A DP based Search Using Monotone Alignments in Statistical Translation

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga
Lehrstuhl für Informatik VI, RWTH Aachen
D-52056 Aachen, Germany
{tillmann,ney}@informatik.rwth-aachen.de

Abstract

In this paper, we describe a Dynamic Programming (DP) based search algorithm for statistical translation and present experimental results. The statistical translation uses two sources of information: a translation model and a language model. The language model used is a standard bigram model. For the translation model, the alignment probabilities are made dependent on the differences in the alignment positions rather than on the absolute positions. Thus, the approach amounts to a first-order Hidden Markov model (HMM) as they are used successfully in speech recognition for the time alignment problem. Under the assumption that the alignment is monotone with respect to the word order in both languages, an efficient search strategy for translation can be formulated. The details of the search algorithm are described. Experiments on the EuTrans corpus produced a word error rate of 5.1%.

1 Overview: The Statistical Approach to Translation

The goal is the translation of a text given in some source language into a target language. We are given a source ('French') string $f_1^J = f_1...f_j...f_J$, which is to be translated into a target ('English') string $e_1^I = e_1...e_i...e_I$. Among all possible target strings, we will choose the one with the highest probability which is given by Bayes' decision rule (Brown et al., 1993):

$$\begin{aligned} \hat{e}_1^I &= \arg\max_{e_1^I} \{Pr(e_1^I|f_1^J)\} \\ &= \arg\max_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)\} \quad . \end{aligned}$$

$Pr(e_1^I)$ is the language model of the target language, whereas $Pr(f_1^J|e_1^I)$ is the string translation model. The argmax operation denotes the search problem. In this paper, we address

- the problem of introducing structures into the probabilistic dependencies in order to model the string translation probability $Pr(f_1^J|e_1^I)$.
- the search procedure, i.e. an algorithm to perform the argmax operation in an efficient way.
- transformation steps for both the source and the target languages in order to improve the translation process.

The transformations are very much dependent on the language pair and the specific translation task and are therefore discussed in the context of the task description. We have to keep in mind that in the search procedure both the language and the translation model are applied *after* the text transformation steps. However, to keep the notation simple we will not make this explicit distinction in the subsequent exposition. The overall architecture of the statistical translation approach is summarized in Figure 1.

2 Alignment Models

A key issue in modeling the string translation probability $Pr(f_1^J|e_1^I)$ is the question of how we define the correspondence between the words of the target sentence and the words of the source sentence. In typical cases, we can assume a sort of pairwise dependence by considering all word pairs (f_j, e_i) for a given sentence pair $[f_1^J; e_1^I]$. We further constrain this model by assigning each source word to *exactly one* target word. Models describing these types of dependencies are referred to as *alignment models* (Brown et al., 1993), (Dagan et al., 1993), (Kay & Röscheisen, 1993), (Fung & Church, 1994), (Vogel et al., 1996).

In this section, we introduce a monotone HMM based alignment and an associated DP based search algorithm for translation. Another approach to statistical machine translation using DP was presented in (Wu, 1996). The notational convention will be as follows. We use the symbol $Pr(.)$ to denote general

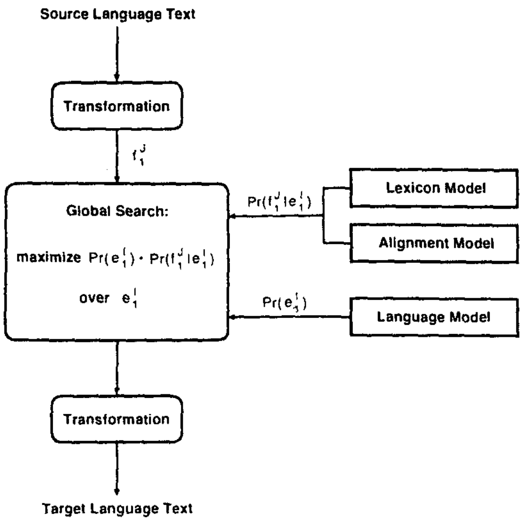

Figure 1: Architecture of the translation approach based on Bayes decision rule.

probability distributions with (nearly) no specific assumptions. In contrast, for model-based probability distributions, we use the generic symbol $p(.)$.

2.1 Alignment with HMM

When aligning the words in parallel texts (for Indo-European language pairs like Spanish-English, German-English, Italian-German,...), we typically observe a strong localization effect. Figure 2 illustrates this effect for the language pair *Spanish-to-English*. In many cases, although not always, there is an even stronger restriction: the difference in the position index is smaller than 3 and the alignment is essentially monotone. To be more precise, the sentences can be partitioned into a small number of segments, within each of which the alignment is monotone with respect to word order in both languages.

To describe these word-by-word alignments, we introduce the mapping $j \rightarrow a_j$, which assigns a position j (with source word f_j) to the position $i = a_j$ (with target word e_i). The concept of these alignments is similar to the ones introduced by (Brown et al., 1993), but we will use another type of dependence in the probability distributions. Looking at such alignments produced by a human expert, it is evident that the mathematical model should try to capture the strong dependence of a_j on the preceding alignment a_{j-1}. Therefore the probability of alignment a_j for position j should have a dependence on the previous alignment position a_{j-1}:

$$p(a_j|a_{j-1}) \quad .$$

A similar approach has been chosen by (Dagan et al., 1993) and (Vogel et al., 1996). Thus the problem formulation is similar to that of the time alignment problem in speech recognition, where the so-called Hidden Markov models have been successfully used for a long time (Jelinek, 1976). Using the same basic principles, we can rewrite the probability by introducing the 'hidden' alignments $a_1^J := a_1...a_j...a_J$ for a sentence pair $[f_1^J; e_1^I]$:

$$\begin{aligned} Pr(f_1^J|e_1^I) &= \sum_{a_1^J} Pr(f_1^J, a_1^J|e_1^I) \\ &= \sum_{a_1^J} \prod_{j=1}^{J} Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) \quad . \end{aligned}$$

To avoid any confusion with the term *'hidden'* in comparison with speech recognition, we observe that the model states as such (representing words) are *not* hidden but the actual alignments, i.e. the *sequence* of position index pairs $(j, i = a_j)$.

So far there has been no basic restriction of the approach. We now assume a first-order dependence on the alignments a_j only:

$$\begin{aligned} Pr(f_j, a_j|f_1^{j-1}, a_1^{j-1}, e_1^I) &= p(f_j, a_j|a_{j-1}, e_1^I) \\ &= p(a_j|a_{j-1}) \cdot p(f_j|e_{a_j}), \end{aligned}$$

where, in addition, we have assumed that the lexicon probability $p(f|e)$ depends only on a_j and not on a_{j-1}.

To reduce the number of alignment parameters, we assume that the HMM alignment probabilities $p(i|i')$ depend only on the jump width $(i - i')$. The monotony condition can than be formulated as:

$$p(i|i') = 0 \quad \text{for} \quad i \neq i' + 0, i' + 1, i' + 2.$$

This monotony requirement limits the applicability of our approach. However, by performing simple word reorderings, it is possible to approach this requirement (see Section 4.2). Additional countermeasures will be discussed later. Figure 3 gives an illustration of the possible alignments for the monotone hidden Markov model. To draw the analogy with speech recognition, we have to identify the states (along the vertical axis) with the positions i of the target words e_i and the time (along the horizontal axis) with the positions j of the source words f_j.

2.2 Training

To train the alignment and the lexicon model, we use the maximum likelihood criterion in the so-called maximum approximation, i.e. the likelihood criterion covers only the most likely alignment rather than the set of all alignments:

$$\begin{aligned} Pr(f_1^J|e_1^I) &= \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j})] \\ &\cong \max_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j})] \quad . \end{aligned}$$

```
days   |.   .   .   .   .   .   .   o
two    |.   .   .   .   .   .   o   .
for    |.   .   .   .   .   o   .   .
room   |.   .   .   o   .   .   .   .
double |.   .   .   .   o   .   .   .
a      |.   .   o   .   .   .   .   .
is     |.   o   .   .   .   .   .   .
much   |.   .   .   .   .   .   .   .
how    |o   .   .   .   .   .   .   .
       |--------------------------------
        c   v   u   h   d   p   d   d
         u   a   n   a   o   a   o   '
          '   l   a   b   b   r   s   i
           a   e       i   l   a       a
            n           t   e           s
             t           a
              o           c
                           i
                            '
                             o
                              n
```

```
room |.   .   o   .   .   .
the  |.   o   .   .   .   .
in   |o   .   .   .   .   .
cold |.   .   .   .   .   o
too  |.   .   .   .   o   .
is   |.   .   .   .   .   .
it   |.   .   .   o   .   .
     |------------------------
      e   l   h   h   d   f
       n   a   a   a   e   r
                b   c   m   '
                 i   e   a   i
                  t       s   o
                   a       i
                    c       a
                     i       d
                      '       o
                       o
                        n
```

```
night    |.   .   .   .   .   .   .   .   .   .   .   .   o
a        |.   .   .   .   .   .   .   .   .   .   .   o   .
for      |.   .   .   .   .   .   .   .   .   .   o   .   .
tv       |.   .   .   .   .   .   .   .   .   o   .   .   .
a        |.   .   .   .   .   .   .   .   .   .   .   .   .
and      |.   .   .   .   .   .   .   .   o   .   .   .   .
safe     |.   .   .   .   .   .   o   o   .   .   .   .   .
a        |.   .   .   .   .   .   .   .   .   .   .   .   .
telephone|.   .   .   .   .   o   .   .   .   .   .   .   .
a        |.   .   .   .   .   .   .   .   .   .   .   .   .
with     |.   .   .   .   o   .   .   .   .   .   .   .   .
room     |.   .   .   o   .   .   .   .   .   .   .   .   .
a        |.   .   o   .   .   .   .   .   .   .   .   .   .
booked   |.   o   .   .   .   .   .   .   .   .   .   .   .
have     |.   .   .   .   .   .   .   .   .   .   .   .   .
we       |o   .   .   .   .   .   .   .   .   .   .   .   .
         |----------------------------------------------------
          t   r   u   h   c   t   c   f   y   t   p   u   n
           e   e   n   a   o   e   a   u       e   a   n   o
            n   s   a   b   n   l   j   e       l   r   a   c
             e   e       i       '   a   r       e   a       h
              m   r       t       e       t       v           e
               o   v       a       f       e       i
                s   a       c       o               s
                     d       i       n               i
                      a       '       o               '
                               o                       o
                                n                       n
```

Figure 2: Word alignments for Spanish-English sentence pairs.

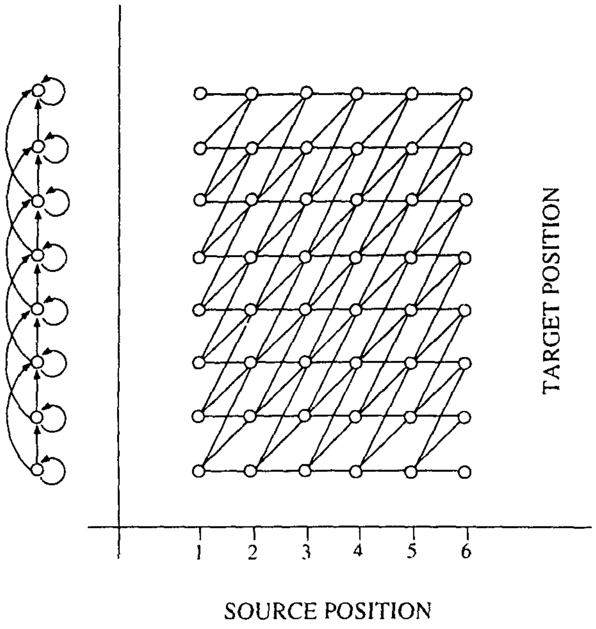

Figure 3: Illustration of alignments for the monotone HMM.

To find the optimal alignment. we use dynamic programming for which we have the following typical recursion formula:

$$Q(i,j) = p(f_j|e_i)\max_{i'}[p(i|i')\cdot Q(i',j-1)]$$

Here. $Q(i,j)$ is a sort of partial probability as in time alignment for speech recognition (Jelinek, 1976). As a result. the training procedure amounts to a sequence of iterations. each of which consists of two steps:

- *position alignment:* Given the model parameters, determine the most likely position alignment.
- *parameter estimation:* Given the position alignment. i.e. going along the alignment paths for all sentence pairs, perform maximum likelihood estimation of the model parameters; for model-free distributions. these estimates result in relative frequencies.

The IBM model 1 (Brown et al., 1993) is used to find an initial estimate of the translation probabilities.

3 Search Algorithm for Translation

For the translation operation. we use a bigram language model. which is given in terms of the conditional probability of observing word ϵ_i given the predecessor word ϵ_{i-1}:

$$p(\epsilon_i|e_{i-1}) \quad .$$

Using the conditional probability of the bigram language model. we have the overall search criterion in the maximum approximation:

$$\max_{e_1^I}\left\{\prod_{i=1}^{I}p(\epsilon_i|e_{i-1})\max_{a_1^J}\prod_{j=1}^{J}[p(a_j|a_{j-1})p(f_j|e_{a_j})]\right\}.$$

Here and in the following, we omit a special treatment of the start and end conditions like $j = 1$ or $j = J$ in order to simplify the presentation and avoid confusing details. Having the above criterion in mind. we try to associate the language model probabilities with the alignments $j \rightarrow i \equiv a_j$. To this purpose. we exploit the monotony property of our alignment model which allows only transitions from a_{j-1} to a_j if the difference $\delta \equiv a_j - a_{j-1}$ is $0, 1, 2$. We define a modified probability $p_\delta(e|e')$ for the language model depending on the alignment difference δ. We consider each of the three cases $\delta = 0, 1, 2$ separately:

- $\delta = 0$ (horizontal transition = alignment repetition): This case corresponds to a target word with two or more aligned source words and therefore requires $\epsilon = e'$ so that there is no contribution from the language model:

  $$p_{\delta=0}(e|e') = \begin{cases} 1 & \text{for } \epsilon = e' \\ 0 & \text{for } \epsilon \neq \epsilon' \end{cases} \quad .$$

- $\delta = 1$ (forward transition = regular alignment): This case is the regular one, and we can use directly the probability of the bigram language model:

  $$p_{\delta=1}(e|e') = p(e|e') \quad .$$

- $\delta = 2$ (skip transition = non-aligned word): This case corresponds to skipping a word. i.e. there is a word in the target string with no aligned word in the source string. We have to find the highest probability of placing a non-aligned word $\tilde{e}$ between a predecessor word ϵ' and a successor word ϵ. Thus we optimize the following product over the non-aligned word $\tilde{e}$:

  $$p_{\delta=2}(e|e') = \max_{\tilde{e}}[p(\epsilon|\tilde{e})\cdot p(\tilde{e}|e')] \quad .$$

  This maximization is done beforehand and the result is stored in a table.

Using this modified probability $p_\delta(e|\epsilon')$, we can rewrite the overall search criterion:

$$\max_{\epsilon_1^I, a_1^J}\prod_{j=1}^{J}[p(a_j|a_{j-1})p_{[a_j-a_{j-1}]}(\epsilon_{a_j}|\epsilon_{a_{j-1}})p(f_j|\epsilon_{a_j})] \, .$$

The problem now is to find the unknown mapping:

$$j \rightarrow (a_j, \epsilon_{a_j})$$

which defines a path through a network with a uniform trellis structure. For this trellis. we can still use Figure 3. However. in each position i along the

Table 1: DP based search algorithm for the monotone translation model.

```
input: source string $f_1 ... f_j ... f_J$
initialization
for each position $j = 1, 2, ....J$ in source sentence do
    for each position $i = 1, 2, ..., I_{max}$ in target sentence do
        for each target word $e$ do
            $Q(i, j, e) = p(f_j|e) \cdot \max_{\delta, e'}\{p(i|i-\delta) \cdot p_\delta(\epsilon|\epsilon') \cdot Q(i-\delta, j-1, \epsilon')\}$
traceback:
    - find best end hypothesis: $\max_{i,\epsilon} Q(i, J, \epsilon)$
    - recover optimal word sequence
```

vertical axis, we have to allow *all* possible words ϵ of the target vocabulary. Due to the monotony of our alignment model and the bigram language model, we have only first-order type dependencies such that the local probabilities (or costs when using the negative logarithms of the probabilities) depend *only* on the arcs (or transitions) in the lattice. Each possible index triple (i, j, ϵ) defines a grid point in the lattice, and we have the following set of possible transitions from one grid point to another grid point:

$$\delta \in \{0, 1, 2\} : \quad (i - \delta, j - 1, \epsilon') \rightarrow (i, j, e) \quad .$$

Each of these transitions is assigned a local probability:

$$p(i|i-\delta) \cdot p_\delta(\epsilon|\epsilon') \cdot p(f_j|\epsilon) \quad .$$

Using this formulation of the search task, we can now use the method of dynamic programming (DP) to find the best path through the lattice. To this purpose, we introduce the auxiliary quantity:

> $Q(i, j, \epsilon)$: probability of the best partial path which ends in the grid point (i, j, ϵ).

Since we have only first-order dependencies in our model, it is easy to see that the auxiliary quantity must satisfy the following DP recursion equation:

$$Q(i, j, \epsilon) = p(f_j|\epsilon) \cdot \\ \max_{\delta} \{p(i|i-\delta) \cdot \max_{\epsilon'} p_\delta(\epsilon|\epsilon') \cdot Q(i-\delta, j-1, \epsilon')\}.$$

To explicitly construct the unknown word sequence ϵ_1^I, it is convenient to make use of so-called backpointers which store for each grid point (i, j, ϵ) the best predecessor grid point (Ney et al., 1992).

The DP equation is evaluated recursively to find the best partial path to each grid point (i, j, ϵ). The resulting algorithm is depicted in Table 1. The complexity of the algorithm is $J \cdot I_{max} \cdot E^2$, where E is the size of the target language vocabulary and I_{max} is the maximum length of the target sentence considered. It is possible to reduce this computational complexity by using so-called pruning methods (Ney et al., 1992); due to space limitations, they are not discussed here.

4 Experimental Results

4.1 The Task and the Corpus

The search algorithm proposed in this paper was tested on a subtask of the "Traveler Task" (Vidal, 1997). The general domain of the task comprises typical situations a visitor to a foreign country is faced with. The chosen subtask corresponds to a scenario of the human–to–human communication situations at the registration desk in a hotel (see Table 4).

The corpus was generated in a semi-automatic way. On the basis of examples from traveller booklets, a probabilistic grammar for different language pairs has been constructed from which a large corpus of sentence pairs was generated. The vocabulary consisted of 692 Spanish and 518 English words (including punctuation marks). For the experiments, a training corpus of 80,000 sentence pairs with 628,117 Spanish and 684,777 English words was used. In addition, a test corpus with 2,730 sentence pairs different from the training sentence pairs was constructed. This test corpus contained 28,642 Spanish and 24,927 English words. For the English sentences, we used a bigram language model whose perplexity on the test corpus varied between 4.7 for the original text and 3.5 when all transformation steps as described below had been applied.

Table 2: Effect of the transformation steps on the vocabulary sizes in both languages.

Transformation Step	Spanish	English
Original (with punctuation)	692	518
+ Categorization	416	227
+ 'por_favor'	417	-
+ Word Splitting	374	-
+ Word Joining	-	237
+ Word Reordering	-	-

4.2 Text Transformations

The purpose of the text transformations is to make the two languages resemble each other as closely as possible with respect to sentence length and word order. In addition, the size of both vocabularies is reduced by exploiting evident regularities; e.g. proper names and numbers are replaced by category markers. We used different preprocessing steps which were applied consecutively:

- **Original Corpus:** Punctuation marks are treated like regular words.
- **Categorization:** Some particular words or word groups are replaced by word categories. Seven non-overlapping categories are used: three categories for names (surnames, male and female names), two categories for numbers (regular numbers and room numbers) and two categories for date and time of day.
- **Treatment of `'por favor'`:** The word `'por favor'` is always moved to the end of the sentence and replaced by the one-word token *'por_favor'*.
- **Word Splitting:** In Spanish, the personal pronouns (in subject case and in object case) can be part of the inflected verb form. To counteract this phenomenon, we split the verb into a verb part and pronoun part, such as *'darnos'* — *'dar _nos'* and *'pienso'* — *'_yo pienso'*.
- **Word Joining:** Phrases in the English language such as *'Would you mind doing ...'* and *'I would like you to do ...'* are difficult to handle by our alignment model. Therefore, we apply some word joining, such as *'would you mind'* — *'would_you_mind'* and *'would like '* — *'would_like'*.
- **Word Reordering:** This step is applied to the Spanish text to take into account cases like the position of the adjective in noun-adjective phrases and the position of object pronouns. E.g. *'habitación doble'* — *'doble habitación'*. By this reordering, our assumption about the monotony of the alignment model is more often satisfied.

The effect of these transformation steps on the sizes of both vocabularies is shown in Table 2. In addition to all preprocessing steps, we removed the punctuation marks before translation and resubstituted them by rule into the target sentence.

4.3 Translation Results

For each of the transformation steps described above, all probability models were trained anew, i.e. the lexicon probabilities $p(f|e)$, the alignment probabilities $p(i|i-\delta)$ and the bigram language probabilities $p(e|e')$. To produce the translated sentence in normal language, the transformation steps in the target language were inverted.

The translation results are summarized in Table 3. As an automatic and easy-to-use measure of the translation errors, the Levenshtein distance between the automatic translation and the reference translation was calculated. Errors are reported at the word level and at the sentence level:

- word level: insertions (INS), deletions (DEL), and total number of word errors (WER).
- sentence level: a sentence is counted as correct *only* if it is identical to the reference sentence.

Admittedly, this is not a perfect measure. In particular, the effect of word ordering is not taken into account appropriately. Actually, the figures for sentence error rate are overly pessimistic. Many sentences are acceptable and semantically correct translations (see the example translations in Table 4).

Table 3: Word error rates (INS/DEL, WER) and sentence error rates (SER) for different transformation steps.

Transformation Step	Translation Errors [%]		
	INS/DEL	WER	SER
Original Corpora	4.3/11.2	21.2	85.5
+ Categorization	2.5/9.6	16.1	81.0
+ `'por_favor'`	2.6/8.3	14.3	75.6
+ Word Splitting	2.5/7.4	12.3	65.4
+ Word Joining	1.3/4.9	7.3	44.6
+ Word Reordering	0.9/3.4	5.1	30.1

As can be seen in Table 3, the translation errors can be reduced systematically by applying all transformation steps. The word error rate is reduced from 21.2% to 5.1%; the sentence error rate is reduced from 85.5% to 30.1%. The two most important transformation steps are categorization and word joining. What is striking, is the large fraction of deletion errors. These deletion errors are often caused by the omission of word groups like *'for me please'* and *'could you'*. Table 4 shows some example translations (for the best translation results). It can be seen that the semantic meaning of the sentence in the source language may be preserved even if there are three word errors according to our performance criterion. To study the dependence on the amount of training data, we also performed a training with only 5 000 sentences out of the training corpus. For this training condition, the word error rate went up only slightly, namely from 5.1% (for 80,000 training sentences) to 5.3% (for 5 000 training sentences).

To study the effect of the language model, we tested a zerogram, a unigram and a bigram language model using the standard set of 80 000 training sentences. The results are shown in Table 5. The

Table 4: Examples from the EuTrans task: O= original sentence, R= reference translation. A= automatic translation.

O:	He hecho la reserva de una habitación con televisión y teléfono a nombre del señor Morales.
R:	I have made a reservation for a room with TV and telephone for Mr. Morales.
A:	I have made a reservation for a room with TV and telephone for Mr. Morales.
O:	Súbanme las maletas a mi habitación, por favor.
R:	Send up my suitcases to my room, please.
A:	Send up my suitcases to my room, please.
O:	Por favor, querría que nos diese las llaves de la habitación.
R:	I would like you to give us the keys to the room, please.
A:	I would like you to give us the keys to the room, please.
O:	Por favor, me pide mi taxi para la habitación tres veintidós?
R:	Could you ask for my taxi for room number three two two for me, please?
A:	Could you ask for my taxi for room number three two two, please?
O:	Por favor, reservamos dos habitaciones dobles con cuarto de baño.
R:	We booked two double rooms with a bathroom.
A:	We booked two double rooms with a bathroom, please.
O:	Quisiera que nos despertaran mañana a las dos y cuarto, por favor.
R:	I would like you to wake us up tomorrow at a quarter past two, please.
A:	I want you to wake us up tomorrow at a quarter past two, please.
O:	Repáseme la cuenta de la habitación ochocientos veintiuno.
R:	Could you check the bill for room number eight two one for me, please?
A:	Check the bill for room number eight two one.

WER decreases from 31.1% for the zerogram model to 5.1% for the bigram model.

The results presented here can be compared with the results obtained by the finite-state transducer approach described in (Vidal, 1996; Vidal, 1997), where the same training and test conditions were used. However the only preprocessing step was categorization. In that work, a WER of 7.1% was obtained as opposed to 5.1% presented in this paper. For smaller amounts of training data (say 5 000 sentence pairs), the DP based search seems to be even more superior.

Table 5: Language model perplexity (PP), word error rates (INS/DEL, WER) and sentence error rates (SER) for different language models.

Language		Translation Errors [%]		
Model	PP	INS/DEL	WER	SER
Zerogram	237.0	0.6/18.6	31.1	98.1
Unigram	74.4	0.9/12.4	20.4	94.8
Bigram	4.1	0.9/3.4	5.1	30.1

4.4 Effect of the Word Reordering

In more general cases and applications, there will always be sentence pairs with word alignments for which the monotony constraint is not satisfied. However even then, the monotony constraint is satisfied *locally* for the lion's share of all word alignments in such sentences. Therefore, we expect to extend the approach presented by the following methods:

- more systematic approaches to local and global word reorderings that try to produce the same word order in both languages.
- a multli-level approach that allows a small (say 4) number of *large* forward and backward transitions. Within each level, the monotone alignment model can still be applied, and only when moving from one level to the next, we have to handle the problem of different word orders.

To show the usefulness of global word reordering, we changed the word order of some sentences by hand. Table 6 shows the effect of the global reordering for two sentences. In the first example, we changed the order of two groups of consecutive words and placed an additional copy of the Spanish word "cuesta" into the source sentence. In the second example, the personal pronoun "me" was placed at the end of the source sentence. In both cases, we obtained a correct translation.

5 Conclusion

In this paper, we have presented an HMM based approach to handling word alignments and an associated search algorithm for automatic translation. The characteristic feature of this approach is to make the alignment probabilities explicitly dependent on the alignment position of the previous word and to assume a monotony constraint for the word order in both languages. Due to this monotony constraint, we are able to apply an efficient DP based search algorithm. We have tested the model successfully on the EuTrans traveller task, a limited domain task with a vocabulary of 200 to 500 words. The result-

Table 6: Effect of the global word reordering: O= original sentence, R= reference translation, A= automatic translation, O'= original sentence reordered, A'= automatic translation after reordering.

O:	Cuánto cuesta una habitación doble para cinco noches incluyendo servicio de habitaciones ?
R:	How much does a double room including room service cost for five nights ?
A:	How much does a double room including room service ?
O':	Cuánto cuesta una habitación doble incluyendo servicio de habitaciones cuesta para cinco noches ?
A':	How much does a double room including room service cost for five nights ?
O:	Explíque _me la factura de la habitación tres dos cuatro.
R:	Explain the bill for room number three two four for me.
A:	Explain the bill for room number three two four.
O':	Explíque la factura de la habitación tres dos cuatro _me.
A':	Explain the bill for room number three two four for me.

ing word error rate was only 5.1%. To mitigate the monotony constraint, we plan to reorder the words in the source sentences to produce the same word order in both languages.

Acknowledgement

This work has been supported partly by the German Federal Ministry of Education, Science, Research and Technology under the contract number 01 IV 601 A (Verbmobil) and by the European Community under the ESPRIT project number 20268 (EuTrans).

References

A. L. Berger, P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, J. R. Gillett, J. D. Lafferty, R. L. Mercer, H. Printz, and L. Ures. 1994. "The Candide System for Machine Translation". In *Proc. of ARPA Human Language Technology Workshop*, pp. 152-157. Plainsboro, NJ, Morgan Kaufmann Publishers, San Mateo, CA, March.

P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, and R. L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation". *Computational Linguistics*, Vol. 19, No. 2, pp. 263-311.

I. Dagan, K. W. Church, and W. A. Gale. 1993. "Robust Bilingual Word Alignment for Machine Aided Translation". In *Proc. of the Workshop on Very Large Corpora*, pp. 1-8, Columbus, OH.

P. Fung, and K. W. Church. 1994. "K-vec: A New Approach for Aligning Parallel Texts". In *Proc. of the 15th Int. Conf. on Computational Linguistics*, pp. 1096-1102, Kyoto.

F. Jelinek. 1976. "Speech Recognition by Statistical Methods". *Proc. of the IEEE*, Vol. 64, pp. 532-556, April.

M. Kay, and M. Röscheisen. 1993. "Text-Translation Alignment". *Computational Linguistics*, Vol. 19, No. 2, pp. 121-142.

H. Ney, D. Mergel, A. Noll, A. Paeseler. 1992. "Data Driven Search Organization for Continuous Speech Recognition". *IEEE Trans. on Signal Processing*, Vol. SP-40, No. 2, pp. 272-281, February.

E. Vidal. 1996. "Final report of Esprit Research Project 20268 (EuTrans): Example-Based Understanding and Translation Systems". Universidad Politécnica de Valencia, Instituto Tecnológio de Informática, October.

E. Vidal. 1997. "Finite-State Speech-to-Speech Translation". In *Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing*, Munich, April.

S. Vogel, H. Ney, and C. Tillmann. 1996. "HMM Based Word Alignment in Statistical Translation". In *Proc. of the 16th Int. Conf. on Computational Linguistics*, pp. 836-841, Copenhagen, August.

D. Wu. 1996. "A Polynomial-Time Algorithm for Statistical Machine Translation". In *Proc. of the 34th Annual Conf. of the Association for Computational Linguistics*, pp. 152-158, Santa Cruz, CA, June.