



Integration of ASR and Machine Translation Models in a Document Translation Task

Aarathi Reddy¹, Richard Rose¹, Alain Désilets²

¹Department of Electrical and Computer Engineering, McGill University, Montreal, Canada

²National Research Council - Institute for Information Technology, Ottawa, Canada

aarathi.reddy@mail.mcgill.ca, rose@ece.mcgill.ca, alain.desilets@nrc-cnrc.gc.ca

Abstract

This paper is concerned with the problem of machine aided human language translation. It addresses a translation scenario where a human translator dictates the spoken language translation of a source language text into an automatic speech dictation system. The source language text in this scenario is also presented to a statistical machine translation system (SMT). The techniques presented in the paper assume that the optimum target language word string which is produced by the dictation system is modeled using the combined SMT and ASR statistical models. These techniques were evaluated on a speech corpus involving human translators dictating English language translations of French language text obtained from transcriptions of the proceedings of the Canadian House of Commons. It will be shown in the paper that the combined ASR/SMT modeling techniques described in the paper were able to reduce ASR WER by 26.6 percent relative to the WER of an ASR system that did not incorporate SMT knowledge.

1. Introduction

One class of techniques that addresses the problem of machine aided language translation involves the combination of models for statistical machine translation (SMT) and automatic speech recognition (ASR) [1, 2, 3, 4]. These techniques are intended to be applied to scenarios where human translators speak the spoken language translation of a source language text into an automatic speech dictation system while the text is simultaneously presented to a SMT system. The work presented in this paper attempts to reduce ASR word error rate (WER) on speech utterances from the human translator in the target language by incorporating knowledge acquired from SMT in decoding the optimum word string.

The paper makes two contributions. First, an alternative optimization criterion is investigated for ASR decoding that integrates machine translation models with ASR acoustic and language models. This is referred to below as *tight integration* of SMT and ASR systems and was motivated by previous work [1, 4]. Second, scenarios for integrating SMT and ASR systems are investigated where target language text generated by applying SMT to the entire source language text document is used as a source of information for updating the target language model in ASR. This latter approach is implemented by estimating N -gram probabilities from the target language strings generated by the SMT system and combining the resulting statistical language model (LM) with the N -gram LM used in the

ASR system. This combination is referred to in the paper as *loose integration*. Both LM interpolation and lattice re-scoring techniques are investigated depending on the assumptions made when integrating the two sources of target language probability.

The language translation industry is largely driven by the requirement that official documents for many international organizations must exist in all official languages of that organization. The European Union has twenty official languages while the government of Canada has two, French and English. The Canadian government requires that legislation exists in both languages. Maintaining archives of multilingual documents has created a large translation industry in Canada alone. It is estimated that the size of the translation industry in Canada may approach 500 million dollars per year and may employ well over 13,500 translators [5]. This provides tremendous financial motivation to develop tools for making this large population of human translators more efficient. Studies conducted over 40 years ago demonstrated that the productivity of human translators increases by a factor of four when the translations are dictated by voice as compared to written or typed translations [1]. This general observation has driven interest over the past 12 years in creating powerful systems for dictation of translation utterances. More recently, machine-aided human translation (MAHT) systems have been built to help professional translators improve their efficiency, by giving them the ability to create, modify, export, and import lexical and terminological databases [6]. Other MAHT paradigms include interactive machine translation with target text prediction [7].

The task domain that is of interest in this work involves the translation of the proceedings from the Canadian House of Commons. In Canada, it is required that the records of the parliamentary debates be published in both French and English. This report of *in extenso* debates which take place in the House is commonly referred to as Hansard, and has been in existence for the past 125 years. For many years the translation of reports has been handled by the Translation Bureau of Canada, which employs a team of professional translators. ASR results will be reported in Section 5 for a pilot corpus consisting of English language speech utterances arising from translations of French language source text. However, a larger corpus consisting of spoken language translations of Hansard texts by a larger population of professional translators employed by the Translation Bureau is currently under development [8].

The paper is organized as follows. To provide background, a brief summary of previous approaches to machine aided human translation is given in Section 2.1. An overview of the ASR and SMT systems that are used here is given in Section 2.2. In Section 3, the pilot speech corpus that was used for the experimental study presented in this paper is described along with a

This work was performed in collaboration with the DIVINES FP6 project and supported under NSERC Program Number 307188-2004

description of the larger corpus that is currently under development. In Section 4, the approaches for integrating statistical models from SMT and ASR, namely *loose integration* and *tight integration* are described. Finally, the implementation and associated results are presented in Section 5.

2. Background

2.1. Related Work in MAHT

One of the earliest efforts made to directly combine translation models and LMs was performed by Brown et al. [1]. In that work, the optimum target language string in a stack decoder based ASR system was obtained from the joint probability of the source language and target language strings computed from both LM and translation model parameters. While they did not report ASR results, they demonstrated that the perplexity of the combined translation/language model was significantly less than the original trigram LM for utterances taken from the Canadian Hansard corpus. At about the same time, Brousseau et al. presented two methods for combining ASR and SMT models as part of the TransTalk project which involved English to French translation in the Canadian Hansard domain [2]. Of particular interest was a method for re-scoring N -best lists of French word hypotheses generated by a large vocabulary continuous speech recognizer (LVCSR) with a language translation model.

More recently, Paulik et al. applied a number of techniques to achieve a closer coupling between text based SMT and acoustic ASR on an English to Spanish travel-phrase language translation task [3]. Integration was accomplished both through re-scoring of N -best word hypotheses and by incorporating candidates from SMT into cache and interpolated LMs for ASR. Khadivi et al., demonstrated a decrease in WER on an English-German technical document translation task when using different translation models to re-score the N -best lists obtained from the recognizer [4]. An interesting result of both of these recent papers was that the largest increase in ASR performance was obtained not from the very best performing translation models, but instead from translation information that incorporated limited word context information.

2.2. ASR and SMT Systems

The large vocabulary continuous speech recognition system used in this work was developed at the Centre de Recherche Informatique de Montréal (CRIM) [2], and is based on a weighted finite state machine (FSM) approach to ASR [9]. In general, FSM based ASR systems represent speech as a cascade of independent models for language, pronunciation, acoustic context, and hidden Markov model (HMM) topology. FSM representations of each of these models are composed to create a single network, and decoding the optimum word string in ASR is performed by expanding this network during search. Word lattices are generated for each input utterance in the AT&T FSM format [2, 9]. One of the techniques described under “loose integration” in Sections 4 and 5 involves re-scoring these lattices with new LMs derived from the target language text generated by the SMT system.

The PORTAGE machine translation system (SMT) used in this work was developed at the NRC Institute for Information Technology and is based on a phrase based statistical approach [10]. It relies on a decoder that obtains an optimum target language word string, e , from a source language sentence, f , by maximizing $P(e|f)$ which is the log-linear com-

ination of four components. These include a target language trigram LM, a phrase translation model, a distortion model, and a word length model. The configuration and use of PORTAGE for the French-English language translation task has been discussed in [10]. The “tight integration” procedure given in Sections 4 and 5 involves incorporating the PORTAGE decoder for re-evaluating ASR string hypotheses.

3. PAT Speech Corpus

The PAT (Paroles Aux Traducteurs) project has been initiated by the Interactive Language Technologies group at the National research Council (NRC) of Canada. The project involves developing and evaluating automatic speech dictation based on machine aided human translation tools, in an effort to make human translators more efficient. There are three goals: evaluating productivity gains using ASR systems for this task, identifying ergonomic issues encountered by translators while interacting with speech interfaces, and improving ASR accuracy based on knowledge acquired from applying SMT to source language text [8].

As an initial study, a PAT corpus was collected from 3 bilingual subjects who were asked to translate two Hansard French texts each. Apart from these recordings, each subject was also asked to provide enrollment data which consisted of reading a short paragraph. All of the simulations reported in Section 5 are based on the utterances in the PAT corpus. The enrollment data was used to perform MAP and MLLR acoustic adaptation for each of the speakers. The audio obtained from the three speakers reading English Hansard text was used as development data. Another set, where the speakers are translating a French Hansard text, was used as the test data.

The enrollment data consisted of approximately 1400 words spoken by each of the speakers amounting to a total of 18 minutes of speech. The development data consisted of read text with 1520 words amounting to a total of 26 minutes. The evaluation data consists of dictated translations of French text into English utterances. A larger corpus is currently under development involving professional translators at the Canadian Translation Bureau. The goal is for this corpus to include audio data collected from 16 French-English translators and 16 English-French translators.

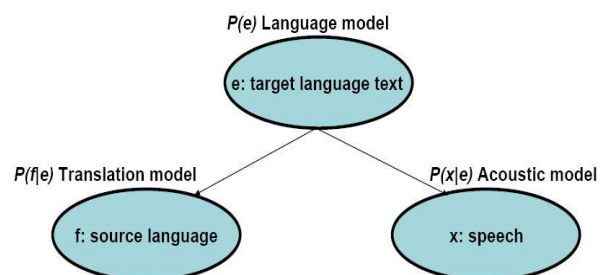


Figure 1: Combined ASR/SMT Model

4. Integrating ASR and SMT Models

A general decoding procedure for obtaining an optimum word string \hat{e} , in the target language according to a combined ASR/SMT framework is shown in Figure 1. The translation probability, $P(f|e)$, describing how the source language text, f , is generated from the target language text, e , is obtained from the component models in the SMT system as summarized in

Section 2.2. The acoustic probability, $P(x|e)$, where x is the speech uttered, is obtained from a hidden Markov model based ASR system.

The generative model in Figure 1 implies that the optimum target language string is obtained as:

$$\hat{e} = \operatorname{argmax}_e \{P(e)P(f|e)P(x|e)\}. \quad (1)$$

It is assumed in this paper that the language probability, $P(e)$, has two components. The first component, $P_S(e)$, characterizes those aspects of language that can be acquired from large text corpora in the target language. A trigram LM trained from the English language Canadian Hansards was used for $P_S(e)$. The second component, $P_M(e)$, represents the effects that can be acquired from the source language text. The model for $P_M(e)$ was obtained by training statistical N -gram models directly from the translated text taken from the SMT system.

These two component LMs can be used to form $P(e)$ by making two different assumptions. First, it can be assumed that $P(e)$ is a mixture of the two sources,

$$P(e) = \lambda_M P_M(e) + \lambda_S P_S(e). \quad (2)$$

In this case a new LM can be obtained by interpolating the N -gram probabilities of the two component models, $P_M(e)$ and $P_S(e)$. A second assumption would be that $P(e)$ is actually the joint probability of two independent component probabilities:

$$P(e) = P_M(e)^{\lambda_M} P_S(e)^{\lambda_S}. \quad (3)$$

To realize this case, the N -gram LM associated with $P_M(e)$ is used to re-score ASR lattices that were generated using the LM associated with $P_S(e)$. These two methods are collectively referred to in the paper as *loose integration* of ASR and SMT.

Another method of integrating the ASR and SMT models, referred to as *tight integration*, involves using SMT to re-evaluate ASR hypotheses. Here, each string hypothesis appearing in the ASR N -best list is re-scored using the language translation probability, $P(f|e)$, obtained from the SMT. The score for each string is computed from a log-linear combination of acoustic, language model, and translation model probabilities,

$$\hat{e} = \operatorname{argmax}_e \{\lambda_1 \log(P(e)) + \lambda_2 \log(P(f|e)) + \lambda_3 \log(P(x|e))\}. \quad (4)$$

5. Experimental Study

This section describes the implementation and evaluation of the three strategies for ASR/SMT integration that were presented in Section 4. First, the loose integration strategies, including the ASR/SMT LM interpolation and LM lattice re-scoring implementations of the models given in Equations 2 and 3, are described. Second, the tight integration strategy, that involves a scheme for re-scoring N -best ASR string hypotheses using the SMT translation model, is described. Finally, the ASR word error rates (WERs) for these strategies are reported and summarized in Table 1.

5.1. Loose Integration of ASR/SMT Models

The combined ASR/SMT LM approaches are implemented under a scenario where it is assumed that the SMT system operates on the entire source language text before the text is presented to the human translator. In the following, the SMT system generates a set of N -best English translation hypotheses for each

sentence in the French language text where $N = 100$. For each speaker, statistical LMs characterizing the probability, $P_M(e)$, in Equations 2 and 3 are then estimated from the resulting text strings.

A statistical trigram LM trained from a 34 million word subset of the English language Canadian Hansard corpus with a 22 thousand word vocabulary was used to represent $P_S(e)$. The perplexity of this LM measured on held out development text was found to be 82, whereas the perplexity measured on the test set was found to be 112.

5.1.1. Interpolated Language Models

The LMs derived from the English Canadian Hansards corpora and the SMT output are combined as in Equation 2. The trigram language model is estimated from the Hansard text and is interpolated with the language model trained from the N -best strings obtained from the SMT. The two weights λ_S and λ_M indicated in Equation 2 are estimated empirically on the development set. This interpolated LM is then used in single pass ASR. In these experiments it was found that recognition was best when $P_M(e)$ is represented by a simple unigram LM trained from N -best strings obtained from the SMT.

5.1.2. Lattice Re-scoring

A set of combined LMs of the form given in Equation 3 are implemented by re-scoring ASR output lattices with N -gram LMs representing $P_M(e)$. The LM representing $P_S(e)$, obtained from the English language Hansard corpus, was incorporated in the first recognition pass. The exponents λ_S and λ_M in Equation 3 are applied to the target language LM and the SMT LM respectively. In this case, the SMT LM was applied by re-scoring the ASR lattices. Again, both values were fixed and were estimated on the development set. Although unigram, bigram, and trigram LMs were generated from the SMT output, it was found in these experiments that the best recognition results were obtained when a bigram model for $P_M(e)$ was used to re-score the output of the ASR.

5.2. Tight Integration of ASR/SMT Models

In order to evaluate the effect of incorporating the translation model probabilities, $P(f|e)$, for decoding the optimum string hypothesis, an N -best re-scoring strategy was used. The speech recognizer generates a word lattice for all utterances from which N -best word strings can be obtained. In this experiment, these N -best word strings are re-scored using the translation model probabilities, $P(f|e)$, obtained from the SMT for that utterance. The translation model probabilities for each N -best string was obtained from the SMT alignment of the string with the corresponding French language string for that utterance. The log linear interpolation weights in Equation 4 were also estimated on the development set described in Section 3. The best string after re-scoring the N -best hypotheses is therefore the one that maximizes the argument in Equation 4.

5.3. Experimental Results

The ASR results for all of the techniques described in Sections 5.1 and 5.2 are presented in Table 1 as the word error rate (WER) measured on the PAT corpus evaluation set. All system parameters including the interpolation weights were adjusted to optimize WER on the development set. The three columns in Table 1 display WER separately for each of the three speakers. Both MAP and MLLR acoustic speaker adaptation was

performed on the combined enrollment/development data described in Section 3. The first and second rows of Table 1

Experiment	Speaker 1	Speaker 2	Speaker 3
Baseline WER	23.98	21.7	23.49
MAP/MLLR	23.58 (1.6)	17.12 (21.10)	21.57 (8.17)
Interp LM	21.68 (9.59)	16.86 (22.30)	20.84 (11.28)
Lat Re-score	19.11 (20.30)	14.12 (34.93)	17.71 (24.60)
SMT Re-score	22.43 (6.46)	15.03 (30.73)	21.69 (7.66)

Table 1: Experimental Results on Pilot data using Translation Model Probabilities and Combined ASR/SMT LMs

show the WER obtained before and after acoustic adaptation. It is clear that adaptation reduces WER for all speakers, though the degree of improvement varies substantially across speakers. All of the techniques displayed in rows three to five of Table 1 are applied separately to the MAP and MLLR adapted system shown in the second row.

The WERs obtained when interpolated LMs are used as part of the ASR engine are shown in the third row of Table 1. The average relative decrease in WER for the test set using this method was found to be 14.39%. This result was obtained using a unigram model to represent the SMT component of the LM described in Equation 2. Attempts at incorporating more long term structure by training bigram and trigram models to represent $P_M(e)$ resulted in negligible reductions in WER.

The fourth row of Table 1 displays the WER for the SMT LM lattice re-scoring approach described in Section 5.1. This was by far the most successful of the techniques implemented in this work. The results given in Table 1 correspond to re-scoring the lattices produced from the first recognition pass with a bigram LM trained from the 100 best translations produced for that speaker. This was implemented by composing the finite state models representing the lattice and the SMT LM. It was found that weighted bigram LMs from SMT output gave the lowest WERs as compared to either unigram or trigram LMs obtained from the same SMT output.

The fifth row of Table 1 displays the WER obtained for re-scoring N -best word strings using the SMT translation model probabilities as described in Section 5.2. A list of $N = 10$ strings were re-scored for each utterance. A small reduction in WER was obtained for two of the speakers, and essentially no change in performance was observed for the third speaker. The limited improvements obtained may be explained by the fact that the phrase based SMT system had low coverage for the set of ASR N -best hypotheses. As a result, it obtained very low probabilities or no alignment at all for many of the N -best strings. It may be the case that this problem may be alleviated somewhat when a less constrained word based SMT model is used.

6. Conclusions and Future Work

This paper has presented a set of techniques for improving the performance of ASR systems within a combined ASR/SMT framework. The larger goal is to improve the efficiency of human language translators that dictate their translations to an automatic dictation system. The assumed working scenario for all of the techniques presented is that the text based SMT system produces target language translations for an entire source language working document.

The most significant performance improvement was obtained by re-scoring ASR lattices from an initial recognition

pass with a LM trained from the SMT output for the given document and speaker. Using this strategy, an average reduction of 26.6% ASR WER was obtained.

A method for re-scoring recognition string hypotheses using the translation model probability obtained for that string was also investigated. This resulted in a small reduction in WER. Future work will attempt to improve on this result by applying less constrained SMT models that have larger coverage and are more appropriate for this re-scoring task. These techniques will also be applied to a larger PAT corpus that is being collected from employees of the Canadian Translation Bureau.

7. Acknowledgments

The authors would like to thank Patrick Cardinal and Gilles Boulianne from CRIM for providing valuable assistance with the CRIM speech recognition engine and for many helpful conversations. The authors would also like to thank the NRC-LTRC group: George Foster, Roland Khun and Samuel Larkin for assistance with the PORTAGE machine translation engine and for helpful advice.

8. References

- [1] Brown, P. F., Chen, S. F., Pietra, S. A. D., Pietra, V. D., Kehler, A. S., and Mercer, R. L. "Automatic speech recognition in machine aided translation", *Computer Speech and Language*, 3(8), 1994.
- [2] Brousseau, J., Drouin, D., Foster, G., Isabelle, P., Kuhn, R., Normandin, Y., and Plamondon, P. "French speech recognition in an automatic dictation system for translators: The TransTalk project", *Eurospeech*, 1995.
- [3] Paulik, M., Fügen, C., Stüker, S., Schultz, T., Schaaf, T., and Waibel, A. "Document driven machine translation enhanced ASR" *European Conference on Speech Communication and Technology*, *Interspeech*, 2005
- [4] Khadivi, S., Zolnay, A., and Ney, H. "Automatic text dictation in computer assisted translation", *European Conference on Speech Communication and Technology*, *Interspeech*, 2005.
- [5] Duchaine, M. "Financing the language industry in Canada, AILIA", 2006.
- [6] Kay, M., Boitet, C., Fluhr, C., Waibel, A., Muthuswamy, Y. K., and Spitz, A. L. "Multilinguality", in *Survey of the State of the Art in Human Language Technology*, Cole, R. A., Editor. Chapter 8, Cambridge University Press, 1996.
- [7] Vidal, E., Casacuberta, F., Rodriguez, L., Civera, J., and Martínez-Hinarejos, C. D. "Computer-assisted translation using speech recognition", *IEEE Transactions on Audio, Speech & Language Processing* 14(3), 2006.
- [8] Désilets, A. "PAT: Evaluation and improvement of speech recognition systems for translators", *Project Description for Ottawa Research Ethics Board (O-REB) Review*, 2006.
- [9] Mohri, M., Pereira, F., Riley, M. "Weighted automata in text and speech processing", 12th biennial European Conference on Artificial Intelligence (ECAI-96), Workshop on Extended Finite State Models of Language, 1996.
- [10] Sadat, F., Johnson, H., Agbago, A., Foster, G., Kuhn, R., Martin, J., Tikuisis, A. "PORTAGE: A phrase based machine translation system", *ACL* 2005.