EFFICIENT INTEGRATION OF TRANSLATION AND SPEECH MODELS IN DICTATION BASED MACHINE AIDED HUMAN TRANSLATION

Luis Rodríguez[1], *Aarthi Reddy*[2], *Richard Rose*[2]

[1]Departamento de Sistemas Informáticos, University of Castilla La Mancha
[2]Dept. of Electrical and Computer Engineering, McGill University, Montreal, Quebec
luis.rruiz@uclm.es, aarthi.reddy@gmail.com, rose@ece.mcgill.ca

ABSTRACT

This paper is concerned with combining models for decoding an optimum translation for a dictation based machine aided human translation (MAHT) task. Statistical language model (SLM) probabilities in automatic speech recognition (ASR) are updated using statistical machine translation (SMT) model probabilities. The effect of this procedure is evaluated for utterances from human translators dictating translations of source language documents. It is shown that computational complexity is significantly reduced while at the same time word error rate is reduced by 30%.

Index Terms: speech recognition, machine translation, speech input interfaces

1. INTRODUCTION

There are many language translation applications which place high cognitive load on human translators and also pose rigorous standards of quality on the resulting translation. As a result, it is not expected that fully automated machine translation (MT) approaches will be able to meet the high standards associated with tasks like professional document translation as performed in many translation bureaus. However, there is an extensive literature on machine aided human translation (MAHT) which is motivated by the inability of existing automated systems to meet these demands. This literature describes a large number of scenarios where human translators interact with a machine through a variety of modalities including typing, handwriting, and speaking to improve the efficiency and accuracy of the translation process[1, 2, 3].

Most of these scenarios are addressed by approaches that integrate statistical machine translation (SMT) models with statistical models representing one or more of these interactive modalities. The goal in this integration is to either present suggestions to the human translator based on previous input from the translator or to constrain the solution space when decoding input from the translator. In one particular MAHT scenario, it is assumed that the translator's input is in the form of speech dictation of a target language translation given the text of a source language document. From a Bayesian perspective, integration in this scenario involves decoding the translation, a target language text string, $\hat{\mathbf{e}}$, from a speech feature vector sequence, $\mathbf{x}$, and a source language text string, $\mathbf{f}$, by maximizing the posterior probability:

$$\begin{aligned}\hat{\mathbf{e}} &= \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{e}|\mathbf{x},\mathbf{f}) = \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{e})\cdot p(\mathbf{e}|\mathbf{f}) \\ &= \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{e})\cdot p(\mathbf{f}|\mathbf{e})\cdot p(\mathbf{e}), \qquad (1)\end{aligned}$$

THIS WORK WAS SUPPORTED BY NSERC, THE QUEBEC MDEIE AND THE SPANISH MEC/MICINN PROJECTS MIPRCV (CSD2007-00018) AND ITRANS2 (TIN2009-14511)

where it is assumed in Eq. (1) that $\mathbf{x}$ and $\mathbf{f}$ are conditionally independent given $\mathbf{e}$ [1, 4, 3]. This assumption is considered reasonable since the acoustic model representing the speech vector sequence probability, $p(\mathbf{x}|\mathbf{e})$, and the n-gram statistical language model (LM) representing the target language text, $p(\mathbf{e})$, are trained when configuring an automatic speech recognition (ASR) system, and the translation model, representing the relationship between the source and target language text, $p(\mathbf{f}|\mathbf{e})$, is trained using a parallel corpus associated with configuring an SMT system.

There are many different scenarios that can be applied to combining models for decoding the optimum translation as given by Eq. (1) where each implies different assumptions about system implementation and constraints on computational complexity. One scenario is to assume that, for each source language sentence, the language model of the ASR system can be updated using information derived from the text of that individual sentence [1]. Updating the LM assumes that it is practical to update the n-gram statistical language model (SLM) probabilities prior to decoding the translation speech utterance, $\mathbf{x}$, using translation model probabilities, $p(\mathbf{f}|\mathbf{e})$. This first pass SMT / SLM integration approach is investigated in this paper. A second scenario is to re-rank hypotheses generated by the ASR system using the translation model probabilities as part of a multi-pass decoding scenario [4]. This can be done either by re-ranking the list of m-best string hypotheses or re-scoring word lattices produced by the ASR system.

If complexity is not a factor, both of these approaches should have a similar impact on the word error rate (WER) of the combined system. However, the advantage of the first pass approach is that applying the translation model to constraining search in ASR allows for more aggressive pruning strategies to be applied during search without having to sacrifice WER. This is important if ASR decoding time and lattice size are important issues as they often are in MAHT scenarios. Of course, response times must always be minimized in human interactive scenarios and the size of lattices must be minimized to reduce the overhead associated with re-scoring with other knowledge sources. In fact, this work was initially motivated by the need to generate more compact lattices to reduce the memory associated with operating on the lattices [3].

A method for implementing the combined decoding strategy in Eq. (1) by sentence level updating of the ASR tri-gram LM using the SMT derived translation probabilities is presented in Section 2. Both the efficiency and the performance of this procedure are evaluated in Section 5.2 using a corpus collected from human translators dictating their first draft English language translation of French language Hansard documents [3].

2. FIRST PASS INTEGRATION OF SMT AND SLM

The dictation based MAHT scenario presented in Section 1 involves decoding the optimum target language translated text string from a translator's target language utterance as described in Eq. (1) while taking advantage of the fact that the source language word string, $\mathbf{f} = f_1 \dots f_J$, is assumed to be known prior to recognition. The first pass integration scenario involves updating the ASR SLM probabilities, $p(\mathbf{e})$, with SMT model probabilities, $p(\mathbf{f}|\mathbf{e})$, so that ASR decoding is performed with a new SLM $p_w(\mathbf{e})$. Hence, the decoding problem in Eq. (1) reduces to

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{e}) p_w(\mathbf{e}). \tag{2}$$

This section describes the estimation of $p_w(\mathbf{e})$ and the implications of the first pass integration procedure for the handling out-of-vocabulary words.

2.1. Incorporating Translation Probabilities

In order to use source language information, translation models $p(\mathbf{f}|\mathbf{e})$ are used to modify the generic LM, $p(\mathbf{e})$, and create a modified LM represented by, $p_w(\mathbf{e})$. An n-gram LM, $p(\mathbf{e})$, can be defined simply as the probability of seeing a word given a history of previously occurring words. That is

$$p(\mathbf{e}) = \prod_{i=1}^{I} p(e_i|e_{i-n+1}^{i-1}) \tag{3}$$

where e_i is the current word and e_{i-n+1}^{i-1} is its history. Therefore, re-scoring an n-gram model implies re-scoring every one of these elements. In the method described in this paper, the translation model is used as a re-scoring function. This implies that the probability, $p(\mathbf{f}|e_i)$, for each n-gram conditional probability producing the word e_i needs to be obtained.

It is important to note that statistical translation models [5] usually deal with whole source and target sentences and not with the probability of generating a whole sentence from a single target word. Hence, the usual concepts on which word statistical models rely (alignments, fertility, distortion, etc.) cannot be applied here and a simpler model needs to be adopted. We propose a translation model where each word e_i in the target sentence can be produced only by a single word f_j in the source sentence (similar to what is described in [6]). Specifically, we propose to define the probability, $p(\mathbf{f}|e_i)$, as:

$$p(\mathbf{f}|e_i) \approx \underset{j}{\operatorname{argmax}}\, p(f_j|e_i) \tag{4}$$

That is, we choose for each word e_i the most likely word in the source sentence that can be generated from e_i. From this, we can approach $p(\mathbf{f}|\mathbf{e}) \cdot p(\mathbf{e})$ as:

$$p_w(\mathbf{e}) = p(\mathbf{f}|\mathbf{e}) \cdot p(\mathbf{e}) \approx \prod_{i=1}^{I} p(e_i|e_{i-n+1}^{i-1}) \cdot \prod_{i=1}^{I} \underset{j}{\operatorname{argmax}}\, p(f_j|e_i). \tag{5}$$

Nevertheless, more complex translation models could be also applied in a later stage, following, for instance, a similar strategy as described in [4].

2.2. Updating n-gram LM Probabilities

In this work, the process of n-gram re-scoring involves modifying all the n-gram probabilities using the translation probabilities as described in Eq. (5). Updating the n-gram LM should result in a a new LM with a valid probability distribution. Given an initial smoothed back-off LM, the procedure described in this section is used for updating both the n-gram probabilities and the back-off weights.

The n-gram probabilities are updated as follows. First, the overall probability mass for all the n-gram probabilities sharing a history e_{i-n+1}^{i-1} is stored. Since we are dealing with a smoothed model, this probability mass will be less than one. Second, all these n-gram probabilities are updated as shown in Eq. (5). Finally, these probabilities are normalized according to the previous stored probability. This allows probability to be redistributed according to the translation model, but leaves the discounted probability mass estimated during the original smoothed n-gram training unchanged.

Let e be a target language word and let h be a word sequence corresponding to the history of e for the set E of n-grams in the LM. Let $p(e|h)$ be the original n-gram probability and let $p(f_j|e)$ be the probability of producing the word f_j in the source sentence from the target word e. The new n-gram probability, $p_w(e|h)$, is computed as:

$$p_w(e|h) = \frac{p(e|h) \cdot \underset{j}{\operatorname{argmax}}\, p(f_j|e)}{norm(h)}. \tag{6}$$

The normalization factor is obtained as follows:

$$norm(h) = \frac{\sum_{e'/he' \in E} p_w(e'|h)}{\sum_{e'/he' \in E} p(e'|h)}. \tag{7}$$

Where e' denotes all words seen after history h in the LM.

Once all of the relevant probabilities have been updated, the back-off weights in the n-gram are re-normalized to obtain a true probability distribution. For a history h, the new back-off weight $bow(h)$ is updated as:

$$bow(h) = \frac{1 - \sum_{e'/he' \in E} p(e'|h)}{\sum_{e''/he'' \notin E} p(e''|h)}. \tag{8}$$

Note that the denominator in Eq (8) denotes the lower-order probability estimate.

The translation model probabilities, $p(f_j|e_i)$, are obtained by training an IBM 3 SMT model [5] from a parallel corpus as described in Section 3. The translation table, $t(\mathbf{f}|\mathbf{e})$, from this model was used to represent $p(f_j|e_i)$. Given the source sentence and all the n-grams in the LM, only a small number of target language words are observed in the table as suitable translations for the source language words. As a consequence, the computation of the updated probabilities is actually performed on a small fraction of the n-grams in the LM. This allows for an efficient implementation of the LM update procedure.

There are clearly occasional instances where the translation probabilities derived from the source language text string do not provide significant information for updating the ASR LM. An obvious example would be when many words in the source language text are not included in the SMT vocabulary. It has been found in practice that overall performance can be improved by estimating a sentence level measure of confidence for the translation model and incorporating it in weighting the translation model's contribution to the final combined score. Therefore, the LM is updated according to a log linear model as it is shown in Eq. (9).

$$\log p_w(\mathbf{e}) = \lambda_1 \log p(\mathbf{f}|\mathbf{e}) + \lambda_2 \log p(\mathbf{e}). \tag{9}$$

Where the weights in Eq. (9) are estimated dynamically for each sentence. The estimation of the sentence level translation confidence and its role in updating the weights in Eq. (9) is discussed in [7].

2.3. OOV Words, Pass-Throughs, and Named Entities

It is extremely important to account for words in the source language text string, $\mathbf{f}$, which are out-of-vocabulary (OOV) with respect to the SMT system and words in the target language utterance, $\mathbf{x}$, which are OOV with respect to the ASR system. In both cases, these OOV words often correspond to person names, locations, and other named entity (NE) categories. In SMT, it is often advantageous to allow for OOV words and phrases associated with NE's in the source language text to be "passed-through" and included un-translated in the target language text. Pass-through words for a given source language sentence can be incorporated in the updated LM probabilities, $p_w(\mathbf{e})$ as follows. First, NEs are tagged in the source language sentence using a named entity recognizer (NER). Second, OOV words which are tagged as NEs are included as unigrams in the updated LM. The probability assigned to those words is $p_w(NE) = 1/|\hat{\mathbf{e}}|$, where $|\hat{\mathbf{e}}|$ is an estimator for the target sentence length obtained from the source sentence and the fertility models used in the IBM 3 translation model. Since almost all of the SMT OOV words are also ASR OOV words, this approach was found to dramatically reduce the impact of ASR OOV words in the combined decoder.

3. COMPONENT ASR, SMT AND NER SYSTEMS

In this Section a brief description of the various systems used in the experiments is given. The large vocabulary speech recognition system used in this work is the HTK Toolkit from Cambridge University [8]. The acoustic models used in these experiments were trained from 80 hours of read speech collected from 988 speakers [9]. The models consisted of 6015 clustered states and 96,240 Gaussian densities. The baseline LM used for the experiments conducted here was built from more than 350 million words obtained from Broadcast news [10], North American news corpus [11] and Canadian English language Hansard corpus [12]. The dictionary was built from the 20000 most frequently occurring words in the aforementioned database.

The statistical machine translation system used in this work was obtained through the GIZA++ tool [6] from the Canadian Hansard corpus. About 1 million French/English parallel sentences of the Hansard corpus were used to train the translation model.

The NER system used in this paper was built at the University of Tours [13]. It consists of a series of Finite State Transducer cascades that implement syntactic analysis and information extraction in order to tag every word occurring in the source language text. For this system, the NER system was configured such that it tags words into the following categories: Organization, Person, Product, Location, Classifier, Event, Time/Date, and Other. For the source language test set used in this paper, the NER system gave a 95% recall and 61.2% precision when used to detect NEs.

4. TRANSLATION DICTATION TASK DOMAIN

The experiments were performed on a test corpus that was acquired from translators dictating a first draft translation of a source language document. Each translator is given an excerpt from the Canadian French Hansards and was asked to dictate a first draft translation in English. Prior to dictation, the translator goes through the source language document to mark any unfamiliar words or phrases to be looked up in a dictionary or terminology database. Under this scenario, speech data was collected from 9 bilingual speakers, 3 male and 6 female. Each translator was given a 700-2000 word excerpt from non-overlapping sections of the Canadian French Hansards. The translators dictated the translations of the source language documents, amounting to a total of 456 dictated sentences with an average of 25 words per sentence. Of the 456 sentences thus collected, 200 were used as development data and 256 were used as test data.

5. EXPERIMENTAL SETUP AND RESULTS

This section describes the experiments performed to evaluate the techniques presented in Section 2. First the implementation of the evaluation scenario where document level utterances are aligned with sentence level source language text is described. Second, performance is presented both as the perplexity of the updated ASR LM, word error rate (WER) and oracle WER (OWER) on the test utterances taken from the spoken dictation task domain described in Section 4.

5.1. Implementation of LM Updating Scenario

As mentioned in Section 2.2, the LM used in the ASR system is modified such that it incorporates information from the translation model. In the scenario discussed in this paper each translator was asked to dictate translation of an entire source language document. However, update of the LM is performed here for each sentence in the source language document using translation probabilities derived at the sentence level. In order to do this, first the entire dictated translation was segmented into smaller utterances based on silence intervals in the utterance. Then, ASR was performed on these utterances giving a baseline transcription of the dictated translation. This transcription not only gave baseline word error rate (WER) results, but also provided a means for performing sentence alignment. The sentence alignment for this corpus is described in greater detail in [3].

Once the sentence alignment is performed, translation model probabilities, $p(\mathbf{f}|\mathbf{e})$ are obtained for each sentence in the source language document. These are used to update the baseline LM, $p(\mathbf{e})$, for that sentence in order to obtain the updated LM, $p_w(\mathbf{e})$, as described in Section 2. Hence, for each sentence in the source language document, an updated LM is produced to incorporate translation model probabilities for that sentence. This modified LM is used to perform ASR on translation utterances corresponding to that source language sentence. It is important to note that, using the procedure for local updating and normalizing of n-gram probabilities described in Section 2.2, the sentence level updating can be very efficient. Only a very small percentage of LM parameters are updated for a given sentence.

The baseline LM used in the experiments is different from the updated LMs in two ways. First, as discussed in Section 2.3, NEs that are OOVs to the original LM are included in the unigram model of the updated LM as part of the updating process. Consequently, a closed vocabulary was employed for the updated LMs, whereas the baseline LM is an open vocabulary model. This inclusion of NEs in the unigram model of the closed vocabulary updated LMs causes a significant decrease in WER as discussed in Section 5.2. Second, as mentioned above, the updated LMs are specific for each sentence in the document. The baseline LM on the other hand is a generic model that is the same for every sentence.

5.2. Results

The performance of the updated LM in terms of both perplexity and ASR WER are reported in Table 1. The first row of Table 1 displays the LM test set perplexity for the baseline and sentence updated LMs. The perplexity for the sentence updated LM is obtained by accumulating the sentence level perplexity estimates obtained for

Performance measure	Baseline LM	Updated LMs
Perplexity	64.7	32.6
WER	16.5	11.7
Oracle WER	6.3	3.8

Table 1. Test set perplexity and WER performance for translation dictation domain test

each sentence specific updated LM. As can be seen, a 50% relative decrease in perplexity is achieved by updating the LM. The second row of Table 1 displays the WER obtained using the baseline LM and the updated LM. A 29.1% relative decrease in WER is obtained for updated LM relative to the baseline LM.

The LM updating procedure is also evaluated in terms of the OWER which provides a measure of the best possible WER that could be obtained by re-scoring the ASR lattices generated using the updated LM. The OWER is computed by aligning the reference string for each utterance with the word lattice generated for that utterance. The third row of Table 1 shows that the updated LM provides a 39% reduction in OWER relative to the baseline LM. It should be noted here that the experiments to compute OWER were performed after ensuring that the lattices generated by baseline LM and updated LMs were of the same size. The results in the third row of Table 1 show that even though the lattices generated by the two systems are of the same size, the lattices generated using the updated LMs are richer and contain more relevant data than lattices generated by the baseline system. This implies potentially more efficient implementations of procedures for lattice re-scoring in this domain since the updated LM could produce much smaller lattices with equivalent OWERs.

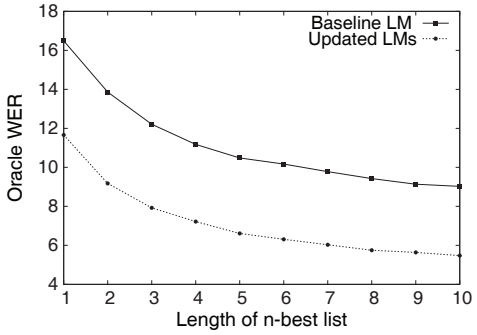

Fig. 1. Oracle WER for Lattices with N-best hypotheses

The plot in Figure 1, provides a description of how the OWER depends on the lattice size. The horizontal axis of Figure 1 represents the size of the lattices after they were pruned to retain only the top N highest scoring paths. The solid line in the plot shows the lattice OWER obtained using the baseline LM and the dotted line in the plot shows the lattice OWER obtained using the updated LM. The OWER scores for 1-best list in Figure 1 correspond exactly to the WER results in the second row of Table 1. This is because the WER is computed by using the first best hypothesis generated by the ASR system. If the solid and dotted lines in Figure 1 are extended to include all hypotheses in the lattice, then the OWER values would correspond exactly with the values in line 3 of Table 1. The plot in Figure 1 demonstrates improved efficiency in that, for a given lattice size, the OWER decreases for the updated LM by approximately 3-4 absolute percentage points.

6. SUMMARY AND CONCLUSIONS

The language model updating approach presented here for a translation domain dictation task provides a 50% decrease in test set perplexity and around a 29% decrease in ASR WER with respect to performance obtained using a baseline trigram LM. The approach provides an efficient way to incorporate knowledge of the source language text directly in the ASR decoder by updating target language word n-gram probabilities based on the probability of those words as predicted by the translation model. It also provides a potential mechanism for introducing new words in the ASR LM with the help of named entity tags derived from the source language text.

Combining multiple information sources in a ASR lattice re-scoring scenario for improving overall MAHT performance has been investigated in [3]. It was found there that the memory and computational requirements of these techniques can become very large unless the size of the lattices can be constrained through pruning. The 39% improvement in oracle WER for the lattices produced by the updated LM relative to the baseline LM suggests that, even when these lattices have been aggressively pruned, they may provide a solution space that is rich enough for these lattice re-scoring techniques to be both efficient and effective. This is proved by the plot in Figure 1, which shows that for any given N-best hypotheses lattice, the OWER for lattices obtained using updated LMs will be lower than the OWER obtained when using baseline LM.

7. REFERENCES

[1] P. Brown, S. Chen, S. Della Pietra, V. Della Pietra, S. Kehler, and R. Mercer, "Automatic speech recognition in machine aided translation," *Computer Speech and Language*, vol. 8, pp. 177–187, 1994.

[2] E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera, and C. Martínez, "Computer-assisted translation using speech recognition," *IEEE Transaction on ASLP*, vol. 14, no. 3, pp. 941–951, 2006.

[3] A. Reddy and R. C. Rose, "Integration of statistical models for dictation of document translations in a machine-aided human translation task," *IEEE Trans. on ASLP*, vol. 18, no. 8, pp. 2015 –2027, nov. 2010.

[4] S. Khadivi and H. Ney, "Integration of automatic speech recognition and machine translation in computer-assisted translation," *IEEE Trans. on ASLP*, vol. 16, no. 8, pp. 1551–1564, Nov. 2008.

[5] Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert. L. Mercer, "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, pp. 263–311, 1993.

[6] Franz Josef Och and Hermann Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[7] Luis Rodríguez-Ruiz, *Interactive Pattern Recognition Applied to Natural Language Processing*, Ph.D. thesis, Universidad Politécnica de Valencia, Valencia (Spain), June 2010, Advisors: Enrique Vidal and Ismael García-Varea.

[8] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[9] John Garofalo, David Graff, Doug Paul, , and David Pallett, "CSR-I (WSJ0) other," 1993.

[10] David Graff, John Garofolo, Jonathan Fiscus, William Fisher, and David Pallett, "1996 English broadcast news speech (HUB4)," LDC Catalog No.: LDC97S44, ISBN: 1-58563-109-4, 1997.

[11] David Graff, "North American News text corpus," LDC Catalog No.:LDC95T21, 1995.

[12] Salim Roukos, David Graff, and Dan Melamed, "Hansard French/English," LDC Catalog No.:LDC95T20, 1995.

[13] N. Friburger and D. Maurel, "Finite-state transducer cascades to extract named entities in texts," *Theoretical Computer Science*, vol. 313, no. 1, pp. 93–104, 2004.