

Automatic Text Dictation in Computer-Assisted Translation

Shahram Khadivi, András Zolnay, and Hermann Ney

Lehrstuhl für Informatik VI - Computer Science Department
RWTH Aachen University, Aachen, Germany
{khadivi, zolnay, ney}@cs.rwth-aachen.de

Abstract

In this paper, we study the incorporation of statistical machine translation models to automatic speech recognition models in the framework of computer-assisted translation. The system is given a source language text to be translated and it shows the source text to the human translator to translate it orally. The system captures the user speech which is the dictation of the target language sentence. Since the system has simultaneous access to the source language text and the speech signal of the target language text, it is possible to improve the speech recognition accuracy by incorporating the statistical machine translation models. We show that statistical translation models have a high impact on improving the speech recognition results. Using these models, we achieve a relative word error rate reduction of 17%.

1. Introduction

Professional translators can translate a given text faster by dictation rather than directly typing the translation. Due to this fact, one desired feature of a computer-assisted translation system (CAT) is to provide an environment to accept the translators speech signal of the target language to speed up the translation process. In such a system, two sources of information are available to recognize the speech input; the target language speech and the given source language text. The target language speech is just a human-produced translation of the source language text. Machine translation models are used only to take into account the source text in order to increase the speech recognition accuracy. The overall schematic of automatic text dictation in computer-assisted translation is depicted in Figure 1.

The idea of incorporating statistical machine translation and speech recognition models was independently initiated about one decade ago by two groups: researchers at the IBM Thomas J. Watson Research Center [1], and researchers involved in the TransTalk project [2, 3].

In [1], the authors described the statistical speech recognition models and statistical translation models. Then, they proposed a method for combining those models, but they did not report any recognition or translation results. Instead, they just reported the perplexity reduction when the translation models were combined to recognition models.

In the TransTalk project [2, 3], the authors reported three different combination methods between translation and recognition models. The first method is capable only of isolated-word recognition. In the second method, the speech recognition system generates a list of the most probable word sequence hypotheses. Then the statistical translation models rescore them and select the best word sequence hypothesis. The idea behind the third method is the *dynamic vocabulary* for a speech

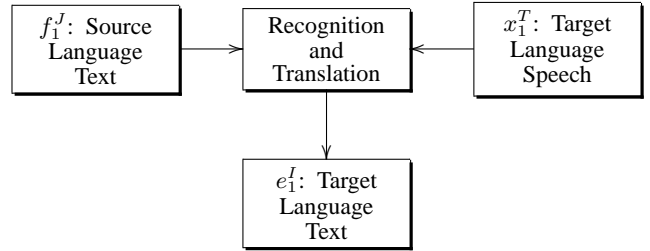


Figure 1: Schematic of automatic text dictation in computer-assisted translation

recognition system which translation models generate for each source language sentence. The best recognition results have been achieved with the second method, while the third method was faster. The authors have shown the promising results of combining the translation models to speech recognition models. However, they neither described the details of the utilized translation model nor studied the impact of different translation models.

In this paper, we describe an automatic text dictation system in the computer-assisted translation framework for translating English text to German text. Also, the incorporation of different state-of-the-art translation models to the speech recognition model will be investigated and analyzed.

In Section 2, we describe a general model for an automatic text dictation system in the computer-assisted translation framework. Section 3 explains the machine translation models. Section 4 describes the utilized speech recognition system, and Section 5 shows the experimental results.

2. Automatic Text Dictation Models in CAT

In a speech-enabled computer-assisted translation system, we are given a source language sentence $f_1^J = f_1 \dots f_j \dots f_J$, which is to be translated into a target language sentence $e_1^I = e_1 \dots e_i \dots e_I$, and an acoustic signal $x_1^T = x_1 \dots x_t \dots x_T$, which is the speech of the target language sentence. Among all possible target language sentences, we will choose the sentence with the highest probability:

$$\hat{e}_1^I = \operatorname{argmax}_{e_1^I} \{Pr(e_1^I | f_1^J, x_1^T)\} \quad (1)$$

$$= \operatorname{argmax}_{e_1^I} \{Pr(e_1^I, f_1^J, x_1^T)\} \quad (2)$$

$$= \operatorname{argmax}_{e_1^I} \{Pr(e_1^I, f_1^J) \cdot Pr(x_1^T | e_1^I)\} \quad (3)$$

$$= \operatorname{argmax}_{e_1^I} \{Pr(e_1^I) \cdot Pr(f_1^J | e_1^I) \cdot Pr(x_1^T | e_1^I)\} \quad (4)$$

Equation 2 is decomposed into Equation 3 by considering that there is no direct dependence between x_1^T and f_1^J . The decomposition into three knowledge sources in Equation 4 allows an independent modeling of the target language model $Pr(e_1^I)$, the translation model $Pr(f_1^J|e_1^I)$ and the acoustic model $Pr(x_1^T|e_1^I)$.

The target language model describes the well-formedness of the target language sentence. The translation model links the source language sentence to the target language sentence. The acoustic model links the acoustic signal to the target language sentence. The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. We have to maximize over all possible target language sentences.

Another approach for modeling the posterior probability $Pr(e_1^I|f_1^J, x_1^T)$ is direct modeling by the use of a log-linear model. The direct posterior probability is given by:

$$Pr(e_1^I|f_1^J, x_1^T) = \frac{\exp[\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J, x_1^T)]}{\sum_{e_1^I} \exp[\sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J, x_1^T)]} \quad (5)$$

This approach has been suggested by [4, 5] for a natural language understanding task, by [6] for automatic speech recognition, and by [7] for statistical machine translation. The time-consuming renormalization in Equation 5 is not needed in the search. Therefore we obtain the following decision rule:

$$\hat{e}_1^I = \operatorname{argmax}_{e_1^I} \left\{ \sum_{m=1}^M \lambda_m h_m(e_1^I, f_1^J, x_1^T) \right\}$$

Each of the terms $h_m(e_1^I, f_1^J, x_1^T)$ denotes one of the various models which are involved in the recognition process. Each individual model is weighted by its model scaling factor λ_m . As there is no direct dependence between f_1^J and x_1^T , the $h_m(e_1^I, f_1^J, x_1^T)$ is in one of these two forms: $h_m(e_1^I, x_1^T)$ and $h_m(e_1^I, f_1^J)$.

This approach is a generalization of Equation 4. The direct modeling has the advantage that additional models or feature functions can be easily integrated into the overall system. Based on Equation 4, the principal models which will contribute to the final system are the acoustic model, the language model, and the translation model(s). We may use one or more translation models in the final system. A set of possible translation models consists of *HMM*, *IBM-1*, *IBM-2*, *IBM-3*, *IBM-4*, *IBM-5*, and *Alignment Template* models, which will be described in Section 3. The details of utilized acoustic and language models will be explained in Section 4.

The model scaling factors λ_1^M in Equation 5 are trained according to the maximum entropy principle, e.g. using the GIS algorithm. Alternatively, one can train them with respect to the final recognition quality measured by the word error rate [8].

The development of an efficient search algorithm for integrating automatic speech recognition and statistical machine translation models is very complicated. Thus, in order to facilitate the implementation of the above log-linear model, we use the principle of N -best rescoring instead of implementing a new search algorithm. The N -best rescoring approach helps us to quickly examine many different dependencies and models for the combination of automatic speech recognition and statistical machine translation.

The recognition process is performed in two steps. In the first step, the baseline speech recognition system creates an N -best list of length N for every utterance x_1^T of the given corpus. In the second step, the translation models rescore every sentence pair (the entries in the N -best list with their corresponding

source sentence). For each utterance, the decision about the best recognized sentence is made according to the recognition and the translation models. Then the implementation approach is very similar to the second method explained in [3].

3. Translation Models

A key issue in modeling the string translation probability $Pr(f_1^J|e_1^I)$ is the question of how we define the correspondence between the words of the target sentence and the words of the source sentence. In typical cases, we can assume a sort of pairwise dependence by considering all word pairs (f_j, e_i) for a given sentence pair (f_1^J, e_1^I) . A family of such *alignment models* (IBM-1,...,IBM-5) was developed in [9]. Using the similar principles as in Hidden Markov models (HMM) for speech recognition, we re-write the translation probability by introducing the *hidden alignments* \mathcal{A} for each sentence pair (f_1^J, e_1^I) :

$$Pr(f_1^J|e_1^I) = \sum_{\mathcal{A}} Pr(f_1^J, \mathcal{A}|e_1^I)$$

IBM-1,2 and Hidden Markov Models. The first type of alignment models is virtually identical to HMMs and is based on a mapping $j \rightarrow i = a_j$, which assigns a source position j to a target position $i = a_j$. Using suitable modeling assumptions [9, 10], we can decompose the probability $Pr(f_1^J, \mathcal{A}|e_1^I)$ with $\mathcal{A} = a_1^J$:

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

with the length model $p(J|I)$, the alignment model $p(i|i', I, J)$ and the lexicon model $p(f_j|e_i)$. The alignment models IBM-1 and IBM-2 are obtained in a similar way by allowing only zero-order dependencies.

IBM-3,4 and 5 Models. For the generation of the target sentence, it is more appropriate to use the concept of *inverted alignments* which perform a mapping from a target position i to a set of source positions j , i.e. we consider mappings \mathcal{B} of the form:

$$\mathcal{B} : i \rightarrow \mathcal{B}_i \subset \{1, \dots, j, \dots, J\}$$

with the constraint that each source position j is covered exactly once. Using such an alignment $\mathcal{A} = \mathcal{B}_1^I$, we re-write the probability $Pr(f_1^J, \mathcal{A}|e_1^I)$:

$$Pr(f_1^J, \mathcal{B}_1^I|e_1^I) = p(J|I) \cdot \prod_{i=1}^I [p(\mathcal{B}_i|\mathcal{B}_i^{i-1}) \cdot \prod_{j \in \mathcal{B}_i} p(f_j|e_i)]$$

By making suitable assumptions, in particular first-order dependencies for the inverted alignment model $p(\mathcal{B}_i|\mathcal{B}_i^{i-1})$, we arrive at what is more or less equivalent to the alignment models IBM-3, 4 and 5 [10].

Alignment Template Model. In all the above models, the single words are taken into account. In [11, 12], the authors show significant improvement in translation quality by modeling *word groups* rather than *single words* in both the alignment and lexicon models. The method is known as the *alignment template* (AT) approach.

3.1. Training

The unknown parameters of the alignment and lexicon models are estimated from a corpus of bilingual sentence pairs. The training criterion is the maximum likelihood criterion. As usual, the training algorithms can guarantee only local convergence. In order to mitigate the problems with poor local optima, we apply the following strategy [9]. The training procedure is started with

a simple model for which the problem of local optima does not occur or is not critical. The parameters of the simple model are then used to initialize the training procedure of a more complex model, in such a way that a series of models with increasing complexity can be trained [10]. To train the above models except for the alignment template model, we use the GIZA++ software [10]. The alignment template model training scheme, and also the description of our translation system which is based on the alignment template approach are explained in [12].

4. Speech Recognition System

The speech recognition system is trained on the VerbMobil II corpus [13]. The corpus consists of German large-vocabulary conversational speech: 36k training-sentences (61.5h) from 857 speakers. The test corpus is created from the German part of the bilingual English-German XEROX corpus. The corpus consists of technical manuals describing various aspects of Xerox hardware and software installation, administration, usage, etc. Sentences taken from the German XEROX corpus have been read by 10 speakers where every speaker uttered on an average 16 minutes of test data. Recording sessions were carried out in a quiet office room. The data was recorded at a sampling rate of 16kHz.

The remaining part of the XEROX corpus is used to train the translation models as well as the language model. We make a trigram language model by using the SRI language modeling toolkit [14]. The perplexity of the speech recognition test corpus is about 83. The other statistics of the speech recognition test corpus are shown in Table 1.

Table 1: Statistics of the speech recognition test corpus

	Test
Overall Duration	2.6 h
Silence Fraction[%]	20%
# Speakers	10
# Sentences	1 562
# Running Words	18 144
# Running Phonemes	111 916
3-gram LM perplexity	83

The baseline recognition system (acoustic model) can be characterized as follows:

- recognition vocabulary of 16716 words;
- 3-state-HMM topology with skip;
- 2501 decision tree based within-word triphone states including noise plus one state for silence;
- 237k gender independent Gaussian densities with global pooled diagonal covariance;
- 33 acoustic features after applying LDA;
- max. likelihood training using Viterbi approximation;
- trigram language model, test set perplexity: 83.

5. Results

To train the translation models we use the remaining part of the English-German XEROX corpus which has not been used in the speech recognition test corpus. To carry out the integration experiments, we also need a development corpus for optimizing the scaling factors (explained in Section 2) and an evaluation corpus to report the results. We split the test corpus of speech recognition (Table 1) into two parts, the first 700 utterances as the development corpus and the rest as the evaluation corpus.

The statistics of the corpus are depicted in Table 2. The term OOVs in the table denotes the total number of occurrences of unknown words, the words which have not been seen in the training corpus. The German part of the training corpus is also used for training the language model.

Table 2: Statistics of machine translation corpus

	English	German
Train: Sentences	47 619	
Running Words	528 779	467 633
Vocabulary	9 816	16 716
Singletons	2 302	6 064
Dev: Sentences	700	
Running Words	8 823	8 050
Vocabulary	1 323	1 356
OOVs	56	108
Eval: Sentences	862	
Running Words	11 019	10 094
Vocabulary	1 181	1 197
OOVs	58	100

When we use the N -best list approach for integrating automatic speech recognition and statistical machine translation models, we have to analyze the quality of the N -best list. The N -best list has an upper limit for the possible recognition quality improvement, which can be measured by extracting the best hypothesis for each sentence. We call the best set of hypotheses of an N -best list *oracle recognition*. Different characteristics of the generated N -best list are shown in Table 3.

Table 3: Development and evaluation N -best lists statistics

Feature	Development	Evaluation
# utterances	700	862
Average of N per utterance	216	236
Single best WER[%]	18.4	23.4
Oracle recognition WER[%]	10.6	14.8

Experiments. In order to rescore the N -best list generated by the automatic speech recognizer, we make use of the translation models described in Section 3.

To train the alignment template system, we make use of a chain of word-based alignment models, as described in Section 3. Different combination of models, from simple models to complicated models, are possible. We choose the sequence of models which generate the best translation results. We obtain the following sequence of models: IBM-1, HMM, IBM-4, IBM-5 and AT.

We show the recognition results when the speech recognition models are combined with different statistical machine translation models. The recognition results are summarized in Table 4. In this table, the recognition results of the automatic speech recognition (ASR) system are shown first, then the translation results of the machine translation (MT) system, which is obtained by the alignment template approach. Then, the results of combined speech recognition and translation models are presented. For each translation model, we calculate the translation probability in both directions: $p(e_1^I | f_1^J)$ and $p(f_1^J | e_1^I)$. Then we have two log-linear model for each translation model, e.g. the row specified with IBM-1 shows the recognition results when the $p_{\text{IBM-1}}(e_1^I | f_1^J)$ and $p_{\text{IBM-1}}(f_1^J | e_1^I)$ translation models are used in addition to the speech recognition and language model. The last row (indicated by ALL), shows the recognition results when all models are combined: speech recognition, IBM-1, HMM, IBM-4, IBM-5, and AT models. The model scaling

factors are trained with respect to the final recognition quality measured by the word error rate.

Table 4: Recognition word error rates [%] using translation model rescoring

	Models	Development	Evaluation
ASR	acoustic	30.8	38.3
	acoustic + LM	18.4	23.4
MT	best system	61.1	59.7
ASR+MT	IBM-1	16.8	21.1
	HMM	16.8	21.0
	IBM-4	15.9	20.0
	IBM-5	15.8	19.8
	AT	17.4	21.6
	ALL	15.3	19.4

All improvements of the combined models are statistically significant at the 99% level with respect to the speech recognition system only (acoustic+LM) [15]. The above experiments were not designed for real-time (or close to real-time) performance. In the present implementation, the processing time mainly depends on the size of N -best list and the complexity of the translation model. For real-time operation, a redesign of the search organization might be appropriate.

As we can see in Table 4, the more sophisticated translation models result in larger recognition quality. But one surprising result of these experiments is that the AT model, which is the best translation model, has the least contribution in improving the recognition results. One possible explanation for this is that the AT works on word groups, and working on word groups causes better translation quality but in this task the translations are already generated. In addition, due to the nature of speech recognition error which is a single word error (not a word-group error) and is basically independent from the context, the single word based translation models are more suitable.

6. Conclusion

The goal of this paper was to evaluate if the accuracy of a speech recognition system could be improved by incorporating translation models. We introduced a general framework for integrating the speech recognition and translation models for automatic text dictation in the context of computer-assisted translation. The most interesting characteristic of the introduced model was its flexibility to handle as many features (models) as we desire. The main idea of the implementation was to use N -best list in the interface between the speech recognizer and the translation system. In the experiments, we showed a relative 17% improvement in the recognition results when the translation models were combined with the speech recognition model.

7. Acknowledgements

This work has been partly funded by the European Union under the RTD project TransType2 (IST 2001 32091). We also thank Tibor Szilassy for his efforts in providing the orthographic transcription of the speech recognition test corpus. This paper has greatly profited from discussions with Richard Zens.

8. References

- [1] P. F. Brown, S. F. Chen, S. A. D. Pietra, V. D. Pietra, A. S. Kehler, and R. L. Mercer, "Automatic speech recognition in machine-aided translation," *Computer Speech and Language*, vol. 8, no. 3, pp. 177–187, July 1994.
- [2] M. Dymetman, J. Brousseau, G. Foster, P. Isabelle, Y. Normandin, and P. Plamondon, "Towards an automatic dictation system for translators: the TransTalk project," in *Proceedings of ICSLP-94*, Yokohama, Japan, 1994, pp. 193–196.
- [3] J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Kuhn, Y. Normandin, and P. Plamondon, "French speech recognition in an automatic dictation system for translators: the transtalk project," in *Proceedings of Eurospeech*, Madrid, Spain, 1995, pp. 193–196.
- [4] K. A. Papineni, S. Roukos, and R. T. Ward, "Feature-based language understanding," in *EUROSPEECH*, Rhodes, Greece, September 1997, pp. 1435–1438.
- [5] —, "Maximum likelihood and discriminative training of direct translation models," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Seattle, WA, May 1998, pp. 189–192.
- [6] P. Beyerlein, "Discriminative model combination," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, Seattle, WA, May 1998, pp. 481–484.
- [7] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, July 2002, pp. 295–302.
- [8] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, July 2003, pp. 160–167.
- [9] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, June 1993.
- [10] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, March 2003.
- [11] F. J. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in *Proc. Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, University of Maryland, College Park, MD, June 1999, pp. 20–28.
- [12] F. J. Och and H. Ney, "The alignment template approach to statistical machine translation," *Computational Linguistics*, vol. 30, no. 4, pp. 417–449, December 2004.
- [13] A. Sixtus, S. Molau, S. Kanthak, and H. N. R. Schlüter, "Recent improvements of the RWTH large vocabulary speech recognition system on spontaneous speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Istanbul, Turkey, June 2000, pp. 1671–1674.
- [14] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *In Proc. of the Int. Conf. on Speech and Language Processing (ICSLP)*, vol. 2, Denver, CO, September 2002, pp. 901–904.
- [15] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, pp. 409–412.