carried out with the EM (Expectation Maximisation) algorithm [20], our system allows using previous alignments that can be difficult to find. All the alignment hypotheses that can be obtained with different methods will serve as a reference for future alignment tasks. It is important to mention that the system can start the analysis using purely linguistic data similar to some proposed methods [21], if it is configured to do so.

4 General Architecture

In order to test the proposed ideas, we have implemented a highly configurable and flexible system that we called HWA (Hybrid Word Aligner), which allows our method to be adapted to the implementation environment and the availability of resources. The combination of the statistical and linguistic approaches has the purpose of obtaining a parameterizable algorithm that can be used in different ways depending on the requirements of the expected results.

The architecture of the alignment system is shown in Fig. 2. The alignment process is subdivided into three main phases: preprocessing, main algorithm, and output generation.

5 Preliminary Results

While statistic processing can be applied to any language pairs, the linguistic processing module requires specific grammatical resources, with the exception of the cognate detection[2] and the learning information. We have chosen for our experiments Spanish–English parallel texts from five novels (*Dracula*, *Don Quixote of la Mancha*, *The Shop of Ghosts*, *Little Red Riding Hood* and *The Haunted House*). The selection of fragments was made randomly by the paragraph number. It is important to emphasize that in every case the paragraphs

[2] Providing the languages have the same type of characters.

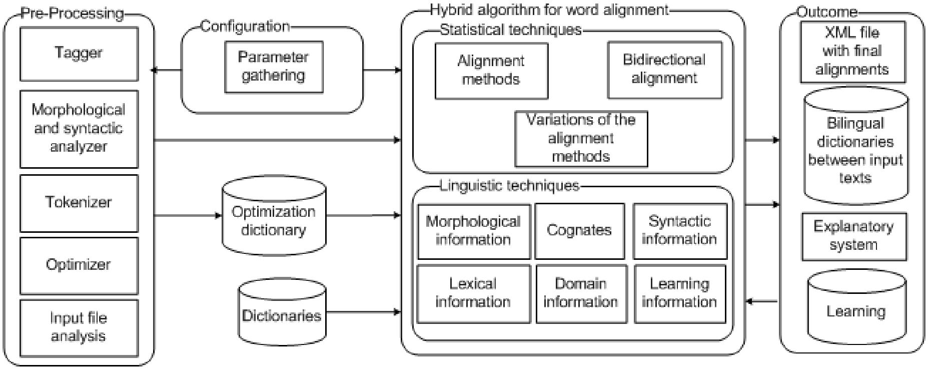

Fig. 2. General architecture of the aligner

Table 1. Results

Aligner	% successes	generation of dictionary?	generation of alignment file?
GIZA++	52	Yes	Yes
Uplug	45	No (not as a result)	Yes
K-Vec++	35	Yes (not for all the words)	No
HWA	53	Yes	Yes

have been previously aligned at sentence level. While this is required for other systems but not for our system, yet this influences the alignment quality.

Table 1 shows the obtained results in terms of precision: the percentage of established correct alignment correspondences. We compare our results with those of three other aligners: GIZA++ [22], original K-Vec [23], and Uplug [24]. The results of the proposed algorithm (HWA) were obtained by executing the modified K-Vec procedure during the statistical processing and by applying the cognates during the linguistic processing to reinforce the alignments. For the moment, we did not apply other linguistic modules, that is why we call the obtained results "preliminary".

To obtain the results, each aligner was provided with the parallel texts of the specified novels. The output of each aligner was manually verified to determine the correct alignments, and an average percentage of the correct alignments was obtained for the five input data sets. Due to the differences in the aligners, similar parameters were used in each test: – input parallel texts were the same for each alignment program, – texts had no labels, – no previous training was applied, – learning bases were not applied, – bidirectional alignments were not performed, – text segmentation was performed with the best consideration of each algorithm, and – the same association test was used. It is important to note that the results from the aligners can vary depending on the configuration parameters and on the size of the input texts.

6 Conclusions

We have presented an alignment algorithm that combines two main approaches, statistical and linguistic-based, in order to improve the alignment between words of bilingual parallel texts. We conducted experiments with short texts fragments. The results obtained by the proposed algorithm are better than those of the existing alignment algorithms.

The statistical techniques provides a good starting point for the alignment processes; however, incorporation of linguistic techniques increases the efficiency of the system by involving intrinsic characteristics of the implied languages. The main disadvantage of linguistic processing is the need in the linguistic resources given their limited availability. In addition, this has an impact on the algorithm speed. Nonetheless, employing databases of optimization has proven to minimize this disadvantage. The cost-benefit trade-off of linguistic techniques implies a great emphasis on the particular configuration of the algorithm so as to obtain the best alignments. This is due to the fact that the system allows free incorporation or exclusion of linguistic resources at any given moment during the process.

Combining statistical and linguistic techniques is a viable option thanks to the current computing capacities and will be more acceptable as the speed of computers grows, costs of hardware (memory and storage) decreases, and more resources become available to natural language processing community.

References

1. Langlais, P., Simard, M., Vronis, J.: Methods and practical issues in evaluating alignment techniques. In: Proceedings of the 17th International Conference on Computational Linguistics, Montréal, pp. 711–717 (1998)
2. Veronis, J.: Parallel Text Processing: Alignment and Use of Translation Corpora. Kluwer Academic Publishers, Dordrecht (2001)
3. McEnery, T., Xiao, R., Tonio, Y.: Corpus-based language studies: An advanced resource book. Routledge, London (2006)
4. Kit, C., Webster, J., Kui, K., Pan, H., Li, H.: Clause alignment for hong kong legal texts: A lexical-based approach. International Journal of Corpus Linguistics 9, 29–51 (2004)
5. Agirre, E., Díaz de Ilarraza, A., Labaka, G., Sarasola, K.: Uso de información morfológica en el alineamiento español-euskara. In: Actas del XXII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, Zaragoza, pp. 257–264 (2006)
6. Dale, R., Moisl, H., Somers, H.L.: Handbook of natural language processing. Marcel Dekker Inc., New York (2000)
7. Moore, R.: A discriminative framework for bilingual word alignment. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, pp. 81–88 (2005)
8. Ma, Y., Ozdowska, S., Sun, Y., Way, A.: Improving word alignment using syntactic dependencies. In: Proceedings of the ACL 2008:HLT Second Workshop on Syntax and Structure in Statistical Translation, Ohio, pp. 69–77 (2008)

9. Pianta, E., Bentivogli, L.: Knowledge intensive word alignment with knowa. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, pp. 1086–1092 (2004)
10. Mihalca, R., Pedersen, T.: An evaluation exercise for word alignment. In: Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: data driven machine translation and beyond, Edmonton, vol. 3, pp. 1–10 (2003)
11. Ayan, N., Borr, B., Habash, N.: Multi-align: Combining linguistic and statistical techniques to improve alignments for adaptable mt. In: Proceedings of the 6th Conference of the Association for Machine Translation in the Americas, Washington DC, pp. 17–26 (2004)
12. De Gispert, A., Mario, J., Crego, J.: Linguistic knowledge in statistical phrase-based word alignment. Natural Language Engineering 12, 91–108 (2006)
13. Hermjakob, U.: Improved word alignment with statistics and linguistic heuristics. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 229–237 (2009)
14. Hwang, Y., Finch, A., Sasaki, Y.: Improving statistical machine translation using shallow linguistic knowledge. Computer Speech and Language 21, 350–372 (2007)
15. Fung, P., Church, K.: K-vec: A new approach for aligning parallel text. In: Proceedings of the 15th Conference on Computational Linguistics, Kyoto, vol. 2, pp. 1096–1102 (1994)
16. Och, F., Ney, H.: A systematic comparison of various statistical alignment models. In: Proceedings of the 18th Conference on Computational Linguistics, vol. 2, pp. 19–51 (2003); Computational Linguistics 29(1), 19–51 (2003)
17. Tiedeman, J.: Word to word alignment strategies. In: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, pp. 221–218 (2004)
18. Bentivogli, L., Forner, P., Magnini, B., Pianta, E.: Revising wordnet domains hierarchy: Semantics, coverage, and balancing. In: Proceedings of COLING 2004 Workshop on Multilingual Linguistic Resources, Geneva, pp. 101–108 (2004)
19. Wu, H., Wang, H., Liu, Z.: Alignment model adaptation for domain-specific word alignment. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Michigan, pp. 467–474 (2005)
20. Och, F., Ney, H.: Improved statistical alignment models. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, Hong Kong, pp. 440–447 (2000)
21. De Gispert, A., Mario, J., Crego, J.: Phrase-based alignment combining corpus cooccurrences and linguistic knowledge. In: Proceedings of the International Workshop on Spoken Language Translation, Kyoto, pp. 85–90 (2004)
22. GIZA++: Training of statistical translation models, `http://www.fjoch.com/GIZA++html`
23. K-Vec++: Approach for finding word correspondences, `http://www.d.umn.edu/tpederse/Code/Readme.K-vec++.v02.txt`
24. Uplug: The home page, `http://stp.ling.uu.se/~joerg/uplug/`