# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | |
| Title(English) | Improvement of language model adaptation using machine-translated text |
| 著者(和文) | ｼﾞｪﾝｿﾝ ｱｰﾅｰ, ウィッタッカー エドワード, 岩野 公司, 古井 貞熙 |
| Authors(English) | Arnar Jensson, Edward Whittaker, Koji Iwano, Sadaoki Furui |
| 出典(和文) | 日本音響学会2005年秋季講演論文集, Vol. , No. 2-1-4, pp. 43-44 |
| Citation(English) | , Vol. , No. 2-1-4, pp. 43-44 |
| 発行日 / Pub. date | 2005, 9 |

# Improvement of Language Model Adaptation Using Machine-Translated Text *

◎ Arnar Thor Jensson, Edward W. D. Whittaker, Koji Iwano, Sadaoki Furui
(Tokyo Institute of Technology)

## 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.

However, using text translated from other languages may possibly improve the resource deficient LM either using sentence-by-sentence ($SBS$) translation or word-by-word ($WBW$) translation. $WBW$ translation only requires a dictionary whereas $SBS$ machine translation ($MT$) needs a large sentence-aligned parallel corpus, which is expensive to obtain, to train the $MT$ system. The $WBW$ approach is expected to be successful only for closely related languages.

LM adaptation with target task machine-translated text is addressed in [3] but without speech recognition experiments.

In this paper, we expand our $WBW$ experiments presented in [4] by adding more speech evaluation data. The technique described in [4] improves the LM built on a task-dependent corpus using $MT$ which is similar to [3]. Adaptation of $WBW$ translation from English to Icelandic is presented using word error rate ($WER$) obtained by speech recognition experiments.

## 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 translation of the rich corpus ($RT$), in the same domain area as the task. This involves two steps shown graphically in Fig. 1. First 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 $SD$ set is used to optimize the weight ($\lambda$) used in Step 2.

Step 2 involves first optionally combining the $ST$ and the $SD$ corpora and building a new language model, LM3 from them. LM3 and LM2 are then linearly interpolated using Equation (1),

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

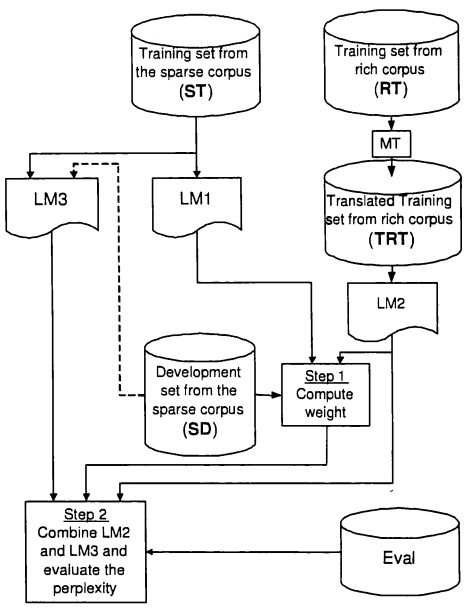


Fig. 1 Data diagram

where $h$ is the history, $P_1$ is the probability from either LM1 or LM3 and $P_2$ is the probability from LM2.

The final perplexity value is calculated using the evaluation set ($Eval$) which is a disjoint from all other data sets.

## 3 Experimental Data

The weather information domain was chosen for the Icelandic experiments and translation from English ($rich$) to Icelandic ($sparse$) using $WBW$. For the experiments the Jupiter corpus [5] was used. It consists of 67116 unique sentences gathered from actual users' utterances. A set of 1500 sentences were manually translated from English to Icelandic and split into $SD$ (300 sentences), $Eval$ (200 sentences) and $ST$ (1000 sentences). 63116 sentences were used as $RT$.

A list of all unique words was then created from the Jupiter corpus and manually translated. Names of places were identified and then replaced randomly with Icelandic place names since the task is in the weather information domain. The English to Icelandic word list was then used to automatically translate $RT$ to create $TRT$.

* 機械翻訳されたテキストを用いた言語モデル適応の改良
アーナー ジェンソン，エドワード ウィッタッカー，岩野公司，古井貞熙（東工大）

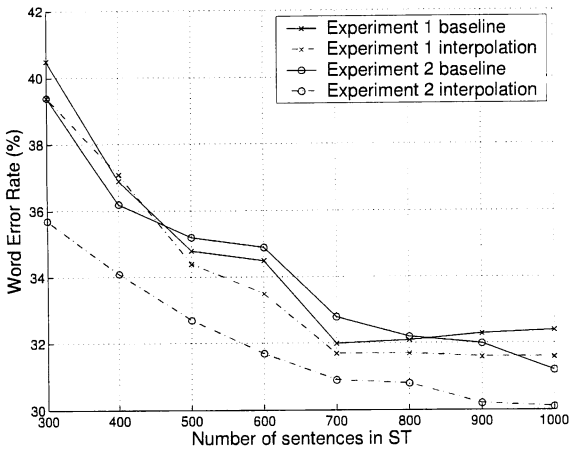


Fig. 2 Word error rate results from Experiment 1 and Experiment 2.

The acoustic model was trained on 13 male and 7 female speakers, a total of 3.8 hours of spoken data. The speech evaluation corpus consists of 200 sentences spoken by 3 male and 2 female speakers. Total length of the spoken evaluation corpus is 32 minutes. Tri-gram LMs were used throughout.

## 4 Results

Two different experiments were performed. The $SD$, $Eval$ and $TRT$ sets were common for both the experiments but the $ST$ set size varied from 300 to 1000 sentences and the vocabulary varied. Interpolation of the language models was done slightly differently to that explained in Section 2. If the $SD$ corpus were added to the $ST$ corpus to make LM3, the weights calculated in Step 1 would be inaccurately optimized for the combined set especially when the $ST$ corpus is small. Therefore LM1 was used instead of LM3. The optimization of the weights when $ST$ and $SD$ are combined into LM3 is postponed for future work.

**Experiment 1** used the unique words found in the $ST$ set as the vocabulary, $V_{ST}$. The results are shown in Fig. 2. The $WER$ improvement is positive for all the $ST$ conditions except when $ST$ comprises 400 sentences.

**Experiment 2** used the vocabulary from the $TRT$ set combined with $V_{ST}$. The results are shown in Fig. 2. As expected the $WER$ improvement is gradually reduced as more manually transcribed data is added to the $ST$ set.

The improvement of the Icelandic LM with translated English data was confirmed by reduction in $WER$. "Experiment 1 baseline" indicates that, when 300 sentences were used as $ST$, the $WER$ was 40.5%. "Experiment 2 interpolation" indicates that, when 300 sentences were used as $ST$, the $WER$ was 35.7%, which is 12% relative improvement from "Experiment 1 baseline". The relative improvement decreased to 7% when 1000 sentences were used as $ST$.

## 5 Conclusions

The results presented in this paper show that a LM can be improved considerably by using $WBW$ translation. The $WBW$ translation is especially important for resource deficient languages such as Icelandic that do not have sentence-by-sentence $MT$ tools available. It is possible to create a dictionary for many language pairs and the work for applying $WBW$ translated text is reduced if the translated corpus is large and the manually created dictionary needed is small.

Future work involves evaluation with more speakers and solving the weight calculation when the $ST$ and the $SD$ corpora are added together. Adapting $WBW$ translated class models will also be examined.

## 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. This work is supported in part by 21st Century COE Large-Scale Knowledge Resources Program.

## References

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

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

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

[4] Jensson, A., Whittaker, E., Iwano, K. Furui, S., "Language Model Adaptation for ASR Using Machine-Translated Data", *IEICE Technical Report*, Morioka, Japan, 2005, vol 105, no 132, pp. 19-23.

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

[6] Jensson, A., Whittaker, E., Furui, S., "Improving a Language Model Using Machine Translated Data", *ASJ Spring Meeting*, Tokyo, Japan, 2005, vol 1, pp. 51-52.