HOKKAIDO UNIVERSITY

Title	Development of a WFST based Speech Recognition System for a Resource Deficient Language Using Machine Translation
Author(s)	Jensson, Arnar Thor; Oonishi, Tasuku; Iwano, Koji; Furui, Sadaoki
Citation	Proceedings : APSIPA ASC 2009 : Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, 50-56
Issue Date	2009-10-04
Doc URL	http://hdl.handle.net/2115/39642
Type	proceedings
Note	APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. 4-7 October 2009. Sapporo, Japan. Oral session: Speech and Music Processing (5 October 2009).
File Information	MA-L2-4.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Development of a WFST based Speech Recognition System for a Resource Deficient Language Using Machine Translation

Arnar Thor Jensson*, Tasuku Oonishi*, Koji Iwano* and Sadaoki Furui*
* Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: {arnar, oonishi, iwano, furui}@furui.cs.titech.ac.jp Tel: +81-3-5734-3481

***Abstract*—Text corpus size is an important issue when building a language model (LM) in particular where insufficient training and evaluation data are available. In this paper we continue our work on creating a speech recognition system with a LM that is trained on a small amount of text in the target language. In order to get better performance we use a large amount of foreign text and a dictionary mapping between the languages. A dictionary is used since we are assuming that the target language is resource deficient and therefore statistical machine translation (MT) is not available. In this paper we take a step forward from our previous published method by using a coupling of the speech recognition part and the translation part rather than pre-translating the foreign text. The coupling is achieved with a weighted finite state transducer ($WFST$) network which as well makes it possible to easily switch between the output language, i.e. that the output text is in the format of the resource deficient language or in the resource rich language. Our method outperforms the resource-deficient Icelandic speech recognition $baseline$, 82.6% keyword accuracy (KA), when the system is trained on 1500 Icelandic sentences, both for the English output (2.6% absolute KA improvement) and for the Icelandic output (1.6% absolute KA improvement) where the English text corpus consists of 63003 sentences.**

I. Introduction

State of the art speech recognition has advanced greatly for several languages [1]. Extensive databases, both acoustic and text, have been collected in these languages in order to develop the speech recognition systems. Collection of large databases is time and labor intensive and requires the existence of relevant data resources for each of the target languages and target task domains. More than 6000 living languages are spoken in the world today [2]. Developing a speech recognition system for each of these languages seems unimaginable but since one country or a region can quickly gain political and economical importance a quick approach to developing a speech recognition system for resource deficient languages is important. A resource deficient language is a language where data is non existent or the amount of data available is severely limited. Resource deficiency can also apply to the rich languages for specific domains since speech recognition for a specific domain is more robust if the data used for training the models required comes from the same domain. Since data, for the purpose of developing a speech recognition system, is sparse or non-existent for resource deficient languages, it may be possible to use data from the other resource rich languages.

Development of speech recognizers for resource deficient languages using spoken utterances in a different language has already been reported in [3] where phonemes are identified in several different languages and used to create or aid an acoustic model for the target language. Text for creating the language model (LM) is on the other hand assumed to exist in a large quantity and therefore sparseness of text is not addressed in [3].

In order to create robust speech recognition system a large amount of text is needed. The more text used for the training the better. It depends though on whether the text is of good quality and addresses the target domain. Assuming that only a small amount of text is available for the training of the LM we are obviously facing a problem. This can happen if the language is resource deficient or if little or if no text is available for a new domain in the target language. Several approaches have been proposed in the literature to improve statistical language modeling. In [4] sentences are selected from a large corpus in the same language and a new LM is trained from the selected sentences that may be relevant for the target domain. This method obviously does not apply to resource deficient languages since a large corpus is needed to select sentences from. It might be possible to use machine translation (MT) if a large text corpus is available in another language for the specific target domain. An approach using a cross-lingual information retrieval method to aid an LM in a different language is addressed in [5]. News stories are translated from a resource $rich$ language to a resource $sparse$ language using a statistical MT system trained on a sentence-aligned corpus in order to improve the LM used to recognize similar or the same story in the resource $sparse$ language. A method described in [6] uses ideas from latent semantic analysis for cross lingual modeling to develop a single low-dimensional representation shared by words and documents in both languages. It uses automatic speech recognition transcripts and aligns each with the same or similar story in another language. Using this parallel corpus a statistical MT system is trained. The MT system is then used to translate a text in order to aid the LM used to recognize the same or similar story in the original language. LM adaptation with target task machine-translated text is addressed in [7] but without speech recognition experiments. A system that uses an

automatic speech recognition system for human translators is improved in [8] by using a statistical machine translation of the source text. It assumes that the content of the text translated is the same as in the target text recognized. The above mentioned systems all use statistical MT often expensive to obtain and unavailable for resource deficient languages since a statistical MT system is trained on a large bilingual parallel corpus.

In order to handle the sparseness of the resources available we proposed a method in [9], that instead of using an expensive or unavailable statistical MT system to translate the foreign text a simple dictionary was used to do the translation. In this paper we make a step forward and instead of pre-translating the large text corpora as we did in [9] we couple the speech recognition process and the MT process together in one network. In order to do the coupling we use a weighted finite state transducer (WFST) based speech recognition system where the input speech is in Icelandic and the output text is in either Icelandic or English using a word-by-word (WBW) translation transducer. A WBW translation transducer is in other words a simple rule based MT system based on a dictionary. A coupling of a speech recognition system and a MT system has already been reported in [10] where the input and output languages are the same, i.e. Japanese, but the input speech is in a different style from the output text, i.e. the input speech is in a spoken style and the output text is in a more polite written style. A coupling of ASR (automatic speech recognition) and MT was also done in [11] but with a statistical machine translation system.

Having recognition output in a resource rich language (such as English) when the input is in a resource deficient language (such as Icelandic) 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.

In this paper we investigate the effectiveness of the on-the-fly WBW translation that does not require a preparation of a large translated text in beforehand as done in [9] nor does it require a statistical machine translation system. Here we also investigate the effects of having the output language in a resource rich language when the input speech is in a resource deficient language.

In Section II, we explain the method we use. Section III explains the experimental corpora. Section IV explains the experimental setups. Experimental results are reported in Section V followed by a discussion in Section VI, and Section VII concludes this paper.

II. Method

Our method involves a coupling of the ASR and the MT systems into one network where the MT system is based on a simple WBW translation. The method involves two different kinds of setups demonstrated in Eq. (1) and Eq. (3) depending on the target output language. Eq. (1) demonstrates when the output language is different from the input language.

$$\begin{aligned}\widetilde{T} &= \underset{T}{\operatorname{argmax}}\, P(T|O) \\ &= \underset{T}{\operatorname{argmax}}\, P(O|T)\cdot P(T) \\ &= \underset{T}{\operatorname{argmax}} \sum_{W} P(O|W)\cdot P(W|T)\cdot P(T) \\ &\cong \underset{T}{\operatorname{argmax}} \max_{W} P(O|W)\cdot P(W|T)\cdot P(T) \end{aligned} \quad (1)$$

Here $P(O|W)$ is an acoustic probability of speech input sequence vectors O given a word sequence W from the resource deficient language, $P(W|T)$ is the translation model, where T is a word sequence in the resource rich language and $P(T)$ is the language probability of T. For the translation model probability $P(W|T)$ we use the following approximation

$$P(W|T) \approx P(W)\delta_S(W,T), \quad (2)$$

where $P(W)$ is a prior probability of W, given by a resource deficient language model, and $\delta_S(W,T)$ takes binary 0 or 1 values depending on whether it is possible to substitute W with T or not, which is given by a set of substitution rules of a word or words.

Eq. (3) demonstrates when the output language is the same as the input language.

$$\begin{aligned}\widetilde{W} &= \underset{W}{\operatorname{argmax}}\, P(W|O) \\ &= \underset{W}{\operatorname{argmax}}\, P(O|W)\cdot P(W) \\ &= \underset{W}{\operatorname{argmax}}\, P(O|W)\cdot \sum_{T} P(W|T)\cdot P(T) \\ &\cong \underset{W}{\operatorname{argmax}} \max_{T} P(O|W)\cdot P(W|T)\cdot P(T) \end{aligned} \quad (3)$$

Recently the WFST approach has become a promising alternative formulation to the traditional decoding approach [12]. It offers a unified framework representing various knowledge sources and producing the full search network optimized up to the HMM states. WFST representation of Eq. (1) and Eq. (3) are demonstrated in Eq. (4) and Eq. (5) respectively.

$$H \circ C \circ L \circ G_{ST} \circ Tr \circ G_{RT} \quad (4)$$

$$H \circ C \circ L \circ G_{ST} \circ \pi(Tr \circ G_{RT}) \quad (5)$$

Here H maps HMM states to context-dependent phones. C represents a transduction from context-dependent phones to context-independent phones. L is a lexicon converted to a WFST that will map context-independent phone sequences 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. Tr is a WBW 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 $Tr \circ G_{RT}$ 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 $Tr \circ G_{RT}$.

Whether the setup explained with Eq. (4) or Eq. (5) is selected depends on the purpose of the system and the output language required. The resource rich language output is especially interesting since it can speed up the development of a system if the backend system already exists for the resource rich language.

III. Experimental Corpora

A. Experimental Data: LM

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 [13] was used. It consists of unique sentences gathered from actual users' utterances. A set of 2160 sentences were manually translated from English to Icelandic and split into 1500 sparse training (ST) sentences and 660 evaluation ($Eval$) sentences. Table I shows some attributes of the ST and $Eval$ corpora. A set of 63003 sentences were used as RT database. A unique word list was made out of the RT corpus and machine translated to Icelandic using [14] in order to create an English to Icelandic dictionary. A unique list was also made from the ST corpus and translated to English to create an Icelandic to English dictionary. These two dictionaries were then combined to create the translation transducer, Tr, used in the WFST network. Names of places were identified and then replaced randomly with Icelandic place names for the RT corpus.

A 1-gram, 2-gram, 3-gram and 4-gram translation evaluation using BLEU [21] was performed on 100 sentences created from a simple WBW translation using the previously described dictionary. Two human translators were used to provide manual translations for use as references for each of the 100 sentences. Table II shows the BLEU evaluation results. In the BLEU evaluation the n-gram represents an n-gram match between the machine translation output and the human translation reference text. The higher the n-gram the more difficult it is to match the words in the reference text. It is a known fact that even human translators do not get full mark (1.0) using the BLEU evaluation [21].

TABLE I
Datasets of Icelandic text (also used to create evaluation utterances)

Corpus Set	Sentences	Words	Unique Words
ST	1500	8591	805
$Eval$	660	3767	554

TABLE II
BLEU evaluation of the WBW machine translation.

Translation Method	BLEU			
	1-gram	2-gram	3-gram	4-gram
WBW	0.47	0.28	0.19	0.15

B. Experimental Data: Acoustic Model

A bi-phonetically balanced (PB) Icelandic text corpus was used to create an acoustic training corpus. A text-to-phoneme translation dictionary was created for this purpose based on [15] using 257 pronunciation rules. The whole set of 30 Icelandic phonemes used to create the corpus, consisting of 13 vowels and 17 consonants, are listed in IPA format in Table III.

TABLE III
Icelandic phonemes in IPA format

Vovel	/ ɪ, i, ɛ, a, ʏ, œ, u, ɔ, au, ou, ei, ai, œy /
Consonant	/ p, p^h, t, t^h, c, c^h, f, v, ð, s, j, ç, ɣ, m, n, l, r /

Some attributes of the PB corpus are given in Table IV. The acoustic training corpus was then recorded in a clean environment to minimize external noise. Table V describes some attributes of the acoustic training corpus.

25-dimensional feature vectors consisting of 12 MFCCs, their delta, and delta energy were used to train gender independent acoustic model. Phones were represented as context-dependent, 3-states, left-to-right hidden markov models (HMM). The HMM states were clustered by a phonetic decision tree. The number of leaves was 1000. Each state of the HMMs was modeled by 16 Gaussian mixtures. No prosodic information was incorporated. HTK [16] version 3.2 was used to train the acoustic model and then converted to the format used by the T^3 decoder (Tokyo Tech Transducer-based decoder) [17].

C. Evaluation Speech Corpus

An evaluation corpus was recorded using sentences from the previously explained $Eval$ set. There were 660 sentences

TABLE IV
Some attributes of the phonetically balanced Icelandic text corpus.

Attribute	Text Corpus
No. sentences	290
No. words	1375
No. phones	8407
PB unit	biphone
No. unique PB units	916
Avg. no of words / sentence	4.7
Avg. no of phones / word	6.1

TABLE V
Some attributes of the Icelandic acoustic training corpus.

Attribute	Acoustic Corpus
No. male speakers	13
No. female speakers	7
Time (hours)	3.8

TABLE VI
Some attributes of the Icelandic evaluation speech corpus.

Attribute	Evaluation Speech Corpus
No. utterances	4000
No. male speakers	10
No. female speakers	10
Time (hours)	2.0

TABLE VII
Keyword detection example.

Icelandic utterance	já ég myndi vilja **hitastigið** í **Osló**
Icelandic recognition	ég mun vilja **hitastig Osló**
English recognition	i would like **temperature Oslo**

TABLE VIII
Keyword group example.

Icelandic keywords	English keywords
hitastigið, hiti stig, hitastiginu, hiti, hitastig, hlýtt, heitur, heitt	the temperature, temperature the temperature, heat temperature, warm hot, hot
Osló	Oslo

in total, divided into five sets of 220 sentences for each speaker overlapping every 110 sentences. The final speech evaluation corpus was stripped down to 200 sentences for each speaker since several utterances were deemed unusable. Some attributes of the corpus are presented in Table VI. None of the speakers in the evaluation speech corpus is included in the acoustic training corpus described in Section III-B.

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. In addition keyword detection is often used for applications such as the weather information system described in this paper. A keyword set was therefore created for each utterance in the $Eval$ set for both Icelandic and English output, in total 8693 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 possibility of several matches applies especially to the Icelandic keyword detection since Icelandic is an inflected language.

An example sentence in Icelandic with its corresponding Icelandic and English speech recognition results is provided in Table VII with the keywords in bold. The keyword groups used for detection in the example are provided in Table VIII. A keyword is counted if any of the keywords in a specific group is detected. Therefore in the example in Table VII it does not matter if $temperature$ or $heat$ is detected for the English case since the meaning is similar.

IV. Experimental Setup

Four different experiments were performed that involved two translation methods, Method 1 for English output and Method 2 for Icelandic output. Each method had two different types of vocabulary. The experimental conditions of the methods can be found in Table IV. To be exact Method

TABLE IX
Experimental conditions.

		Vocabulary base	
		ST	ST + RT
Output Language	English	Method 1.1	Method 1.2
	Icelandic	Method 2.1	Method 2.2

1.1 output is in English using an English translation of the ST vocabulary, V_{ST_e} while Method 2.1 output is in Icelandic using the vocabulary from ST, V_{ST_i}. Method 1.2 output is in English and uses V_{ST_e} combined with the vocabulary from RT, V_{RT_e} while Method 2.2 output is in Icelandic and uses V_{ST_i} combined with an Icelandic translation of the RT vocabulary, V_{RT_i}. All the language models used were trigram. Weight factors were added to the language models for each method. Eq. 1 and Eq. 3 explained in Section II are respectively modified with the weights in Eq. 6 and Eq. 7.

$$\widetilde{T} \cong \underset{T}{\operatorname{argmax}} \max_{W} P(O|W) \cdot P(W|T)^{\lambda_{ST}} \cdot P(T)^{\lambda_{RT}} \quad (6)$$

$$\widetilde{W} \cong \underset{W}{\operatorname{argmax}} \max_{T} P(O|W) \cdot P(W|T)^{\lambda_{ST}} \cdot P(T)^{\lambda_{RT}} \quad (7)$$

Here λ_{ST} and λ_{RT} use the following relationship

$$\lambda_{ST} + \lambda_{RT} = 1. \quad (8)$$

Interpolation of the language models were all run with increments of the variable λ_{ST} in step size of 0.1 from 0.0 to 1.0 for all experiments. The weights were optimized using speech recognition evaluation. The $Eval$ set was used and run on the T^3 decoder.

The following four experiments were conducted to investigate the effects for the size and contents of ST / RT corpora.

Experiment 1 uses 1500 ST sentences and all 63003 RT sentences. The experimental setup with the corresponding vocabulary sizes can be viewed in Table X where OOK represents out of keywords, i.e. the ratio of keywords which cannot be constructed.

Experiment 2 is a subset of Experiment 1 where only Method 1.2 is used. The experiment has three different evaluation sets. The $Eval$ (data set of 4000 utterances) was split into two 2000 utterances, Set1 and Set2. Speech recognition was performed on the two subsets of $Eval$ as well as the $Eval$ set. The purpose was to find out the consistency of the evaluation set i.e. to find out if more realistic environment would be needed having either Set1 or Set2 as a development set and the other as evaluation set.

Experiment 3 expands Experiment 1 where ST comprises of either 500, 1000 or 1500 sentences. The experiment was performed in order to show the effects of increasing 1500 sentences in the ST set.

Experiment 4 uses Method 1.2 and Method 2.2 with 1500 ST sentences and increasing the number of RT sentences from 1000 sentences to 63003 sentences. The vocabulary was fixed to be the same as explained in Table X. This experiment was performed to find out the impact of the RT corpus without having the effect of increasing vocabulary size.

TABLE X
Experimental setup.

Experiment nr.	Language Output	Vocabulary	Vocabulary Size	OOK
Method 1.1	English	V_{ST_e}	482	6.4%
Method 1.2	English	V_{ST_e} + V_{RT_e}	3057	0.6%
Method 2.1	Icelandic	V_{ST_i}	805	3.2%
Method 2.2	Icelandic	V_{ST_i} + V_{RT_i}	2996	2.3%

V. Results

Experiment 1: Keyword Accuracy (KA) results are shown in Figure 1. All methods performed better for some $\lambda_{ST} < 1.0$ than if only the ST corpus was used, i.e. when $\lambda_{ST} = 1.0$. A $baseline$ (82.6%) is obtained when only ST information is used for the Icelandic output, i.e. when λ_{ST} is 1.0 for Method 2.1. When the vocabulary base is ST the best Icelandic output in Method 2.1 is 83.0%. The best English output in Method 1.1 is 82.1%. When the vocabulary base is a combination of V_{ST_i} + V_{RT_i} the best Icelandic output in Method 2.2 (84.2%) is outperformed by the English output when the vocabulary is V_{ST_e} + V_{RT_e} in Method 1.2 (85.2%) which gives the best results for all the experiments.

Experiment 2: KA results are shown in Figure 2. Consistency between Set1 and Set2 is considered sufficient and therefore the whole $Eval$ set is used for other experiments instead of using either Set1 or Set2 as development set and the other as evaluation set.

Experiment 3: The results for Method 1.1, 1.2, 2.1 and 2.2, when ST comprises of either 500, 1000 or 1500 sentences, are shown in Figure 3 optimized on λ_{ST} for each experiment. In addition a $baseline$ is also demonstrated in the figure. Method 2.1, i.e. when the output text is Icelandic and the vocabulary is V_{ST_i}, gives 79.4, 81.7 and 83.0 for 500, 1000 and 1500 ST sentences respectively. Method 1.2, i.e. when the output text is English and the vocabulary is V_{ST_e} + V_{RT_e}, gives 82.3, 84.4 and 85.2 for 500, 1000 and 1500 ST sentences respectively. The word accuracy (WA) results are similarly displayed in Figure 4 for Method 2.1 and Method 2.2 only in addition with the $baseline$, since an English reference file needed to evaluate the WA is not available for the Icelandic input speech.

Experiment 4: The results for Method 1.2 and 2.2 with increasing set of RT sentences are shown in Figure 5 optimized on λ_{ST} for each experiment. The experiments have all the same vocabulary explained in Table X in order to investigate the impact of RT without changing the vocabulary. Method 2.2, i.e. when the output is in Icelandic, gives 82.3%, 83.6%, 83.8% and 84.2% for 1000, 4000, 10000 and 63003 RT sentences respectively. Method 1.2, i.e. when the output is in English gives, 82.8%, 84.2%, 84.3% and 85.2% for 1000, 4000, 10000 and 63003 RT sentences respectively.

VI. Discussion

The KA difference between the Icelandic $baseline$, when no foreign text is introduced and Method 1.2 when the output

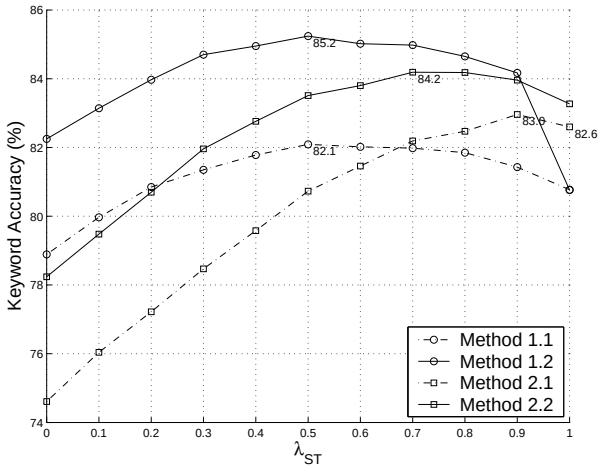

Fig. 1. Experiment 1: Keyword accuracy results for ST = 1500 for each method.

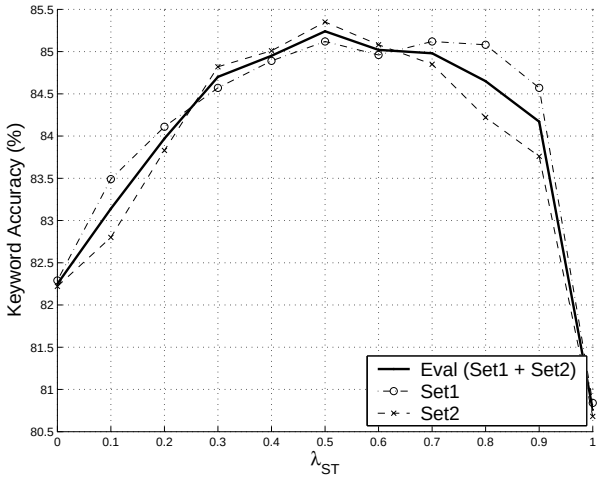

Fig. 2. Experiment 2: Keyword accuracy results for Method 1.2 with different evaluation sets.

is English using a combination of the ST and the RT vocabularies in Experiment 3 is 3.3%, 2.8% and 2.6% for 500, 1000 and 1500 ST sentences respectively and therefore the keyword accuracy difference decreases as more ST sentences are introduced into the system. When Method 1.2 and Method 2.2 are compared then the differences are 1.5%, 1.3% and 1.0 for 500, 1000 and 1500 ST sentences respectively. It is interesting to observe that the KA difference increases when Method 1.1 and Method 1.2 are compared. This is probably because the output is in English and the translated Icelandic ST vocabulary alone does not match the impact of the combined vocabularies with increasing ST sentences. When Methods 2.1 and 2.2 are compared, a similar trend is observed for both KA and WA, i.e. the accuracy difference decreases with larger ST sets. The difference is 1.4%, 1.4% and 1.2% for KA demonstrated in Figure 3 and 1.0%, 0.7% and 0.9% for WA demonstrated in Figure 4 for 500, 1000

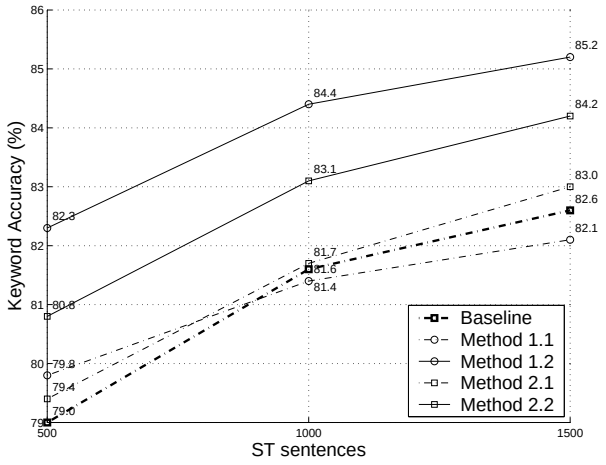

Fig. 3. Experiment 3: Keyword accuracy results for increasing set of ST sentences optimized on λ_{ST} for each method.

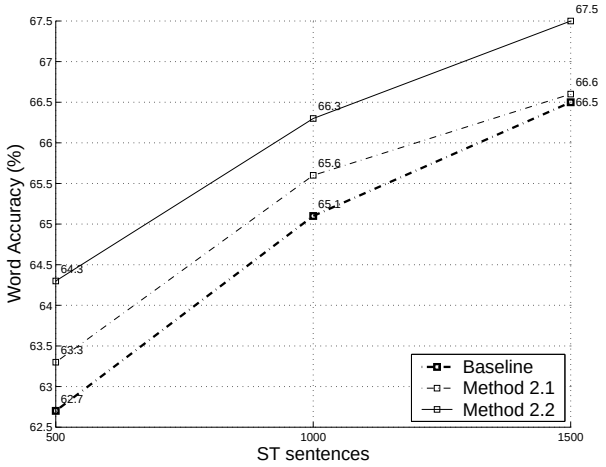

Fig. 4. Experiment 3: Word accuracy results for increasing set of ST sentences optimized on λ_{ST} for each method.

and 1500 ST sentences respectively. This reducing trend is more clearly observable when Method 2.2 is compared to the $baseline$ for either KA or WA where the difference for KA is 3.3%, 2.8% and 2.6% and the difference for WA is 1.6%, 1.2% and 1.0% for 500, 1000 and 1500 ST sentences respectively.

Figure 4 also demonstrates the $baseline$ WA. When it is compared against the best Icelandic output in Method 2.2 the WA difference is 1.6%, 1.2% and 1.0% for 500, 1000 and 1500 ST sentences respectively. The advantage of using the RT corpus has almost all vanished using 1500 ST sentences for Method 2.1 when only the ST vocabulary is used compared to the $baseline$.

Experiment 4 shows that as more RT sentences are introduced to the system the better the results for either Method 1.2 (English output) or Method 2.2 (Icelandic output). Both methods had fixed vocabulary for the experiment in order to investigate if the improvement was mainly through a larger vocabulary or not. The experiment clearly supports that a

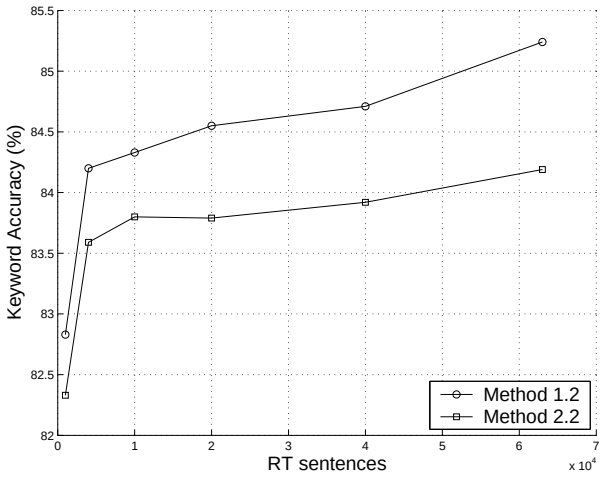

Fig. 5. Experiment 4: Keyword accuracy results for increasing set of RT sentences optimized on λ_{ST} for each method.

larger RT set improves the recognition accuracy.

The English output always outperforms the Icelandic output for the experiments performed. Manual result analysis has shown that this is mainly because of the translation model. In theory the difference between the keyword accuracy for either language should be the same or very close to each other since they are using the same translation model but the real case is that several translations in the translation model used made the difference larger. These several translations had for example one English word translated into a group of words in Icelandic such as "today" into "í dag". When an Icelandic utterance that included "'í dag" was recognized, the "í" was sometimes lacking in the output which makes that keyword detection unclear for the Icelandic case. The main point is that the method used with the WBW translation transducer inside the WFST network optimized on some λ_{ST} for either Icelandic or English output outperforms the Icelandic $baseline$.

VII. Conclusion

The results shown in this paper indicate that improvement can be obtained with the proposed models over the $baseline$ using only a simple WBW translation transducer easily obtainable for resource deficient languages through dictionaries. This especially applies when developing a prototype system where the amount of target domain sentences is very limited. 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 1.0% and 2.6% absolute keyword detection improvements were observed for English output over the best performed Icelandic output and the $baseline$ respectively when Icelandic ST comprises of 1500 sentences and the English RT corpus comprises of 63003 sentences. Also note that 63003 sentences is indeed not very large.

Even though English and Icelandic are quite different languages the structure of the grammar is somewhat similar which

makes it possible to get such improvement with the WBW translation transducer. It is our belief that other more closely related languages could get better improvements with the described method. Confirming the effectiveness of the WBW translation method for other language pairs is left as future work as well as applying the WBW translation methods to a larger domain, for example broadcast news. Future work also involves an investigation of other maximum a posteriori adaptation methods such as the unsupervised language model adaptation by Bacchiani and Roark [18] and methods like the ones described by Sarikaya et al in [19], Sethy et al in [20] and Klakow in [4] 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 RT corpus. Since the acoustic model is only built from 3.8 hours of acoustic data which gives rather poor results we would like to either collect more Icelandic acoustic data or use data from foreign languages to aid current acoustic modeling. Probabilistic translation, i.e. when a word can have multiple translation output adding a probabilistic value to the word translation trained on either a source or a target text is also left as future work.

ACKNOWLEDGMENT

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] Adda-Decker, M., "Towards multilingual interoperability in automatic speech recognition", *Speech Communication*, Vol 35, Issue 1-2, pp 5-20, August 2001.

[2] Gordon, R. G., "Ethnologue: Languages of the World", Fifteenth edition. Dallas, Tex, 2005.

[3] Schultz, T., Waibel, A., "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition", *Speech Communication*, Vol 35, Issue 1-2, pp 31-51, August 2001.

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

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

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

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

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

[9] 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, 2008.

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

[11] Matusov, et al., "Integrating Speech Recognition and Machine Translation: Where Do We Stand?", *Proc. ICASSP*, pp. 1217-1220, 2006.

[12] Mohri, M. "Finite-State Transducers in Language and Speech Processing", *Computational Linguistics*, vol 23, pp. 269-311, 1997.

[13] 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.

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

[15] Rognvaldsson, E., "Islensk hljodfraedi", Malvisindastofnun Haskola Islands, Reykjavik, 1989.

[16] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., *The HTK Book (for version 3.2.1)*.

[17] 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.

[18] Bacchiani, M., Roark, B., "Unsupervised Language Model Adaptation", *Proc. ICASSP*, vol. 1, pp. 224-227, 2003.

[19] 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.

[20] 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.

[21] 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.