DEVELOPMENT OF A SPEECH RECOGNITION SYSTEM FOR ICELANDIC USING MACHINE TRANSLATED TEXT

Arnar Jensson, Koji Iwano, Sadaoki Furui

Tokyo Institute of Technology

arnar@furui.cs.titech.ac.jp; iwano@furui.cs.titech.ac.jp; urui@cs.titech.ac.jp

ABSTRACT

Text corpus size is an important issue when building a language model (LM). This is a particularly important issue for languages where little data is available. This paper introduces an LM adaptation technique to improve an LM built using a small amount of task dependent text with the help of a machine-translated text corpus. Icelandic word error rate experiments were performed using data, machine translated (MT) from English to Icelandic on a sentence-by-sentence and word-by-word basis. The baseline word error rate was 49.6%. LM interpolation using the baseline LM and an LM built from sentence-by-sentence translated text reduced the word error rate significantly to 41.9%.

Index Terms— Language Model Adaptation, Automatic Speech Recognition, Machine Translation, Sparse Text Corpus, Resource Deficient Languages.

1. INTRODUCTION

Statistical language modeling is well known to be very important in large vocabulary speech recognition but creating a robust language model (LM) typically requires a large amount of training text. Therefore it is difficult to create a statistical LM for resource deficient languages. In our case we would like to build an Icelandic speech recognition dialogue system in the weather information domain. Since Icelandic is a resource deficient language there is no large text data available for building a statistical LM, especially for spontaneous speech.

Methods have been proposed in the literature to improve statistical language modeling using machine translated text from another source language such as in [1], [2], [3] and [4]. The applications presented in [1], [2] and [4] are all different from our target application while [3] is similar but represents results only with perplexity values and no speech recognition results. The above mentioned systems all use statistical machine translation (MT) trained on a parallel text corpus often expensive to obtain and unavailable for resource deficient languages.

MT methods other than statistical MT are also available, such as rule-based MT systems. A rule based MT system can be based on a sentence-by-sentence (SBS) translation or word-by-word (WBW) translation. WBW translation only requires a dictionary, already available for many language pairs, whereas rule based SBS MT needs more extensive rules and therefore more expensive to obtain. The WBW approach is expected to be successful only for closely related languages.

In [5], we proposed a method to improve the LM built on a task-dependent corpus using MT which is similar to [3]. This paper extends our machine translation experiments. The dictionary used to translate WBW is now created automatically by an MT system. This paper also introduces a rule based SBS machine translated texts from English to Icelandic. The evaluation speech corpus is extended in this paper from 0.1 hour to 2.0 hours.

2. ADAPTATION METHOD

Our method involves adapting a task dependent LM that is created from a sparse amount of text using a large translated text (TRT), where TRT denotes the machine translation of the rich corpus (RT), preferably in the same domain area as the task. This involves two steps shown graphically in Figure 1. First of all the sparse text is split into two, a training text corpus (ST) and a development text corpus (SD). A language model LM1 is created from ST, and LM2 from TRT. The TRT can either be obtained from SBS or WBW translation. The SD set is used to optimize the weight (λ) used in Step 2. Step 2 involves interpolating LM1 and LM2 linearly using Equation (1),

$$P_{comb}(\omega_i|h) = \lambda \cdot P_1(\omega_i|h) + (1-\lambda)P_2(\omega_i|h), \qquad (1)$$

where h is the history. P_1 is the probability from LM1 and P_2 is the probability from LM2.

The final perplexity or word error rate (WER) value is calculated using an evaluation text set or speech evaluation set ($Eval$) which is disjoint from all other data sets.

3. EXPERIMENTAL WORK

3.1. Experimental Data

The weather information domain was chosen for the Icelandic experiments and translation from English ($rich$) to Icelandic ($sparse$) using WBW and SBS. For the experiments the Jupiter corpus [6] was used. It consists of 67116 unique sentences gathered from actual users' utterances. A set of 2460 sentences were manually translated from English to Icelandic and split into ST, SD and $Eval$ sets as shown in Table 1. 63116 sentences were used as RT.

A unique word list was made out of the Jupiter corpus and machine translated using [7] in order to create a dictionary. This MT is a rule based system. The dictionary was then used to translate RT into TRT_{WBW}. Another translation TRT_{SBS} was created by SBS machine translation using [7]. Names of places were identified and then replaced randomly with Icelandic place names for both

Table 1. *Data sets*

Corpus Set	Sentences	Words	Unique Words
ST	1500	8591	805
SD	300	1870	342
$Eval$	660	3767	554

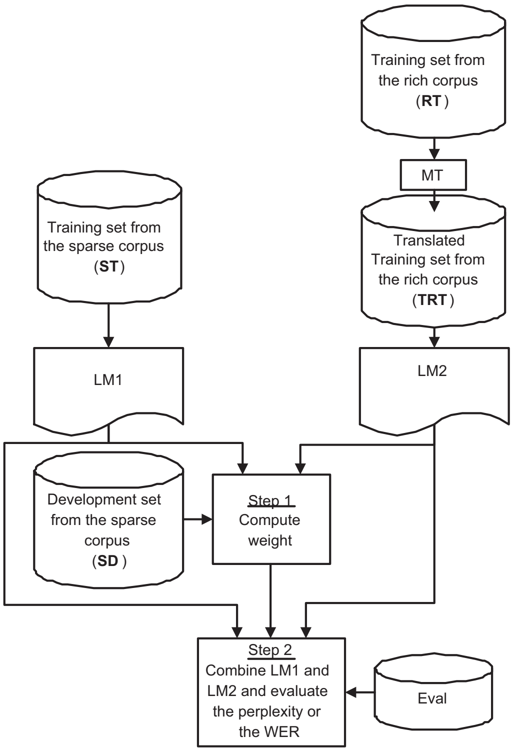

Fig. 1. Data diagram

Table 2. *Translated data sets*

Corpus Set	Sentences	Words	Unique Words
TRT_{WBW}	62962	440347	3396
TRT_{SBS}	62996	406814	7312

TRT_{WBW} and TRT_{SBS}, since the task is in the weather information domain. Table 2 shows some attributes of the WBW and SBS translated Jupiter texts. The reason why the number of sentences in Table 2 does not match the number of sentences found in the RT set is because of empty translations.

A 1-gram and 2-gram translation evaluation using BLEU [8] was performed on 20 sentences created from both the SBS and the WBW machine translators, using two human references. The 1-gram and 2-gram BLEU evaluation was 0.47 and 0.23 for WBW MT respectively. The 1-gram and 2-gram BLEU evaluation was 0.61 and 0.43 for SBS MT respectively.

A phonetically balanced (PB) Icelandic text corpus, the Jensson PB corpus [5], was used to create an acoustic training corpus. The training corpus consists of 3.8 hours of speech from 13 male and 7 female speakers. An evaluation corpus was recorded using sentences from the previously explained $Eval$ set. 2 hours of read speech was recorded from 10 male and 10 female speakers. None of the speakers in the evaluation speech corpus are in the acoustic training corpus.

Table 3. *Experimental setup*

Experiment nr.	TRT Corpus	Vocabulary
Experiment 1	None	V_{ST}
Experiment 2	None	$V_{ST} + V_{TRT_{WBW}}$
Experiment 3	TRT_{WBW}	V_{ST}
Experiment 4	TRT_{WBW}	$V_{ST} + V_{TRT_{WBW}}$
Experiment 5	None	$V_{ST} + V_{TRT_{SBS}}$
Experiment 6	TRT_{SBS}	V_{ST}
Experiment 7	TRT_{SBS}	$V_{ST} + V_{TRT_{SBS}}$
Experiment 8	$TRT_{WBW} + TRT_{SBS}$	$V_{ST} + V_{TRT_{WBW}} + V_{TRT_{SBS}}$

3.2. Experimental Setup

In total eight different experiments were performed. The experimental setup can be viewed in Table 3. Experiment 1 used no translation and its vocabulary consisted only from the unique words found in the ST set, creating V_{ST}, and is therefore considered as the $baseline$. Experiments 2 to 4 used WBW machine translated data. Experiment 2 used no TRT corpus but used the unique words found in TRT_{WBW}, creating the vocabulary $V_{TRT_{WBW}}$. This was done in order to find the impact of including only WBW translated vocabulary. Experiment 3 used the WBW machine translated corpus along with the V_{ST} vocabulary. Experiment 4 used the WBW MT along with the combined vocabulary from the ST and TRT corpora.

Experiments 5 to 8 used SBS machine translated data. Experiment 5 used no TRT corpus but used the unique words found in TRT_{SBS}, creating the vocabulary $V_{TRT_{SBS}}$. This was done in order to find the impact of including only SBS translated vocabulary. Experiment 6 used TRT_{SBS} as the TRT corpus without adding translated words to the vocabulary. Experiment 7 used the SBS MT along with the combined vocabulary found from the ST and TRT corpora. Experiment 8 used both information from the SBS and WBW MT. Using WBW translated data along with SBS MT can be done since the dictionary used to create the WBW MT was created using the SBS MT.

The ST set size varied from 100 to 1500 sentences for all the experiments. In the following text ST^n corresponds to a subset of the ST set where n is the number of sentences used. Experiments with no ST set included, ST^0, was also performed on Experiment 4, Experiment 7 and Experiments 8. All LMs were built using 3-grams with Kneser-Ney smoothing. The WER experiments were performed three times with different, randomly chosen sentences, creating each ST and SD set, in order to increase the accuracy of the results. An average WER was calculated over the three experiments. This increases accuracy when comparing different experiments especially when the ST set is very sparse.

3.3. Results

The WER results from Experiment 1, Experiment 2, Experiment 3 and Experiment 4 are shown in Figure 2. When no manual ST sentences are present and only WBW machine translated data is used, Experiment 4 gives WER of 67.6%. When 100 ST sentences are used in Experiment 1, the WER $baseline$ is 49.6%. Experiment 4 reduces the WER to 46.6% when adding the same number of ST sentences. As more ST sentences are added, the improvement in Experiment 4 reduces and converges with the $baseline$ when 500 ST sentences are added to the system. Experiment 2 and Experi-

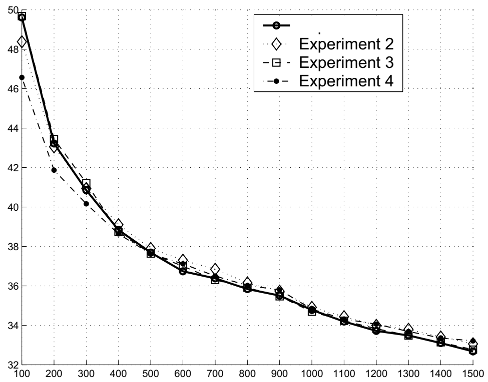

Fig. 2. Word error rate results using the $baseline$ from Experiment 1 and the interpolated WBW machine translated results from Experiment 2, Experiment 3 and Experiment 4.

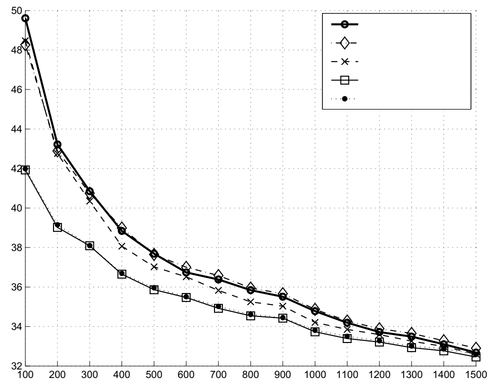

Fig. 3. Word error rate results using the $baseline$ from Experiment 1 and the interpolated SBS machine translated results from Experiment 5, Experiment 6, Experiment 7 and Experiment 8.

Table 4. *Perplexity results*

	ST^n				
Experiment nr.	ST^0	ST^{100}	ST^{500}	ST^{1000}	ST^{1500}
Experiment 1	NA	30.7	26.4	26.3	26.5
Experiment 3	NA	29.4	26.0	26.1	26.3
Experiment 6	NA	26.6	25.3	25.3	25.4
Experiment 2	NA	58.2	34.2	31.9	30.8
Experiment 4	664.6	50.2	32.6	30.7	29.9
Experiment 5	NA	88.9	43.5	37.7	35.3
Experiment 7	287.0	61.1	38.4	34.1	32.5
Experiment 8	274.8	61.6	38.5	34.4	32.6

Table 5. *OOV results*

	ST^n				
Vocabulary	ST^0	ST^{100}	ST^{500}	ST^{1000}	ST^{1500}
V_{ST^n}	NA	14.0	6.8	5.5	4.6
$V_{ST^n} + V_{TRT_{WBW}}$	26.8	8.4	4.8	4.0	3.4
$V_{ST^n} + V_{TRT_{SBS}}$	9.2	4.4	2.6	2.5	2.2
$V_{ST^n} + V_{TRT_{WBW}} + V_{TRT_{SBS}}$	9.0	4.4	2.6	2.4	2.2

ment 3 give a small improvement over the $baseline$ when the ST set is small but converges quickly as more ST sentences are added.

The WER results from Experiment 5, Experiment 6, Experiment 7 and Experiment 8 along with the $baseline$ in Experiment 1, are shown in Figure 3. When no ST sentences are present and only SBS or SBS and WBW machine translated data is used, Experiment 7 and Experiment 8 gives WER of 56.5% and 56.8% respectively. When 100 ST sentences are added to the system and interpolated with the TRT corpus in Experiment 7, the WER is 41.9%. Experiment 8 gives a 42.0% WER when 100 ST sentences are added to the system. As more ST sentences are added the relative improvement reduces. When 1500 ST sentences are used, the WER in Experiment 7 gives 32.5% compared to 32.7% when the $baseline$ is used. When the translated vocabulary is alone added, Experiment 5 does not give any significant improvement over the $baseline$. When the vocabulary is fixed to the ST set and TRT_{SBS} is used as the TRT set, Experiment 6 gives a small improvement over the $baseline$. When ST composes of 1500 sentences, the interpolation in Experiment 6 gives a WER of 32.6%. Each experiment was performed three times with different ST and SD set, and the average WER calculated, as explained before. For example, Experiment 7 shown in Figure 3 gives WER 41.8%, 41.9% and 42.1%, with an average of 41.9%, when 100 ST sentences are used.

Perplexity and out-of-vocabulary (OOV) results are shown in Table 4 and Table 5 respectively for some ST values. The perplexity results for Experiment 1, Experiment 3 and Experiment 6 should be compared together since the vocabulary is the same for those experiments, V_{ST}. Experiment 2 and Experiment 4 have the same vocabulary, V_{ST} combined with $V_{TRT_{WBW}}$ and should be compared together. For the same reason Experiment 5 and Experiment 7 should be compared toghether having the same vocabulary, V_{ST} combined with $V_{TRT_{SBS}}$. As shown in Table 4 all perplexity results get improved when a TRT corpus is introduced and interpolated with the corresponding ST set. The OOV rate shown in Table 5 is reduced by adding the unique words found in the TRT set to V_{ST} as expected. When the system corresponds of 100 ST sentences, the OOV rate is reduced from 14.0% to either 8.4% or 4.4% using WBW or SBS MT respectively. Not applicable (NA) are displayed in Table 4 and Table 5 for experiments that have no ST sentences and are based solely on the V_{ST} vocabulary and/or are not using any TRT corpus, and therefore do not have data to carry out the experiment.

4. DISCUSSION

The improvement of the Icelandic LM with translated English text/data was confirmed by reduction in WER by using either WBW or SBS MT. Experiment 1 should be compared with the other experiments since Experiment 1 does not assume any foreign translation. When the $baseline$ in Experiment 1 is compared with the interpolated results using WBW MT in Experiment 4, we get a WER 49.6% reduced to 46.6% respectfully, a 6.0% relative improvement when using 100 ST sentences. The relative improvement reduces as more ST sentences are added to the system and converges to the $baseline$ when 500 ST sentences are added to the system. Neither Experiment 2 nor Experiment 3 gives any significant improvement over the $baseline$. This along with the results in Experiment 4 suggests that when WBW translated data is available, both the translated corpus and its vocabulary should be added to the system when the ST sentences are sparse.

When the $baseline$ is compared with the interpolated results using SBS MT in Experiment 7, we get a WER 49.6% reduced to 41.9% respectfully, a 15.5% relative improvement when 100 ST sentences are added to the system. Improvements by merging the vocabulary from the TRT_{SBS} and V_{ST} is confirmed by comparing Experiment 6 and Experiment 7 for all ST sets. The WER improvement of the SBS MT over the WBW MT is confirmed for all the ST sets as the BLEU evaluation results in Section 3.1 suggests. This can be seen by comparing Experiment 4 in Figure 2 with Experiment 7 in Figure 3. The improvement is as well confirmed with perplexity results when Experiment 3 and Experiment 6 are compared in Table 4.

5. CONCLUSIONS

The results presented in this paper show that an LM can be improved considerably using either WBW or SBS translation. In order to get significant improvement a good (high BLEU score) MT system is needed. The WBW translation is especially important for resource deficient languages that do not have SBS machine translation tools available. It is believed that a high BLEU score can be obtained with WBW MT for very closely related language pairs and between dialects. Future work involves applying the rule based WBW and SBS translation methods to a larger domain such as broadcast news. Future work also involves an investigation of methods such as the ones described in [9], [10] and [11] that selects a relevant subset from a large text collection such as the World Wide Web to aid sparse target domain. These methods assume that a large text collection is available in the target language but we would like to apply these methods to extract sentences from the TRT corpus.

6. ACKNOWLEDGEMENTS

We would like to thank Drs. J. Glass and T. Hazen at MIT and all the others who have worked on developing the Jupiter system. We also would like to thank Dr. Edward W. D. Whittaker for his valuable input. Special thanks to Stefan Briem for his English to Icelandic machine translation tool and allowing us to use his machine translation results. This work is supported in part by 21st Century COE Large-Scale Knowledge Resources Program.

7. REFERENCES

[1] Khudanpur, S. and Kim, W., "Using Cross-Language Cues for Story-Specific Language Modeling", *Proc. ICSLP*, Denver, CO, vol. 1, pp. 513-516, 2002.

[2] Kim, W. and Khudanpur, S., "Cross-Lingual Latent Semantic Analysis for Language Modeling", *Proc. ICASSP*, Montreal, Canada, vol. 1, pp. 257-260, 2004.

[3] Nakajima, H., Yamamoto, H., Watanabe, T., "Language Model Adaptation with Additional Text Generated by Machine Translation", *Proc. COLING*, vol. 2, pp. 716-722, 2002.

[4] Paulik, M., Stuker S., Fugen C., Schultz T., Schaaf T. and Waibel A., "Speech Translation Enhanced Automatic Speech Recognition", ASRU, San Juan, Puerto Rico, 2005.

[5] Removed for blind review.

[6] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T. and Hetherington, L., "JUPITER: A Telephone-Based Conversational Interface for Weather Information", *IEEE Trans. on Speech and Audio Processing*, 8(1):100-112, 2000.

[7] Briem, S., "Machine Translation Tool for Automatic Translation from English to Icelandic", *http://www.simnet.is/stbr/*, Iceland, 2007.

[8] Papineni, K., Roukos, S., Ward T. and Zhu W., "BLEU: a Method for Automatic Evaluation of Machine Translation", *Proc. ACL*, PA, pp. 311-318, 2002.

[9] Sarikaya, R., Gravano, A. and Gao, Y., "Rapid Language Model Development Using External Resources for New Spoken Dialog Domains", *Proc. ICASSP*, vol. 1, pp. 573-576, 2005.

[10] Sethy, A., Georgiou, P. and Narayanan, S., "Selecting Relevant Text Subsets from Web-Data for Building Topic Specific Language Models", *Proc. ACL*, pp. 145-148, 2006.

[11] Klakow, D., "Selecting articles from the language model training corpus", *Proc. ICASSP*, vol. 3, pp. 1695-1698, 2000.