論文 / 著書情報
Article / Book Information

Title	Language model adaptation for resource deficient languages using translated data
Authors	Arnar Thor Jensson, Edward W. D. Whittaker, Koji Iwano, Sadaoki Furui
Citation	Interspeech2005, Vol. , No. , pp. 1329-1332,
Pub. date	2005, 9
Copyright	(c) 2005 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Language Model Adaptation for Resource Deficient Languages Using Translated Data

Arnar Thor Jensson, Edward W. D. Whittaker, Koji Iwano and Sadaoki Furui

Department of Computer Science
Tokyo Institute of Technology, Tokyo, Japan
{arnar, edw, 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 technique to improve a LM built using a small amount of task dependent text with the help of a machine-translated text corpus. Perplexity experiments were performed using data, machine translated (MT) from English to French on a sentence-by-sentence basis and using dictionary lookup on a word-by-word basis. Then perplexity and word error rate experiments using MT data from English to Icelandic were done on a word-by-word basis. For the latter, the baseline word error rate was 44.0%. LM interpolation reduced word error rate significantly to 39.2%.

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 this paper, we propose a method to improve the LM built on a task-dependent corpus using MT which is similar to [3]. Two language pairs are used. First perplexity experiments using French to English translation are presented using both SBS and WBW translation. Then WBW translation from English to Icelandic is presented using both LM perplexity and word error rate 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), 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.

Figure 1: Data diagram for the experimental setup.

Figure 2: Perplexity results (French to English scenario).

3. Experimental Work

In Section 3.1 we describe French to English perplexity experiments using SBS and WBW MT. Section 3.2 then describes English to Icelandic LM perplexity and speech recognition word error rate experiments using WBW MT.

3.1. French to English

3.1.1. Experimental Data: LM

For the French to English experiments the Hansard corpus was used, which is a parallel text corpus in English and Canadian French. In the following text TRT_{WBW} corresponds to a French to English WBW translation of RT and TRT_{SBS} corresponds to a French to English SBS translation of RT . T_e corresponds to an English corpus that was extracted and its evaluation simulates a perfect translation.

A vocabulary V_{TDE} of 1711 unique English words was chosen randomly from the corpus, and sentences, where each word in the sentence is the vocabulary, were extracted and used to create ST , SD and $Eval$. TDE is an abbreviation for the combination of ST , SD and $Eval$. An intentionally large $Eval$ set was chosen so as to minimize variation in the perplexity results. Table 1 shows the size of these sets. Another vocabulary, V_{TD} , was defined as well using the ST and the SD sets. A different set of randomly chosen French sentences (*rich*) was machine-translated to English using the Google machine translation tool [4] and used either as is (*RAN*) or with only those selected sentences (*SS*) in which all words were from the vocabulary V_{TD} . A translation of the large French set (RT) was also done on a WBW basis using [4]. To allow comparison the vocabulary for perplexity calculation was fixed using the vocabulary V_{TDE} .

All perplexity results were evaluated using tri-gram models since these gave better results than uni-gram and bi-gram models.

Table 1: Data sets (French to English scenario).

Corpus Set	Words	Unique Words
ST	6690	815
SD	6704	810
$Eval$	334375	1593

Table 2: Translated data characteristics (French to English scenario). *SS results only*

Corpus Set	Words	Perplexity (improvement)
$ST + SD$ (TD)	13K	54.7 (-)
$TD + TRT_{WBW-SS}$	200K	52.8 (3.5%)
$TD + TRT_{SBS-SS}$	200K	48.3 (11.7%)
$TD + T_e-SS$	6K	53.0 (3.1%)
$TD + T_e-SS$	26K	48.6 (11.2%)
$TD + T_e-SS$	200K	40.1 (26.7%)

3.1.2. Results

Figure 2 shows perplexity performance versus number of words in TRT and the top half of Table 2 shows perplexity results for approximately the same number of words (200K). The bottom half of Table 2 shows perplexity results for T_e close to the TRT_{WBW} and TRT_{SBS} results which reflect the number of sentences that must be manually translated to get the same improvement as for the machine-translated data. Table 2 shows only *SS* results while Figure 2 shows both *SS* and *RAN* results.

As Table 2 shows, translation from French to English improves the perplexity of the LM between 3.5% and 11.7% when 200,000 words are used. If a *perfect* translation was per-

Table 3: Data sets (English to Icelandic scenario).

Corpus Set	Sentences	Words
ST_i^{300}	300	1736
ST_i^{400}	400	2277
ST_i^{500}	500	2869
ST_i^{600}	600	3416
ST_i^{700}	700	4009
SD	300	1729
$Eval$	200	1147

Table 4: Translated data set (English to Icelandic scenario).

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

formed, using SS the improvement would increase to 26.7%. Table 2 also shows that eight times more words are needed if the same improvement from manual translation is to be expected from SBS MT whereas thirty times more words are needed to obtain the same improvement with WBW translation.

3.2. English to Icelandic

3.2.1. Experimental Data: LM

The French to English results in Section 3.1.2 showed that perplexity could indeed be reduced using both SBS and WBW translation. 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 1200 sentences were manually translated from English to Icelandic and split into ST_i , SD and $Eval$ sets as shown in Table 3, where ST_i corresponds to the Icelandic version of 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 4 shows some attributes of the WBW translated text TRT_{ei} .

3.2.2. 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 5. The acoustic training corpus was then recorded. Table 6 describes some attributes of the Jensson acoustic training corpus.

An evaluation corpus was recorded using the sentences in the $Eval$ set. The recording includes one speaker only, who was not part of the acoustic training corpus.

Table 5: 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 6: Some attributes of the Jensson Icelandic acoustic training corpus.

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

3.2.3. Results

In total three 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 700 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 optimized for the combined set especially since the ST_i corpus is small. Therefore LM1 was used instead of LM3. The optimization of the weights when the ST_i and the SD are combined into LM3 is postponed for future work. 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 $Eval$ set as the vocabulary. This scenario reflects the case when the vocabulary of the system is known. The results are shown in Table 7. The WER improvement is 4.2% when the TRT_{ei} corpus is interpolated with 300 ST_i sentences. As more ST_i data is added the improvement from adding the MT data decreases to 0.3% for 700 sentences.

Experiment 2 used the unique words found in the ST_i set as the vocabulary, V_{ST_i} . The results are shown in Table 8. The WER improvement is positive when ST_i comprises 300 and 400 sentences but as more sentences are added to the ST_i combination leads to the WER improvement becoming negative.

Table 7: Results for experiment 1 (English to Icelandic scenario). The unique words found in the $Eval$ corpus was used as a Voc .

Attribute	ST_i^{300}	ST_i^{400}	ST_i^{500}	ST_i^{600}	ST_i^{700}
Voc size	276	276	276	276	276
PP ST_i	36.0	34.4	31.5	31.4	29.3
PP Int	31.0	30.3	28.1	27.9	27.2
PP Imp (%)	13.9	11.9	10.8	11.1	7.2
OOV (%)	0.0	0.0	0.0	0.0	0.0
WER ST_i (%)	38.0	36.8	36.8	34.9	32.7
WER Int (%)	36.4	36.0	35.8	33.9	32.6
WER Imp (%)	4.2	2.2	2.7	2.9	0.3

Table 8: Results for experiment 2 (English to Icelandic scenario). The unique words found in the ST_i^n corpus was used as a Voc .

Attribute	ST_i^{300}	ST_i^{400}	ST_i^{500}	ST_i^{600}	ST_i^{700}
Voc size	333	395	451	489	535
PP ST_i	21.0	20.2	19.5	20.0	19.3
PP Int	18.9	18.6	17.9	18.2	18.1
PP Imp (%)	10.0	7.9	8.2	9.0	6.2
OOV (%)	9.3	8.5	7.4	6.8	6.0
WER ST_i (%)	44.0	43.0	40.2	39.9	36.7
WER Int (%)	41.6	41.1	40.2	40.1	38.3
WER Imp (%)	5.5	4.4	0	-0.5	-4.4

Table 9: Results for experiment 3 (English to Icelandic scenario). The unique words found in both the ST_i^n corpus and the TRT_{ei} corpus was used as a Voc .

Attribute	ST_i^{300}	ST_i^{400}	ST_i^{500}	ST_i^{600}	ST_i^{700}
Voc size	1987	2015	2043	2063	2089
PP ST_i	32.1	30.0	26.9	26.7	24.5
PP Int	25.6	24.8	23.1	22.9	22.2
PP Imp (%)	20.2	17.3	14.1	14.2	9.4
OOV (%)	4.2	3.9	3.7	3.4	3.3
WER ST_i (%)	43.3	41.8	38.3	38.7	36.6
WER Int (%)	39.2	38.0	37.7	36.7	35.9
WER Imp (%)	9.5	9.1	1.6	5.2	1.9

Experiment 3 used the set of unique words from the TRT_{ei} set combined with V_{ST_i} . These results are shown in Table 9. As expected the WER improvement gradually reduces for most experiments as more manually transcribed data is added to the ST_i set. Comparing Table 8 with Table 9 shows that a lower WER can be obtained using the combined vocabulary as performed in Experiment 3. Figure 3 shows the WER from Experiments 1, 2 and 3 graphically.

The improvement of the Icelandic LM with translated English text/data was confirmed by reduction in PP and WER. As Table 9 shows, PP improvement with WBW translation varies from 20.2% to 9.4% when 300 and 700 manually 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 9.5% in experiment 3 when using ST_i^{300} and by 1.9% when 400 more manually translated sentences are used in ST_i^{700} in the same experimental set.

Figure 3: Word error rate results (English to Icelandic scenario).

4. Conclusions

The results presented in this paper show that a LM can be improved considerably using either SBS translation or WBW translation. The WBW translation is especially important for resource deficient languages such as Icelandic that do not have SBS machine translation 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 multiple speakers and solving the weight calculation when the ST and the SD corpora are added together.

5. 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.

6. References

- [1] Khudanpur, S. and Kim, W., "Using Cross-Language Cues for Story-Specific Language Modeling", *Proc. IC-SLP*, 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] http://www.google.com/language_tools/
- [5] 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*, 2000, 8(1):100-112.
- [6] Rognvaldsson, E., "Islensk hljodfraedi", 1989, Malvisindastofnun Haskola Islands, Reykjavik.
- [7] Wutiwatchai, C., Cotsomrong, P., Suebisai, S., Kanokphara, S., "Phonetically Distributed Continuous 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.