Translation memories enrichment by statistical bilingual segmentation

Francisco Nevado[*], **Francisco Casacuberta**[*], **Josu Landa**[†]

[*]Dept. de Sistemas Informáticos y Computación
Camino de Vera s/n, 46022 Valencia, Spain
{fnevado, fcn}@dsic.upv.es

[†]Ametzagaiña AIE
Zirkuitu Ibilbidea 2-1, 20160 Lasarte-Oria, Spain
jlanda@ametza.com

Abstract

A majority of Machine Aided Translation systems are based on comparisons between a source sentence and reference sentences stored in Translation Memories (TMs). The translation search is done by looking for sentences in a database which are similar to the source sentence. TMs have two basic limitations: the dependency on the repetition of complete sentences and the high cost of building a TM. As human translators do not only remember sentences from their preceding translations, but they also decompose the sentence to be translated and work with smaller units, it would be desirable to enrich the TM database with smaller translation units. This enrichment should also be automatic in order not to increase the cost of building a TM. We propose the application of two automatic bilingual segmentation techniques based on statistical translation methods in order to create new, shorter bilingual segments to be included in a TM database. An evaluation of the two techniques is carried out for a bilingual Basque-Spanish task.

1. Introduction

The majority of Machine Aided Translation systems (Trados, Wordfast, ...) are based on comparisons between a source sentence and reference sentences stored in Translation Memories (TMs). All of the systems based on TMs have a common principle: a text has many sentences which are similar to sentences that occur in other texts and can be reused in new translations. The division of the source text in sentences is usually based on punctuation marks and other rules. Commercial systems usually do not use translation units that are shorter than a sentence.

The translation search is done by similarity: the system is able to look for sentences in a database which are similar to the source sentence (typically, using fuzzy logic). Most of the current translation tools based on TMs have two basic limitations:

- *Dependency on whole sentences*: Practice has demonstrated that the similarity rate of complete sentences is relatively low, even when using a low similarity threshold. The performance decreases when the length of the sentence to be translated increases. Therefore, to achieve an adequate performance, these systems should be applied to strictly structured documents.

- *High cost of building a TM*: The initial construction of a TM can be done using a corpus of previous translations. This corpus can be extended by adding new translations performed by a human expert with tools based on TMs. However, both these processes are relatively slow. Furthermore, to achieve a good performance with such a TM, there must be a huge database, which increases the construction cost of the TM.

Typically, the degree of similarity and the repetition of segments which are smaller than a sentence are higher than the similarity or repetition of complete sentences. It would be interesting for TMs to be able to deal with subsegments of sentences which are inside the TM database. Recent research works (Macklovitch et al., 2000; Simard and Langlais, 2001; Simard, 2003) have tried to work with TMs at a level that is smaller than the sentence level.

A human translator usually decomposes the sentence to be translated and works with smaller units; therefore, it would be desirable to enrich the TM database with smaller translation units as well. It would also be desirable for this enrichment to be automatic in order not to increase the cost of building a TM. The automatic obtention of these new translation units is the main purpose of this work.

Our proposal is the application of two automatic *bilingual segmentation* (*bisegmentation*) techniques based on statistical translation methods to create new, shorter *bilingual segments* (*bisegments*) to be included in a TM database. We think that the application of this technique can outperform the translation results of systems based on current TMs. We describe the two bisegmentation techniques briefly in section 2. An evaluation of the two techniques was carried out on a bilingual Basque-Spanish task; this evaluation is shown in section 3. Additional lines for future work are given in section 4.

2. Bilingual segmentation

A first approach to the formal description of the bisegmentation concept is described in (Langlais et al., 1998a; Langlais et al., 1998b; Simard and Plamondon, 1998), using the term *Alignment*. We prefer to use *bisegmentation* because *alignment* has been widely used in the last few years in machine translation with a different meaning (this will be described below). As the purpose of our bisegmentation techniques is to obtain translation units at a subsentence level, we redefine the formal definition of (Simard and Plamondon, 1998), which was used for the alignment of sentences in parallel texts.

This work has been partially supported by the Spanish project TIC2003-08681-C02-02 and the project INTEK-CN03AD02.

Let $\mathbf{f} = \{f_1, f_2, \ldots, f_J\}$ be the source sentence and $\mathbf{e} = \{e_1, e_2, \ldots, e_I\}$ be its corresponding target sentence in the bilingual corpus of the TM. A bisegmentation S of **f** and **e** is defined as a set of ordered pairs included in $\mathcal{P}(\mathbf{f}) \times \mathcal{P}(\mathbf{e})$, where $\mathcal{P}(\mathbf{f})$ and $\mathcal{P}(\mathbf{e})$ are the set of all subsets of **f** and **e**, respectively. Each of the ordered pairs of the segmentation is a bisegment. In the following sections we describe the two proposed bisegmentation techniques.

2.1. GIATI-based bilingual segmentation

The GIATI technique is an automatic method to infer statistical finite-state transducers described in (Casacuberta, 2000). As a first step, this technique carries out a labelling of the words of the source sentence with the words of the output sentence from a word alignment between both sentences. The concept of word alignment is formally described in (Brown et al., 1993). Intuitively, an alignment shows the words in a target sentence that are connected to specific words of the source sentence. The alignments can be automatically obtained with the software tool GIZA++ (Och and Ney, 2000).

This kind of labelling can produce a bisegmentation if we consider that the bisegments are composed of the source words and their corresponding labels of target words. Basically, the method labels every source word with its connected target words except when a reordering is done in the alignment. In this case, the method groups all the necessary source and target words in order to consider the reordering inside the bisegment.

An example of bisegmentation based on the GIATI labelling is shown in Figure 1. This figure shows the word alignment obtained with GIZA++ as a dot matrix and the bisegments obtained with the GIATI labelling as boxes.

2.2. Recursive bilingual segmentation

Now, in order to obtain a bisegmentation of a bisentence, we look for a bilingual recursive alignment of this bisentence. Basically, a recursive alignment is an alignment between phrases[1] of a source sentence and phrases of a target sentence. A recursive alignment represents the translation relations between two sentences, but it also includes information about the possible reorderings needed in order to generate the target sentence from the source sentence. A recursive alignment can be represented using a binary tree, where the internal nodes store the reordering directions and the leaf nodes store the translation relations. A formal description of bilingual recursive alignments can be found in (Vilar, 1998). From a recursive alignment, a bisegmentation can be obtained by considering only the segments in the leaf nodes of the output trees. Figure 2 shows an example of recursive alignment. The corresponding bisegmentation for this tree is constructed using the bisegments of the leaf nodes.

A greedy algorithm is proposed to compute recursive alignments from a bilingual corpus aligned at the sentence level. In this algorithm, the probability of translating a sequence of words from the source language into a sequence of words in the target language will be approximated using the IBM Model 1 (Brown et al., 1993). In an intuitive manner, Model 1 computes the probability of a sequence of words being translated into another sequence of words without taking into account the word order. The software tool GIZA++ (Och and Ney, 2000) estimates the IBM statistical translation models automatically. The proposed greedy algorithm computes a recursive alignment based on Model 1 for a source and a target sentence in this way: given the two sentences, it computes the most probable breakpoint in each sentence using Model 1 by exploring all possible breakpoints. Now, if the translation probability of Model 1 for the whole sentences is higher than the translation probability of dividing them, it creates a leaf node where the output sequence is considered to be the translation of the input sequence and it stops. Otherwise, it creates a new inner node of the tree, and, depending on the most probable direction of the alignment in the breakpoint, recursively applies the algorithm to the left and the right children. As this algorithm computes recursive alignments and the bisegmentations are obtained as a byproduct, we call this system *Ralign*.

Other similar approaches to bisegmentation also use statistical translation models; however, they impose additional restrictions on the breakpoint selection and do not consider reorderings. These are described in (Simard and Plamondon, 1998) and (Nevado et al., 2003).

3. Experiments

3.1. Corpus description

The bisegmentation techniques described above have been applied to the DFB[2] Basque-Spanish bilingual corpus. The characteristics of the corpus are shown in Table 1.

	Basque	Spanish
Sentences	283,277	
Words	4,239,528	5,717,907
Vocabulary	212,815	109,898

Table 1: DFB Basque-Spanish corpus characteristics.

The large vocabulary denotes the difficulty of the task. This difficulty is also increased due to the dissimilarity between the two languages, where preliminary translation results tend to show that machine translation between Basque and Spanish is not as simple as between other language pairs which are more similar to each other, e.g., Catalan-Spanish or English-Spanish.

In order to test the goodness of the bisegmentation technique, a bilingual test corpus was constructed using 20 bisentences. A reference bisegmentation was made manually for these 20 bisentences.

3.2. Bilingual segmentation evaluation

In order to evaluate the bisegmentations achieved by the proposed techniques, we will use the method described in (Langlais et al., 1998a; Langlais et al., 1998b; Simard and Plamondon, 1998).

[1] Here, the term *phrase* refers to a consecutive sequence of words, not necessarily with a linguistic structure or an independent meaning.

[2] Acronym for *Diputación Foral de Bizkaia*.

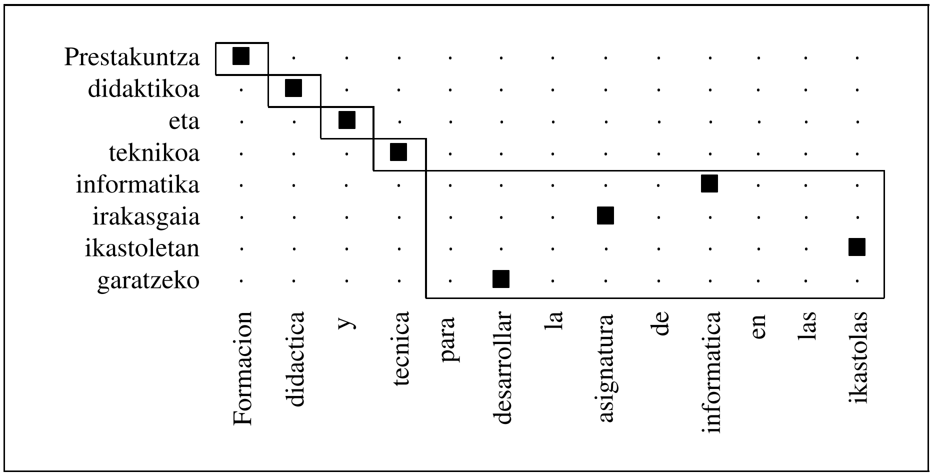


Figure 1: Example of bisegmentation based in the GIATI labelling.

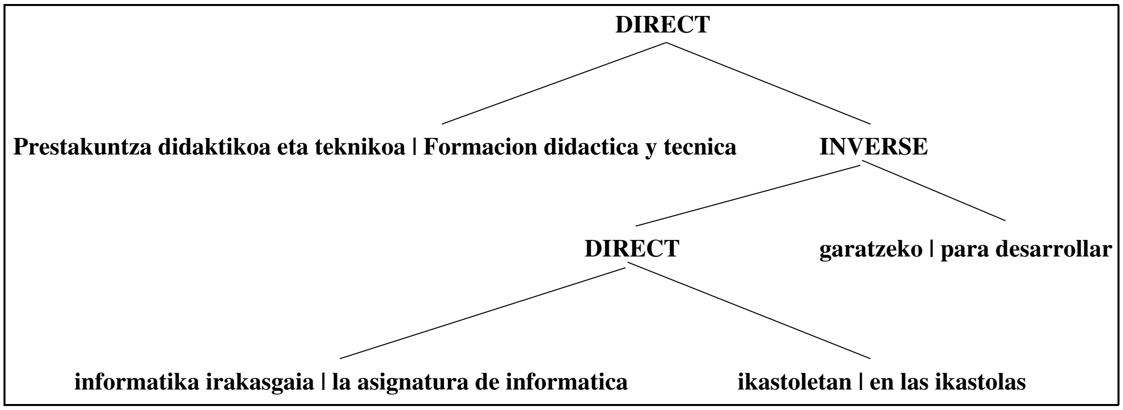

DIRECT
Prestakuntza didaktikoa eta teknikoa | Formacion didactica y tecnica
INVERSE
DIRECT
garatzeko | para desarrollar
informatika irakasgaia | la asignatura de informatica
ikastoletan | en las ikastolas

Figure 2: Example of recursive alignment (and the corresponding bisegmentation).

Prestakuntza didaktikoa	Formacion didactica
eta teknikoa	y tecnica
informatika irakasgaia	la asignatura de informatica
ikastoletan	en las ikastolas
garatzeko	para desarrollar

Table 2: Reference bisegmentation for the example bisentence of Figures 1 and 2.

Table 2 shows S_r, the *Reference* bisegmentation for the example bisentence in Figures 1 and 2. If the bisegmentation achieved by the system is S, the *recall* with respect to the reference, S_r, is defined as: *recall* $= |S \cap S_r|/|S_r|$. It represents the proportion of bisegments in S that are correct with respect to the reference S_r. The *precision* with respect to the reference S_r is defined as: *precision* $= |S \cap S_r|/|S|$. It represents the proportion of bisegments in S that are correct with respect to the number of bisegments proposed.

In this way, recall and precision do not take into account that some bisegments could be partially correct. To consider partial correctness, we need to compute recall and precision at a lower level than the segment level. The two measures can be easily redefined in order to use words as granularity units (Langlais et al., 1998a; Langlais et al., 1998b). Figure 3 represents the two system bisegmentations and the reference bisegmentation extending the segment correspondence into a word correspondence (each source word connects to all target words of the corresponding segment). As can be seen in Figure 3, the recall and precision at word level for the GIATI-based bisegmentation with respect to the reference are $17/21 = 0.81$ and $17/40 = 0.43$, respectively; the recall and precision at the word level for the bisegmentation computed by the Ralign system are $21/21 = 1$ and $21/29 = 0.72$, respectively.

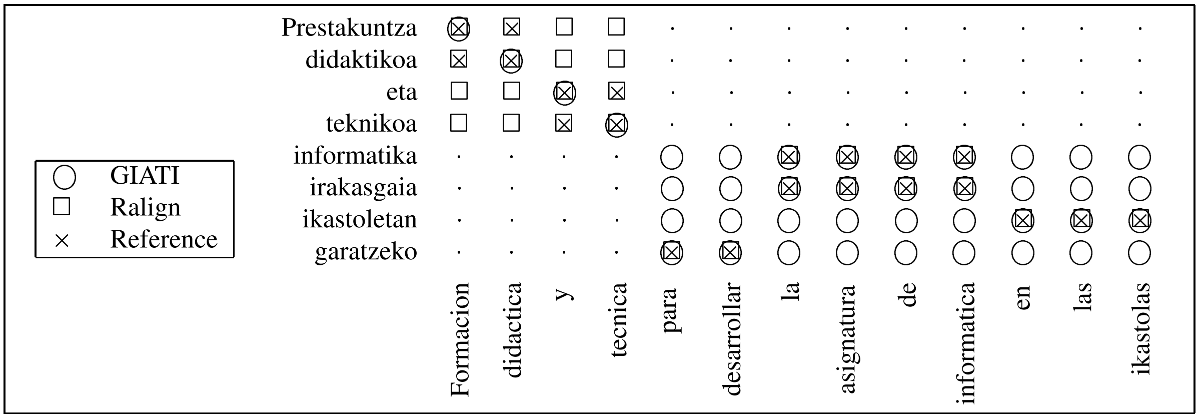
3.3. Results

Table 3 shows the recall and precision results at word level for the bisegmentations obtained with the two techniques described in section 2.

Bisegmentation Technique	Recall	Precision
GIATI-based	83.42	15.96
RALIGN-based	81.43	28.27

Table 3: Recall and precision results for the two proposed techniques for bilingual segmentation.

The recall values for the two techniques are very similar, but the precision increases significantly for the Ralign



	Formacion	didactica	y	tecnica	para	desarrollar	la	asignatura	de	informatica	en	las	ikastolas
Prestakuntza	○☒	☒	□	□	·	·	·	·	·	·	·	·	·
didaktikoa	☒	○☒	□	□	·	·	·	·	·	·	·	·	·
eta	□	□	○☒	☒	·	·	·	·	·	·	·	·	·
teknikoa	□	□	☒	○☒	·	·	·	·	·	·	·	·	·
informatika	·	·	·	·	○	○	○☒	○☒	○☒	○☒	○	○	○
irakasgaia	·	·	·	·	○	○	○☒	○☒	○☒	○☒	○	○	○
ikastoletan	·	·	·	·	○	○	○	○	○	○	○☒	○☒	○☒
garatzeko	·	·	·	·	○☒	○☒	○	○	○	○	○	○	○

Figure 3: GIATI, Ralign and Reference bisegmentation representation using a word level correspondence.

bisegmentation. Both precision results are very low, which denotes a great difference between the systems bisegmentations and the reference. However, this does not imply that the bisegmentations are bad from a human expert perspective. Human experts also say that it is not easy to construct a *good* reference translation manually. A better evaluation could be done by using various reference bisegmentations, but this will also increase the construction cost of the references.

4. Future work

We want to explore new proposals in order to obtain bisegmentations from bilingual corpora:

- A modification of the proposed Ralign system is to restrict the bisegmentations of a bisentence to be compatible with a Viterbi alignment obtained automatically with GIZA++.
- A new approach is to obtain monolingual segments in one of the two languages using linguistic segmentation tools, and then to obtain the corresponding monolingual segments in the other language using statistical translation models. The use of syntactic trees can restrict the type of possible segments to those with linguistic meaning.

Another line of future work is the implementation of a statistical phrase-based translation system which takes into account possible reorderings of the phrases. This system will be based directly on a target language model and a statistical phrase-based translation dictionary; this dictionary will be constructed from the bisegments obtained with the techniques proposed in section 2. In the last few years, the application of statistical phrase-based translation has given significant results (Och, 2003); these techniques can improve the translation systems based on TMs.

5. References

Brown, P., S. Della Pietra, V. Della Pietra, and R. Mercer, 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Casacuberta, F., 2000. Inference of finite-state transducers by using regular grammars and morphisms. In *Grammatical Inference: Algorithms and Applications*, volume 1891 of *Lecture Notes in Computer Science*. Springer-Verlag. 5th International Colloquium Grammatical Inference.

Langlais, P., M. Simard, and J. Véronis, 1998a. Methods and practical issues in evaluating alignment techniques. In *Proc. od COLING-ACL 98*. Montréal, Canada.

Langlais, P., M. Simard, J. Véronis, S. Armstrong, P. Bonhomme, F. Débili, P. Isabelle, E. Souissi, and P. Théron, 1998b. Arcade: A cooperative research project on parallel text alignment evaluation.

Macklovitch, E., M. Simard, and P. Langlais, 2000. Transsearch: A free translation memory on the world wide web. In *Second International Conference On Language Resources and Evaluation, LREC2000*.

Nevado, F. F. Casacuberta, and Enrique Vidal, 2003. Parallel corpora segmentation by using anchor words. In *Proceedings of the 7th International EAMT Workshop on MT and other language technology tools*.

Och, F. J. and H. Ney, 2000. Improved statistical alignment models. In *Proceedings of ACL00*. Hongkong, China.

Och, F.J., 2003. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, Department of Computer Science, RWTH-Aachen.

Simard, M., 2003. *Mémoires de Traduction Sous-Phrastiques*. Ph.D. thesis, Dépt. d'Informatique et de Recherche Opérationnelle, Université de Montréal.

Simard, M. and P. Langlais, 2001. Sub-sentential exploitation of translation memories. In *Proc. of MT Summit VIII*.

Simard, M. and P. Plamondon, 1998. Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation*, 13(1):59–80.

Vilar, J.M., 1998. *Aprendizaje de Transductores Subsecuenciales para su empleo en tareas de Dominio Restringido*. Ph.D. thesis, Dept. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia.