Towards an Automatic Dictation System for Translators : the TransTalk Project*

Marc Dymetman††, Julie Brousseau‡, George Foster†,
Pierre Isabelle†, Yves Normandin‡, Pierre Plamondon†

† Centre d'Innovation en Technologies de l'Information (CITI)
1575, Boul. Chomedey, Laval H7V 2X2, Quebec, Canada

‡ Centre de Recherche Informatique de Montreal (CRIM)
1801, McGill College, Montreal H3A 2NA, Quebec, Canada

September 16, 1994

Abstract

Professional translators often dictate their translations orally and have them typed afterwards. The TransTalk project aims at automating the second part of this process. Its originality as a dictation system lies in the fact that both the acoustic signal produced by the translator and the source text under translation are made available to the system. Probable translations of the source text can be predicted and these predictions used to help the speech recognition system in its lexical choices. We present the results of the first prototype, which show a marked improvement in the performance of the speech recognition task when translation predictions are taken into account.

1 Introduction

The integration of machine translation and speech technology is currently the focus of major projects in several countries [5, 6, 9]. Usually, the aim of these efforts is some type of speech-to-speech translation, where speech recognition, machine translation and speech synthesis are performed sequentially. However, both speech recognition and machine translation are tasks that can at present

---

[*] Published in proceedings of ICSLP 94

[†] Present address: Rank Xerox Research Centre, 6 chemin de Maupertuis, 38240 Meylan, France.

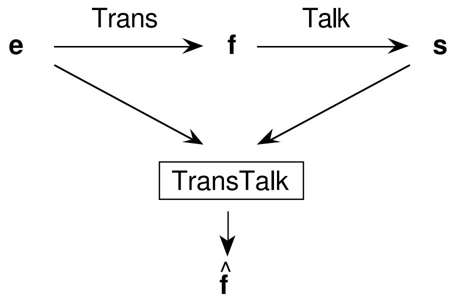

Figure 1: TransTalk's underlying model. Starting from an English sentence $\mathbf{e}$, the translator mentally formulates its French translation $\mathbf{f}$, then produces its acoustic rendering $\mathbf{s}$. The system's aim is to find $\hat{\mathbf{f}} = \mathrm{argmax}_f\, p(\mathbf{f}\,|\,\mathbf{e},\mathbf{s})$, or equivalently, from Bayes's formula, $\hat{\mathbf{f}} = \mathrm{argmax}_f\, p(\mathbf{s}\,|\,\mathbf{e},\mathbf{f})\cdot p(\mathbf{f}\,|\,\mathbf{e})$. By neglecting the influence of $\mathbf{e}$ on $\mathbf{s}$ once $\mathbf{f}$ is known, we can take $\hat{\mathbf{f}} = \mathrm{argmax}_f\, p(\mathbf{s}\,|\,\mathbf{f})\cdot p(\mathbf{f}\,|\,\mathbf{e})$.

be reliably accomplished only under stringent lexical, syntactic and semantic restrictions, and consequently developers of speech-to-speech translation systems need to find application domains for which narrow sub-languages can be naturally defined.

In the TransTalk project, we attempt to integrate speech recognition and machine translation in a way which, instead of compounding the weaknesses of both technologies, makes maximal use of their complementary strengths. We do not try to replace the human translator by a machine (a hopeless endeavor, in general), but undertake instead the more realistic task of providing a *dictation tool* to the translator. Our aim is to use machine translation to make probabilistic predictions of the possible target language verbalizations freely produced by the translator, and to use these predictions to reduce the difficulty of the speech recognition task to such an extent that complete recognition of the translator's utterances can be achieved.[1]

For example, suppose that, in the case of English-to-French translation, the translator decides to render the sentence “what splendid horses you have” as “tes chevaux sont vraiment magnifiques”. A speech recognition system without access to the source text might have difficulty distinguishing *chevaux* (horses) from the acoustically close, and contextually more likely, *cheveux* (hair). On the other hand, the presence in the English source of the word *horses* serves as a strong indicator that the correct choice should be *chevaux*, and it is on such knowledge of probable translations that TransTalk attempts to capitalize.

Conceptually, the main difference between a conventional “noisy channel” speech recognition system for French and TransTalk is that, instead of maxi-

---

[1] The idea was independently advanced by us [3] and by researchers at the IBM Thomas J. Watson Research Center [1].

mizing in $\mathbf{f}$ the product $p(\mathbf{s} \,|\, \mathbf{f}) \cdot p(\mathbf{f})$ of an "acoustic model" and a "language model" for French (where $\mathbf{s}$ stands for the acoustic signal and $\mathbf{f}$ for the French sentence), we maximize the product $p(\mathbf{s} \,|\, \mathbf{f}) \cdot p(\mathbf{f} \,|\, \mathbf{e})$ of an acoustic model and a "translation model" from English to French (where $\mathbf{e}$ stands for the English sentence under translation). See figure 1.

We have implemented a prototype version of TransTalk that operates in an isolated-word dictation mode over a vocabulary of 20,000 French word forms. It is specialized for the domain of Canadian Parliamentary debates, which are transcribed in bilingual form in the Canadian Hansard corpus. Two years of Hansard transcripts (approximately 10M French words and 10M English words) were used as training data for the translation model.

2 Acoustic model

We use an HMM based on context-independent phone models to describe $p(\mathbf{s} \,|\, \mathbf{f})$. The TransTalk vocabulary is represented with a set of 47 phonemes including 20 vowels and 27 consonnants. The base pronunciations were obtained using a set of grapheme-to-phoneme rules which take into account phonetic particularities found in the French spoken in Quebec such as assibilation and vowel laxing.

Recognition is performed with an n-best search of a compressed phonetic graph representing the entire 20,000 word vocabulary [7]. This graph is such that no two paths produce the same phone sequence and every path corresponds to a valid phonetic representation in the dictionary. A given path will therefore correspond to all lexicon entries sharing the same phonetics. The search yields a list containing the 20 most acoustically probable words for each (isolated) acoustic token.

3 Translation model

The aim of the translation model is to describe $p(\mathbf{f} \,|\, \mathbf{e})$, the probability that a translator will produce a French translation $\mathbf{f}$ for an English sentence $\mathbf{e}$.

3.1 Modelling Approaches

There are at least two distinct approaches to modelling this distribution. In [2], Brown et al. expand it as the product $p(\mathbf{f}) \cdot p(\mathbf{e} \,|\, \mathbf{f})$, to which it is proportional under maximization over $\mathbf{f}$. The main advantage of this arrangement is that it provides for a division of labour in which $p(\mathbf{f})$ is responsible for the well-formedness of $\mathbf{f}$, and $p(\mathbf{e} \,|\, \mathbf{f})$ for ensuring that $\mathbf{e}$ and $\mathbf{f}$ are acceptable translations without having to be unduly preoccupied with the internal structure of either. Although this is a powerful technique, it has one drawback that makes it unsuitable for our purposes: it does not easily lend itself to efficient searches over large sets of French sentence candidates.

Because of this, we have chosen to model $p(\mathbf{f}\,|\,\mathbf{e})$ more directly as a family of parameterized Markov language models $p_{\lambda(e)}(\mathbf{f})$, where each $\mathbf{e}$ specifies a parameter vector λ, not necessarily uniquely. This approach presents the challenge of incorporating information from $\mathbf{e}$ in a way that does not interfere with the language model's knowledge of the structure of French—particularly for language models that are accurate to begin with. In the work reported here we have largely avoided this difficulty by using a fairly weak language model; our aim is mainly to investigate to what exent the performance of such a model can be improved without substantially increasing its low run-time cost.

3.2 Derivation

The translation model is based on a standard tri-class language model conditioned on $\mathbf{e}$. The first key assumption we make is that the sequence $\mathbf{c}$ of word classes for $\mathbf{f}$ is independent of $\mathbf{e}$, which allows us to write:

$$p(\mathbf{f}\,|\,\mathbf{e}) = \sum_{\mathbf{c}} p(\mathbf{c}) \cdot p(\mathbf{f}\,|\,\mathbf{c},\mathbf{e}) \tag{1}$$

This approximation is motivated by the intuition that $\mathbf{e}$ will be most informative about the actual words in $\mathbf{f}$, and only weakly informative about gross syntactic structure of the sort that $\mathbf{c}$ captures. Because it is most valid when $\mathbf{c}$ consists of broad classifications[2] we use a minimal set of 15 classes which correspond to the major grammatical categories (noun, verb, etc).

To incorporate translation information, we suppose, following Brown et al. [2], that $\mathbf{f}$ and $\mathbf{e}$ are related via an *alignment* (see figure 2) in which each French word is connected to either a single English word in $\mathbf{e}$ or none at all. An alignment can be represented as a vector $\mathbf{a}$ of length $|\mathbf{f}|$ which contains, for each French word, the position in $\mathbf{e}$ of the English word to which it connects, or zero if it is not connected. We assume that all $A_{\mathbf{f},\mathbf{e}}$ possible alignments are equally likely, with probability $p(\mathbf{a}\,|\,\mathbf{c},\mathbf{e}) = 1/A_{\mathbf{f},\mathbf{e}}$, so we have:[3].

$$p(\mathbf{f}\,|\,\mathbf{c},\mathbf{e}) = \sum_{\mathbf{a}} \frac{1}{A_{\mathbf{f},\mathbf{e}}} \cdot p(\mathbf{f}\,|\,\mathbf{a},\mathbf{c},\mathbf{e}) \tag{2}$$

This is a rough approximation which runs contrary to our knowledge that some alignments—such as those in which all French words connect to a single English word, or those in which French verbs connect only to English prepositions—will be much less likely than others. Its purpose is simplification, and we justify it on the grounds that a reasonable model for $p(\mathbf{f}\,|\,\mathbf{a},\mathbf{c},\mathbf{e})$ will minimize the contribution from (most) poor alignments in any case.

---

[2] This assumption becomes increasingly untenable for finer classification schemes; in the limit when classes are identical to words, the model collapses into a pure tri-gram with no translation component whatsoever.

[3] Where $A_{\mathbf{f},\mathbf{e}} = (|\mathbf{e}| + 1)^{|\mathbf{f}|}$

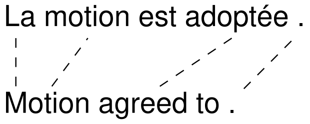

Figure 2: An example of an alignment, one of 5^5 which are possible for this sentence pair.

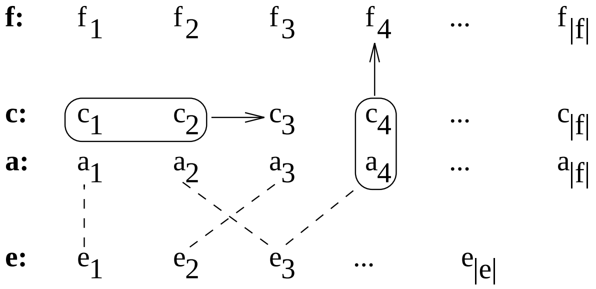

Figure 3: The structure of the Markov source underlying the translation model. First, **c** is established by choosing each class based on the previous two with probability given by the appropriate contextual parameter. Next **a** is established by picking a position in **e** at random for each position in **c**. Finally, **f** is generated by choosing each word based on its class and its English partner, with probability given by the appropriate bi-lexical parameter.

The final step is to assume that the words in $\mathbf{f}$ are conditionally independent given $\mathbf{a}$, $\mathbf{c}$, and $\mathbf{e}$, and furthermore that each word depends only on its class and the English word to which it connects in the alignment:

$$p(\mathbf{f} \mid \mathbf{a}, \mathbf{c}, \mathbf{e}) = \prod_{i=1}^{|\mathbf{f}|} p(f_i \mid c_i, e_{a_i}) \tag{3}$$

Our complete model is a Markov source (see figure 3) which depends on two sets of parameters: *contextual* parameters of the form $p(c_i \mid c_{i-2}, c_{i-1})$, which predict a class from its two predecessors; and *bi-lexical* parameters of the form $p(f \mid c, e)$, which predict a French word from its class and its English partner.

It is possible to rearrange the straightforward combination of equations 1, 2, and 3 in a way which permits more efficient calculations. The key observation is that the sum over all alignments can be reorganized into a product of sums over English words. The result is the equation

$$p(\mathbf{f} \mid \mathbf{e}) = \sum_{\mathbf{c}} \prod_{i=1}^{|\mathbf{f}|} p(c_i \mid c_{i-2}, c_{i-1}) p(f_i \mid c_i, \mathbf{e}) \tag{4}$$

where $p(f_i \mid c_i, \mathbf{e}) = \sum_{j=0}^{|\mathbf{e}|} p(f_i \mid c_i, e_j)/(|\mathbf{e}| + 1)$. From this it should be obvious that our translation model is nothing more that a standard tri-class model in which the lexical parameters $p(f \mid c)$ have been replaced by $p(f \mid c, \mathbf{e})$.

3.3 Parameter Estimation

The two families of parameters in the translation model were estimated separately. Contextual parameters were estimated as part of a pure tri-class language model for French, which was trained on the French half of our bilingual corpus via the EM algorithm, using a dictionary to identify valid classes for each word.

Bi-lexical parameters were estimated as part of a simplified translation model in which contextual information was assumed to be explicit:

$$p(\mathbf{f}, \mathbf{c} \mid \mathbf{e}) = \frac{1}{A_{\mathbf{f},\mathbf{e}}} \sum_{\mathbf{a}} \prod_{i=1}^{|\mathbf{f}|} p(f_i, c_i | e_{a_i}) \tag{5}$$

To train this model, we first aligned the training corpus to the sentence level using the method described in [8]. To improve the quality of our training data, we filtered out alignments which involved more than one sentence in either language as well as those which contained more than 40 tokens in either language—this reduced the size of the training set by approximately 20%, to about 8M tokens in each language. Next, we used the pure language model to tag each word in the French part of the reduced corpus with its most likely class. Finally,

f	$p(f \mid c, e)$
gouvernement	0.7363
m.	0.0227
monsieur	0.0134
prsident	0.0109
canada	0.0081
faon	0.0033
mesure	0.0024
part	0.0023
ministre	0.0023
dcision	0.0022

Figure 4: A sample of TransTalk's bi-lexical parameters. These are the ten most probable French words, given the class NOUN and the English word *government*.

we used the EM algorithm to estimate parameters $p(f, c \mid e)$ from the aligned, tagged corpus. These were transformed into bi-lexical parameters as follows:

$$p(f \mid c, e) = \frac{p(f, c \mid e)}{\sum_f p(f, c \mid e)} \tag{6}$$

Figure 4 shows a sample of the results.

Because many valid bi-lexical combinations do not occur in our training corpus, it was necessary to smooth the bi-lexical parameters. Rather than modifying the empirical distribution $p(f \mid c, e)$ directly, we chose to dynamically smooth the more robust quantity $p(f \mid c, \mathbf{e})$ involved in calculations based on equation 4. We experimented with three simple methods of combining this with the less precise but more reliable lexical parameters $p(f \mid c)$ from the pure language model: linear interpolation; using the maximum of $p(f \mid c, \mathbf{e})$ and $p(f \mid c)$; and using $p(f \mid c)$ iff $\max_e p(f \mid c, e)/p(f \mid c, \mathbf{e})$ did not exceed some threshold. The rationale for the second method is that we expect higher probabilities to be more reliably estimated on average than lower ones. The third method is intended to reject translation information when there is no English word that is strongly associated with the current French word. Because the last two methods result in unnormalized distribitions, they can be compared only in terms of recognition performance and not by means of the perplexity measure (see section 5).

4 Search

The aim of the search component is to find an approximation to the sentence $\hat{\mathbf{f}}$ that maximizes the product of acoustic and translation scores $p(\mathbf{s} \mid \mathbf{f}) \cdot p(\mathbf{f} \mid \mathbf{e})$. Our search algorithm is divided into two stages, both of which are suboptimal.

The first stage involves using the acoustic model to prune the list of word hypotheses for each acoustic token from 20,000 to some number n (currently

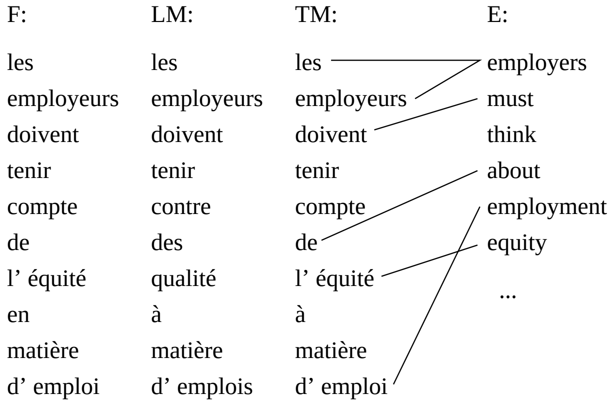

Figure 5: Comparison of language model (LM) and translation model (TM) results for a sentence pair (F,E) from the test corpus. (This pair has been truncated for space reasons.) Lines indicate salient parts of the most probable alignment between the output sentence and E. The presence of *equity* in the English source allowed the translation model to correctly choose *quit* instead of *qualit.*

20). Since this pruning is performed without reference to the translation model, there is no guarantee that $\hat{\mathbf{f}}$ is among the $n^{|\mathbf{f}|}$ sentence candidates retained.

The second stage is a Viterbi search through the remaining sentence candidates using the translation model. This permits us to find the pair $(\tilde{\mathbf{f}}, \tilde{\mathbf{c}})$ that maximizes the product $p(\mathbf{s}\,|\,\mathbf{f}) \cdot p(\mathbf{f}, \mathbf{c}\,|\,\mathbf{e})$ in time which is proportional to $nC^3|\mathbf{f}||\mathbf{e}|$, where C is the number of word classes in the translation model (currently 15). Given the coarse nature of our word classes, we feel that $\tilde{\mathbf{f}}$ is a reasonable approximation to $\hat{\mathbf{f}}$.

5 Results

We tested TransTalk on a small corpus of 50 French/English sentence pairs from the Hansard corpus which were not used as training data. The French sentences were all between 15 and 20 tokens in length (counting punctuation) and were selected so as not to contain words outside our 20,000 word vocabulary. They were dictated in isolated-word mode by two different speakers.

Figure 5 illustrates the results for a single sentence pair. Overall statistics are given in figure 6. The translation model yielded an average error-rate decrease of 24% over the pure language model. For errors which involved "content" words (eg, *action* for *section*) the decrease was 42%. The perplexity of the test corpus was reduced by more than half by the use of the translation model.

Model	Words Correct (/918)		Perplexity
	Speaker 1	Speaker 2	
pure language	686 (74.7%)	677 (73.8%)	385
interpolated (.85)	735 (80.1%)	734 (80.0%)	180
maximum	735 (80.1%)	732 (79.7%)	–
e-testing (.30)	742 (80.8%)	734 (80.0%)	–

Figure 6: Summary of TransTalk results. The first line contains statistics for the pure language model; the remaining lines contain statistics for the translation model with each of the three different smoothing methods described in section 3.3 (where .85 was the optimum the weighting factor for bi-lexical parameters, and .30 was the optimum confidence threshold).

6 Conclusions

Our initial results demonstrate that it is possible to cheaply and effectively make use of translation information for speech recognition. We feel that the simple approach described in this paper barely begins to tap the potential of the TransTalk idea, and we are currently investigating a number of ways of improving on it.

References

[1] Peter F. Brown, Stanley F. Chen, Stephen A. Della Pietra, Vincent Della Pietra, Andrew S. Kehler, and Robert L. Mercer. Automatic speech recognition in machine aided translation. *Computer Speech and Language*, 8:177–187, 1994.

[2] Peter F. Brown, Stephen A. Della Pietra, Vincent Della J. Pietra, and Robert L. Mercer. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–312, June 1993.

[3] Marc Dymetman, Steven Krauwer, and Herbert Ruessink. Robustness: A project proposal, October 1992. Unpublished EEC-Canada joint project proposal.

[4] Pierre Isabelle, Marc Dymetman, George Foster, Jean-Marc Jutras, Elliott Macklovitch, Franois Perrault, Xiaobo Ren, and Michel Simard. Translation analysis and translation automation. In *Proceedings of the Fifth Internation Conference on Theoretical and Methodological Issues in Machine Translation*, pages 201–217, July 1993.

[5] Martin Kay, Jean Mark Gawron, and Peter Norvig. *Verbmobil. A Translation System for Face-to-Face Dialog. Lecture Note No. 33.* Center for the Study of Language and Information, Stanford, 1994.