A Statistical Approach to Machine Translation

Peter F. Brown John Cocke Stephen A. Della Pietra
Vincent J. Della Pietra Fredrick Jelinek John D. Lafferty
Robert L. Mercer Paul S. Roossin

Abstract

In this paper, we present a statistical approach to machine translation. We describe the application of our approach to translation from French to English and give preliminary results.

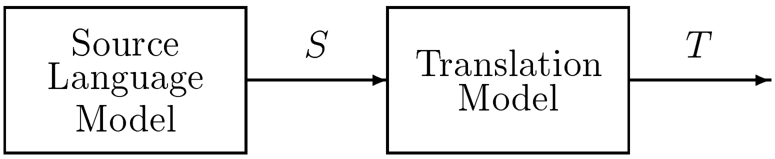
1 Introduction

The field of machine translation is almost as old as the modern digital computer. In 1949 Warren Weaver suggested that the problem be attacked with statistical methods and ideas from information theory, an area which he, Claude Shannon and others were developing at the time [12]. Although researchers quickly abandoned this approach, advancing numerous theoretical objections, we believe that the true obstacles lay in the relative impotence of the available computers and the dearth of machine-readable text from which to gather the statistics vital to such an attack. Today, computers are five orders of magnitude faster than they were in 1950 and have hundreds of millions of bytes of storage. Large, machine-readable corpora are readily available. Statistical methods have proven their value in automatic speech recognition [1] and have recently been applied to lexicography [11] and to natural language processing [2, 6, 7, 9, 10]. We feel that it is time to give them a chance in machine translation.

The job of a translator is to render in one language the meaning expressed by a passage of text in another language. This task is not always straightforward. For example, the translation of a word may depend on words quite far from it. Some English translators of Proust's seven volume work *A la Recherche du Temps Perdu* have striven to make the first word of the first volume the same as the last word of the last volume because the French original begins and ends with the same word [4]. Thus, in its most highly developed form, translation involves a careful study of the original text and may even encompass a detailed analysis of the author's life and circumstances. We, of course, do not hope to reach these pinnacles of the translator's art.

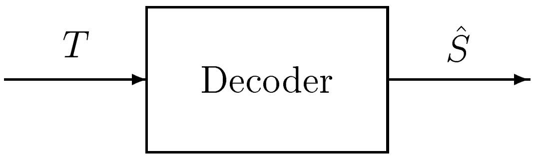
In this paper, we consider only the translation of individual sentences. Usually, there are many acceptable translations of a particular sentence, the choice among them being

A Statistical Machine Translation System



$$\Pr(S) \qquad \times \qquad \Pr(T \mid S) \;=\; \Pr(S,T)$$

A *Source Language Model* and a *Translation Model* furnish a probability distribution over source-target sentence pairs (S,T). The joint probability $\Pr(S,T)$ of the pair (S,T) is the product of the probability $\Pr(S)$ computed by the language model and the conditional probability $\Pr(T \mid S)$ computed by the translation model. The parameters of these models are estimated automatically from a large database of source-target sentence pairs using a statistical algorithm which optimizes, in an appropriate sense, the fit between the models and the data.



$$\hat{S} = \operatorname*{argmax}_{S} \Pr(S \mid T) = \operatorname*{argmax}_{S} \Pr(S,T)$$

A *Decoder* performs the actual translation. Given a sentence T in the target language, the decoder chooses a viable translation by selecting that sentence $\hat{S}$ in the source language for which the probability $\Pr(S \mid T)$ is maximum.

Figure 1

largely a matter of taste. We take the view that every sentence in one language is a possible translation of any sentence in the other. We assign to every pair of sentences (S, T) a probability, $\Pr(T \mid S)$, to be interpreted as the probability that a translator will produce T in the target language when presented with S in the source language. We expect $\Pr(T \mid S)$ to be very small for pairs like (*Le matin je me brosse les dents* | *President Lincoln was a good lawyer*) and relatively large for pairs like (*Le president Lincoln était un bon avocat* | *President Lincoln was a good lawyer*). We view the problem of machine translation then as follows. Given a sentence T in the target language, we seek the sentence S from which the translator produced T. We know that our chance of error is minimized by choosing that sentence S which is most probable given T. Thus, we wish to choose S so as to maximize $\Pr(S \mid T)$. Using Bayes' theorem, we can write

$$\Pr(S \mid T) = \frac{\Pr(S)\Pr(T \mid S)}{\Pr(T)}.$$

The denominator on the right of this equation does not depend on S and so it suffices to choose the S which maximizes the product $\Pr(S)\Pr(T \mid S)$. Call the first factor in this product the language model probability of S and the second factor the translation probability of T given S. Although the interaction of these two factors can be quite profound, it may help the reader to think of the translation probability as suggesting words from the source language which might have produced the words that we observe in the target sentence and to think of the language model probability as suggesting an order in which to place these source words.

Thus, as illustrated in Figure 1, a statistical translation system requires a method for computing language model probabilities, a method for computing translation probabilities, and, finally, a method for searching among possible source sentences S for the one which gives the greatest value for $\Pr(S)\Pr(T \mid S)$.

In the remainder of this paper we describe a simple version of such a system which we have implemented. In the next section we describe our language model for $\Pr(S)$, and in Section 3 we describe our translation model for $\Pr(T \mid S)$. In Section 4 we describe our search procedure. In Section 5 we explain how we estimate the parameters of our models from a large database of translated text. In Section 6 we describe the results of two experiments we performed using these models. Finally, in Section 7 we conclude with a discussion of some improvements which we intend to implement.

2 The Language Model

Given a word string, $s_1 s_2 \ldots s_n$, we can, without loss of generality, write

$$\Pr(s_1 s_2 \ldots s_n) = \Pr(s_1) \Pr(s_2 \mid s_1) \ldots \Pr(s_n \mid s_1 s_2 \ldots s_{n-1}).$$

Thus, we can recast the language modelling problem as one of computing the probability of a single word given all of the words that precede it in a sentence. At any point in the sentence, we must know the probability of an object word, s_j, given a history, $s_1 s_2 \ldots s_{j-1}$. Because there are so many histories, we cannot simply treat each of these probabilities as a separate parameter. One way to reduce the number of parameters is to place each of the histories into an equivalence class in some way and then to allow the probability of an object word to depend on the history only through the equivalence class into which that history falls. In an n-gram model, two histories are equivalent if they agree in their final $n-1$ words. Thus, in a bigram model, two histories are equivalent if they end in the same word and in a trigram model, two histories are equivalent if they end in the same two words.

While n-gram models are linguistically simple-minded, they have proven quite valuable in speech recognition and have the redeeming feature that they are easy to make and to use. We can see the power of a trigram model by applying it to something that we call bag translation from English into English. In bag translation we take a sentence, cut it up into words, place the words in a bag, and then try to recover the sentence given the bag. We use the n-gram model to rank different arrangements of the words in the bag. Thus, we treat an arrangement S as better than another arrangement S' if $\Pr(S)$ is greater than $\Pr(S')$. We tried this scheme on a random sample of sentences. From a collection of 100 sentences, we considered the 38 sentences with fewer that 11 words each. We had to restrict the length of the sentences because the number of possible rearrangements grows exponentially with sentence length. We used a trigram language model that had been

Bag model examples

Exact reconstruction (24 of 38)

Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)

Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)

In our organization research has two missions.
⇒ In our missions research organization has two.

FIGURE 2

constructed for a speech recognition system. We were able to recover 24 (63%) of the sentences exactly. Sometimes, the sentence that we found to be most probable was not an exact reproduction of the original, but conveyed the same meaning. In other cases, of course, the most probable sentence, according to our model was just garbage. If we count as correct all of the sentences that retained the meaning of the original, then 32 (84%) of the 38 were correct. Some examples of the original sentences and the sentences recovered from the bags are shown in Figure 2. We have no doubt that if we had been able to handle longer sentences, the results would have been worse and that the probability of error grows rapidly with sentence length.

3 The Translation Model

For simple sentences, it is reasonable to think of the French translation of an English sentence as being generated from the English sentence word by word. Thus, in the sentence pair (*Jean aime Marie* | *John loves Mary*) we feel that *John* produces *Jean*, *loves* produces *aime*, and *Mary* produces *Marie*. We say that a word is *aligned* with the word that it produces. Thus *John* is aligned with *Jean* in the pair that we just discussed. Of course, not all pairs of sentences are as simple as this example. In the pair (*Jean n'aime personne* | *John loves nobody*), we can again align *John* with *Jean* and *loves* with *aime*, but now, *nobody* aligns with both *n'* and *personne*. Sometimes, words in the English sentence of the pair align with nothing in the French sentence, and similarly occasionally words in the French member of the pair do not appear to go with any of the words in the English sentence. We refer to a picture such as that shown in Figure 3 as an alignment. An alignment indicates the origin in the English sentence of each of the words in the French sentence. We call the number of French words that an English word produces in a given alignment its *fertility* in that alignment.

If we look at a number of pairs, we find that words near the beginning of the English sentence tend to align with words near the beginning of the French sentence and that words near the end of the English sentence tend to align with words near the end of the French sentence. But this is not always the case. Sometimes, a French word will appear quite far from the English word that produced it. We call this effect *distortion.* Distortions will, for example, allow adjectives to precede the nouns which they modify in English but to follow them in French.

It is convenient to introduce the following notation for alignments. We write the French sentence followed by the English sentence and enclose the pair in parentheses. We separate

Alignment example

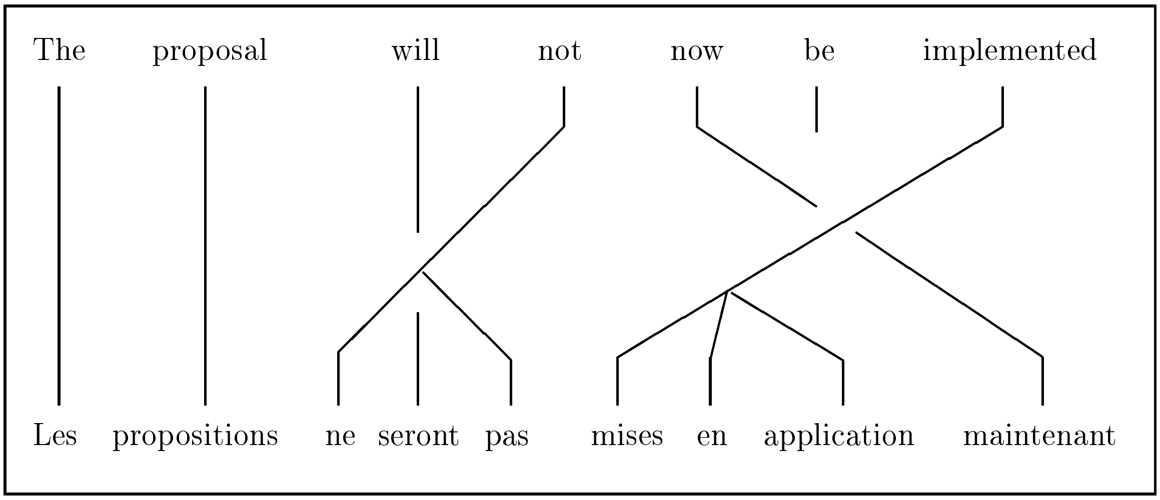


FIGURE 3

the two by a vertical bar. Following each of the English words, we give a parenthesized list of the positions of the words in the French sentence with which it is aligned. If an English word is aligned with no French words, then we omit the list. Thus (*Jean aime Marie* | *John(1) loves(2) Mary(3)*) is the simple alignment with which we began this discussion. In the alignment (*Le chien est battu par Jean* | *John(6) does beat(3,4) the(1) dog(2)*), *John* produces *Jean*, *does* produces nothing, *beat* produces *est battu*, *the* produces *Le*, *dog* produces *chien*, and *par* is not produced by any of the English words.

Rather than describe our translation model formally, we present it by working an example. To compute the probability of the alignment, (*Le chien est battu par Jean* | *John(6) does beat(3,4) the(1) dog(2)*), begin by multiplying the probability that *John* has fertility 1 by $\Pr(Jean \mid John)$. Then multiply by the probability that *does* has fertility 0. Next, multiply by the probability that *beat* has fertility 2 times $\Pr(est \mid beat) \Pr(battu \mid beat)$, and so on. The word *par* is produced from a special English word which is denoted by $\langle null \rangle$.

The result is

$$\begin{aligned}
&\Pr(fertility = 1 \mid John) \times \Pr(Jean \mid John) \times \\
&\Pr(fertility = 0 \mid does) \times \\
&\Pr(fertility = 2 \mid beat) \times \Pr(est \mid beat) \Pr(battu \mid beat) \times \\
&\Pr(fertility = 1 \mid the) \times \Pr(Le \mid the) \times \\
&\Pr(fertility = 1 \mid dog) \times \Pr(chien \mid dog) \times \\
&\Pr(fertility = 1 \mid \langle null \rangle) \times \Pr(par \mid \langle null \rangle)\,.
\end{aligned}$$

Finally, factor in the distortion probabilities. Our model for distortions is, at present, very simple. We assume that the position of the target word depends only on the length of the target sentence and the position of the source word. Therefore, a distortion probability has the form $\Pr(i \mid j, l)$ where i is a target position, j a source position, and l the target length.

In summary, the parameters of our translation model are a set of fertility probabilities $\Pr(n \mid e)$ for English word e and for each fertility n from 0 to some moderate limit, in our case 25; a set of translation probabilities $\Pr(f \mid e)$, one for each element f of the French vocabulary and each member e of the English vocabulary; and a set of distortion probabilities $\Pr(i \mid j, l)$ for each target position i, source position j, and target length l. We limit i, j, and l to the range 1 to 25.

4 Searching

In searching for the sentence S which maximizes $\Pr(S)\Pr(T \mid S)$ we face the difficulty that there are simply too many sentences to try. Instead, we must carry out a suboptimal search. We do so using a variant of the *stack search* that has worked so well in speech recognition [1]. In a stack search, we maintain a list of partial alignment hypotheses. Initially, this list contains only one entry corresponding to the hypothesis that the target sentence arose in some way from a sequence of source words that we do not know. In the alignment notation introduced earlier, this entry might be (*Jean aime Marie* | $*$) where the asterisk is a place holder for an unknown sequence of source words. The search proceeds by iterations, each of which extends some of the most promising entries on the list. An entry is extended by adding one or more additional words to its hypothesis. For example, we might extend the initial entry above to one or more of the following entries:

(*Jean aime Marie* | *John(1)* $*$),

(*Jean aime Marie* | $*$ *loves(2)* $*$),

(*Jean aime Marie* | $*$ *Mary(3)*),

(*Jean aime Marie* | *Jeans(1)* $*$).

The search ends when there is a complete alignment on the list that is significantly more promising than any of the incomplete alignments.

Sometimes, the sentence S' that is found in this way is not the same as the sentence S that a translator might have been working on. When S' itself is not an acceptable translation, then there is clearly a problem. If $\Pr(S')\Pr(T \mid S')$ is greater than $\Pr(S)\Pr(T \mid S)$, then the problem lies in our modelling of the language or of the translation process. If, however, $\Pr(S')\Pr(T \mid S')$ is less than $\Pr(S)\Pr(T \mid S)$, then our search has failed to find the most likely sentence. We call this latter type of failure a search error. In the case of a search error, we can be sure that our search procedure has failed to find the most probable source sentence but we cannot be sure that were we to correct the search we would also correct the error. We might simply find an even more probable sentence which nonetheless is incorrect. Thus, while a search error is a clear indictment of the search procedure, it is not an acquittal of either the language model or the translation model.

5 Parameter Estimation

Both the language model and the translation model have many parameters which must be specified. In order to estimate these parameters accurately, we need a large quantity of data. For the parameters of the language model, we need only English text, which is available in computer-readable form from many sources; but for the parameters of the translation model, we need pairs of sentences which are translations of one another.

By law, the proceedings of the Canadian parliament are kept in both French and English. As each member rises to address the question before the house or otherwise express himself, his remarks are jotted down in whichever of the two languages he has used. After the meeting adjourns, a collection of translators begins working to produce a complete set of the proceedings in both French and English. These proceeding are called Hansards in remembrance of the publisher of the proceedings of the British parliament in the early 1800's. All of these proceedings are available in computer-readable form and we have been able to obtain about 100 million words of English text and the corresponding French text from the Canadian government. Although the translations are not made sentence by sentence, we have been able to extract about 3 million pairs of sentences by using a statistical algorithm based on sentence length. Approximately 99% of these pairs are made up of sentences that are actually translations of one another. It is this collection of sentence pairs, or more properly various subsets of this collection, from which we have estimated the parameters of the language and translation models.

In the experiments which we describe later, we use a bigram language model. Thus, we have one parameter for every pair of words in the source language. We estimate these parameters from the counts of word pairs in a large sample of text from the English part of our Hansard data using a method described by Jelinek and Mercer [8].

In section 3 we discussed alignments of sentence pairs. If we had a collection of aligned pairs of sentences, then we could estimate the parameters of the translation model by counting, just as we do for the language model. However, we do not have alignments but only the unaligned pairs of sentences. This is exactly analogous to the situation in speech recognition where one has the script of a sentence and the time waveform corresponding to an utterance of it, but no indication of just what in the time waveform corresponds to what in the script. In speech recognition, this problem is attacked with the EM algorithm [3, 5]. We have adapted this algorithm to our problem in translation. Briefly, it works like this. Given some initial estimate of the parameters, we can compute the probability of any particular alignment. We can then re-estimate the parameters by weighing each possible alignment according to its probability as determined by the initial guess of the parameters. Repeated iterations of this process lead to parameters which assign ever greater probability to the set of sentence pairs that we actually observe. This algorithm leads to a local maximum of the probability of the observed pairs as a function of the parameters of the model. There may be many such local maxima. The particular one at which we arrive will, in general, depend on the initial choice of parameters.

English: the

French	**Probability**	**Fertility**	**Probability**
le	.610	1	.871
la	.178	0	.124
l'	.083	2	.004
les	.023		
ce	.013		
il	.012		
de	.009		
à	.007		
que	.007		

FIGURE 4

6 Two Pilot Experiments

In our first experiment, we test our ability to estimate parameters for the translation model. We chose as our English vocabulary the 9000 most common words in the English part of the Hansard data and as our French vocabulary the 9000 most common French words. For the purposes of this experiment, we replaced all other words with either the *unknown English word* or the *unknown French word*, as appropriate. We applied the iterative algorithm discussed above in order to estimate some 81 million parameters from 40,000 pairs of sentences comprising a total of about 800,000 words in each language. The algorithm requires an initial guess of the parameters. We assumed that each of the 9000 French words was equally probable as a translation of any of the 9000 English words; we assumed that each of the fertilities from 0 to 25 was equally probable for each of the 9000 English words; and finally, we assumed that each target position was equally probable given each source position and target length. Thus, our initial choices contained very little information about either French or English.

English: not

French	**Probability**	**Fertility**	**Probability**
pas	.469	2	.758
ne	.460	0	.133
non	.024	1	.106
pas du tout	.003		
faux	.003		
plus	.002		
ce	.002		
que	.002		
jamais	.002		

FIGURE 5

Figure 4 shows the translation and fertility probabilities which we estimated for the English word *the*. We see that, according to the model, *the* translates most frequently into the French articles *le* and *la*. This is not surprising, of course, but we emphasize that it is determined completely automatically by the estimation process. In some sense, this correspondence is inherent in the sentence pairs themselves. Figure 5 shows these probabilities for the English word *not*. As expected, the French word *pas* appears as a highly probable translation. Also, the fertility probabilities indicate that *not* translates most often into two French words, a situation consistent with the fact that negative French sentences contain the auxiliary word *ne* in addition to a primary negative word such as *pas* or *rien*.

For both of these words, we could easily have discovered the same information from a dictionary. In Figure 6, we see the trained parameters for the English word *hear*. As we would expect, various forms of the French word *entendre* appear as possible translations, but the most probable translation is the French word *bravo*. When we look at the fertilities here,

English: hear

French	**Probability**	**Fertility**	**Probability**
bravo	.992	0	.584
entendre	.005	1	.416
entendu	.002		
entends	.001		

FIGURE 6

we see that the probability is about equally divided between fertility 0 and fertility 1. The reason for this is that the English speaking members of parliament express their approval by shouting *Hear, hear!*, while the French speaking ones say *Bravo!* The translation model has learned that usually two *hear*s produce one *bravo* by having one of them produce the *bravo* and the other produce nothing.

A given pair of sentences has many possible alignments since each target word can be aligned with any source word. A translation model will assign significant probability only to some of the possible alignments and we can gain further insight about the model by examining the alignments that it considers most probable. We show one such alignment in Figure 3. Observe that, quite reasonably, *not* is aligned with *ne* and *pas* while *implemented* is aligned with the phrase *mises en application*. We can also see here a deficiency of the model since intuitively we feel that *will* and *be* act in concert to produce *seront* while the model aligns *will* with *seront* but aligns *be* with nothing.

In our second experiment, we used the statistical approach to translate from French to English. In order to have a manageable task, we limited the English vocabulary to the 1000 most frequently used words in the English part of the Hansard corpus. We chose the French vocabulary to be the 1700 most frequently used French words in translations of sentences that were completely covered by the 1000-word English vocabulary. We estimated the 17 million parameters of the translation model from 117,000 pairs of sentences which were completely covered by both our French and English vocabularies. We estimated the parameters of the bigram language model from 570,000 sentences from the English part of the Hansard data. These sentences contain about 12 million words altogether and are not restricted to sentences completely covered by our vocabulary.

We used our search procedure to decode 73 new French sentences from elsewhere in the Hansard data. We assigned each of the resulting sentences a category according to the following criteria. If the decoded sentence was exactly the same as the actual Hansard translation, we assigned the sentence to the *exact* category. If it conveyed the same meaning as the Hansard translation but in slightly different words, we assigned it to the *alternate* category. If the decoded sentence was a legitimate translation of the French sentence but did not convey the same meaning as the Hansard translation, we assigned it to the *different* category. If it made sense as an English sentence but could not be interpreted as a translation of the French sentence, we assigned it to the *wrong* category. Finally, if the decoded sentence was grammatically deficient, we assigned it to the *ungrammatical* category. An example from each category is shown in Figure 7, and our decoding results are summarized in Figure 8.

Only 5% of the sentences fell into the exact category. However, we feel that a decoded sentence which is in any of the first three categories (exact, alternate, or different) represents a reasonable translation. By this criterion, the system performed successfully 48% of the time.

Translation examples

Exact

	Ces ammendements sont certainement nécessaires.
Hansard:	These amendments are certainly necessary.
Decoded as:	These amendments are certainly necessary.

Alternate

	C'est pourtant très simple.
Hansard:	Yet it is very simple.
Decoded as:	It is still very simple.

Different

	J'ai reçu cette demande en effet.
Hansard:	Such a request was made.
Decoded as:	I have received this request in effect.

Wrong

	Permettez que je donne un example à la Chambre.
Hansard:	Let me give the House one example.
Decoded as:	Let me give an example in the House.

Ungrammatical

	Vous avez besoin de toute l'aide disponible.
Hansard:	You need all the help you can get.
Decoded as:	You need of the whole benefits available.

FIGURE 7

As an alternate measure of the system's performance, one of us corrected each of the sentences in the last three categories (different, wrong, and ungrammatical) to either the exact or the alternate category. Counting one stroke for each letter that must be deleted

Translation results

Category	Number of sentences	Percent
Exact	4	5
Alternate	18	25
Different	13	18
Wrong	11	15
Ungramatical	27	37
Total	73	

FIGURE 8

and one stroke for each letter that must be inserted, he needed 776 strokes to repair all of the decoded sentences. This compares with the 1916 strokes required to generate all of the Hansard translations from scratch. Thus, to the extent that translation time can be equated with key strokes, the system reduces the work by about 60%.

7 Plans

There are many ways in which the simple models described in this paper can be improved. We expect some improvement from estimating the parameters on more data. For the experiments described above, we estimated the parameters of the models from only a small fraction of the data we have available: for the translation model, we used only about one percent of our data and for the language model, only about ten percent.

We have serious problems in sentences in which the translation of certain source words depends on the translation of other source words. For example, the translation model produces *aller* from *to go* by producing *aller* from *go* and nothing from *to*. Intuitively we feel that *to go* functions as a unit to produce *aller*. While our model allows many target words to come from the same source word, it does not allow several source words to work together to produce a single target word. In the future, we hope to address the problem of identifying groups of words in the source language that function as a unit in translation. This may take the form of a probabilistic division of the source sentence into groups of words.

At present, we assume in our translation model that words are placed into the target sentence independently of one another. Clearly, a more realistic assumption must account for the fact that words form phrases in the target sentence which are translations of phrases in the source sentence and that the target words in these phrases will tend to stay together even if the phrase itself is moved around. We are working on a model in which the positions of the target words produced by a particular source word depend on the identity of the source word and on the positions of the target words produced by the previous source word.

We are preparing a trigram language model which we hope will substantially improve the performance of the system. A useful information-theoretic measure of the complexity of a language with respect to a model is the perplexity as defined by Bahl, Jelinek and Mercer [1]. With the bigram model that we are currently using, the source text for our 1000-word translation task has a perplexity of about 78. With the trigram model that we are preparing, the perplexity of the source text is about 9. In addition to showing the strength of a trigram model relative to a bigram model, this also indicates that the 1000-word task is very simple.

We treat words as unanalyzed wholes, recognizing no connection, for example, between *va*, *vais*, and *vont*, or between *tall*, *taller*, and *tallest*. As a result, we cannot improve our statistical characterization of *va*, say, by observation of sentences involving *vont*. We are working on morphologies for French and English so that we can profit from statistical regularities that our current word-based approach must overlook.

Finally, we treat the sentence as a structureless sequence of words. Sharman, Jelinek, and Mercer discuss a method for deriving a probabilistic phrase structure grammar automatically from a sample of parsed sentences [10]. We hope to apply their method to construct grammars for both French and English and to base future translation models on the grammatical constructs thus defined.

References

[1] L.R. Bahl, F. Jelinek, and R.L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190, March 1983.

[2] J.K. Baker. Stochastic modeling for automatic speech understanding. In R.A. Reddy, editor, *Speech Recognition*, pages 521–541, Academic Press, New York, 1979.

[3] L.E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8, 1972.

[4] R. Bernstein. Howard's Way. *The New York Times Magazine*, 40 ff., 1988. September 25, 1988.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[6] J.D. Ferguson. Hidden Markov analysis: an introduction. In J.D. Ferguson, editor, *Hidden Markov Models for Speech*, pages 8–15, IDA-CRD, Princeton, N.J., October 1980.

[7] R.G. Garside, G.N. Leech, and G.R. Sampson. *The Computational Analysis of English: A Corpus-Based Approach.* Longman, 1987.

[8] F. Jelinek and R.L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, Amsterdam, The Netherlands: North-Holland, May 1980.

[9] G.R. Sampson. A stochastic approach to parsing. In *Proceedings of the 11th International Conference on Computational Linguistics*, pages 151–155, Bonn, GDR, 1986.

[10] R.A. Sharman, F. Jelinek, and R.L. Mercer. Generating a grammar for statistical training. In *Proceedings of the IBM Conference on Natural Language Processing*, Thornwood, NY, October 1988.

[11] J.M. Sinclair. Lexicographic evidence. In R. Ilson, editor, *Dictionaries, Lexicography and Language Learning*, pages 81–94, Pergamon Press, New York, N.Y., 1985.

[12] W. Weaver. Translation(1949). In *Machine Translation of Languages*, MIT Press, Cambridge, MA, 1955.