[*Proceedings of a Workshop on Machine Translation*, July 1990, UMIST]

An Introduction to Statistical Machine Translation

R.A.Sharman

IBM UK Scientific Centre
Athelstan House
St. Clements St.
Winchester
England

The automatic machine translation of a text in one natural language into another natural language is a task which been considered for at least as long as computers have existed. A wide variety of techniques has been considered for this task, although none has proved outstandingly successful, at least judging by the performance of current translating systems. It seems worthwhile, therefore to consider the relevance of techniques not seriously considered before, just in case some additional insight can be obtained into the translation process. The approach of treating language translation as a task in *decoding,* familiar in many applications of Information Theory, is briefly described here. This approach has been found very successful in Machine Speech Recognition, and it may be that some of the techniques used there can also carry over to the Machine Translation domain. Some preliminary results suggest that the approach may be capable one day of performing certain translation tasks, although it is by no means a commercially practical proposition at the moment.

An Historical Note

On the 4th March, 1947 Warren Weaver, conscious of the tremendous advances in both deciphering messages, and in telecommunications, which had been made during World War II, wrote to Norbert Weiner, the father of Cybernetics, as follows [1]:

> "... one naturally wonders if the problem of translation could conceivably be treated as a problem in *cryptography.* When I look at an article in Russian, I say: This is really written in English, but it has been coded in some strange symbols. I will proceed to decode."

He recounts that Weiner replied on April 30th, 1947 in the following fashion:

> "...I frankly am afraid the boundaries of words in different languages are too vague and the emotional and international connotations are too extensive to make any quasimechanical translation scheme very hopeful."

Judging from his reply, it is not clear that Weiner really addressed himself to the main point of Weaver's argument: that the use of novel decoding techniques might be useful in Machine Translation. While the idea of treating statements in a language as a form of communication received some attention at the time it was lost in accusations that it was naive and inappropriate to natural language tasks. Subsequently, the mainstream Linguistic world turned its back on statistical and numerical studies of language in the wake of the Chomsky revolution [2], a situation that lasted until nearly the present day [3,4],

Simultaneously, Weaver was interested in the capability of computers to perform tasks of this complexity, and mentions a reply to an enquiry of his to an early researcher in the field of computer construction, A.D.Booth of Birkbeck College, London on Feb. 12th, 1948, who said:

> "... we have considered this problem (translation) in some detail, and it transpires that a machine of the type envisaged (an electronic computer) could perform this function without any modification in its design."

Since that date, few people have seriously suggested that computers, as we currently understand them, are inadequate to perform at least simple translation tasks. Indeed most modern work on machine translation takes this point for granted. The significant debate is on *how* a computer can be made to perform the translation, not *whether* such a task can be performed. In summary, Weaver suggested four main ideas which he believed relevant to the consideration of translation by machine, and which are still of interest, today. They were:

1. Meaning and Context

   Resolve the ambiguity of a word by using the context of N words before it and N words after it. How big should the window of 2N+ 1 be, to ensure x% accuracy? We would like to be able to specify x in advance, and then deduce the N which is required to satisfy it.

2. Language and Logic

   A computer performs logical deductions and consequently, we would expect to model *algorithmic* aspects of language relatively easily, but find the *alogical* aspect of language difficult to handle. This will probably prevent high-quality, literary translations from being performed adequately, but there will be many simpler tasks, such as translating scientific literature, where these aspects of language are absent, or simplified, and in which simple methods would be adequate.

3. Translation and Cryptography

   Communication systems are fundamentally *statistical* in nature, so perfect decoding will often be unattainable. But a process which decodes at some specified level of error *is* attainable. This is the model for the approximate translation of one language into another, which we hope will approach arbitrarily closely to the ultimate performance, that of the best human translators.

4. Language and Invariants

   It may be that the direct approach of translating from one language to another, without any intervening transition step, is possible. It may also be the case that some intermediate processing is desirable. In what appears to be an early reference to the use of an *interlingua* Weaver summed this up as Follows:

   > "Descend from each language to the common base of human communication -- universal language -- and then re-emerge by whatever particular route is convenient."

It can be seen that these early ideas were by no means simplistic and naive. What has changed in the interim is, of course the amazing growth in the power of computers (by a factor of 5 or 6 orders of magnitude in both storage and computing power), and in the availability of suitable data on which to base models of language (see below). In addition, there have been important advances in the theory and practice of communications, to the extent that signals can be received from distant spacecraft with very low loss of information, and human speech can be recognised by computer.

An example of translation

The following example is taken from the proceedings of the Canadian parliament, as recorded in Hansard [5]. The sentence, which is not untypical of the material in the proceedings, was spoken in either Canadian English, or Canadian French (we do not know which), and then translated by official translators into the the other language. The separate sentences are:

```
Mr Speaker, I rise on a question of privilege affecting the
the rights and prerogatives of parliamentary committees
and one which reflects on the word of two ministers of the Crown.

Monsieur l'Orateur, je souleve la question de privelege
apropos des droit et des prerogatives des comites parlement
et  pour mettre en doute les propos de deux ministres de la Couronne.
```

Let us assume for the sake of argument that it is sentences of this length, subject matter, and syntactic complexity, which we are interested in translating. The absolute quality, or desirability, of the particular translation given in the example does not really concern us here, the significant point being that the translation was produced by a professional translator, and is of at least minimal quality for acceptance by its intended users. Can we learn how to translate by machine by reference to parallel texts of this kind? Even if we do so, the resulting machine translation is unlikely to exceed the performance of the human translations, but hopefully it might be nearly as good, or at least within the range of performance of a group of human translators.

When roughly aligned, so that the phrases of one sentence correspond approximately with the phrases of the other, it is apparent that there is a great similarity between the two sentences, in the example above, as follows:

```
Mr       Speaker    , I  rise on a question  of privilege
Monsieur l'Orateur  , je souleve la question de privelege

affecting the rights and prerogatives     of parliamentary committees
apropos   des droit  et  des prerogatives des  comités parlement

and one  which reflects on the word of   two  ministers of the Crown
et  pour mettre en doute   les propos de deux ministres de la  Couronne.
```

This is by no means a word-for-word, or literal, translation, and we would not expect every sentence to align so neatly, but nevertheless there is a strong similarity between the two sentences. It is this similarity that gives us hope that some kind of decoding from one to the other will be possible. Decoding of much more scrambled messages is widespread practice in cryptography, and decoding speech signals to determine the words which gave rise to the signal is a practical possibility in automatic speech recognition. So on the face of it, it would appear not unreasonable that the rather straightforward mapping from words to words, shown in the example, might be plausible.

It seems that a respectable task for a machine translation system to aspire to is the translation of text taken from a specific domain (here: parliament proceedings); with an unrestricted vocabulary; for sentences of arbitrary length; and perform about as well as a human translator would in the same circumstances.

A Naive Approach to Translation

From the preceding example it would appear that a simple method of translation could be constructed by mapping the words and phrases in the *source language* into words and phrases in the *target language* by the use of a *glossary,* and then *re-arranging* the phrases in the target language into a sentence in the target language. The precise suggestion is due to Brown, et al [6,7], and their presentation is followed here, fairly faithfully. The questions of how we know what the glossary contains, and how we perform the re-arrangement, are details to which we shall return below.

It should be emphasised that we are looking for a sensible means of translating which will result in true sentences in the target language, not just jumbles of plausible looking words from which a clever reader might deduce the intent of the original sentence. Thus, of course, a word-by-word translation would not be adequate. In general, our criterion for an adequate translation is formed by reference to what a skilled human translator would have done in the same circumstance.

The simple method can be summarised, as follows:

1. Partition the source text into a set of fixed locutions.

2. Use a glossary, plus contextual information, to select a corresponding set of fixed locutions in the target language.

3. Arrange the words of the target fixed locutions into a sequence that forms a sentence in the target language.

Our basic requirements for this method are, then, a glossary of corresponding locutions and a mechanism of re-ordering locutions. First, however, it will be useful to formalise the concepts of translation from source to target language.

A General Model of Translation

Assume that a *source* emits sentences in a *source language,* and that these are *encoded* into sentences in a *target language.* The sentences in the target language are transmitted to an observer, who passes them through a *decoder* to reveal the hidden message in the source language, which is then passed to user of the messages, a *sink.* This is the general, and widely known, model of communication which can be depicted as in Figure 1 on page 5.

It should be noted that from the *decoders* point of view, only the encoded, target language sentence can be observed. The question to be solved by the decoder is: what was the sentence in the source language which most probably gave rise to the observed target language sentence, the output of the encoder?

To make the discussion a little more specific, if we wish to translate from French into English, then we consider that English is the *source* language, and that the English sentence has been encoded in French before we see it. French is the *target* language, and the job of the decoder is to recover the original, unknown, source language sentence, in this case, the English sentence. (This may seem a little backwards, but it is the standard way of looking at these sort of decoding tasks, and soon becomes quite natural). A formulation of English to French decoding could be obtained by reversing the concepts.

When we observe a French sentence we have to ask what English sentence could possibly have been used, which when encoded into French yielded the sentence we actually observed? Of course, there will be number of different sentences which could all have resulted in the observed French sentence, but some will be less likely than others. In principle, we consider that *any* English sentence could have given rise to the observed French sentence, some more probably than others. In particular, we would like to know the English sentence which was most likely to have given rise to the French sentence and we will consider this one to have been the *best,* in some sense.

If, for example, we observe the French sentence *President Lincoln etait un bon avocat,* we would think it unlikely that it arose from the English sentence *This morning I brushed my teeth,* but rather more likely that it arose from the English sentence *President Lincoln was a good lawyer.* This is equivalent to saying that the conditional probability *P(This morning I brushed my teeth | President Lincoln etait un bon avocat)* will be small, whereas the P*(President Lincoln was a good lawyer | President Lincoln etait un bon avocat)* will be large. It seems reasonable to assume that the higher probability the better the translation.

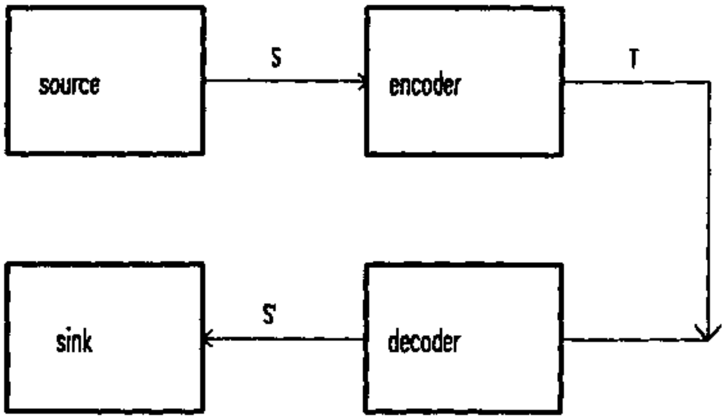

Figure 1. A simple model of translation as encoding/decoding:

The *source* produces a sentence, S, in the source language, with probability *P(S)*. The *encoder* uses S to create the encoded sentence, T, in the target language with probability *P(T|S)*. The joint probability of both S and T occurring together is therefore *P(S)* • *P(T|S)* = *P(S,T)*. The sentence, T, is transmitted to the receiver over some communication channel, the details of which do not concern us here. The *decoder* seeing only the sentence, T, constructs the sentence, S', with probability *P(S'| T),* which is its best guess of what the source originally produced. For successful decoding it is required that S' should approximate 5 as closely as possible. The *sink* is the final recipient of the transmitted message.

Let *S be* a sentence in the *source language,* then *P(S)* is the *a priori* probability that *S* occurs. Let *T* be a sentence in the *target language,* then *P(T)* is the *a priori* probability that *T* occurs. The probability that a *particular* source language sentence, *S*, was used to generate the *given* target language sentence, *T*, is (according to Bayes Theorem):

$$P(S|T) = \frac{P(S) \bullet P(T|S)}{P(T)}$$

Here, the *P(S|T)* is the *conditional probability* that source language sentence, *S,* is the one we seek, given that we actually observed the target language sentence, *T*. Similarly, the *P(T|S)* is the *conditional probability* that target language sentence, *T,* could have arisen, given that source language sentence, *S,* was specified. In general, we seek the source sentence that *most likely* gave rise to the observed target sentence, or

$$\max_{S} P(S|T) = \underset{S}{\text{argmax}}\, [\text{P(S,T)}]$$

This is equivalent to saying that we seek to construct pairs of sentences which are highly correlated *in exactly the way that human translated sentences are correlated.* For this we need to construct a model of translation which will allow us to calculate values for the terms *P(S)* and *P(T|S)* , and consequently perform a selection by choosing the *S* which achieves this maximisation. It is by no means obvious what the desirable model should look like, and a large number of possible models could be examined. In as far as this is similar to other problems in, for example, speech recognition, there is be a wealth of experience to draw on to guide the selection of model. If a particular model performs poorly, this may only be a reflection of the assumptions of that model, not of the method in general.

We turn now to an elaboration of the naive model of translation sketched above, and look at the two terms *P(S|T)* , for which we need a probabilistic glossary, and *P(S),* for which we need a probabilistic language model.

Creating a Simple Glossary

In the Hansard text, referred to above, the corresponding phrases are not identified as in our earlier example. In fact, the sentences are not conveniently paired either, but with some effort it is possible to produce a parallel text where each sentence in the English proceedings is identified, and paired with exactly one sentence in the French proceedings. For the remainder of the paper it is assumed that a large parallel text of this nature is a fundamental pre-requisite of any statistically based model, and that some millions of words, if not hundreds of millions of words, of text are available in this way. In order to consider this approach for another language pair a suitable corpus of parallel sentences would be required. If this seems unreasonable consider that at the very worst it would only(!) require some actual translators to work away at translating for some months, or years, to create the data. It would not necessarily require new theoretical insights or practical tools and techniques to be developed. In fact, the creation of large machine-readable corpora is already happening in the EC and the UN where at least the major languages are readily translated in large quantities.

Let us assume, for the moment, that we have not only paired sentences, but that the actual phrase-by-phrase correspondences have also been marked. An example of a sentence in English which has been paired with a sentence in French, and for which the corresponding words have been identified, is shown in Figure 2 on page 7. This is called an *aligned* sentence.

How can such a glossary of translations be constructed automatically? A simple method suggests itself, as follows:

Consider, for the moment, the simple case that one English word translates to one French word, and vice-versa (the method can be generalised for multi-word to multi-word translation). Let e_1, $e_2,...,e_E$ be an English vocabulary, and let $f_1,f_2,...f_F$ be the corresponding French vocabulary. Now, let $C(e_i,\ f_j)$ be a count, the number of times that that the English word e_i translates to the French word f_j, as observed from the alignments between the two sentences, and let $C(f_j)$ *b*e the count of the total number of occurrences of f_j. Then, the probability of e_i being the required word, given only f_j can be estimated from the frequency of the joint event of both words occurring together relative to the frequency of f_j in all its contexts, or

$$P(e_i|f_j) = \frac{C(e_i,f_j)}{C(f_j)}$$

In the limit, as the size of corpus from which these relative frequencies are obtained, grows, the ratio approximates to the true probability. However, two difficulties exist. One is that insufficient data can be observed to gain reliable estimates of the relative frequencies, but this can only be overcome by using a very large corpus. A more serious problem is that, in practice, it is *not* known which words actually correspond, since the alignments we assumed to exist are not usually visible.

An alternative solution can, however, be found [6] from the knowledge of the paired, but unaligned, sentences, in the following way: assume that each of the words in the source sentence contributes equally to the derived word in the target sentence (what else could one assume without more knowledge?).

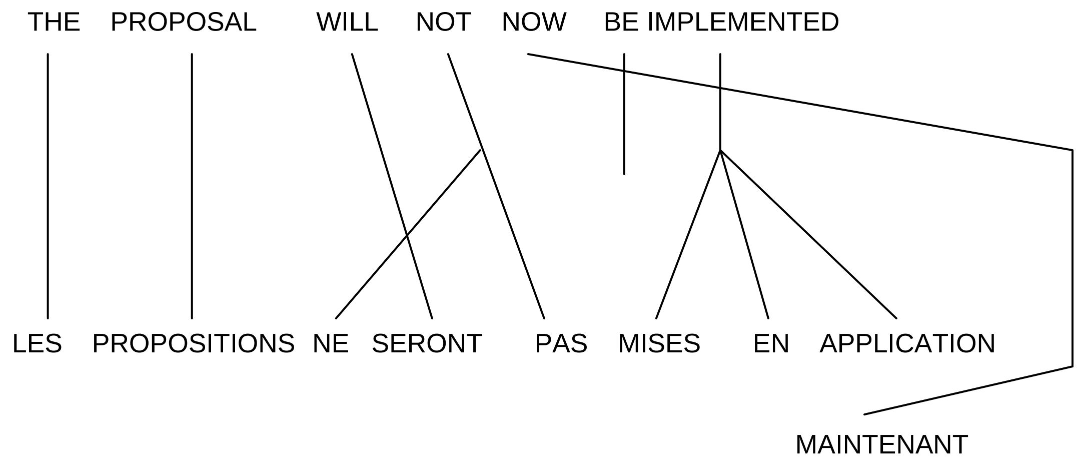

Figure 2. A sample alignment for two sentences: The alignment between the two parallel sentences is shown by linking the words which correlate. If these correlations are known for a suitably large enough text then it may be possible to learn idiomatic translations, such as the use of *mises en application* for *implemented* by reference solely to the fact that human translators tend to use this to some degree, rather than looking up the words in a dictionary.

Let $f_{i1}, f_{i2}, \ldots f_{in}$ be the French words of some sentence, say the k-th sentence, in the corpus. Now, let $C_k(e_i,f_j)$ be the count of the number of times that that the English word e_i in the k-th English sentence translates to some French word, f_j, in the k-th French sentence, for $q = 1,\ldots,k_n$. Then, the counts of the joint occurrences of e_i, and f_j in the k-th sentence can be formed from:

$$C_k(e_i,f_j) = \sum_{q=1}^{k_n} C_k\ (e_i,f_j)\ /\ k_n$$

The counts of the joint occurrences of e_i and f_j in the entire corpus can be formed by summing all these terms to form

$$C\ (e_i,f_j) = \sum_{k} C_k\ (e_i,f_j)$$

Then, the $P(e_i|f_j)$ can be estimated by relative frequency as before. This estimate will, however, unfairly favour high frequency words, and thus should be normalised by the probability of e_i. This normalisation is well supported by arguments from Information Theory [8] since the ratio of $P(e_i|f_j)/P(e_i)$ is related to $I(e_i{:}f_j)$, the mutual information of the words e_i and f_j, in the following way:

$$I(e_i{:}f_j) = \log \frac{P(e_i|f_j)}{P(e_i)}$$

The result of finding the pairs of words which have the highest mutual information is shown in Figure 3 on page 9. It must be stressed that it is not being claimed that this technique yields words which are translations of each other, just that they are highly correlated. If there are also useful candidates for translation, then we may have the basis for refining a technique which really could build a proper glossary.

Creating a Probabilistic Glossary

What is really needed is a system in which the relative likelihoods of one word translating as another are given, so it might be possible to select the appropriate word in a given context. For this, we need to know the probability with which each source word translates to each target word. While there are several heuristic ways to adapt the procedure of the previous section, there is unfortunately no optimal procedure to calculate the complete list of French words to which an English word translates, and the relative likelihood of each of those possibilities.

However, let us postulate that there is a model of the production of English words from French words, in the following way. Each French word gives rise to one (principal) English word, and that word then (optionally) gives rise to a number of other English words. If we had such a model we would be able to look at a pair of sentences, and deduce the alignments which map between them.

Let $P(e_i|f)$ be the probability of e_i being generated by *f*. Let $Q(e_j|e_i)$ be the probability of e_j being generated if e_i has already been generated. Let $R(k|e_i)$ be the probability of *k* items being generated if e_i has already been generated. The probability of a particular pattern of alignments, λ, between the French word, *f*, and the corresponding English word(s), e_{i1},..., can be now be modelled as:

$P(e,\lambda|f) = P(f \rightarrow e_{i1}, e_{i1 \rightarrow} e_{i2} \ldots, e_{ik}) =$

$$P(e_i|f) \bullet R(k\text{-}1|e_{i1}) \bullet Q(ei_2|e_{i1}) \bullet Q(e_{i3}\,|e_{i1}) \bullet \ldots\ldots \bullet Q(e_{ik} \setminus e_{i1})$$

The probability of a particular pattern of alignments, Λ, between the French sentence *F* and the corresponding English sentence *E* can be modelled as:

$$P(E,\, \Lambda|F) = \prod_j P(e,\lambda|f_j)$$

Thus, given the relations, *P, Q,* and *R,* we can calculate the probabilities of various alignments, and thus select the best alignment. (In fact, this was how the alignment shown in Figure 2 on page 7 was generated.) There is of course, only one problem -- we do not have the values of these functions! Fortunately, a method exists of finding a good approximation to the values of the functions by iteratively refining some arbitrary initial estimate [9]. The method used is called the *forward-backward* algorithm, and is a special case of a well-known class of optimisation problems satisfied by the EM-algorithm. The general approach is as follows:

1. Make an initial estimate of parameters,

2. Train the model on some large body of text,

3. Re-estimate parameters iteratively.

The result of performing such a process on a large body of training text is shown for a few sample words in Figure 4 on page 10. The general pattern of association will be seen to be very encouraging, although some peculiarities are evident. For example the idiosyncratic correlation of the English word *hear* with the French word *bravo* is a direct consequence of the behaviour in the Canadian Parliament!

eau	water	toujours	always
lait	milk	trois	three
banque	bank	monde	world
banques	banks	pourquoi	why
hier	yesterday	aujourd'hui	today
janvier	January	sans	without
jours	days	lui	him
votre	your	mais	but
enfant	children	suis	am
trop	too	seulement	only
peut	cannot	ceintures	seat
bravo	!	ceintures	belt

Figure 3. High mutual information word pairs:

Obtaining the correct word order

We now turn to the second major problem in the generation of a translated text from a source text: the problem of re-arranging the translated words in a sensible order to form a true sentence in the target language. A simple way of doing this is to evaluate the term *P(S)* by using a language model, much in the same way as is done in speech recognition. There, it is assumed that $P(S) = P(s_1, s_2, ..., s_n)$ or,

$$P(S) = P(s_1) \bullet P(s_2|s_1) \bullet P(s_3|s_1,s_2) \bullet ....$$

or, in general,

$$P(S) = \prod_{i=1}^{m} P(s_i|s_1, ..., s_{m-1})$$

A convenient approximation is to stop after a number of terms, since it is assumed that the past history contributes little to the conditioning. In practice a model based on groups of three words (a tri-gram model) is the limit of what is practically feasible, and for which reliable estimates of the probabilities can be obtained. The probabilities are already difficult to compute for tri-grams, since $m = 3$ (past histories of only 2 words are used), and $L = 5000$ (a vocabulary of only 5000 words), means that $L^i = 1.25^{11}$, or over a billion, parameters are needed for the model!

A text could now be re-constructed using such a language model in the following way: Assume that the *n* words of the source sentence are known, but not the order they occurred in. From all the possible *n*! sentences which could have occurred, select the one for which the language model gives the maximum probability. For example, given the following words:

```
as,as,give,me,please,possible,response,soon,your
```

we would like to obtain the sequence:

```
please give me your response as  soon  as  possible
```

When this procedure was tested on a small test corpus [7], the following results were obtained:

1. 63% sentences were recovered exactly,
2. 21% sentences preserved the meaning of the original,
3. 16% were wrongly reconstructed.

This leads to the conclusion that the trigram model is not a wonderful model of the way sentence orders *differ* between the two languages. This is hardly surprising, since it contains only information about English word orders, and has no knowledge about French

WHICH		QUI	
qui	0.380	who	0.188
que	0.177	which	0.161
dont	0.082	that	0.084
de	0.060	'.'	0.038
d'	0.035	to	0.032
laquelle	0.031	of	0.027
ou	0.027	the	0.026
et	0.022	what	0.018

THE		HEAR	
le	0.610	bravo	0.992
la	0.178	entendre	0.005
1'	0.083	entendu	0.002
les	0.023	entends	0.001
ce	0.013		
il	0.012		
de	0.009		
a	0.007		
que	0.007		

Figure 4. Some words from a probabilistic glossary:

at all. What is required is a model of the *distortion* which a sentence undergoes when it is translated from the source language into the target language. Distortion will, of course, be different for different language pairs. The results of a proper model of distortion, described in [7] are shown in Figure 5 on page 11.

Conclusion

A methodology has been proposed which could potentially perform some of the tasks required in machine translation, and some results presented which suggest that the method is capable of some development. In summary, it is to:

1. Generate progressively more sophisticated models of glossary, and word order,
2. Train the models on parallel texts,
3. Do translation as a task in decoding.

The advantages of this methodology are:

1. Computers now fast enough and big enough to do the required calculations.
2. Large corpora of machine-readable texts are now available.
3. Advanced techniques are available from Speech Recognition.
4. A *methodology* can be established for creating a translator between any two languages, in either direction.

Some of the disadvantages of this approach are, of course, that a number of theoretical problems are still to be addressed. It is also clear that any eventual solution will be very compute-intensive. The same situation is true in the use of these methods in Speech Recognition, but there special purpose hardware is used to accelerate the computation, and it is certainly possible that a similar approach will be required to do, say, real-time translation. It should also be observed that training the models by using the forward-backward algorithm is also a costly process. However, the training is principally a one-time cost, setting up a working model. On the other hand, simply decoding is typically rather faster, which is fortunate. Finally, the methodology should be adaptable to other language pairs, special domains, and particular vocabularies, all of which are desirable properties.

Exact (5%)

T: Ces amendements sont certainement necessaire.
S: These ammendments are certainly necessary.
H: These amendements are certainly necessary.

Alternate(25%)

T: C'est pourtant tres simple.
S: It is still very simple.
H: Yet it is very simple.

Different (18%)

T: J'ai recu cette demande en effet.
S: I have received this request in effect.
H: Such a request was made.

Wrong (15%)

T: Permettez que je donne un example a la Chambre.
S: Let me give an example in the Mouse.
H: Let me give the House one example.

Ungrammatical (37%)

T: Vous avez besoin de toute l'aide disponible.
S: You need of the the whole benefits available.
H: You need all the help you can get.

Figure 5. **Some preliminary translation results**.: A statistical model or translation derived from the Hansard corpus was used on a test set of short sentences (10 words or less in length). The form of the sentence prefixed by T: is the observed target language sentence. The form preceded by S: is the output of the decoder. The form preceded by H: is the actual sentence in the corpus which corresponds to the observed sentence. The current model translates from French into English, although a model which translates the other way could of course be constructed.

References

1. W.Weaver, *Translation,* in *Machine. Translation of Languages,* ed. W.N.Locke and A.D.Booth, Wiley, New York, 1955.

2. J.Lyons, *Chomsky,* Fontana, 1970.

3. R.A.Sharman, *Observational Evidence for a Statistical Model of Language,* IBM UKSC Report 205, Winchester, 1989.

4. R.A.Sharman, *Evaluating a grammar as a language model for Speech Recognition,* Proc. Eur. Sig. Proc. Conf., Barcelona, 1990

5. Hansard, *Official Proceedings of the House of Commons of Canada,* Canadian Government Printing Bureau, Hull Quebec Canada, 1974-78

6. P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P.Roosin *A Statistical Approach to French/English Translation,* IBM Research Report, Yorktown Heights, NY 1987

7. P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P.Roosin *A Statistical Approach to Machine Translation,* IBM Research Report #14773, Yorktown Heights, NY 1989

8. R.G.Gallagher, *Information Theory and Reliable Communication,* Wiley, New York, 1968.

9. L.R.Bahl, F.Jelinek, and R.L.Mercer *A Maximum likelihood approach to continuous speech recognition* IEEE Trans. Patt. An. & Mach. Int. vol PAMI-5, no 2, March 1983