Incorporating Knowledge of Source Language Text in a System for Dictation of Document Translations

Aarthi Reddy and **Richard Rose**
McGill University
Dept. of Electrical Engineering
Montreal, Canada
aarthi.reddy, rose@mcgill.ca

Hani Safadi
University of Toronto
Dept. of Computer Science
hani@cs.toronto.edu

Samuel Larkin
IIT, National Research Council
Gatineau, Canada
Samuel.Larkin@cnrc-nrc.gc.ca

Gilles Boulianne
CRIM
Montreal, Canada
Gilles.Boulianne@crim.ca

Abstract

This paper describes methods for integrating source language and target language information for machine aided human translation (MAHT) of text documents. These methods are applied to a language translation task involving a human translator dictating a first draft translation of a source language document. A method is presented which integrates target language automatic speech recognition (ASR) models with source language statistical machine translation (SMT) and named entity recognition (NER) information at the phonetic level. Information extracted from a source language document including translation model probabilities and translated named entities are combined with acoustic-phonetic information obtained from phone lattices produced by the ASR system. Phone-level integration allows the combined MAHT system to correctly decode words that are either not in the ASR vocabulary or would have been incorrectly decoded by the ASR system. It is shown that the combined MAHT system results in a decrease in word error rate on the dictated translations of 32% relative to a stand alone baseline ASR system.

1 Introduction

The goal of MAHT systems in document translation is to provide tools to human translators for increasing their performance and productivity. Many different scenarios for MAHT have been proposed over the last two decades [Brown et al. 1994,Brousseau et al. 1995]. The focus in this work is on MAHT scenarios where a human translator dictates the translation of a source language document. This scenario differs from a traditional dictation task in two ways. The first is the difficulty associated with the utterances themselves. In addition to the potential disfluencies that may be associated with other dictation tasks, the translation utterances may contain higher level errors with respect to what might be considered a "correct" translation of the source language text. The second difference and potential advantage is the large amount of side information, mostly in the form of lexical information and named entities, contained in the source language text. The goal of this work is to exploit a noisy version of this side information to improve the quality of the ASR transcription and to improve the first draft of the translated document.

One of the earliest discussions of this MAHT scenario appeared in [Brown et al. 1994] which involved incorporating translation model probabilities obtained from SMT into the statistical language model used in ASR. A general model was posed where the optimum target language word string, $\hat{e}$, decoded from the speech utterance, x, for a source language string, f, is expressed as:

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, p(e|f,x) = p(x|e,f)p(e|f) \quad (1)$$

$$= \underset{e}{\operatorname{argmax}}\, p(x|e)p(f|e)p(e) \quad (2)$$

In Equation 2, $p(x|e)$ is the ASR acoustic model probability, $p(f|e)$ is the translation model probability obtained from a SMT system and, $p(e)$, is the ASR language model (LM) probability. Equation 2 is obtained by applying Bayes' rule and assuming that the speech utterance is independent of f given e. This model has been used for re-scoring ASR

string hypotheses generated from translation utterances [Reddy et al. 2007].

The document translation scenario followed by the human translator is assumed to involve the following passes. First, the translator reads through and summarizes the source language document. Second, the translator identifies unfamiliar terminology and phrases and resolves them using available reference tools. The first draft of the target language translation is then dictated by the translator. The combined MAHT system described here produces the decoded word string that best explains the dictated utterance and the source language text. Finally, the translator updates the text of this draft translation to correct errors introduced by the MAHT system as well as errors associated with decisions made by the translator.

2 Combined MAHT System

This section begins with a description of the MAHT model. Section 2.1 describes how an optimum target language word string is obtained by combining statistical models of language, acoustics, translation, and named entities. Discriminative minimum error rate training techniques are used to obtain an objective function based on the weighted log probabilities associated with these models. These are described in Section 2.2. Finally, Section 2.3 describes the implementation of the ASR, SMT, and NER systems used in this work.

2.1 System Description

Figure 1 is a block diagram representation of the document translation MAHT scenario in terms of finite state machines (FSMs). A French language document is presented to a human translator. The same document is also processed by a statistical machine translation (SMT) system and a named entity recognition (NER) system.

Translation model probabilities, target language decoded text, and named entity tags are generated for each sentence in the source language document. This information is stored as a set of weighted hypothesized transcriptions along with the associated phone level pronunciations in the phone/word transducer, L. A set of hypothesized transcriptions of the translator's utterance are generated by the ASR system in the form of a phone lattice, R. The phone sequence R, that best explains the source language derived information in L is obtained in this case using a string edit distance. This is implemented by composing with an edit transducer, T, as $W = R \circ T \circ L$.

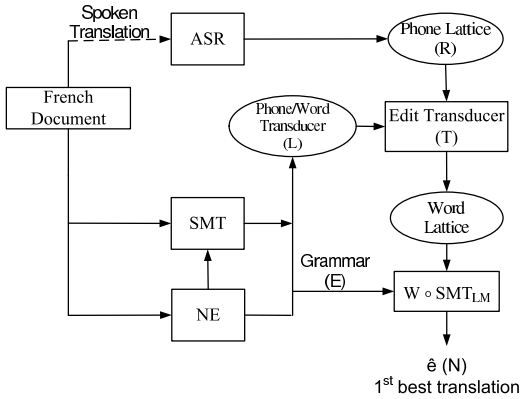

Figure 1: Block diagram of combined MAHT system

The final decoded output, $\hat{e}$, is generated after rescoring W with a language model, SMT_{LM} which incorporates statistics derived from translated word strings obtained from the SMT system.

The approach described in Figure 1 is motivated by the following probabilistic model. The ASR system produces a phone sequence, $\mathbf{r}$, that maximizes the likelihood of the input utterance, $\mathbf{x}$. A phone string, $\mathbf{q}$, is obtained through pronunciation rules that best explain the translated text string, $\mathbf{e}$, of the source language text string $\mathbf{f}$. It is assumed that $\mathbf{x}$ is indirectly dependent on $\mathbf{f}$ and $\mathbf{e}$ through the phone sequences $\mathbf{r}$ and $\mathbf{q}$.

To describe this dependency, let the phone sequence hypothesized by the ASR and the phone sequence hypothesized from the translated text serve as latent variables in defining $p(\mathbf{x}|\mathbf{e},\mathbf{f})$ from Equation 2. Under these assumptions, the conditional probability of the input utterance given a source language / target language sentence pair can be written as:

$$p(\mathbf{x}|\mathbf{e},\mathbf{f}) = \sum_{\mathbf{r}}\sum_{\mathbf{q}} p(\mathbf{x},\mathbf{r},\mathbf{q}|\mathbf{e},\mathbf{f}) \quad (3)$$

$$= \sum_{\mathbf{r}}\sum_{\mathbf{q}} p(\mathbf{x}|\mathbf{r})p(\mathbf{r},\mathbf{q}|\mathbf{e},\mathbf{f}) \quad (4)$$

$$= \sum_{\mathbf{r}}\sum_{\mathbf{q}} p(\mathbf{x}|\mathbf{r})p(\mathbf{r}|\mathbf{q})p(\mathbf{q}|\mathbf{e},\mathbf{f}). \quad (5)$$

Equation 4 is obtained by assuming that $\mathbf{x}$ depends on $\mathbf{f}$ only through the phone string $\mathbf{r}$. In Equation 5, $p(\mathbf{x}|\mathbf{r})$, represents phone level acoustic probability obtained from the ASR system. Furthermore, $p(\mathbf{r}|\mathbf{q})$ is the relationship between the phone string associated with the input utterance and the phone string associated with the translated utterance obtained from the SMT and NER systems. In the experiments conducted for this paper, $p(\mathbf{r}|\mathbf{q})$ is approximated by a string edit distance. The probability $p(\mathbf{q}|\mathbf{e},\mathbf{f})$ is the pronunciation model used to obtain $\mathbf{q}$ from the hypothesized translated text.

The expression for $p(\mathbf{x}|\mathbf{e},\mathbf{f})$ given in Equation 5 can be incorporated into Equation 1 to obtain the optimum target language word sequence, $\hat{\mathbf{e}}$:

$$\begin{aligned}\hat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{e}|\mathbf{f},\mathbf{x}) \qquad (6)\\ &= \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{f}|\mathbf{e})p(\mathbf{e})\sum_{\mathbf{r}}\sum_{\mathbf{q}} p(\mathbf{x}|\mathbf{r})p(\mathbf{r}|\mathbf{q})p(\mathbf{q}|\mathbf{e},\mathbf{f})\\ &\approx \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{f}|\mathbf{e})p(\mathbf{e})\max_{\mathbf{r},\mathbf{q}} p(\mathbf{x}|\mathbf{r})p(\mathbf{r}|\mathbf{q})p(\mathbf{q}|\mathbf{e},\mathbf{f}).\end{aligned}$$

The translation model probabilities, $p(\mathbf{f}|\mathbf{e})$, in Equation 6 are derived from the SMT system. The SMT system is also used to generate N-best lists of English language translations, which are then used for training the LM described in Section 3.

The NE strings extracted from the source language text as shown in Figure 1 are used to provide additional side information to the MAHT system. This is important because NEs often correspond to rarely occurring words which are likely to be decoded incorrectly by both the ASR and SMT systems. An NE tag associated with a source language word can be used to help decode the optimum word string in the target language. If source language words associated with an NE tag are translated to the target language independent from the surrounding context, then the NE tag sequence, $\mathbf{t} = t_1,\ldots,t_N$, in the source language sentence maps directly to the words in the target language sentence. This can be made more clear by re-writing Equation 2 by incorporating the NE tag sequence, $\mathbf{t}$:

$$\begin{aligned}\hat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{e}|\mathbf{f},\mathbf{x}) \qquad (7)\\ &= \underset{\mathbf{e}}{\operatorname{argmax}}\sum_{\mathbf{t}} p(\mathbf{e},\mathbf{t}|\mathbf{f},\mathbf{x}) \qquad (8)\\ &= \underset{\mathbf{e}}{\operatorname{argmax}}\sum_{\mathbf{t}} p(\mathbf{x}|\mathbf{e},\mathbf{t},\mathbf{f})p(\mathbf{e},\mathbf{t}|\mathbf{f}) \qquad (9)\\ &= \underset{\mathbf{e}}{\operatorname{argmax}}\sum_{\mathbf{t}} p(\mathbf{x}|\mathbf{e},\mathbf{f})p(\mathbf{f}|\mathbf{e},\mathbf{t})p(\mathbf{e},\mathbf{t}). \qquad (10)\end{aligned}$$

Equation 10 is the same as Equation 2, except that the LM probability is replaced by the joint probability of the word and NE tag string, $p(\mathbf{e},\mathbf{t})$. The joint probability, $p(\mathbf{e},\mathbf{t})$, can be expressed using a chain rule similar to the one used to express n-gram LM probabilities, $p(\mathbf{e},\mathbf{t}) = \prod_i p(e_i,t_i|e_1,t_1,\ldots,e_{i-1},t_{i-1})$. This expression can be simplified by assuming that a given t_i depends only on the associated word e_i so that $p(\mathbf{e},\mathbf{t})$ can be approximated as the product of language model and first order NE tag probabilities

$$p(\mathbf{e},\mathbf{t}) = \prod_i p(e_i|e_1,\ldots,e_{i-1})\prod_i p(t_i|e_i). \qquad (11)$$

The procedure for incorporating NE tags in decoding the optimum target language string by the combined MAHT system involves three steps. First, the NER system tags each word in the source language string. Second, each word in the source language tagged as an NE is translated to the target language. The only exception to this step is when a source language word is OOV with respect to the SMT system vocabulary. In that case, the target language word is assumed to be the same as the source language word. Finally, the n-gram LM probability associated with each word $\mathbf{e_i}$ is weighted by the probability of the NE tag for that word, $p(t_i,e_i)$. If e_i is OOV, the probability $p(t_i,e_i)$ is assumed to be a constant which is empirically derived.

2.2 Minimum Error Rate Training

The decoding algorithm for the combined MAHT system described above can be expressed as Equation 6, and it includes translation model, $p(\mathbf{f}|\mathbf{e})$, language model, $p(\mathbf{e})$, acoustic model, $p(\mathbf{x}|\mathbf{r})$,and pronunciation model, $p(\mathbf{q}|\mathbf{e},\mathbf{f})$. When augmented to include NE tags as shown in Equation 10, it includes the probability of a NE tag sequence, $p(\mathbf{e},\mathbf{t})$. The optimum target language sequence, $\hat{\mathbf{e}}$, is obtained by maximizing the log of the expression given in Equations 6 and 10. Each of the log probability terms in this expression are represented as model $\mathbf{h}_m$. The optimum string is chosen according to the following criterion:

$$\hat{\mathbf{e}}(\mathbf{f},\mathbf{x}) = \underset{e}{\operatorname{argmax}}\sum_{m=1}^{M}\lambda_m\mathbf{h}_m(\mathbf{e},\mathbf{f},\mathbf{x}), \qquad (12)$$

where the weights $\lambda_m, m = 1,\ldots,M$ are estimated using minimum error rate training (MERT) [Och

2003]. The goal of MERT is to directly integrate the final evaluation metric as part of the training procedure.

In the combined MAHT process described in this paper, $M = 7$, and the various models, $\mathbf{h}_m$, are translation models generated by the SMT system - IBM Model 1 and 2, acoustic model score, ASR LM score, LM derived from strings decoded by the SMT and NER systems, the phonetic distance model $p(\mathbf{r}|\mathbf{q})$ and the NE model $p(\mathbf{e}, \mathbf{t})$. Assuming that there are M weights, $\lambda_m, m = 1, \ldots, M$ for M feature models, $h_m, m = 1, \ldots, M$ that are to be estimated, the optimum decoded string is given by:

$$P(\mathbf{e}|\mathbf{f}, \mathbf{x}) = p_{\lambda_1^M}(\mathbf{e}|\mathbf{f}, \mathbf{x}) = \frac{\exp(\sum_{m=1}^{M} \lambda_m \mathbf{h}_m(\mathbf{e}, \mathbf{f}, \mathbf{x}))}{\sum_{\mathbf{e}_1^I} \exp(\sum_{m=1}^{M} \lambda_m \mathbf{h}_m(\mathbf{e}, \mathbf{f}, \mathbf{x}))}. \quad (13)$$

2.3 System Implementation

The MAHT system shown in Figure 1 is comprised of three main components: the ASR system, the SMT system, and the NER system. The combined MAHT system was described above in Section 2.1 and the three main components are described here.

The ASR system used in these experiments is based on a cascade of finite state transducers [Bouillanne et al. 2000]. It is a composition of four independent blocks: hidden Markov model topology, acoustic context information, lexicon, and word based trigram LM, each of which are represented as a weighted finite state machine (FSM). These four FSMs are then composed together to create a single network, and decoding the utterance dictated by the translator involves expanding this network during search. This yields a word lattice for each utterance. Phone lattices required for each utterance are generated by composing the word lattices with the same pronunciation lexicon transducer used in the ASR system.

The SMT system, PORTAGE, used in these experiments was developed at the National research Council (NRC), Canada [Sadat et al. 2005]. Translation models based on IBM models 1 and 2 were trained from a corpus of approximately 2.87 million French/English pairs obtained from LDC Hansard French-English corpus. Previous studies have suggested that less constrained word based SMT models can achieve better performance in rescoring ASR string candidates [Reddy et al. 2008, Khadivi et al. 2005].

The finite state automaton based NER system was built at the University of Tours, France [Friburger et al. 2004]. The NER system consists of a series of FST cascades which allow for the implementation of syntactic analysis and information extraction. In this work, NEs from the following NE classes were extracted from the French language text:*Organization*, *Person*, *Product*, *Location*, *Event* and *Time/Date*. The system was not re-trained for this task domain.

3 Evaluation and Results

This section describes an experimental study performed on utterances collected from human translators. The translators are dictating the first draft of translations of 400-2000 word French language documents taken from the Canadian Hansard domain. The speech corpus and evaluation scenario are described in Sections 3.1 and 3.2 respectively. Results and discussion are given in Section 3.3

3.1 MAHT Task Domain and Corpus

Speech data was collected under the scenario from 9 bilingual speakers, 3 male and 6 female. Six of the nine speakers had experience working as translators. Dictated utterances were obtained for the translations of 456 sentences. These utterances contained a total of 11,491 words and were 106 minutes in duration. This corresponds to approximately 25 words per sentence which is quite long in comparison to other MT tasks. Speech data from each of the speakers reading various non-overlapping portions of the English Hansard amounting to a total of 20 minutes was also collected to use for acoustic model adaptation. Of the 456 sentences collected from the translators, 200 sentences were held out and used for MERT 250 sentences were used as test data. The results reported in Sections 3.2 and 3.3 are calculated using these 250 sentences.

In addition to the dictated translations described above, two English language reference translations were obtained. One was supplied with the Hansard corpus and the other was obtained separately from a professional translator.

The speaking style associated with the utterances collected from this "translation dictation" task was more spontaneous in character than other speech translation tasks. About half of the utterances that were collected contained a significant number of disfluencies including filled pauses, word fragments, repetitions, and false starts. In order to study the effect of disfluencies on both speech recognition and translation performance, the test corpus utterances were subdivided into disfluent and well-formed utterances.

3.2 Evaluation Scenario

In this section, the experimental conditions considered for the study of the MAHT system are described. These experiments are designed to understand the separate impact of acoustic, lexical and grammatical information obtained from the ASR, SMT and NER systems.

In the first experiment, the performance of the baseline ASR system is measured. The acoustic models in the ASR system are gender independent and built using the WSJ corpus. The vocabulary used was the 20000 most frequently occurring words in the Broadcast News (BN) corpus and the LM was built from the BN and Hansard French-English parallel corpus.

The second experiment, referred to as "phone level integration" (PLI), studies the impact of lexical information obtained from the SMT and NER systems. The PLI system has a sentence specific lexicon that includes pronunciations of source language words tagged according to NE class, and pronunciations of their translations, in addition to the 20000 word vocabulary used in the baseline system. The LM is the same as the one described in the baseline system. The inclusion of source language words in the lexicon of the PLI system is designed to account for instances in the utterances when the translator chooses not to translate certain words or phrases, and instead dictates them as they would appear in the source language. Examples of such instances occur when the translator encounters words in the source language document belonging to certain NE categories like company names and movie titles.

The third experiment is designed to study the impact of grammatical information obtained from the SMT and NER systems. A bi-gram LM containing both French and English language strings is used to interpolate the LM used in the baseline ASR system. This interpolation is referred to as *Loose Integration* and is described in detail in [Reddy et al. 2007]. The lexicon used is the same as the one in the PLI system.

Table 1: System decoded word strings for French language phrase "... mon collègue, le député de Nepean Carleton, qui disait ..."

Example Dictated Translations	
French Text	... mon collègue, le député de Nepean Carleton, qui disait ...
Dictated Transcription	... my colleague comma the deputy from Nepean Carleton comma who was saying ...
ASR decoded word string $\hat{w}$	... my colleague comma the deputy from **the pin carton** comma who was saying ...
ASR decoded phone string $\hat{r}$	m ay k aa l iy g k aa m ah dh ah d eh p y ah t iy f r ah m **dh ah p ih n k aa r t ah n** k aa m ah hh uw w aa z s ey ih ng ...
Combined system decoded word string	... my colleague comma the deputy from **Nepean Carleton** comma who was saying ...

This third experiment is referred to as "Bi-lingual LM Rescore" and is represented in schematic form in Figure 1. The effect of using a bi-lingual lexicon and LM in this combined MAHT system can be illustrated using the example utterance shown in Tables 1 and 2. The first row of Table 1 shows a French language string presented to the human translator. The second and third rows of Table 1 show the transcription of the utterance as dictated by the translator and the transcription as decoded by the ASR system respectively. The word sequence "Nepean Carleton" in the example utterance is decoded as "dip in carton" by the ASR system, probably due to the fact that like most proper names "Nepean Carleton" does not occur in the training text or vocabulary of the ASR system. The phonetic expansion of the utterance as decoded by the ASR system is shown in row four of Table 1. Row five corresponds to the string decoded by the combined MAHT system.

Table 2 shows the phonetic expansions of the seg-

Table 2: Phonetic expansions of utterance corresponding to "Nepean Carleton" as decoded by Baseline and Combined systems

Baseline System Word/Phone Hypotheses for "Nepean Carleton"														Cost
r	dh	ah	p	ih	n	k	aa	r	-	t	ah	n		-
w	the		pin			carton								
Combined System Word/Phone Hypotheses for "Nepean Carleton"														Cost
$\mathbf{q}_1$	*n*	*eh*	p	*i*y	n	k	aa	r	**l**	t	ah	n		*c*
$\mathbf{e}_1$	nepean					carleton								
$\mathbf{q}_2$	*d*	*ih*	p	ih	n	k	aa	r	-	t	ah	n		*c* + 8
$\mathbf{e}_2$	dip			in		carton								

ment of the utterance in Table 1 corresponding to "Nepean Carleton". The first and second rows show the phone and word strings as decoded by the ASR system and correspond to the bolded characters in rows three and four of Table 1. The third and fifth rows of Table 2 show two hypothesized phonetic expansions as decoded by the combined MAHT system and the fourth and sixth rows show their corresponding word strings. It should be noted that although there are three phone substitutions and a phone insertion in q_1, the word string, e_1 corresponding to q_1 obtains a lower cost than e_2 which has just two phone substitutions. This can be explained by the high probability of the "proper name" NE "Nepean Carleton". This example demonstrates how the combined system can correct and introduce words that are not in the ASR vocabulary but appear in the source language text.

3.3 Results and Discussion

When evaluating MAHT performance, it is important to consider the evaluation metric and the task domain. The most important issue is the evaluation metric itself. At the application level, it is necessary to evaluate the impact of the MAHT system on the productivity of the human translator. This is often measured in terms of the number of translated words per minute (TWM) and a variety of more detailed measures of human interface efficiency [Vidal et al. 2006]. This class of measures is used for evaluating the predictive techniques described above. However, there is no attempt in this paper, or in any of the previous literature on dictation systems that we are aware of for translation tasks, to make any quantitative claims regarding improvements in productivity. It is assumed that a system that produces a text string which is an accurate transcription of the input utterance and an accurate, fluent translation of the source language text will require minimal effort on the part of the translator to produce a final translated document. To address the need for accurate transcription, improvement in ASR WER will be the principal performance metric used here.

A second issue for system evaluation is the task domain. The task domain can determine the degree to which the utterances are well formed. It can also determine the degree to which the correct word string associated with a given translation utterance is itself an accurate translation of the source language text. For example, in the EuroTrans-I corpus used in a Spanish to English speech-to-speech translation task, the utterances were read from "semi-automatically generated phrases obtained from a series of travel books" [Alabau et al. 2007]. It is reasonable to assume in this case that the target language text transcription obtained from an error-free ASR system of these utterances would be judged to be good translations of the source language text.

On the other hand, the utterances used in the experimental study described in Section 3.1 were obtained from human translators dictating the first draft of their English language translations of French language Hansard documents. These are spontaneous speech utterances containing many disfluencies, and, as will be shown in this Section, error-free transcriptions of these utterances do not always correspond to accurate, fluent translations of the source language text. As a result, standard translation evaluation metrics like the metric developed by the National Institute of Standards and Technology (NIST) [Doddington 2002] will also be used when presenting the combined system performance.

In this section the results obtained from the various experimental setups are shown. In addition to that, the results for the combined system are presented with and without the use of the discriminative procedure described in Section 2.2. First, the word error rates are reported on the test speech data described in Section 3.1. Second, the quality of translations decoded by the combined system are also evaluated.

In Table 3, results are reported for 250 test sentences described in Section 1. The 250 sentences

were divided into two categories: well formed and disfluent as described in Section 3.1, in order to study the effect that the occurrence of disfluencies in speech has on WERs and translation evaluation. Row one of Table 3 shows the WER for the sentences decoded by the baseline system. The high WER for both well formed and disfluent utterances can be partly explained by the OOV rate of 5.3% measured on the English language reference translations for this test set.

Row two of Table 3 displays the WERs for sentences decoded by the PLI system. The decrease in WER is due partly to the effect illustrated by the example in Tables 1 and 2. Row three of Table 3 shows the WERs for sentences decoded by the combined MAHT system where the optimum string is decoded as described in Equations 6 and 10. At this stage, the weights assigned to the various log-linear probabilities are empirically derived. Row four shows the results of the combined MAHT system where the weights are determined according to the discriminative procedure described in Section 2.2.

Table 3: Word Error Rates obtained for Baseline (BL) ASR, phone level integration (PLI), and Bilingual LM Re-score (B-LMR)

WER for Speech Utterances		
System	Well Formed	Disfluent
BL	28.3	36.2
BL + PLI	24.4	32.4
BL + PLI + B-LMR (Emp.)	22.8	31.1
BL + PLI + B-LMR (Disc.)	19.2	29.1

The decrease in WER obtained for the combined systems can be attributed to two characteristics of the system. First, the inclusion of French and English language strings in the lexicon and LM allows for the possibility of certain words appearing in the dictated translation as they would in the source language text. Second, the increased weight allotted to the translated strings in the combined ASR/SMT LM have a significant effect on the quality of the decoded strings. These two characteristics of the combined system allow for words that are OOV to the ASR system or simply mis-recognized by the ASR system, to be decoded correctly by the combined system. Additionally, the decrease in WER of strings decoded by the system combined by the discriminative procedure as compared to strings decoded by the system combined empirically, shows the importance of the discriminative model combination procedure.

Table 4: Translation scores obtained for English text stings derived from multiple sources

NIST scores		
Source of English Text	Well-Formed	Disfluent
SMT Output	4.3	3.8
ASR Output	4.4	3.8
Combined System Output	4.9	4.2
Human Transcribed Utterance	5.0	4.3
Human Translation	7.3	6.6

In addition to the WER scores, translation accuracy scores in Table 4 were reported on the same test corpus used in Table 3 and computed using a single reference translation obtained from a professional translator. Row one of Table 4 shows the NIST score associated with sentences decoded by the SMT system, and row two shows the NIST score associated with sentences decoded by the baseline system. As can be seen, the NIST scores for these two systems are similar.

Row three of Table 4 shows the NIST score associated with text decoded by the combined system and corresponds to the WER results shown in row four of Table 3. An average improvement in NIST score of 13.6% and 10.5% for well formed and disfluent utterances over the ASR decoded output is observed. This improvement suggests that the strings decoded by the combined system are more similar to the reference translation than either the ASR or SMT decoded strings.

In row four of Table 4, the NIST scores obtained for human transcriptions of the utterances are shown. These scores are very close to the scores obtained by the combined system. Finally, the NIST scores obtained when evaluating the human translation of the document against a reference translation are shown in row five of Table 4. This final score is an indication of the NIST score that might be obtained if a first draft translations obtained from the combined system is edited to create a final draft translation.

4 Conclusion

A procedure for building a machine aided human translation system that incorporates target language acoustic information derived from dictated translation utterances, NE tags derived from source language text, and prior statistical knowledge of translated text derived from SMT has been presented. The approach was shown to be particularly effective in dealing with the problem of OOV words and infrequently occurring words in ASR.

An experimental study was performed on a document translation task where Canadian Hansard domain documents were translated from French to English. The MAHT problem was to obtain text transcriptions of utterances spoken by human translators dictating first drafts of their translations. A decrease in WER of approximately 32% and 19% was obtained for well formed and disfluent utterances respectively relative to the WER obtained for the baseline ASR system.

References

P. F. Brown, S. F. Chen, S. D. Pietra, V. D. Pietra, A. S. Kehler, and R. L. Mercer. 1994. Automatic speech recognition in machine aided translation. *Computer Speech and Language*, Vol. 8, 177–187.

J. Brousseau, C. Drouin, G. Foster, P. Isabelle, R. Khun, Y. Normandin and P. Plamondon. September 1995. French speech recognition in an automatic dictation system for translators: The TransTalk Project. *Proceedings of the 4^{th} European Conference on Speech, Communication and Technology*, Madrid, Spain.

E. Vidal, F. Casacuberta, L. Rodriguez, J. Civera and C. Hinarejos. May 2006. Computer-assisted translation using speech recognition *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 3, 941-951

1998. Linguistic Data Consortium, CSR HUB4 Language Model., LDC Catalogue Number: LDC98T31

Philippe Langlais, George F. Foster and Guy Lapalme. December 2000. Unit completion for a computer-aided translation typing system. *Machine Translation*, Kluwer Academic Publishers, Vol. 15, No. 4., 267–294.

Aarthi Reddy, Richard C. Rose and Alain Désilets. September 2007. Integration of ASR and machine translation models in a document translation task. *Proceedings of InterSpeech*, Antwerp, Belgium.

Aarthi Reddy and Richard Rose. September 2008. Towards domain independence in machine aided human translation. *Proceedings of InterSpeech*, Brisbane, Australia.

Nathalie Friburger and Denis Maurel. 2004. Finite state transducer cascades to extract named entities in texts. *Theoretical Computer Science*, Vol. 313, No. 1, 93–104.

W. A. Gale and K. W. Church. 1991. A program for aligning sentences in bilingual corpora. *Proceedings of Association for Computational Linguistics*, Morristown, USA.

M. Mohri, F. Pereira and M. Riley. 2002. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, Vol. 16, No. 1, 69–88.

F. Sadat, H. Johnson, A. Agbago, G. Foster, R. Khun, J. Martin and A. Tikuisis. June 2005. PORTAGE: A phrase-based machine translation system. *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, USA.

G. Boulianne, J. Brousseau, P. Ouellet, and P. Dumouchel. June 2000. French Large-Vocabulary Recognition with Cross-Word Phonology Transducers. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey.

K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. July 2002. BLEU: A method for automatic evaluation of machine translation. *ACL '02: Proceedings of the 40^{th} Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA.

V. Alabau and A. Sanchis and F. Casacuberta. September 2007. Improving Speech-to-Speech Translation using Word Posterior Probabilities *Proceedings of the Machine Translation Summit XI*, Copenhagen, Denmark.

S. Khadivi and A. Zolnay and H. Ney. September 2005. Automatic Text Dictation in Computer-Assisted Translation *Proceedings of InterSpeech*, Lisboa, Portugal.

G. Doddington. 2002. NIST: Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. *Proceedings of the Human Language and Technology Conference*, San Diego, USA.

F. J. Och. July 2003. Minimum error rate training in statistical machine translation *Proceedings of the 41^st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.