論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	WFST-based Icelandic ASR Using Machine Translation
著者(和文)	ｼﾞｪﾝｿﾝ ｱｰﾅｰ, 岩野 公司, 古井 貞熙
Authors(English)	Arnar Jensson, Koji Iwano, Sadaoki Furui
出典(和文)	日本音響学会2008年秋季講演論文集, , No. 3-1-3, pp. 83-84
Citation(English)	, , No. 3-1-3, pp. 83-84
発行日 / Pub. date	2008, 9

WFST-based Icelandic ASR Using Machine Translation *

◎ Arnar Thor Jensson, 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 using a word-by-word (WBW) translation. WBW translation only requires a dictionary and is expected to be successful only for closely related languages.

In [1], we proposed a method to improve the LM built on a task-dependent corpus using MT which is similar to [2] where we treat Icelandic as a resource deficient language and English as resource rich language. In this paper we propose a weighted finite state transducer (WFST) based speech recognition system where the input is in Icelandic and the output is in either Icelandic or English using a WBW translation transducer. Our method is similar to the one explained in [3] where the input and output languages are the same but with different speaking styles or dialects. Having a recognition output in resource rich language when the input is in a resource deficient language can be important since if a backend processor already exists in the resource rich language then the same backend system can be used for creating a response for the resource deficient language.

2 Method

Our method involves adapting a task dependent LM that is created from a sparse text (ST) in the resource deficient language with a LM that is created from a rich text (RT) in the resource rich language, preferably in the same domain area as the task. WFST network is used for recognizing the input combining the language models using a WBW translation transducer. Our method involves two different kinds of setup. One, demonstrated in (1), where the output language is the same as the input language. The other, demonstrated in (2), where the output language is different from the input language.

$$H \circ C \circ L \circ G_{ST} \circ \pi(T \circ G_{RT}) \quad (1)$$

$$H \circ C \circ L \circ G_{ST} \circ T \circ G_{RT} \quad (2)$$

Here H maps HMM states to context-dependent phones. C represents a transduction from context-dependent phones to context-independent phones. L is lexicon converted to a WFST that will map context-independent phone sequence to words. G, in general, is a WFST that represents the language model, for example an N-gram model that maps word to N-gram weighted word sequences. G_{ST} represents the G for the sparse text and G_{RT} represents the G for the rich text. T is a translation transducer that maps words from the resource deficient language to the resource rich language. The composition operator (denoted by $\circ$) combines WFSTs together.

$\pi()$ is a projection operator that copies the input symbols of each arc to the output symbols. Although all arcs in the $T \circ G_{RT}$ network have resource deficient language word input and resource rich language word output, resource deficient language N-gram translated from resource rich language N-gram can be obtained by using the projection operator on $T \circ G_{RT}$. In addition to this, weights, λ_{ST} and λ_{RT}, are put on G_{ST} and G_{RT} respectively, where $\lambda_{ST} + \lambda_{RT} = 1$. The weights are optimized using speech recognition evaluation.

3 Experimental Work

3.1 Experimental Data

The weather information domain was chosen for the experiments. English was chosen as a resource rich language and Icelandic as a resource sparse language. For the experiments the Jupiter corpus [4] was used. It consists of unique sentences gathered from actual users' utterances. A set of 1595 sentences were manually translated from English to Icelandic and split into 1500 ST sentences and 95 $Eval$ sentences. 63116 sentences were used as RT. A unique word list was made out of the Jupiter corpus and machine translated to Icelandic using [5] in order to create an English to Icelandic dictionary. A unique list was also made for the ST corpus and translated to English to create a Icelandic to English dictionary. These two dictionaries were then combined to create the translation transducer, T, used in the WFST network. Names of places were identified and then replaced randomly with Icelandic place names for the RT corpus, since the task is in the weather information domain. A phonetically balanced (PB) Icelandic text corpus, the Jensson PB corpus [1], 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 $Eval$ set. 12 minutes, a total of 400 utterances, 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. Evaluation of the speech recognition output is performed with keyword detection since it is difficult to obtain a *correct* reference file for the English output when the input is in Icelandic. A keyword set was therefore created for each utterance in the $Eval$

* 機械翻訳を用いた WFST に基づくアイスランド語音声認識
アーナー ジェンソン，岩野公司，古井貞熙（東工大）

Table 1 *Experimental Setup.*

Experiment nr.	Language Output	Vocabulary	Vocabulary Size	OOK
Experiment 1	Icelandic	V_{ST_i}	809	1.9%
Experiment 2	English	V_{ST_e}	482	3.2%
Experiment 3	Icelandic	V_{ST_i} + V_{RT_i}	2996	0.9%
Experiment 4	English	V_{ST_e} + V_{RT_e}	3057	0.0%

set for both Icelandic and English output, in total 844 keywords for each language. Each keyword had the possibility of several matches since many words can have the same meaning as the following example demonstrates, "tonight" and "this evening". This applies especially to the Icelandic keyword detection since Icelandic is an inflected language.

3.2 Experimental Setup

In total four different experiments were performed. Experiment 1 and Experiment 2 use ST as a base for a vocabulary. Experiment 1 output is in Icelandic using the vocabulary from ST, V_{ST_i} while Experiment 2 output is in English using an English translation of the ST vocabulary, V_{ST_e}. Experiment 3 and Experiment 4 use a combination of ST and RT as a vocabulary. Experiment 3 uses V_{ST_i} combined with an Icelandic translation of the RT vocabulary, V_{RT_i} while Experiment 4 uses V_{ST_e} combined with the vocabulary from RT, V_{RT_e}. The experimental setup with the corresponding vocabulary sizes can be viewed in Table 1 where OOK represents out of keywords, i.e. when a keyword can not be constructed from the vocabulary. The T^3 (Tokyo Tech Transducer-based decoder) [6] was used for all the experiments.

4 Results

The keyword detection results are shown in Figure 1. All the experiments performed better for some $\lambda_{ST} < 1.0$ than if only the ST corpus was used, i.e. when $\lambda_{ST} = 1.0$. A *baseline* (87.6%) is obtained when only ST information is used for the Icelandic output, i.e. when λ_{ST} is 1.0 for Experiment 1. When the vocabulary base is ST the best Icelandic output in Experiment 1 (88.5%) outperforms the best English output in Experiment 2 (87.44%). When the vocabulary base is a combination of ST and TRT the best Icelandic output in Experiment 3 (89.8%) is outperformed by the English output in Experiment 4 (91.0%) which gives the best results for all the experiments.

5 Discussion and Conclusion

The results shown in this paper indicate that improvement can be obtained with the proposed models over the *baseline*. The English output is especially important since if a system has already been developed for English then the same backend system can be used. In addition to this point 1.2% and 3.4% absolute keyword detection improvements were observed for Experiment 4 over the best performed Icelandic speech output and the *baseline* respectively. Future work includes applying the proposed models on larger vocabulary task such as broadcast news.

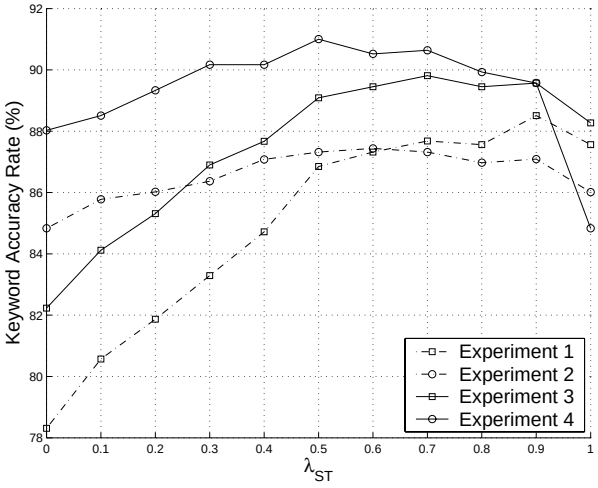

Fig. 1 Keyword Accuracy Results.

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. 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.

References

[1] Jensson, A., Iwano, K., Furui, S., "Development of a speech recognition system for Icelandic using machine translated text", *Proc. SLTU*, Hanoi, Vietnam, pp.18-22, 2007

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

[3] Hori, T., Willet, D., Minami, Y., "Language model adaptation using WFST-based speaking-style translation", *Proc. ICASSP*, vol 1, pp. 228-231, 2003.

[4] 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.

[5] Briem, S., "Machine translation tool for automatic translation from English to Icelandic", *http://www.simnet.is/stbr/*, Iceland, 2007.

[6] Dixon, P. R., Caseiro, D. A., Oonishi, T. and Furui, S., "The TITECH large vocabulary WFST speech recognition system", *Proc. ASRU*, Kyoto, Japan, pp. 443-448, 2007.