論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Improving a language model using machine translated data
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, 古井 貞熙
Authors(English)	Arnar Jensson, Edward Whittaker, Sadaoki Furui
出典(和文)	日本音響学会2005年春季講演論文集, Vol. , No. 1-5-26, pp. 51-52
Citation(English)	, Vol. , No. 1-5-26, pp. 51-52
発行日 / Pub. date	2005, 3

1—5—26 Improving a language model using machine translated data *

©Arnar Thor Jensson, Edward W. D. Whittaker and 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 model requires a large amount of training text. Therefore it is difficult to create a statistical language model (LM) for resource deficient languages.

By using text translated from other languages it may be possible to improve the resource deficient language model either using sentence-by-sentence translation or word-by-word translation. Word-by-word translation only requires a dictionary whereas sentence-by-sentence machine translation needs a large sentence-aligned parallel corpus which is expensive to obtain. The word-by-word 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.

In this paper, we propose a method to reduce the perplexity of a task-dependent corpus using machine translation. Both sentence-by-sentence translation and word-by-word translation results are presented.

2 Method

The idea is to improve a task dependent language model that is created from a sparse amount of text (in this case in English) using a large translated text (in French) 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 (*Train*) and a development text corpus (*Dev*). A language model LM1 is created from *Train*, and LM2 from a large translated text ($Train_{fe}$), where *fe* denotes French-English translation. The translated text can either be obtained from sentence-by-sentence (*sbs*) or word-by-word (*wbw*) translation. The language models are combined using linear interpolation and the *Dev* set is used to optimize the weights used in step 2.

Step 2 involves combining the *Train* and the *Dev* corpora together and building a new language model, LM3 from it. LM3 and LM2 are then mixed using the optimized weights obtained in Step 1. The final perplexity value is calculated using the evaluation set (*Eval*).

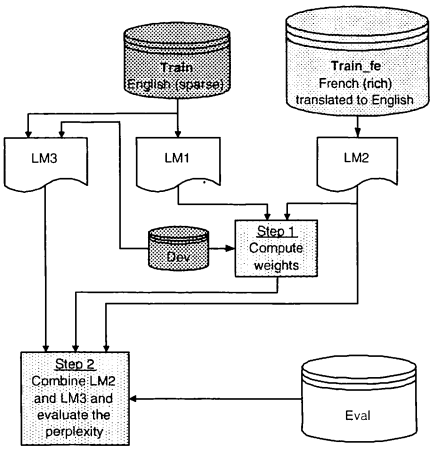

Figure 1. Data diagram

3 Experimental Results

For the experiments the Hansard corpus was used, which is a parallel text corpus in English and Canadian French. A vocabulary V_{TDE} of 1711 unique English words was chosen randomly from the corpus, and sentences that included only those words were extracted and used to create *Train*, *Dev* and *Eval*. Table 1 shows the size of the fixed sets. Another vocabulary, V_{TD}, was defined as well using the *Train* and the *Dev* sets. A different set of randomly chosen French sentences was machine translated to English using [4] and used either as is (*ran*) or with only those selected sentences (*ss*) that included only words from the vocabulary found in V_{TD}.

Table 1. Data Sets

Corpus Set	Words	Unique Words
Train	6690	815
Dev	6704	810
Eval	334375	1593

For comparison reasons the vocabulary for perplexity calculations was fixed using the vocabulary V_{TDE}. A translation of the large French set was also done on a word-by-word basis using [4]. Figure 2 shows perplexity performance versus number

* 機械翻訳データを用いた言語モデルの改良
アーナー ジェンソン，エドワード ウィッタッカー，古井 貞熙（東工大）

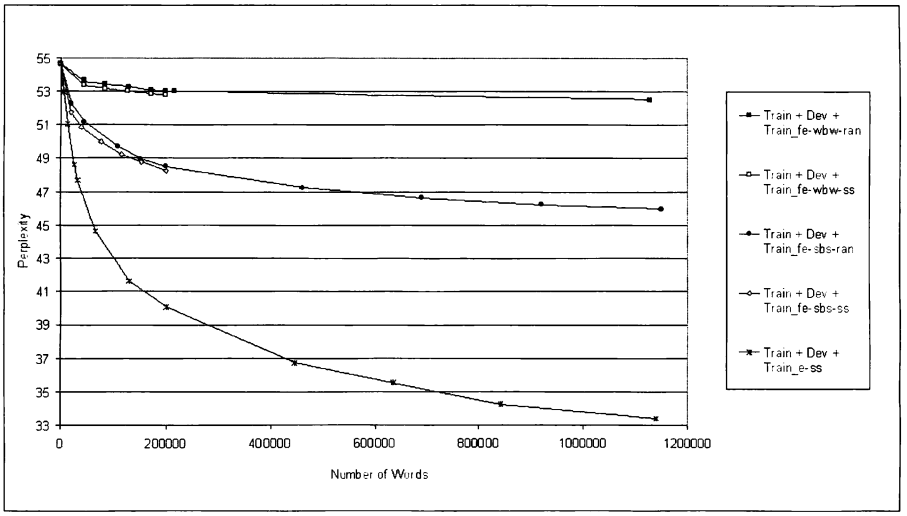

Figure 2. Perplexity results

of words and the top half of Table 2 shows perplexity results for an approximately fixed number of words. The bottom half of Table 2 shows perplexity results for $Train_{e-ss}$ close to the $Train_{fe}$ results which reflects the number of sentences that must be manually translated to get the same improvement as for the machine translated data.

All perplexity results are evaluated using trigram models. Uni-gram and bi-gram models were not found to give better results. $Train_e$ corresponds to an English corpus that was extracted and its evaluation simulates a perfect translation.

Table 2. Translated Data Characteristics

Corpus Set	Words	Perplexity (improvement)
$Train + Dev\ (TD)$	13K	54.7 (-)
$TD + Train_{fe-wbw-ran}$	200K	53.0 (3.1%)
$TD + Train_{fe-wbw-ss}$	200K	52.8 (3.5%)
$TD + Train_{fe-sbs-ran}$	200K	48.5 (11.3%)
$TD + Train_{fe-sbs-ss}$	200K	48.3 (11.7%)
$TD + Train_{e-ss}$	6K	53.0 (3.1%)
$TD + Train_{e-ss}$	26K	48.6 (11.2%)
$TD + Train_{e-ss}$	200K	40.1 (26.7%)

4 Discussion

As Table 2 shows, translation improves the language model between 3.1% and 11.7% when 200,000 words are used. The lowest improvement reflects the perplexity improvement for randomly chosen sentences using a word-by-word translation while the highest improvement is for sentence-by-sentence translation. If a *perfect* translation was performed, using chosen sentences the improvement increased to 26.7%. Table 2 also shows that eight times more data is needed if the same improvement is to be expected with manual translation and sentence-by-sentence machine translation whereas thirty times more data is needed to improve with word-by-word machine translation.

5 Conclusion

The above results show that the perplexity values decrease considerably using sentence-by-sentence machine translated data. It has also been shown that the perplexity can be decreased using word-by-word translation which is important for languages that do not have large enough parallel corpora available to create a machine translation system but do have a dictionary available. Future work will apply the described techniques to an Icelandic dialogue task as a resource deficient language and evaluate performance using speech recognition.

Acknowledge

This work is supported in part by 21st Century COE-LKR Program.

References

[1] S.Khudanpur and W.Kim, "Using cross-language cues for story-specific language modeling," *Proc. ICSLP*, Denver, CO, 2002, vol 1, pp. 513-516.

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

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

[4] http://www.google.com/language_tools/