Parallel Corpora Segmentation Using Anchor Words*

Francisco Nevado and Francisco Casacuberta and Enrique Vidal
Instituto Tecnològico de Informàtica
Universidad Politècnica de Valencia
{fnevado,fcn,evidal}@iti.upv.es

Abstract

A new technique for monotone segmentation of parallel corpora is introduced. This segmentation is based on a set of anchor words which are defined manually. The parallel segments are computed using a dynamic programming algorithm. To assess this technique, finite-state transducers are inferred from both non-segmented and segmented corpora. Experiments have been carried out with Spanish-English and Italian-English translation tasks. This technique has proven useful in improving the results with respect to those obtained with unsegmented corpora.

1 Introduction

In this paper, we present a new technique for improving machine translation systems. This is a heuristic approach for parallel corpora segmentation using anchor words and a dynamic programming algorithm.

In a parallel corpus, the anchor words are specific words that are defined for the two languages of the corpus and that are strongly related.

The goal of parallel corpus segmentation is to segment the source sentence and the target sentence in such a way that the correspondence between segments is monotone and one-to-one.

Using this segmentation, we attempted to improve the word alignments obtained with statistical techniques (Brown et al., 1993; Brown et al., 1990). These models depend on the length of the source and target sentences. The models are better estimated with shorter segments and, consequently, better word alignments are obtained.

The basic scheme of the proposed parallel segmentation is the following:

a) The source and the target sentences are initially segmented in the positions of the anchor words.

b) As the number of source and target segments can be different, a dynamic programming algorithm is applied to find the optimal correspondences between segments.

In section 2, we will show how to segment a bilingual corpus describing the segmentation of a pair of sentences using anchor words. We will then describe the experiments carried out to test this new technique and the obtained results.

2 Segmentation of a parallel corpus

Parallel segmentation is considered from a statistical point of view. Segmentation of a parallel corpus is carried out by segmenting every pair of sentences in this corpus.

2.1 Statistical machine translation

We use a notation which is similar to the one proposed in (Brown et al., 1993), where **f** is a source sentence and **e** is a target sentence.

*This work has been partially supported by TransType 2 (IST-2001-32091) and SisHiTRA (TIC2000-1599-C02).

In order to translate from the source language to the target language in a statistical framework (Brown et al., 1993), we look for the probability of obtaining a sentence **e** from a sentence **f**, that is, **Pr(e | f).** Applying Bayes rule, we have:

$$\Pr(\mathbf{e} \mid \mathbf{f}) = \frac{\Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e})}{\Pr(\mathbf{f})}. \qquad (1)$$

Since we are searching for the target sentence with the best probability of being generated from the source sentence, by maximizing the preceding expression, we have:

$$\begin{aligned} \widehat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e} \mid \mathbf{f}) \\ &= \underset{\mathbf{e}}{\operatorname{argmax}} \Pr(\mathbf{e}) \Pr(\mathbf{f} \mid \mathbf{e}), \qquad (2) \end{aligned}$$

where **Pr(e)** corresponds to the probability of the target language model and **Pr(f e)** is known as the probability of the translation model. This model transforms a sentence in the target language into a sentence in the source language.

2.2 Segmentation of a pair of sentences

We obtain the segmentation as a byproduct of the translation process of a sentence. To start with, a trivial monolingual anchor-point-based **initial segmentation** is assumed on the sentence **f**. A different trivial monolingual anchor-point-based initial segmentation is also assumed on the sentence **e**. Having defined a set of anchor words for the source language and another set of anchor words for the target sentence, the first initial segment for sentence **f** is composed of the sequence of words from the beginning of the sentence until the first anchor word of **f**. The rest of the initial segments are composed of the sequences of words from the first word following the last segment until the next anchor word. The last segment of the sentence may end with the end of the sentence instead of an anchor word. The initial segments of **e** are computed in the same way, but taking into account the anchor words for the target language. Let us suppose that there are *a* initial segments for sentence **f** and there are *b* initial segments for sentence **e**. This initial segmentation is represented as:

$$\begin{aligned} \mathbf{e} &= \bar{e}_1 \bar{e}_2 \cdots \bar{e}_a &= \bar{e}_1^a \\ \mathbf{f} &= \bar{f}_1 \bar{f}_2 \cdots \bar{f}_b &= \bar{f}_1^b \end{aligned}$$

where $\bar{e}_c$ is a segment of consecutive words of and $\bar{f}_d$ is a segment of consecutive words of : See Figure 2 for an example of this kind of in tial segmentation. $\bar{e}_{c_1}^{c_2}$ is the sequence of word constituted by the concatenation of the segment $\bar{e}_{c_1} \bar{e}_{c_1+1} \cdots \bar{e}_{c_2}$, and $\bar{f}_{d_1}^{d_2}$ is the sequence of word constituted by the concatenation of the segment $\bar{f}_{d_1} \bar{f}_{d_1+1} \cdots \bar{f}_{d_2}$.

Processing each initial segment as an atomi block, we can rewrite expression (2) with this nc tation:

$$\widehat{\mathbf{e}} = \underset{\bar{e}_1^a}{\operatorname{argmax}} \Pr(\bar{e}_1^a) \Pr(\bar{f}_1^b \mid \bar{e}_1^a). \qquad (3$$

A parallel segmentation **s** is an ordered set of pairs of sequences of words, where every one of these pairs has a sequence of words from the source sentence and a sequence of words from the target sentence composed by one or more consecutive initial segments of the source sentence or the target sentence, respectively[1].

Given an initial segmentation ($\bar{e}_1^a$,$\bar{f}_1^b$), we can represent a parallel segmentation as:

$$\mathbf{s} \equiv \left([\bar{e}_1^{c_1}, \bar{f}_1^{d_1}], [\bar{e}_{c_1+1}^{c_2}, \bar{f}_{d_1+1}^{d_2}], \ldots, [\bar{e}_{c_{|\mathbf{s}|-1}+1}^{c_{|\mathbf{s}|}}, \bar{f}_{d_{|\mathbf{s}|-1}+1}^{d_{|\mathbf{s}|}}] \right),$$

where $|\mathbf{s}|$ is the number of segments for the parallel segmentation **s**. Clearly we have $c_{|\mathbf{s}|} = a$ and $d_{|\mathbf{s}|} = b$. Therefore, in a segmentation, any segment in the input sentence cannot be left without a corresponding segment of the output sentence, and vice versa. Another restriction is that there cannot be inversions in the order of the initial segments; that is, if $[\bar{e}_{c_1}^{c_2}, \bar{f}_{d_1}^{d_2}]$ is a pair of segments of a segmentation **s**, then $\forall [\bar{e}_{c_3}^{c_4}, \bar{f}_{d_3}^{d_4}] \in \mathbf{s}$:

$$\begin{aligned} &\text{if } c_2 < c_3 \Longrightarrow d_2 < d_3 \\ &\text{if } c_4 < c_1 \Longrightarrow d_4 < d_1 \end{aligned}$$

An example of this kind of segmentation is shown in section 2.3.

The set of possible parallel segmentations for an initial segmentation based on anchor words

[1] Note the difference with an initial segmentation. A segmentation may have joined several consecutive segments of the initial segments, but it has the same number of final segments for the source and target sentences.

$(\bar{e}_1^a, \bar{f}_1^b)$ is denoted by $\mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)$. Now, we can write the probability for the translation model, $\Pr(\bar{f}_1^b \mid \bar{e}_1^a)$:

$$\Pr(\bar{f}_1^b \mid \bar{e}_1^a) = \sum_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \Pr(\bar{f}_1^b, \mathbf{s} \mid \bar{e}_1^a) \quad (4)$$

where $\Pr(\bar{f}_1^b, \mathbf{s} \mid \bar{e}_1^a)$ allows for the interpretation of a segmentation as a generative model. We can say that the segments in the source sentence are generated from the corresponding segments of the target sentence.

Given a sentence $\bar{e}_1^a$, we define the probability for a sentence $\bar{f}_1^b$ and a segmentation $\mathbf{s}$ as:

$$\Pr(\bar{f}_1^b, \mathbf{s} \mid \bar{e}_1^a) = \Pr(\mathbf{s} \mid \bar{e}_1^a) \prod_{q=1}^{|\mathbf{S}|} \Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}), \quad (5)$$

where $\Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q})$ is again the probability of the translation model for a subsequence of the sentence **f** and a subsequence of the sentence **e**.

We do not want to consider the translation model as a recursive model, so we will approximate the probability $\Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q})$ using Model 1 proposed in (Brown et al., 1993). In an intuitive manner, Model 1 computes the probability of a sequence of words to be translated by other sequence of words, without taking into account the word order. Therefore, it can allow translation inversions inside the sequences of words. The translation probability of a sequence of words $\bar{e}_{c_1}^{c_2}$ into a sequence of words $\bar{f}_{d_1}^{d_2}$ using Model 1 is computed by:

$$\Pr(\bar{f}_{d_1}^{d_2} \mid \bar{e}_{c_1}^{c_2}) = M_1(\bar{f}_{d_1}^{d_2} \mid \bar{e}_{c_1}^{c_2}) = \frac{\epsilon}{(p+1)^o} \prod_{j=1}^{o} \sum_{i=0}^{p} t(\langle \bar{f}_{d_1}^{d_2}, j \rangle \mid \langle \bar{e}_{c_1}^{c_2}, i \rangle),$$

where p and o are the lengths in words of the sequences $\bar{e}_{c_1}^{c_2}$ and $\bar{f}_{d_1}^{d_2}$, respectively. $\langle \bar{f}_{d_1}^{d_2}, j \rangle$ is the j-th word of the sequence $\bar{f}_{d_1}^{d_2}$, and $\langle \bar{e}_{c_1}^{c_2}, i \rangle$ is the i-th word of the sequence $\bar{e}_{c_1}^{c_2}$. $t(\langle \bar{f}_{d_1}^{d_2}, j \rangle \mid \langle \bar{e}_{c_1}^{c_2}, i \rangle)$ is a statistical translation dictionary which stores the probability of the target word $\langle \bar{e}_{c_1}^{c_2}, i \rangle$ being translated into the source word $\langle \bar{f}_{d_1}^{d_2}, j \rangle$. This dictionary can be estimated automatically from the bilingual corpus by using the estimation methods described in (Brown et al., 1993). The software used to obtain this statistical dictionary was GIZA++ (Och and Ney, 2000; Knight, 1999). The probability that the sequences of words of the pair have a certain length (number of words) is measured by the ε term.

Now, expanding expression (4) with (5), we have:

$$\Pr(\bar{f}_1^b \mid \bar{e}_1^a) = \sum_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} Pr(\mathbf{s} \mid \bar{e}_1^a) \prod_{q=1}^{|\mathbf{S}|} \Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}). \quad (6)$$

However, we are interested in computing only the best segmentation, so, we define the most probable segmentation probability, $\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a)$, as the maximum of expression (6):

$$\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a) = \max_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \Pr(\mathbf{s} \mid \bar{e}_1^a) \prod_{q=1}^{|\mathbf{S}|} \Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}).$$

Considering $\Pr(\mathbf{s} \mid \bar{e}_1^a)$ to be equally probable for all $\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)$ (that is, $C = \Pr(\mathbf{s} \mid \bar{e}_1^a)$) and assuming that $\Pr(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q})$ is computed using the Model 1:

$$\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a) = C \cdot \max_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \prod_{q=1}^{|\mathbf{S}|} M_1(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}). \quad (7)$$

In order to obtain the segmentation with maximum probability, we want the argument that maximizes expression (7), so, we look for:

$$\widehat{\mathbf{s}} = \operatorname*{argmax}_{\mathbf{s} \in \mathcal{S}(\bar{e}_1^a, \bar{f}_1^b)} \prod_{q=1}^{|\mathbf{S}|} M_1(\bar{f}_{d_{q-1}+1}^{d_q} \mid \bar{e}_{c_{q-1}+1}^{c_q}). \quad (8)$$

To solve the maximization problems (7) and (8), we use a dynamic programming scheme. In order to reduce the computational search cost, we impose a new restriction: no more than k initial segments can be joined for $\bar{f}_1^b$ or $\bar{e}_1^a$.

The algorithm to compute the probability of the best segmentation uses a bidimensional matrix

$\mathbf{s}[d,c]$. A graphical representation of this structure is shown in Figure 1, where the rows correspond to the initial segments of the source sentence and the columns correspond to the initial segments of the target sentence.

The expression which is computed for every position of the matrix **s** in Figure 1 is:

$$\mathbf{s}[d,c] = \max_{\substack{i=0..k\\ j=0..k}} \mathbf{s}[d-j-1, c-i-1] \cdot M_1(\bar{f}_{d-j}^{d} \mid \bar{e}_{c-i}^{c})$$

$\mathbf{s}[d,c]$ is the probability of translating the sequence of words $\bar{e}_1^c$ into the sequence of words $\bar{f}_1^d$, $\Pr(\bar{e}_1^c \mid \bar{f}_1^d)$.

The algorithm for computing the probability of the best parallel segmentation for an initial segmentation based on anchor words is shown in algorithm 1. When the computation of every $\mathbf{s}[d,c]$ is done, the probability of the best segmentation is stored in $\mathbf{s}[b,a]$

$$\begin{array}{c|c||c|c|c|c|c}
 & & & & c = 1 \ldots a & & \\
 & & & \bar{e}_1 & \bar{e}_2 & \ldots & \bar{e}_a \\
\hline
 & & 1 & 0 & 0 & \ldots & 0 \\
\hline
 & \bar{f}_1 & 0 & & & & \\
\hline
d = 1 \ldots b & \bar{f}_2 & 0 & & & & \\
\hline
 & \vdots & \vdots & & & & \vdots \\
\hline
 & \bar{f}_b & 0 & & & \ldots & \\
\end{array}$$

Figure 1: Graphical representation of the matrix $\mathbf{s}[d,c]$ used for the computation of $\widehat{\Pr(\bar{f}_1^b \mid \bar{e}_1^a)}$ in the dynamic programming algorithm 1.

Another matrix can be computed together with the matrix s in order to store the path for the most probable segmentation, that is, to store the groupings of initial segments that are carried out for the most probable segmentation.

2.3 A complete example

Now we offer a complete example of the computation of the segmentation of a pair of sentences. This pair of sentences is extracted from the FUB corpus (Vidal, 2000). This corpus is a bilingual text corpus of Italian-English pairs of sentences with restricted semantic domain. The sentences in the corpus are typical sentences of a tourist in the hotel domain, for example:

- *A che ora é disponibile il servizio navetta per l'aeroporto? /At what time is the shuttle service to the airport available ?.*
- *Avete una stanza libera dal quattro al dieci Settembre? /Do you have a free room from the fourth to the tenth of September?*

Defining the following sets of anchor words for Italian and English, respectively:

Italian Anchors
., ,, :, ;, ?, !, *con*, *per*, *e*, *perché*, *vorrei*, *volevo*

English Anchors
., ,, :, ;, ?, !, *with*, *for*, *and*, *because*, *I would like*, *I wish*

The English expressions *I would like* and *I wish* were treated as atomic anchor words.

This is a pair of sentences extracted from the corpus:

> *buonasera , sono la signora Rossi della camera trecentodue , vorrei disdire per domani mattina la colazione in camera, grazie.*
>
> *good evening , it is Mrs Rossi from room three hundred and two , I would like to cancel breakfast in room for tomorrow morning, thanks.*

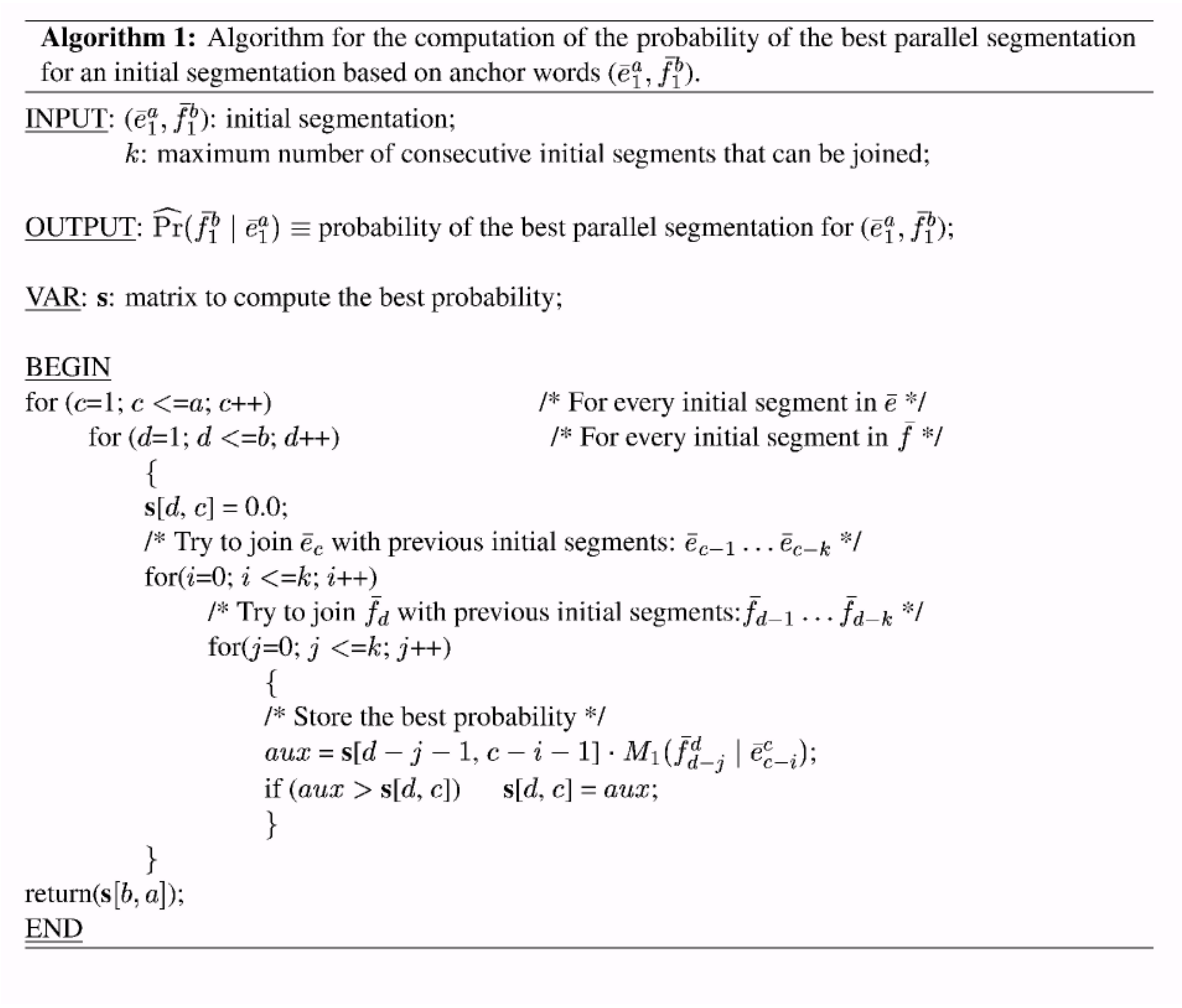
The initial segmentation for the original sentences and the anchor words is shown in Figure 2.

After running Algorithm 1 described in section 2 on the initial segmentation of Figure 2, we obtained the segmentation shown in Figure 3 as the best segmentation.

3 Experiments

3.1 Corpora description

The EuTrans-I corpus (Vidal, 2000) is a Spanish-English corpus which was generated semi-automatically for the EuTrans-I task which is a subtask of the "Traveler Task". The



Algorithm 1: Algorithm for the computation of the probability of the best parallel segmentation for an initial segmentation based on anchor words $(\bar{e}_1^a, \bar{f}_1^b)$.

```
INPUT: $(\bar{e}_1^a, \bar{f}_1^b)$: initial segmentation;
        $k$: maximum number of consecutive initial segments that can be joined;

OUTPUT: $\widehat{\Pr}(\bar{f}_1^b \mid \bar{e}_1^a) \equiv$ probability of the best parallel segmentation for $(\bar{e}_1^a, \bar{f}_1^b)$;

VAR: $\mathbf{s}$: matrix to compute the best probability;

BEGIN
for ($c$=1; $c$ <=$a$; $c$++)                    /* For every initial segment in $\bar{e}$ */
    for ($d$=1; $d$ <=$b$; $d$++)                  /* For every initial segment in $\bar{f}$ */
        {
        $\mathbf{s}[d, c] = 0.0$;
        /* Try to join $\bar{e}_c$ with previous initial segments: $\bar{e}_{c-1} \ldots \bar{e}_{c-k}$ */
        for($i$=0; $i$ <=$k$; $i$++)
            /* Try to join $\bar{f}_d$ with previous initial segments:$\bar{f}_{d-1} \ldots \bar{f}_{d-k}$ */
            for($j$=0; $j$ <=$k$; $j$++)
                {
                /* Store the best probability */
                $aux = \mathbf{s}[d-j-1, c-i-1] \cdot M_1(\bar{f}_{d-j}^d \mid \bar{e}_{c-i}^c)$;
                if ($aux > \mathbf{s}[d, c]$)     $\mathbf{s}[d, c] = aux$;
                }
        }
return($\mathbf{s}[b, a]$);
END
```

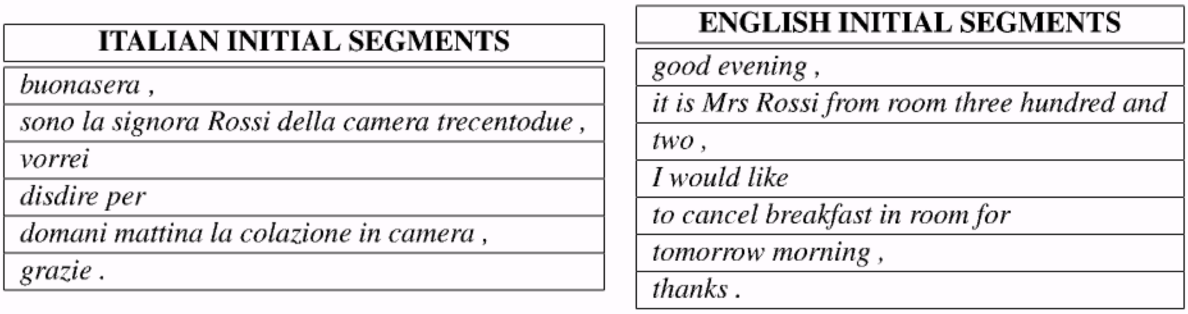
domain of the corpus is a human-to-human communication situation at a reception desk of a hotel. The corpus characteristics are shown in Table 1.

The FUB corpus (Vidal, 2000), is a bilingual Italian-English corpus with a restricted semantic domain. The application is the translation of queries, requests and complaints that a tourist can make at the front desk of a hotel, for example, asking for a booked room, requesting a service of the hotel, etc. The characteristics of the corpus are shown in Table 2.

3.2 Results

There is no standard method for evaluating the quality of a segmentation. One possible method is to compare the segmentation produced by the approach presented here with respect to a reference segmentation produced by hand. However, this is a very expensive procedure which is not error free. Another possible method for assessing the performance of this new segmentation technique is to compare the efficiency of a translation system obtained from the original corpus and another obtained from the segmented corpus on the translations of a test set of sentences.

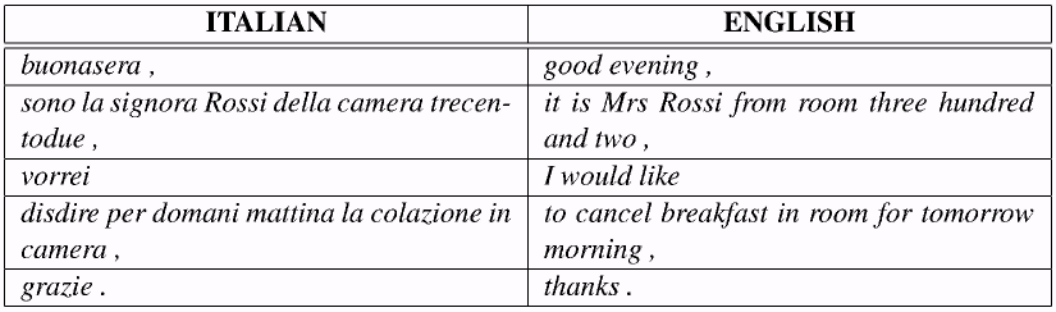
We trained two finite-state transducers: one from the original parallel corpus and one from the segmented parallel corpus. In order to infer the transducers from a parallel corpus we used a technique known as Grammatical Inference and Alignments for Transducer Inference (GIATI) (Casacuberta, 2000). The translation quality was measured for every transducer on the test set by using the translation word error rate (TWER). This is the average number of wrong words in the translations generated by the transducer with respect to fixed reference translations for the source sentences.



ITALIAN INITIAL SEGMENTS
buonasera ,
sono la signora Rossi della camera trecentodue ,
vorrei
disdire per
domani mattina la colazione in camera ,
grazie .

ENGLISH INITIAL SEGMENTS
good evening ,
it is Mrs Rossi from room three hundred and
two ,
I would like
to cancel breakfast in room for
tomorrow morning ,
thanks .

Figure 2: Initial segmentation from the original sentences of the FUB corpus and the sets of anchor words.



ITALIAN	ENGLISH
buonasera ,	*good evening ,*
sono la signora Rossi della camera trecentodue ,	*it is Mrs Rossi from room three hundred and two ,*
vorrei	*I would like*
disdire per domani mattina la colazione in camera ,	*to cancel breakfast in room for tomorrow morning ,*
grazie .	*thanks .*

Figure 3: Final parallel segmentation for the example pair of sentences of the FUB corpus.

Table 1: Training and test data sets of the bilingual corpus EUTRANS-I.

Training

	Spanish	English
N. Sentences	10,000	
N. Words	97,131	99,292
Vocabulary size	686	513
Perplexity(bigram)	8.6	5.2

Test

	Spanish	English
N. Sentences	2,996	
N. Words	35,023	35,590
Vocabulary size	613	469

Table 2: Training and test data sets of the bilingual FUB corpus 5.1.

Training

	Italian	English
N. Sentences	3,038	
N. Words	55,302	64,176
Vocabulary size	2,459	1,712
Perplexity(bigram)	31	25

Test

	Italian	English
N. Sentences	300	
N. Words	6,121	7,243
Vocabulary size	715	547

The number of initial segments that were allowed to be joined in one segment of the final segmentation was five.

In order to infer a finite-state transducer, the GIATI technique needs word-level alignments such as those described in (Brown et al., 1993; Knight, 1999) for every pair of sentences of the training set. Model 4 (Brown et al., 1993) was estimated with the non-segmented corpus and word alignments were obtained. With the segmented corpus, each pair of segments was considered as a pair of sentences, Model 4 was estimated and the corresponding word alignments were computed. These alignments were computed using the software GIZA++ (Och and Ney, 2000; Knight, 1999), obtaining the alignments produced by Model 4 (Brown et al., 1993). The finite-state transducer generated with GIATI is derived from a n-gram model inferred from the source sentences. In these source sentences, the words of every input sentence are labeled with the words of the corresponding target sentence according to the word alignments obtained with Model 4.

Tables 3 and 4 show the average lengths of the source-target sentences, along with the lengths of the segmented sentences obtained by the proposed technique. It is worth noting that on the average, the more complex and long sentences of the FUB corpus are broken down into much shorter (and simpler) segments.

Table 3: Average sentence length (number of words) for the EuTrans-I training set in the non-segmented and segmented versions.

	Spanish	English
Non-segmented	9.71	9.93
Segmented	7.40	7.57

Table 4: Average sentence length (number of words) for the FUB training set in the non-segmented and segmented versions.

	Italian	English
Non-segmented	17.94	21.55
Segmented	4.79	5.76

Table 5 shows the TWER values for the inferred transducers from the EuTrans-I training set using the Model 4 alignments and fourgrams for GIATI. Table 6 shows the TWER values for the corresponding transducers of the FUB training set using the Model 4 alignments and bigrams for GIATI.

The transducer inferred using the segmented EuTrans-I corpus produced a greater error rate than the transducer inferred using the non-segmented corpus. On the other hand, the results for the segmented FUB corpus improved the results over those obtained for the non-segmented version of the corpus.

Table 5: TWER for the EuTRANS-I test set using the transducers inferred with GIATI using fourgrams and the Model 4 alignments.

Non-segemented	8.0
Segmented	10.5

Table 6: TWER for the FUB test set using the transducers inferred with GIATI using bigrams and the Model 4 alignments.

Non-segmented	26.6
Segmented	25.2

4 Conclusions

A new automatic segmentation technique for a parallel corpus has been presented. The method has been tested using the translation results obtained for two tasks: the EuTrans-I task and the FUB task.

The EuTrans-I task is relatively much simpler than the FUB task, and the length of the sentences is significantly shorter. Consequently, alignment models such as Model 4 produce very good results on *unsegmented* pairs of this corpus, thereby directly leading to good translation results with GIATI transducers trained on unsegmented aligned data. The FUB corpus, on the other hand, is much more complex and the lengths of the sentences are much longer. For these (long) pairs of sentences, alignments obtained by alignments models such as Model 4 tend not to be as good as those of EuTrans-I. In this case, using the shorter pairs of sentences obtained by the proposed segmentation technique definitely helps the alignment model to produce better alignments, thereby leading to improved results for the GIATI transducers trained on *segmented* aligned pairs. It should be noted that the FUB task is much more realistic than the EuTrans-I task.

Although the proposed technique has a heuristic component (the selection of the anchor words sets), it improves the translation results with minimum human effort, especially for difficult tasks such as the FUB task.

References

P.F. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Minek, J. Lafferty, R.L. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics,* 16(2):79-85.

P.F. Brown, S.A. Delia Pietra, V.J. Della Pietra, and R.L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics,* 19(2):263-310.

F. Casacuberta. 2000. Inference of finite-state transducers by using regular grammars and morphisms. In *Proceedings of International Conference on Grammatical Inference - ICGI2000,* pages 1-14.

K. Knight. 1999. A statistical MT tutorial workbook. Technical Report prepared in connection with the JHU summer workshop, Johns Hopkins Univ.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. In *ACL00,* pages 440-447, Hongkong, China, October.

E. Vidal. 1997. Finite-state speech-to-speech translation. In *Proceedings of the International Conference on Acoustic, Speech and Signal Processing,* volume 1, pages 111-114.

E. Vidal. 2000. Final report. Technical Report EuTrans project, Technical Report Deliverable D0.lc, Information Technology. Long Term Research Domain. Open Scheme. Project Number 32026.