Language Model Adaptation for Statistical Machine Translation based on Information Retrieval

Matthias Eck, Stephan Vogel, Alex Waibel

Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA, 15213, USA
matteck@cs.cmu.edu, vogel+@cs.cmu.edu, ahw@cs.cmu.edu

Abstract

Language modeling is an important part for both speech recognition and machine translation systems. Adaptation has been successfully applied to language models for speech recognition. In this paper we present experiments concerning language model adaptation for statistical machine translation. We develop a method to adapt language models using information retrieval methods. The adapted language models drastically reduce perplexity over a general language model and we can show that it is possible to improve the translation quality of a statistical machine translation using those adapted language models instead of a general language model.

1. Introduction

Statistical translation systems use the well-known n-gram language models to predict words. Typically, the more data used to estimate the parameters of the language model, the better the translation results.
The main problem is that a general language model does not adapt to the topic or the style of individual texts. It is also obvious that during a longer text the topic of discussion will change.
Experience for speech recognition indicates that even better results might be possible with adapted language models.

In the experiments reported in this paper, the test set consisted of texts from Chinese and Arabic news wires, where different news stories cover different topics. The adaptation of the language models is done by selecting for each news story, or even each sentence, similar stories or sentences from a large English news corpus, using methods of information retrieval, and building smaller, but more specific, language models.

The questions we tried to answer were which kind of adaptation unit (news story or sentence) would be best suited for this idea and how much data should be used for the adapted language models.
Other points to examine are the effect of the amount of data that is used to select the language models from and the effect of the quality of the translations used as query.

It is also interesting to see if some special approaches, that showed good improvements in information retrieval like the usage of stemmers and stopword-lists can have a positive effect in this adaptation task.

1.1. Basic Idea

```
for each test document
    • translate with general language
      model
    • use this translation to select most
      similar documents
    • build adapted language model using
      these similar documents
    • re-translate with adapted language
      model
```

(documents can be stories or individual sentences)

1.2. Previous Work

The main idea is based on the paper by Mahajan, Beeferman and Huang (1999) in which they used similar techniques for language model adaptation. Mahajan, Beeferman and Huang applied the adapted language models on speech recognition and they could significantly reduce perplexity in this task.

Other methods for language model adaptation are presented and reviewed in the paper by DeMori and Federico (1999) and in the paper by Janiszek, DeMori and Bechet (2001).
According to Janiszek, DeMori and Bechet the following basic approaches to language model adaptation exist.

- Training a language model in a new domain if sufficient data is available.
- Pooling data of many domains with the data of the new domain.
- Linear interpolation of a general and a domain specific model (Seymore, Rosenfeld, 1997).
- Back-off of domain specific probabilities with those of a specific model (Besling, Meier, 1995).
- Retrieval of documents pertinent to the new domain and training an language model on-line with those data (Iyer, Ostendorf, 1999).
- Maximum entropy, minimum discrimination adaptation (Chen, Seymore, Rosenfeld, 1998).
- Adaptation by linear transformation of vectors of bigram counts in a reduced space (DeMori, Federico, 1999).

Here we try to apply the approach “Retrieval of documents pertinent to the new domain and training a language model on-line with those data” on a machine translation task.
We also use a local index unlike the approach presented by Zhu and Rosenfeld in (2001), who query web search engines to improve their language models.

2. Language Model Adaptation based on Information Retrieval

2.1. Selecting similar documents with TF-IDF

For the first experiments we used the TF-IDF similarity measure. TF-IDF is a way of weighting the relevance of a query to a document.

TF-IDF is widely used in information retrieval. The main idea of TF-IDF is to represent each document by a vector in the size of the overall vocabulary. Each document D_i is then represented as a vector $\left(w_{i1}, w_{i2}, ..., w_{in}\right)$ if n is the size of the vocabulary. The entry w_{ij} is calculated as:

$$w_{ij} = tf_{ij} * \log(idf_j) .$$

tf_{ij} is the term frequency of the j-th word in the vocabulary in the document D_i i.e. the number of occurrences.

idf_j is the inverse document frequency of the j-th

term, given as

$$idf_j = \frac{\#\,\text{documents}}{\#\,\text{documents containing j - th term}}$$

The similarity between two documents is then defined as the cosine of the angle between the two vectors.

2.2. Test and training data

The test data for all experiments translating from Chinese to English consisted of 993 sentences in 105 news stories. (TIDES test data December 2001)

The data for the information retrieval index is data from Xinhua news service from the years 1991-2001. For the 200 million word index we used all available data and only a part of this data for the 40 million word index. For all information retrieval applications the Lemur-Toolkit (Lemur Toolkit) was used.

2.3. Language Model Adaptation using story-level retrieval

In the first experiment the index contained approximately 180 000 stories with 40 million words. We calculated the perplexity of the adapted language models with the top 10, top 100 and top 1000 stories for each 105 stories and for a general language model using all 40 million words.
The story selection for the adapted language model was done based on the reference translation. This showed a perplexity reduction of up to 39%.

LM type	Average perplexity values	
General LM	199.20	100%
Top 10 LM	125.93	63%
Top 100 LM	147.30	74%
Top 1000 LM	120.68	61%

Table 1: Perplexity results for 40 million word index and reference translation (LM: language model)

Using a larger index of 200 million words (1 million documents) we could reduce the perplexity by 32%.

LM type	Average perplexity values	
General LM	143.30	100%
Top 10 LM	100.89	70%
Top 100 LM	118.79	83%
Top 1000 LM	96.84	68%

Table 2: Perplexity results for 200 million word index and reference translation

It is not realistic to use the reference translation as a query so besides using the reference translation we also used automatic translations with NIST scores (mteval metric) of 7.18 and 7.90 respectively.

LM type	Average perplexity values	
General LM	143.30	100%
Top 10 LM	110.28	77%
Top 100 LM	129.32	90%
Top 1000 LM	117.59	82%

Table 3: Perplexity results for 200 million word index and 7.18 translation

LM type	Average perplexity values	
General LM	143.30	100%
Top 10 LM	102.95	72%
Top 100 LM	128.56	90%
Top 1000 LM	113.72	79%

Table 4: Perplexity results for 200 million word index and 7.90 translation

As expected the perplexity reduction was lower than in the preceding experiments using the reference translation as queries. But 28% perplexity reduction was still possible when using the translation with a NIST score of 7.90.
The translation with a score of 7.18 gave a perplexity reduction of up to 23%.

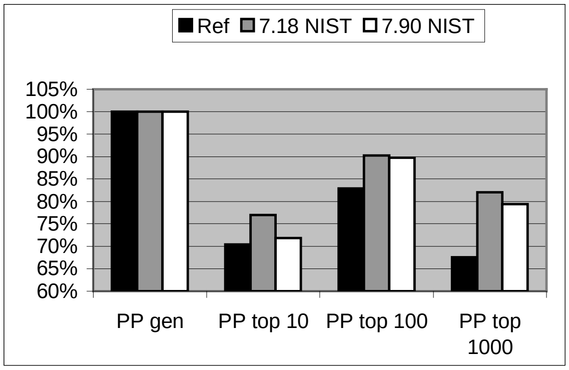

Figure 1: Overview of perplexity reduction using 200 million word index

A final experiment using story-level retrieval showed that the highest reduction of 39% was possible at 9 700 stories. We used only the first 20 test documents and calculated perplexities for top10,20, 30...1000, 1100, 1200...10000, 11000, 12000...100000 language models. Figure 2 shows the sum of the perplexity percentages (compared to the general language model with 1 000 000 documents). The perplexity graphs for each story do not show a clear picture (Figure 2 shows graphs for 4 different stories)

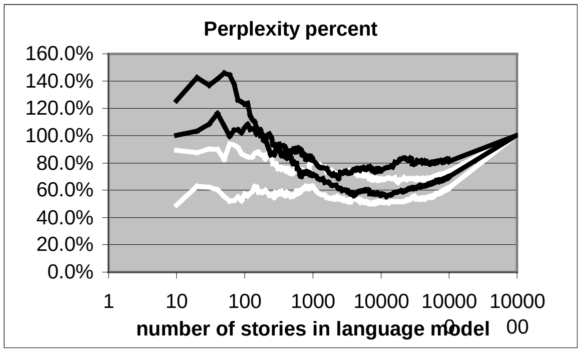

Figure 2: Perplexity percentage for 4 news stories

Figure 3 shows the sum of the perplexity percentages (compared to the general language model with 1 million documents). Here we get a nice graph and clear minimum at around 10 000 (exact number: 9 700) news stories with an average reduction in perplexity of 39%.

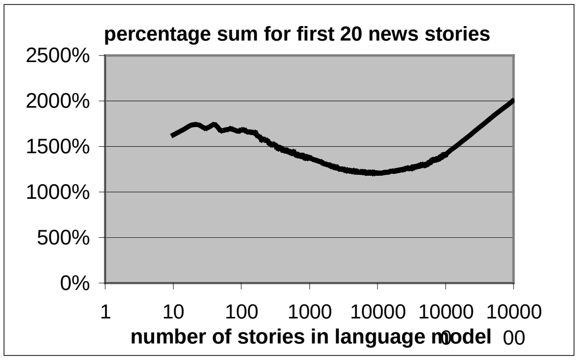

Figure 3: Perplexity percentage sum for first 20 news stories

But even when using those language models with minimal perplexity the translation quality did not improve.

2.4. Language Model Adaptation using sentence-level retrieval

Using the same index of 200 million words (and in this case 7 million sentences) of Xinhua news data we did similar experiments using sentences as documents, i.e. building an individual language model for each sentence to be translated.

We used two different translation systems for these experiments, an Chinese→English system, that had a baseline score of 7.12 (NIST) and an Arabic→English system with a baseline score of 7.32 (NIST).

In this case perplexity reduction of up to 81% was observed at 1000 sentence size of the language model.

Table 5 shows the actual numbers for different language model sizes and the graph in Figure 4 illustrates the relationships.

LM type	Average perplexity values			
	7.12 translation Chinese→English		7.32 translation Arabic→English	
General LM	203.79	100%	380.33	100%
Top 100 LM	52.00	26%	93.39	25%
Top 1000 LM	39.47	19%	188.76	50%
Top 10000 LM	42.61	21%	199.43	52%
Top 100000 LM	83.87	41%	348.60	92%

Table 5: Perplexity results for 200 million word index and 7.12 translation/7.32 translation

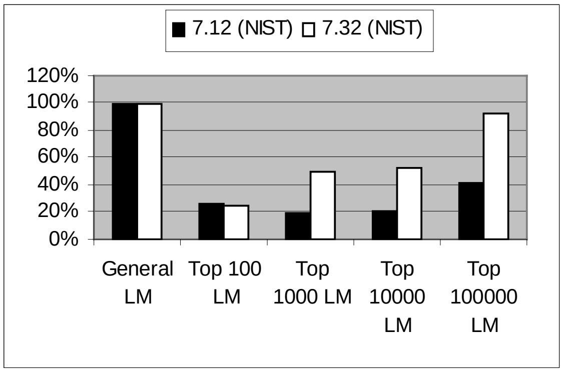

Figure 4: Perplexity results for 7.12 translation and 7.32 translation

As the perplexity results in the document-level experiments did not show any significant correlation to the later translation performance we did not calculate further perplexity values but relied more on actual translation experiments.

In these translation experiments for Chinese to English translation we compared the translation result using a 20 million word general language model with the translations when using specific language models. Starting from an initial translation scoring 7.12 (NIST) the best improvement could be observed when using 15 000 sentences to build the adapted language models.

However, the improvement of 0.06 in NIST score is not statistically significant.

Using the same index data but in this case translating from Arabic to English we could improve the translation score from 7.32 to 7.61 (NIST). In this case, the improvement is highly significant.

LM type	Translation scores	
	7.12 translation Chinese→English	7.32 translation Arabic→English
General LM	7.12	7.32
Top 1000 LM	6.98	7.25
Top 10000 LM	7.17	7.54
Top 15000 LM	7.18	7.61
Top 20000 LM	7.17	7.61
Top 50000 LM	7.06	7.19

Table 5: Translation scores for 200 million word index and 7.12 translation/7.32 translation

2.5. Further experiments and observations

We did further experiments to see if other information retrieval techniques could be used in this task.

An often used method in information retrieval is stemming. Stemming reduces derivative word forms to one root form of a word (stem).

Stemming takes the assumption that different derivative forms of words do not have different meanings so concerning topic they can be treated as the same word.

By using these stems instead of the actual words it should be possible to improve the calculation of the topic similarity.

Stemming was reported to be quite useful in the paper by Mahajan, Beeferman and Huang (1999). In this case the use of Porter´s stemmer when building an index did not show any improvements in translation quality.

A similar idea commonly used in information retrieval is the omission of stopwords.

Stopwords are words of little intrinsic meaning that occur too frequently to be useful in searching text. Typical examples for stopwords are: “a”, “by”, “neither”, “seem”, “those”, “how”, “no” etc.

By just leaving out these words the performance of information retrieval applications can usually be improved

In this case the usage of a stopword-list did not have a positive effect on the translation quality of the resulting language models.

We also tried two other retrieval methods that are offered by the Lemur-Toolkit in addition to TF-IDF. But both Okapi and SimpleKL gave no improvement compared to selecting the language models using TF-IDF.

3. Conclusion

In this paper we have shown that language model adaptation can be successfully applied to statistical machine translation. Not only did the adapted language models drastically reduce perplexity but also a NIST score improvement of up to 0.29 was possible.

The results show that sentence level adaptation outperforms document level adaptation.

The results also indicate that the correlation between perplexity of a language model and actual improvement is rather weak.

When even better translations are possible to use as queries this should further improve.

4. References

Besling S. and Meier H.G. (1995), *Language model speaker adaptation*, Proceedings Eurospeech 1995, Madrid, Spain.

Chen Stanley, Seymore Kristie and Rosenfeld Ronald (1998), *Topic adaptation for language modeling using unnormalized exponential models*, IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA.

DeMori Renato and Federico Marcello (1999), *Language Model Adaptation*, In "Computational Models of Speech Pattern Processing", Keith Pointing (ed.), NATO ASI Series, Springer Verlag.

Iyer R. and Ostendorf M. (1999), *Modeling long distance dependence in language: topic mixtures vs. dynamic cache models*, IEEE Transactions on Speech and Audio Processing, SAP-7(1):30-39.

Janiszek David, DeMori Renato and Bechet Frederic (2001), *Data Augmentation and Language Model adaptation*, IEEE International Conference on Acoustics, Speech and Signal Processing 2001, Salt Lake City, UT, USA.

The Lemur Toolkit for Language Modeling and Information Retrieval http://www-2.cs.cmu.edu/~lemur/

Mahajan Milind, Beeferman Doug and Huang, X.D. (1999), *Improved topic-dependent language modeling using information retrieval techniques*, IEEE International Conference on Acoustics, Speech and Signal Processing, Phoenix, AZ.

Seymore Kristie and Rosenfeld Ronald (1997), *Using story topics for language model adaptation*, Proc. Eurospeech 1997, Rhodes, Greece.

Zhu Xiaojin and Rosenfeld Ronald (2001), *Improving Trigram Language Modeling with the World Wide Web*, IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle 2001, Salt Lake City, UT, USA.