SPEECH TRANSLATION: COUPLING OF RECOGNITION AND TRANSLATION

Hermann Ney

Lehrstuhl für Informatik VI,
RWTH Aachen – University of Technology,
D-52056 Aachen, Germany

ABSTRACT

In speech translation, we are faced with the problem of how to couple the speech recognition process and the translation process. Starting from the Bayes decision rule for speech translation, we analyze how the interaction between the recognition process and the translation process can be modelled. In the light of this decision rule, we discuss the already existing approaches to speech translation. None of the existing approaches seems to have addressed this direct interaction. We suggest two new methods, the local averaging approximation and the monotone alignments.

1. INTRODUCTION

Recently, the statistical approach to text translation has been adopted by a number of research groups [1, 2, 3, 6, 8, 13]. In addition, the translation approach has been extended by using speech input rather than text input [8, 10]. In this case of speech translation, however, there are two processes to be considered: speech recognition of the source language and translation into the target language.

Therefore the question arises: What is the correct decision rule for speech translation? Often, speech translation is simply implemented as a sequential operation by first performing speech recognition and then translation of the recognized text. But then, there is the question of how recognition errors should be handled by the translation process. Ultimately, the problem boils down to the question of how to arrive at a suitable interaction of the recognition process and the translation process. In this paper, we will attempt to derive a suitable decision rule for speech translation and to present suitable implementations.

2. BAYES DECISION RULE

2.1. Review: Text Input

To pave the ground, we review the Bayes decision rule for text translation. 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 $e_1^I = e_1...e_i...e_I$. In this paper, the term *word* always refers to a *full-form* word. Among all possible target strings, we will choose the string with the highest probability which is given by Bayes' decision rule [3]:

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

$Pr(e_1^I)$ 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, i. e. the generation of the output sentence in the target language. The overall architecture of the statistical translation approach is summarized in Fig. 1. Here, we have assumed suitable *transformation* steps [2, 6] and a decomposition of the string translation model into *alignment models* and *lexical models* (see Section 3.1).

The notational convention will be as follows. We use the symbol $Pr(.)$ to denote general probability distributions with (nearly) no specific assumptions. In contrast, for model-based probability distributions, we use the generic symbol $p(.)$.

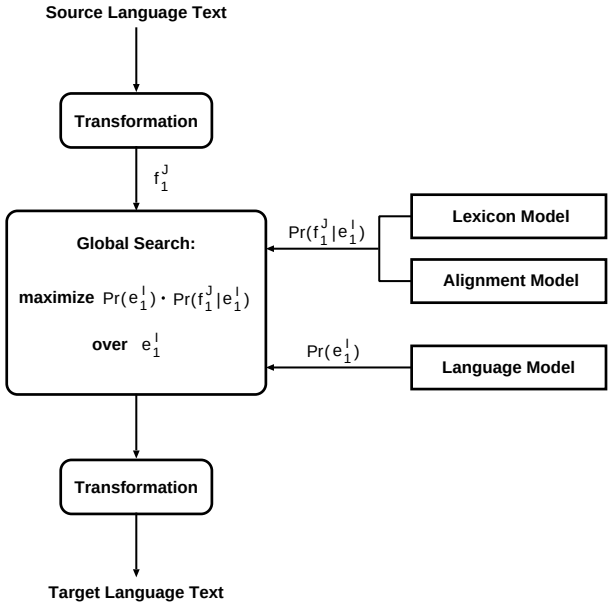

Figure 1. Bayes decision rule for text translation.

2.2. Speech Input

Thus far we have assumed written input, i.e. perfect input with no errors. When trying to apply the same translation concept to spoken input, we are faced with the additional complication of speech recognition errors. So the question comes up of how to integrate the probabilities of the speech recognition process into the translation process. Although there have been activities in speech translation at several

places [1, 8, 10], there has been no work on this question of recognition/translation integration.

Considering the problem of speech input rather than text input for translation, we can distinguish three levels, namely the acoustic vectors $x_1^T = x_1...x_t...x_T$ over time $t = 1...T$, the source words f_1^J and the target words e_1^I:

$$x_1^T \rightarrow f_1^J \rightarrow e_1^I$$

From a strict point of view, the source words f_1^J are not of direct interest for the speech translation task. Mathematically, this is captured by introducing the possible source word strings f_1^J as hidden variables into the Bayes decision rule:

$$\begin{aligned}
&\arg\max_{e_1^I} Pr(e_1^I|x_1^T) \;=\\
&= \arg\max_{e_1^I} \left\{ Pr(e_1^I) \cdot Pr(x_1^T|e_1^I) \right\}\\
&= \arg\max_{e_1^I} \left\{ Pr(e_1^I) \cdot \sum_{f_1^J} Pr(f_1^J, x_1^T|e_1^I) \right\}\\
&= \arg\max_{e_1^I} \left\{ Pr(e_1^I) \cdot \sum_{f_1^J} Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J, e_1^I) \right\}\\
&= \arg\max_{e_1^I} \left\{ Pr(e_1^I) \cdot \sum_{f_1^J} Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J) \right\}\\
&\cong \arg\max_{e_1^I} \left\{ Pr(e_1^I) \cdot \max_{f_1^J} \left\{ Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J) \right\} \right\}
\end{aligned}$$

Here, we have made no special modelling assumption, apart from the reasonable assumption that

$$Pr(x_1^T|f_1^J, e_1^I) = Pr(x_1^T|f_1^J) \;,$$

i. e. the target string e_1^I does not help to predict the acoustic vectors (in the source language) *if* the source string f_1^J is given. In addition, in the last equation, we have used the maximum approximation. Only in that special case of speech translation, at least from a strict point of view, there is the notion of a 'recognized' source word sequence f_1^J. However, this word sequence is very much determined by the combination of the language model $Pr(e_1^I)$ of the target language and the translation model $Pr(f_1^J|e_1^I)$. In contrast, in recognition, there would be only the language model $Pr(f_1^J)$.

It is instructive to re-interpret already existing approaches for handling speech input in a translation task in the light of the Bayes decision rule for speech translation, even if these approaches are not based on stochastic modelling. The key issue in all these approaches is the question of how the requirement of having both a well-formed source sentence f_1^J and a well-formed target sentence e_1^I at the same time is satisfied. From the statistical point of view, this question is captured by finding suitable models for the *joint* probability $Pr(f_1^J, e_1^I) = Pr(e_1^I) \cdot Pr(f_1^J|e_1^I)$.

From the decision rule, it is clear that the translation process will have an effect on the recognition process only if the target language model $Pr(e_1^I)$ is sufficiently strong or, to be more exact, if its strength is comparable to that of the source language model $Pr(f_1^J)$. We mention the following approaches:

- In many systems, the method of n-best lists is used. The recognizer produces a list of n best source sentences, and the translation system works as a filter that selects one out of the n sentences using some suitable criterion. This joint generation and filtering process can be viewed as a crude approximation of the joint probability $Pr(f_1^J, e_1^I)$.
- When using finite-state methodology rather than a fully stochastic approach, the probability $Pr(f_1^J, e_1^I)$ is modelled by the finite-state network of the corresponding transducer, which is typically refined by domain and range restrictions [8].
- In the extreme case, we might be only interested in the *meaning* of the target translation. Such an approach was used in [5] for the Verbmobil task. In Bayes decision rule, this case is captured by putting most emphasis on a *semantically* constrained language model $Pr(e_1^I)$.

However, it is clear that none of these approaches has fully covered the recognition-translation interaction from a statistical point of view.

3. ALIGNMENT AND LEXICON MODELS

To convert the Bayes decision rule derived above into a practical algorithm, we have to introduce specific modelling assumptions.

3.1. 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. To this purpose, alignment models have been introduced [3]. The alignment model used here [6] is similar to the concept of Hidden Markov models (HMM) in speech recognition. The alignment is a mapping $j \rightarrow i = a_j$ from source position j to target position $i = a_j$. Later, we will limit ourselves to so-called monotone HMM alignments as shown in Fig. 2.

Denoting the alignment probabilities by $p(a_j|a_{j-1}, I)$ and the lexicon probability by $p(f_j|e_i)$, we re-write the string translation probability:

$$Pr(f_1^J|e_1^I) \;=\; \sum_{a_1^J} \prod_j \left[p(a_j|a_{j-1}, I) \cdot p(f_j|e_{a_j}) \right]$$

3.2. Speech Input

To simplify the Bayes decision rule for speech translation, we will consider two modelling assumptions:

- Acoustic modelling:
  We assume that the speech recognizer produces a word graph as output. Each arc of the word graph represents a word hypothesis f_j which covers the portion x_j of the acoustic vectors (slightly abusing notation). The

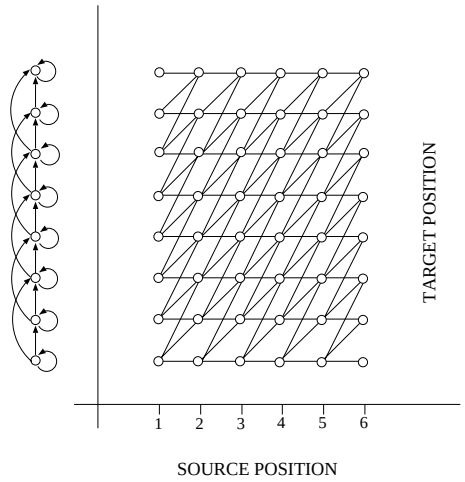

Figure 2. Illustration of monotone HMM alignments.

acoustic probabilities provided by the speech recognizer are denoted by $p(x_j|f_j)$. Thus we have:

$$Pr(x_1^T|f_1^J) = \prod_{j=1}^{J} p(x_j|f_j)$$

This assumption is without serious loss of generality.

- Lexicon modelling:
  When presenting the statistical approach to translation, the tacit assumption had been that the source sentence f_1^J was well formed. However, for speech input, this assumption is no more valid. Therefore, to take into account the requirement of 'well-formedness', we use a more complex translation model by including the dependence on the predecessor word:

$$p(f_j|f_{j-1}, e_{a_j}) \quad \text{in lieu of} \quad p(f_j|e_{a_j})$$

$$Pr(f_1^J|e_1^I) = \sum_{a_1^J} \prod_j \left[p(a_j|a_{j-1}, I) \cdot p(f_j|f_{j-1}, e_{a_j})\right]$$

For the sake of simplicity, here we do not go beyond the bigram dependence.

4. IMPLEMENTATIONS

To reduce the computational complexity, we will present two methods in detail, namely the local averaging approximation and the monotone alignments.

4.1. Local Averaging Approximation

Using the above two modelling assumptions, we re-write the term $Pr(x_1^T|e_1^I)$ in the Bayes decision rule:

$$\sum_{f_1^J} Pr(f_1^J|e_1^I) \cdot Pr(x_1^T|f_1^J) =$$

$$= \sum_{f_1^J} \sum_{a_1^J} \prod_j \left[p(a_j|a_{j-1}, I) \cdot p(f_j|f_{j-1}, e_{a_j}) \cdot p(x_j|f_j)\right]$$

$$\cong \sum_{a_1^J} \prod_j \left[p(a_j|a_{j-1}, I) \cdot p(x_j|e_{a_j})\right] \quad .$$

Here, in the last equation, we have used the approximation that the *averaging* process over f_1^J results in a *local* effect that can be captured by an auxiliary quantity $p(x_j|e)$. This quantity that directly links the acoustic vectors x_j with the target word e of the corresponding source word f. Before justifying and computing the auxiliary quantity $p(x_j|e)$, we study its impact on the system architecture. The ambiguity caused by the acoustic probabilities is captured by the auxiliary quantity $p(x_j|e)$ and replaces the lexicon probabilities $p(f_j|e)$ in the search process [2, 6, 11, 13]. This computation can be done *after* the recognition process and *before* the translation process. Therefore it does not affect the complexity of a full search strategy in translation. However, the new quantities $p(x_j|e)$ may be less focussed than the original lexicon probabilities $p(f|e)$ and thus may result in more search hypotheses when a beam search or some other non-full search method is applied.

The reason why the exact averaging over f_1^J cannot be carried out in a straightforward way is that, for each word hypothesis f_j, there is a dependence on the preceding word f_{j-1} and the subsequent word f_{j+1} (assuming a bigram language model). To avoid this complication, we consider the *single best recognized* sentence $\tilde{f}_1^J := \tilde{f}_1 ... \tilde{f}_j ... \tilde{f}_J$. Then the *joint* effect of the acoustic probabilities $p(x_j|f_j)$ and the bigram LM probabilities $p(f|f')$ can be approximately taken into account by defining:

$$p(x_j|e) := \sum_{f_j} p(x_j|f_j) \cdot p(f_j|e, \tilde{f}_1^J \setminus \tilde{f}_j)$$

$$= \frac{\sum_{f_j} p(x_j|f_j) \cdot p(f_j|\tilde{f}_{j-1}, e) \cdot p(\tilde{f}_{j+1}|f_j)}{\sum_{f_j} p(f_j|\tilde{f}_{j-1}, e) \cdot p(\tilde{f}_{j+1}|f_j)}$$

Note that this quantity is intended to serve as the direct link between the acoustic segment x_j and *any* target word hypothesis e. The approximation can be improved by making use of the so-called *forward-backward* probabilities of the word graph [12] so that not only the single best, but all 'good' word sequences with high enough probability scores can be taken into account. This is achieved by computing 'posterior' probabilities for each word hypothesis f_j that are based on the observations $x_1 ... x_{j-1} x_{j+1} ... x_J$ rather than the recognition hypothesis $\tilde{f}_1 ... \tilde{f}_{j-1} \tilde{f}_{j+1} ... \tilde{f}_J$.

4.2. Monotone Alignments

First, we review the dynamic programming (DP) search for monotone alignments as described in [6, 7]. The monotonicity requirement will be discussed later. In the maximum approximation (applied to the alignments a_1^J) and ignoring the length model $p(I|J)$, we re-write the search criterion for a bigram language model $p(e_i|e_{i-1})$:

$$\arg\max_{e_1^I} Pr(e_1^I|f_1^J) =$$

$$= \arg\max_{e_1^I} \prod_j \left[p(a_j|a_{j-1}, I) \cdot p(e_{a_j}|e_{a_j-1}) \cdot p(f_j|e_{a_j})\right]$$

$$= \arg\max_{e_1^J, \delta_1^J} \prod_j \left[p(\delta_j) \cdot p_{\delta_j}(e_j|e_{j-1}) \cdot p(f_j|e_j)\right] \quad .$$

For the last equation, we have re-formulated the search criterion as described in [4, 6]. Stretching notation, we have

switched from the sequence e_1^I along the target positions to a sequence e_1^J along the source positions using the jump width $\delta_j := a_j - a_{j-1}$ and suitable defined alignment quantities $p(\delta_j)$ and LM quantities $p_{\delta_j}(e_j|e_{j-1})$.

For the above criterion, there is a closed-form solution by the dynamic programming (DP) recursion (with the auxiliary function $Q(j,e)$):

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

For full search, the computational complexity of this recursion is $J \cdot E^2$ (E =vocabulary size of the target language vocabulary), which can be reduced by beam search and accelerated LM recombination [7].

Now, we consider speech input. Again in the maximum approximation (applied to the alignments a_1^J and the source strings f_1^J), we re-write the search criterion:

$$\begin{aligned}
&\arg\max_{e_1^I} Pr(e_1^I|x_1^T) \; = \\
&= \quad \arg\max_{e_1^I} \Big\{ Pr(e_1^I) \; \cdot \max_{f_1^J, a_1^J} \prod_j \Big[p(a_j|a_{j-1}, I) \cdot \\
&\qquad\qquad \cdot p(f_j|f_{j-1}, e_{a_j}) \cdot p(x_j|f_j) \Big] \Big\} \\
&= \quad \arg\max_{e_1^J} \max_{f_1^J, \delta_1^J} \prod_j \big[p(\delta_j) \cdot p_{\delta_j}(e_j|e_{j-1}) \; \cdot \\
&\qquad\qquad \cdot p(f_j|f_{j-1}, e_j) \cdot p(x_j|f_j) \big]
\end{aligned}$$

Here, we have used the same re-formulation as for text input and thus obtain the DP recursion:

$$\begin{aligned}
Q(j,e,f) \; = \; p(x_j|f) \; \cdot \max_{f'} \Big\{ & p(f|f',e) \cdot \\
& \cdot \max_{\delta,e'} \; \{p(\delta) \cdot p_\delta(e|e') \cdot Q(j-1,e',f')\} \Big\}
\end{aligned}$$

For full search, the computational complexity of this recursion is $J \cdot E^2 \cdot F^2$ (F =vocabulary size of the source language vocabulary). In addition to beam search and accelerated LM recombination, the complexity can be reduced by considering only *promising* word pairs (e,f).

The monotonicity requirement can be satisfied by suitably re-ordering the words of either the source or the target language [6]. For speech input, the preferred language is the target language due to the possible errors in recognition and prosodic segmentation. Therefore, re-ordering the target words, maybe in connection with some grammar-based language model, could then be performed as part of the search strategy [9]. As an additional advantage, the monotone DP search could be directly applied to the word graph as provided by the speech recognizer.

Acknowledgment

The problem studied in this paper was the topic of many discussions with E. Vidal when the author was a visiting scientist at Universidad Politecnica de Valencia, Valencia, Spain, in 1996.

This work has been partly carried out in the framework of the Verbmobil project (01 IV 701 T4) funded by the German Federal Ministry of Education, Science, Research and Technology and of the Eutrans project (ESPRIT 30268) funded by the European Community.

REFERENCES

[1] H. Alshawi, F. Xiang: English-to-Mandarin Speech Translation with Head Transducers. *Spoken Language Translation Workshop (SLT-97)*, Madrid, Spain, pp. 54-60, July 1997.

[2] A. L. Berger, P. F. Brown et al.: The Candide System for Machine Translation. *ARPA Human Language Technology Workshop*, Plainsboro, NJ, Morgan Kaufmann Publishers, San Mateo, CA, pp. 152-157, March 1994.

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

[4] F. J. Och, C. Tillmann: Unpublished Results. Computer Science Department, RWTH Aachen, Germany, July 1998.

[5] N. Reithinger, E. Maier: Using Statistical Dialogue Act Processing in Verbmobil. *33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA, pp. 116-121, June 1995.

[6] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga: A DP-based Search Using Monotone Alignments in Statistical Translation. *35th Annual Conf. of the Association for Computational Linguistics*, Madrid, Spain, pp. 289-296, July 1997.

[7] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, H. Sawaf: Accelerated DP Based Search for Statistical Translation. *Fifth European Conf. on Speech Communication and Technology*, pp. 2667-2670, Rhodos, Greece, Sep. 1997.

[8] E. Vidal: Finite-State Speech-to-Speech Translation. *Int. Conf. on Acoustic Speech and Signal Processing*, Munich, Germany, pp. 111-114, April 1997.

[9] J. M. Vilar, E. Vidal, J. C. Amengual: Learning Extended Finite State Models for Language Translation. *12th European Conf. on Artificial Intelligence*, Budapest, Hungary, 1996.

[10] A. Lavie, L. Levin, A. Waibel, D. Gates, M. Gavalda, L. Mayfield: JANUS: Multi-lingual translation of spontaneous speech in a limited domain. *2nd Conf. of the Association for Machine Translation in the Americas* pp. 252-255, Montreal, Quebec, Oct. 1995.

[11] Y.-Y. Wang, A. Waibel: Decoding Algorithm in Statistical Translation. *35th Annual Conf. of the Association for Computational Linguistics*, pp. 366-372, Madrid, Spain, July 1997.

[12] F. Wessel, W. Macherey, R. Schlüter: Using Probabilities as Confidence Measures. *Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 225-228, May 1998.

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