論文 / 著書情報
Article / Book Information

Title(English)	Development of a Speech Recognition System Using a Sparse Training Corpus
Authors(English)	Arnar Thor Jensson, Koji Iwano, Sadaoki Furui
Citation(English)	Symposium on Large-Scale Knowledge Resources(LKR2007), Vol. , No. , pp. 133-136
発行日 / Pub. date	2007, 3

Development of a Speech Recognition System Using a Sparse Training Corpus

Arnar Thor Jensson, Koji Iwano and Sadaoki Furui

Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan,
{arnar, iwano, furui}@furui.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 a LM adaptation technique to improve a LM built using a small amount of task dependent text with the help of a machine-translated text corpus. Icelandic perplexity and word error rate experiments were performed using data, machine translated (MT) from English to Icelandic on a word-by-word basis. The baseline word error rate was 40.2%. LM interpolation reduced word error rate significantly to 36.2%.
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.

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.

Methods have been proposed in the literature to improve statistical language modeling in a resource-deficient language using cross-lingual information retrieval [1]. Another method proposes using latent semantic analysis for cross-lingual modeling which does not require a sentence-aligned corpus [2] but searches for similar types of texts in two languages. LM adaptation with target task machine-translated text is addressed in [3] but without speech recognition experiments.

In [4], 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 speech recognition experiment results for the WBW translation from English to Icelandic where the spoken evaluation corpus is larger. This paper also provides WBW translation results when LM clustering is applied.

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), which is 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 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),

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

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 disjoint from all other data sets.

3. Experimental Work

3.1. Experimental Data: LM

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 1960 sentences were manually translated from English to Icelandic and split into ST_i, SD and $Eval$ sets as shown in Table 1, where ST_i corresponds to the Icelandic version of

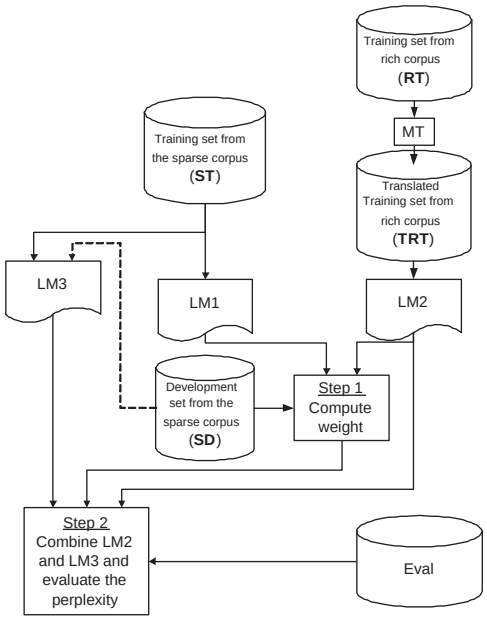

Figure 1: Data diagram

Table 1: *Data sets.*

Corpus Set	Sentences	Words
ST_i^{300}	300	1736
ST_i^{500}	500	2869
ST_i^{700}	700	4009
ST_i^{900}	900	5173
ST_i^{1000}	1000	5820
SD	300	1729
$Eval$	660	3767

ST defined in Section 2. 63116 sentences were used as RT.

A vocabulary of all unique words was then manually created from the Jupiter corpus. 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 vocabulary was then used to automatically translate RT creating TRT_{ei}, where ei corresponds to a translation from English to Icelandic. Table 2 shows some attributes of the WBW translated text TRT_{ei}.

Table 2: *Translated data set.*

Corpus Set	Sentences	Words	Unique Words
TRT_{ei}	60358	392734	1893

Table 3: *Some attributes of the Jensson 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 4: *Some attributes of the Jensson Icelandic acoustic training corpus.*

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

3.2. Experimental Data: Acoustic Model

A phonetically balanced (PB) Icelandic text corpus, the Jensson PB corpus, was used to create an acoustic training corpus. A text-to-phoneme translation tool was created for this purpose based on [6]. Some attributes of the PB corpus are given in Table 3. The acoustic training corpus was then recorded. Table 4 describes some attributes of the Jensson acoustic training corpus.

3.3. Evaluation Speech Corpus

An evaluation speech corpus, Thor, was recorded using sentences from the $Eval$ set. There were 660 sentences in total but divided into 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 5. None of the speakers in the evaluation speech corpus are in the acoustic training corpus described in Section 3.2.

3.4. Results

In total two different experiments were performed. The SD, $Eval$ and TRT_{ei} sets were identical for all the experiments but the ST_i 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_i corpus to make LM3 the weights calculated in Step 1 would be inaccurately

Table 5: *Some attributes of the Thor Icelandic evaluation speech corpus.*

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

optimized for the combined set especially since the ST_i corpus is small. Therefore LM1 was used instead of LM3. In the following text Voc is defined as vocabulary, PP as perplexity, Int as interpolation of the ST_i and the TRT_{ei}, Imp as improvement, OOV as out of vocabulary rate and WER as word error rate.

Experiment 1 used the unique words found in the ST_i set as the vocabulary, V_{ST_i}. The results are shown in Table 6. The WER improvement is positive when ST_i comprises 300 sentences but as more sentences are added to the ST_i combination leads to the WER improvement becoming negative.

Table 6: *Results for experiment 1. The unique words found in the ST_i^n corpus was used as a Voc.*

Attribute	ST_i^{100}	ST_i^{300}	ST_i^{700}	ST_i^{1000}
Voc size	190	333	535	614
PP ST_i	20.2	21.0	19.3	18.6
PP Int	16.8	18.9	18.1	18.0
PP Imp (%)	16.8	10.0	6.2	3.2
OOV (%)	15.5	9.3	6.0	5.2
WER ST_i(%)	47.0	40.2	23.5	23.1
WER Int (%)	45.3	39.1	24.2	23.7
WER Imp (%)	3.6	2.7	-3.0	-2.6

Experiment 2 used the set of unique words from the TRT_{ei} set combined with V_{ST_i}. These results are shown in Table 7. As expected the WER improvement gradually reduces for most experiments as more manually transcribed data is added to the ST_i set. Comparing Table 6 with Table 7 shows that a lower WER can be obtained using the combined vocabulary as performed in Experiment 2. Figure 2 shows the WER from Experiments 1 and 2 graphically. Addition to Table 6 and Table 7, Figure 2 shows the WER for ST_i^{200}, ST_i^{400}, ST_i^{500}, ST_i^{600}, ST_i^{800} and ST_i^{900}.

3.5. Results using LM clustering

Word clustering LM [9] was applied to the ST and TRT and then interpolated to the standard tri-gram versions of the ST and TRT language models. In the following text ST_{cl} is defined as a clustered LM trained from the ST corpus and TRT_{cl} is defined as a clustered LM trained from the TRT corpus. First a baseline PP was calculated for an interpolation of ST, ST_{cl} and TRT. The TRT_{cl} corpus was then introduced and PP calculated from the interpolation of ST, ST_{cl}, TRT and TRT_{cl}. All the models where trained using tri-gram LM. The vocabulary for this experimental set was a combination of the words from ST^n and TRT corpora where n runs from 200 sentences to 900 sentences in an interval of 100 sentences. The results including the relative PP improvements are presented in Table 8.

Table 7: *Results for experiment 2. The unique words found in both the ST_i^n corpus and the TRT_{ei} corpus was used as a Voc.*

Attribute	ST_i^0	ST_i^{100}	ST_i^{300}	ST_i^{700}	ST_i^{1000}
Voc size	1893	1942	1987	2089	2134
PP ST_i	NA	43.6	32.1	24.5	22.8
PP Int	161	29.1	25.6	22.2	21.4
PP Imp (%)	NA	33.3	20.2	9.4	6.1
OOV (%)	10.2	5.8	4.2	3.3	3.0
WER ST_i(%)	73.7	44.6	38.8	24.7	24.2
WER Int (%)	50.8	39.1	36.2	24.9	24.2
WER Imp (%)	31.1	12.3	6.7	-0.8	0

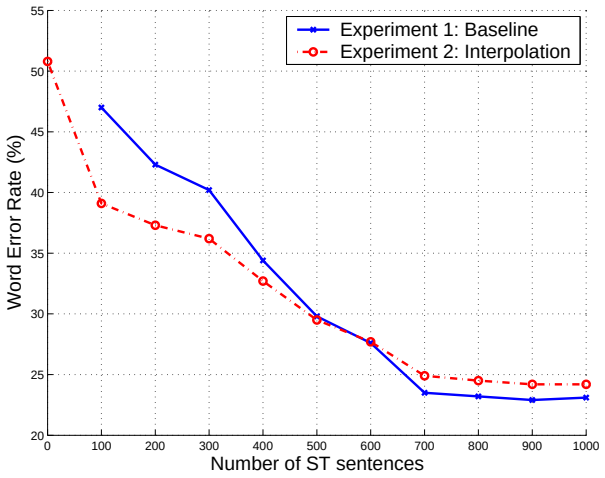

Figure 2: Word error rate results using the baseline from experiment 1 and the interpolation results from experiment 2.

4. Discussion

The improvement of the Icelandic LM with translated English text/data was confirmed by reduction in PP and WER. As Table 7 shows, PP improvement with WBW translation varies from 33.3% to 6.1% when 100 and 1000 manually

Table 8: Perplexity results using LM clustering

Attribute	PP		PP Imp (%)
	ST + ST_{cl} + TRT	ST + ST_{cl} + TRT + TRT_{cl}	
ST_i^{200}	20.94	20.76	0.9
ST_i^{300}	19.95	19.81	0.7
ST_i^{400}	17.51	17.42	0.5
ST_i^{500}	15.28	15.28	0.0
ST_i^{600}	14.23	14.19	0.3
ST_i^{700}	12.84	12.81	0.2
ST_i^{800}	11.88	11.89	-0.1
ST_i^{900}	11.03	11.05	-0.2

translated sentences were used respectively. The OOV rate is reduced as well from 9.3% to 4.2% when the unique translated words are added to the ST_i^{300} set. The speech recognition WER is reduced by 6.7% in experiment 2 when using ST_i^{300} but the improvement reduced to 0% when 700 more manually translated sentences are used in ST_i^{1000} in the same experimental set.

The results from experiment 1 and experiment 2 should be compared together since the baseline in experiment 1 does not assume any foreign translation, while experiment 2 includes the translated words in its vocabulary. When the baseline in experiment 1 is compared with the interpolated results in experiment 2 we get a WER 40.2% reduced to 36.2% respectfully, an 11.0% relative improvement when using ST_i^{300}. The relative improvement reduces as more ST_i sentences are added to the system and reaches a negative improvement when 700 sentences are added to the system.

The improvement of using LM clustering was not considerably large. This could be since the TRT corpus used in the experiment is a large corpus for this domain and already covers most sentence variations.

5. Conclusions

The results presented in this paper show that a LM can be improved considerably using WBW translation. The WBW translation is especially important for resource deficient languages such as Icelandic that do not have SBS machine translation tools available. The work for applying WBW translated text is reduced if the translated corpus is large and the manual work needed to create the dictionary is small.

Future work involves applying the WBW translation method to a larger domain such as broadcast news.

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. 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, 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., "Improvement of Language Model Adaptation Using Machine-Translated Text", The Acoustic Society of Japan Fall Meeting, Sendai, Japan, 2005, pp. 43-44.

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

[7] Wutiwiwatchai, C., Cotsomrong, P., Suebvisai, S., Kanokphara, S., "Phonetically Distributed Continous Speech Corpus for Thai Language", *Proc. LREC2002*, vol. 3, pp. 869-872.

[8] Clarkson, P. R. and Rosendfeld, R., "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proc. Eurospeech*, Rhodes, Greece, 1997, vol 5, pp. 2707-2710.

[9] Ney, H., Essen, U., and Kneser, R., 1994, On structureing probabilistic dependencies in stochastic language modelling, Computer Speech and Language, 8:1-38.