Towards Domain Independence in Machine Aided Human Translation

Aarthi Reddy, Richard C. Rose

Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
aarthi.reddy@mail.mcgill.ca, rose@ece.mcgill.ca

Abstract
This paper presents an approach for integrating statistical machine translation and automatic speech recognition for machine aided human translation (MAHT). It is applied to the problem of improving ASR performance for a human translator dictating translations in a target language while reading from a source language document. The approach addresses the issues associated with task independent ASR including out of vocabulary words and mismatched language models. We show in this paper that by obtaining domain information from the document in the form of labelled named entities from the source language text the accuracy of the ASR system can be improved by 34.5%.

1. Introduction
There have been several approaches proposed for improving the performance of speech dictation systems in machine aided human translation (MAHT) where a human translator is reading from a source language text document and dictating the translation in the target language [1, 2, 3]. In order to improve the performance of an automatic speech recognition (ASR) system in the target language, it is assumed that a noisy target language representation of the source language document is available from a statistical machine translation (SMT) system.

From a Bayesian perspective, decoding the optimum target language word string, $e$, from the input speech, $x$, and the source language text string, $f$, is performed by optimizing the following criterion

$$
\arg\max_{e} p(e|f, x) = p(f|e)p(x|e)p(e).
$$

In Equation 1, $p(f|e)$ represents the probability associated with the SMT translation model, $p(x|e)$ represents the probability associated with the ASR acoustic model, and $p(e)$ represents the probability associated with the ASR language model. The model shown in Equation 1 was first implemented by Brown et al. [1], and although they did not report any ASR scores, they demonstrated a decrease in per-word perplexity by combining language model and translation model probabilities. Brousseau et al., describes a system that implements the paradigm described here [2]. In this system, N-best lists of target language hypotheses generated by the ASR system are re-scored using a translation model. Reddy et al. [4] describes two methods of ASR/SMT integration referred to as loose and tight integration.

Loose integration where ASR target language LM probabilities are updated by using target language translations of source document from SMT and tight integration where machine translation models are integrated with ASR acoustic and language models for optimum ASR decoding.

This paper describes a method for improving the performance of task independent ASR by incorporating knowledge obtained from the source language document. In this case the ASR/SMT integration is performed at the phonetic level in order to facilitate a scenario where the ASR engine can be considered to be “task independent” and is configured separate from the SMT task domain. In general, task independent ASR (ASRTI) performance often suffers from a large number of out-of-vocabulary (OOV) words and, more generally, a language model that is mismatched with respect to the target domain.

The approach presented here augments the model given in Equation 1 in two ways. First, it incorporates knowledge obtained from automatic named entity recognition (NER) performed on the source language text. Incorporating named entity (NE) labels was motivated by the fact that the NE tagged words can in most cases be reliably extracted from well formed text documents in the source language and that they can be reliably translated into the target language. Furthermore, the NE labels themselves can serve as powerful additional side information in dealing with OOVs and LM mismatch associated with task independent ASR. Second, the SMT component is expanded to include a pronunciation lexicon. The ASR system will be treated in this integrated framework as a decoder that produces phone strings that will be aligned with phone strings produced by the SMT system.

The paper is organized as follows. Section 2 describes the SMT, ASR and Named Entity Recognition (NER) systems used in the experiments described here. A brief description of the speech corpus is given in Section 3. The approach for phone level ASR/SMT integration will be presented in Section 4. The experimental results obtained from applying this approach is described in Section 5.

2. System Description
The ASR system used for experiments shown in this paper was developed at the Centre de Recherche Informatique de Montréal (CRIM) [2], and is based on the principles of weighted finite state automata [5]. Each independent block in the ASR system: hidden Markov model topology, acoustic context information, dictionary, and language model is represented as a weighted finite state machine (FSM). Each of these FSMs are then composed to create a single network. Decoding a speech utterance then involves finding the best path through this network that best describes that utterance. It is also possible to obtain a list of N-best hypotheses in the AT&T FSM format for each utterance. All experiments in this paper unless otherwise stated have been performed on these N-best lattices.

The phrase based statistical machine translation system - PORTAGE, used for this work was developed at the National Research Council (NRC) Canada [6]. Finding the target lan-
guage translations $t$ of a source language sentence $f$ involves maximizing $P(e|f)$ which is the log linear combination of four components: target language trigram LM, phrase translation models, distortion model, and word-length feature.

The text based named entity recognizer used in this work is built using a natural language processing tool called Lingpipe that implements a HMM based model [8]. In this case the probability of a tag sequence $G$ over a word sequence $F$ is maximized. Eight main types of named entities are tagged: persons (pers), locations (loc), organizations (org), geo-socio-political groups (gsp), amounts (amount), time (time), and products (prod). Of these only the first four categories are considered for the work in this paper.

### 3. Speech Corpus

The speech data used for this work was collected from three bilingual subjects as an initial study under the Paroles Aux Traducteurs (PAT) project. The main goal of the The PAT project as started by the NRC Interactive Language Technologies group is to develop and evaluate ASR tools allowing translators to efficiently dictate translations [7]. For the pilot corpus three bilingual subjects were asked to record enrollment data, which was used for acoustic adaptation, and to translate two Hansard French texts each. The enrollment data amounted to a total of 18 minutes of speech collected from each of the speakers to perform MAP and MLLR acoustic adaptation. Additionally, development data consisting of 1520 words of read English language text totalling 26 minutes of speech was collected. This was used to estimate log linear weights as described in Section 4. For the evaluation data, the three translators were asked to translate approximately 700 words of text each from the French language Canadian Hansards corpus. More information on the pilot speech corpus can be found in [4].

### 4. Phone Level ASR/SMT Integration

The goal of the techniques described in this section is to improve the performance of a task independent ASR system on target language utterances uttered by a human translator while reading a source language document. Initially ASR is performed on segments of an entire document level utterance that produces both phone lattices as well as word sequence hypotheses for each segment. These segment boundaries are automatically obtained from the speech data by silence detection. The phone lattices and word hypotheses produced by the ASR system for each segment are not initially aligned in any way with the original source language sentences. However, ASR/SMT integration is performed at the sentence level. Therefore, it is necessary to establish a correspondence between the segments processed by the ASR system and the sentences in the source language text. This is accomplished by aligning the source language sentences with the initial target language word string hypotheses produced by the ASR system. This alignment is performed automatically using the Gale and Church algorithm [10].

Text normalization must be performed to address inconsistencies between the ASR output strings, the source language text, and the SMT output. For example, speakers used verbalized punctuation and word expansions of numbers appeared in the ASR output. The word expansions had to be converted to their Non-Standard Words (NSW) form by the process of normalization or conditioning. A modification of the "Text Conditioning Tools" as provided by the LDC [9] was used to perform this text normalization. After normalization, any numbers that were recognized by the ASR system were converted to their numerical form and verbalized punctuation were converted to punctuation marks. For example, "comma" was converted to "," and "nineteen ninety eight", was converted to "1998". However, any errors caused by the ASR system were not converted. This text normalization step was required in order to facilitate the sentence alignment algorithm. After alignment and identifying the utterance segments corresponding to each of the sentences in the source language document, the corresponding phone lattices are concatenated. All experiments described in the following sections are performed on these concatenated phone lattices.

The sentence level process for ASR/SMT integration is performed in three stages. First, a phone level expansion, $q$, is obtained for translated source language text. Second, an optimum alignment between this MT generated phone sequence and the ASR generated phone sequence, $r$, is performed. Third, the most likely target language word sequence, $e$, which reflects the combined acoustic, language, and translation information, is obtained by re-scoring using a target language LM. At the phoneme level, an optimum phone sequence, $q$, is found by maximizing $p(q|f,x)$. The optimum phone sequence, $q$, can be found by applying the naive Bayes approximation. Assuming that the phone string is conditionally independent given the speech, $x$, and the source language text, $f$, then:

$$p(q|f,x) \sim p(q|f)p(q|x), \quad (2)$$

where $p(q|f)$ and $p(q|x)$ are described in Sections 4.1 and 4.2.

### 4.1. Phone Level SMT Decoding

Phone level expansion of translated source language text is performed as follows. NE tags, $g$, are assigned to words in the source language word string, $f$, to optimize the joint probability $p(f,g)$. Then the target language word string, $e$, is obtained by translating the word/tag sequence, < $f$, $g$ >. Finally, the hypothesized phone sequence, $q$, is obtained from $e$ through a pronunciation model, $p(q|e)$. The motivation for producing a phone sequence according to the above steps is that the resulting string is optimum according to the following criterion,

$$\arg\max_q p(q|f) = \arg\max_q \sum_{e',g} p(q,e',g|f)$$

$$\approx \arg\max_q \sum_{e',g} p(q|e')p(e'|g,f)p(g,f) \quad (3)$$

The three terms on the right side of Equation 3 represent the models used in this phonetic expansion. The NE model, $p(g,f)$, was implemented using a hidden Markov model (HMM) based NER system [8] as described in Section 2.

The translation model, $p(e'|g,f)$, was implemented using the NRC PORTAGE phrase based SMT system [6] trained from the Canadian Hansard domain. For a subset of words in $f$ labelled according to NE tags including pers, loc, org, and gsp, translation is performed using the translation phrase tables. This is a simple and reliable way to obtain translations for those words associated with these NE tags. English translation of the entire French language text string is produced using an IBM-2 translation model [6]. Scores derived from the estimated $p(g,f)$ for these words were used to obtain the overall $p(q|f)$ in Equation 3.

The pronunciation model, $p(q|e')$, is based on the frequency of occurrence of pronunciations obtained from a large
pronunciation dictionary or, for words not contained in this dictionary, on a pronunciation model originally developed for a text-to-speech system. It is important to note that the number of words in the overall lexicon are utterance dependent. Dictionary based or rule based word pronunciations exist for all words in the ASR vocabulary and for all words hypothesized for that utterance by the SMT system.

Instead of producing the single most likely phone string that optimizes the model given by Equation 3, a weighted network, $Q$, that transduces hypothesized phone sequences to their associated target language word sequences is created. This network is stored as a weighted finite state transducer in the AT&T FSM format. Section 4.2 describes how this transducer is then used to align the SMT generated phone sequences with the phone sequences, $r$, produced by the ASR system.

### 4.2. SMT/ASR Phone Alignment

It is assumed that for each sentence, a set of hypothesized phone sequences $r$ are produced from the input utterance observation sequence, $x$ by the ASR system. Additionally from the source language text, $f$, another set of hypothesized phone sequences $q$ is generated by the SMT system. The process for augmenting the SMT system with NER and pronunciation model was discussed in Section 4.1. The process for integrating these two separate set of phone lattices generated by SMT and ASR systems by way of aligning the phone string hypotheses is described here.

The goal of this process is to obtain the SMT generated phone sequence, $q$, that optimizes the following criterion,

$$\arg\max_q p(q|x) = \arg\max_q \sum_{r,e} p(q|r)p(r,e|x)$$

$$\approx \arg\max_q \sum_{r,e} p(q|r)p(r,e,x)$$

The probability $p(x|r)$ in the above equation represents the acoustic model probability associated with the ASR engine. The probability $p(q|r)$ represents the probability of generating the target dependent SMT phone sequence, $q$, from the task independent ASR phone sequence, $r$. This corresponds to the alignment probability discussed above and is approximated here by computing a modified Levenshtein distance between the two strings. The probabilities $p(r|e)$ and $p(e)$ are the pronunciation and language models associated with the ASR system. It should be noted that the vocabulary used for the pronunciation model $p(q|e)$ in Equation 3 is larger and includes the vocabulary used for the model $p(r|e)$ in Equation 4. This is in accordance with the fact the ASR system is built for a task independent scenario where as the SMT system is assumed to have prior knowledge of the domain of the document and is therefore task dependent.

An example ASR/SMT phone alignment for a segment of an input utterance is shown in Table 1. The actual word string for this utterance was “The deputy from Nepean Carleton”, of which only a part of the utterance “Nepean Carleton” is shown in the example. The ASR system generated a phone lattice, $R$, as well as the corresponding word sequence hypothesis. Row 2 in Table 1 shows a phone sequence hypothesis, $r$, extracted from lattice $R$, and row 3 shows the corresponding word sequence as decoded by the ASR system. Rows 5 and 7 indicate two phone sequences, $q_1$ and $q_2$ produced by the SMT system and aligned with $r$. The corresponding word sequence hypotheses are shown in rows 6 and 8. The probability of generating either $q_1$ or $q_2$ from $r$ is described by $p(q|r)$ in Equation 4, but is implemented here using a simple Levenshtein distance.

The implementation of SMT/ASR phone alignment is performed by computing the edit distance between the weighted phone lattice, $R$, produced by the ASR system and the weighted phone-to-word transducer, $Q$, produced by the SMT system and described in Section 4.1. This is done here by composing both of these networks with an edit transducer, $T$. The result of this operation is also a phone-to-word transducer:

$$W = R \circ T \circ Q$$

where the path hypotheses are weighted according to the combined probabilities given in Equations 3 and 4.

In this particular example, there are more mismatches between phone sequences $r$ and $q_2$ as compared to $r$ and $q_1$. As a result, the phone sequence $q_1$, and therefore $Word_1$ seems “better” aligned than $Word_2$. However, the combined scores of word/tag sequence, $p(g,f)$, and alignment, $p(q|r)$, assigns a higher score to word sequence $Word_2$ than $Word_1$.

### 4.3. Language Model Re-Scoring

Both SMT and ASR systems incorporate n-gram statistical language models to represent local structure in the target language word sequence, $e$. In [4], a trigram language model, $P_M(e)$, used in the ASR system was combined with a language model, $P_M(e)$, that was estimated using statistics derived from word sequences decoded by the SMT system:

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

The combined LM, $P(e)$, was then used to re-score the word lattices $W$ shown in Equation 5. In Equation 6, $\lambda_M$ and $\lambda_S$ represent interpolation weights that are determined empirically from a development corpus. In the SMT/ASR integration approach described here, the LM described by Equation 6 is used to re-score the word hypotheses contained in $W$. $P_M(e)$ is a bigram LM estimated using the 100 top scoring translation candidates obtained from a given sentence.

### 5. Experimental Results

In this section we describe the implementations of both task dependent and task independent translation dictation scenarios and their results. We show that in the task independent case, by using information from the source document, the ASR word error rate (WER) is reduced to match the WER obtained using a task dependent ASR. Section 5.1 describes the task dependent case and the ASR results. Section 5.2 describes the implementation and results obtained for the task independent scenario. The task domain for these experiments is the transcripts of Canadian parliamentary debates also known as Canadian Hansard.

<table>
<thead>
<tr>
<th>ASR Hypothesis</th>
<th>Word</th>
<th>svm</th>
<th>pin</th>
<th>n</th>
<th>k</th>
<th>e</th>
<th>t</th>
<th>l</th>
<th>ah</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nepean</td>
<td>Car</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SMT Hypothesis</th>
<th>Word</th>
<th>svm</th>
<th>pin</th>
<th>n</th>
<th>k</th>
<th>e</th>
<th>t</th>
<th>l</th>
<th>ah</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nepean</td>
<td>CarTe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Example of hypothesized phone strings and word strings produced by ASR and SMT systems for a segment of the utterance “The deputy from Nepean Carleton”
In all the experiments described here, speech from three translators as collected under the PAT project was used as described in Section 3. All ASR results shown in Table 2 are obtained after acoustic model adaptation and any weights indicated in Section 5.2 were estimated empirically on the development set.

5.1. Task Dependent Baseline Scenario

The task dependent ASR ($ASR_{TD}$) baseline system was created with knowledge of the Hansard task domain. This was achieved by using a 22000 word vocabulary LM trained from the 1996 CSR HUB4 corpus [9] and a 34 million word subset of the English language Canadian Hansards. Although the task domain was known prior to recognition, the actual proceedings used as part of the test data were excluded while building the LM. The $ASR_{TD}$ WER reported in row 1 of Table 2 indicates an average WER of 12.9% across all three speakers.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Spk 1</th>
<th>Spk 2</th>
<th>Spk 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ASR_{TD}$ - Baseline</td>
<td>13.2</td>
<td>12.9</td>
<td>12.6</td>
</tr>
<tr>
<td>Phone Alignment</td>
<td>16.1</td>
<td>20.8</td>
<td>17.3</td>
</tr>
<tr>
<td>SMT-LM Re-score</td>
<td>13.4</td>
<td>11.1</td>
<td>13.0</td>
</tr>
</tbody>
</table>

Table 2: Experimental WER results on pilot data

5.2. Task Independent Scenario

A task independent ASR ($ASR_{TI}$) system having no domain information of the English language Canadian Hansard was built using a vocabulary of 14500 most frequently occurring words obtained only from the HUB4 corpus [9]. The LM used for $ASR_{TI}$ was built from the HUB4 corpus and no Hansard data was used for LM training. The average WER for this system was found to be 19.1% for the three speakers and is shown in row 2 of Table 2.

The phone sequences obtained from the $ASR_{TI}$ were aligned with the phone sequences generated by the SMT as shown in Equations 3 and 4. The resulting phone-to-word transducer, $W$, given in Equation 5 is then re-scored with the HUB4 language model. The WER obtained at this stage is shown in row 3 of Table 2. The average WER across the three speakers was found to be 15.73% corresponding to a decrease in WER of 17.6% relative to $ASR_{TI}$.

The above result can be further improved by re-scoring the phone-to-word transducer with the language model described by the expression in Equation 6. This incorporates the statistics derived from word sequences generated by the SMT system. This results in an average WER of 12.5%. The WER for each of the speakers is shown in row 4 for Table 2. The average WER in this case decreases by 34% relative to the WER obtained in a task independent scenario.

6. Summary

This paper describes a method for improving the performance of task independent ASR in machine aided human translation by integrating ASR and SMT models. In order to incorporate words unknown to the ASR, the ASR and SMT integration is performed at the phone level.

Using the method of SMT phone level decoding followed by ASR/SMT phone alignment, a relative reduction in WER of 17.64% was achieved. By incorporating a LM based on statistics derived from SMT generated word sequences, WER was further reduced by 20.5% relative. This results in an overall relative reduction in WER of 34%. The WER obtained after this stage was nearly identical to the WER obtained from a task dependent system.

Future work will aim at improving upon these results by incorporating translation model probabilities $p(f|e)$ in the decoding process as suggested by Equation 3. In addition, phone transliteration models may also be incorporated which may have particular advantage in the case of proper names. Currently, a larger PAT corpus is also being collected and we intend to apply these techniques to the new PAT corpus.

7. References


