ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

IMPROVING LANGUAGE MODELING BY COMBINING HETEOGENEOUS CORPORA

Zheng-Yu ZHOU

Department of Computer Science
Fudan University, Shanghai
zhou_zhengyu@fudan.edu

Jian-Feng GAO

Microsoft Research Asia

Beijing

jfgao@microsoft.com

Eric CHANG

Microsoft Research Asia

Beijing

echang@microsoft.com

ABSTRACT

In applying statistical language modeling, directly adding training data (e.g. from website) may not always improve the performance of language models because the data may not be suitable for the application or contain errors. This paper presents a method of combining multiple heterogeneous corpora to improve the resulting language models, called compressed context-dependent interpolation scheme. The basic idea behind our method is that we not only want to filter good data, but also want to balance it among all the training data in order to give greater emphasis to data that better matches real usage scenarios or better balances our overall training set. Improvement on the accuracy of phone-character conversion has been observed in our experiments.

1. INTRODUCTION

Statistical language modeling (SLM) has been widely used in many applications, such as speech recognition [1], information retrieval [2], and spoken language understanding [3]. The typical approach of SLM in speech recognition system is to build an n-gram language model (LM), which is relatively easy to be incorporated into a recognizer and has been demonstrated to be highly effective. N-gram models estimate the probability of the word given its preceding word string as conditional probability, $P(w_n \mid w_1,...w_{n-1})$.

N-gram models are very successful when enough text data are used to estimate the model parameters. However, for many real tasks such as speech recognition, n-gram models suffer from the sparse-data problem because large amounts of domain-specific training text are not available. To deal with this problem, several methods [4-8] have been proposed to improve language modeling by combining data sets from different domains, and of mixed quality.

Previous approaches to combining training data can be grouped into two categories: combining counts, and combining models. In the former, n-gram counts from different corpora are combined with different weights. Although the method is simple and efficient, the weights are determined in an *ad hoc* fashion. In the latter approaches, several models are first generated on different corpora, and then combined by linear interpolation. Interpolation weights are tuned by using EM algorithm. In addition to the inefficiency of the EM, this approach suffers from the sub-optimal weight tuning because perplexity, which is used as the metric of tuning weights, does not always have a strong relationship with the performance of the combined models.

To our knowledge, most previous approaches are based on the assumption: the more data the better. That is, more training data will surely improve the performance of an n-gram model. However, this assumption is not always true in real applications. A large portion of available data, such as the web data, is of highly variable quality and contains errors. Directly adding them into training corpus will usually hurt the performance of the resulting LM. How to combine multiple heterogeneous data corpora effectively is the focus of this study.

In this paper, we propose a new approach of combining multiple heterogeneous data corpora, called compressed context-dependent interpolation scheme. Our method first filter good data from the variable-quality corpora, and then combine two data sets by n-gram counts based on context-dependent interpolation. The interpolation weights are tuned in a way to seek the balance among data sets according their relative contributions to the performance. Experimental results show that this approach brought a 16.5% reduction in error rate of phone-character conversion. We also implement traditional methods such as linear interpolation, for the sake of comparison. Unfortunately, these methods achieved no improvements partly because their assumption was not satisfied in our experiment settings. The results also indicate that a corpus of even poor quality is not necessarily useless, and is still possible to be utilized to improve the performance of the LM.

In Section 2, we discuss prior SLM research and related works on LM combination. In Section 3, we propose our new algorithm, the compressed context-dependent interpolation scheme. In Section 4, we compare our method with previous methods, and present experimental results. Finally, we present our conclusions in Section 5.

2. TRIGRAM AND LM COMBINATION

N-gram models, which estimate the probability of a word given its *n-1* previous words, are one of the most successful forms of LM. In practice, *n* is usually set to 2 (bigram), or 3 (trigram). For simplicity, we restrict our discussion to trigram model, $P(w_i|w_{i-2},w_{i-1})$, which assumes that a word depends on the previous two words. But our approach extends to any n-gram.

One of the key issues in language modeling is the lack of training data (i.e. the sparse-data problem). For many specific tasks, it is difficult to collect enough high-quality training data to build a robust language model. A large amount of available data is of variable quality (e.g. out-of-domain, or from website). Naturally, we would like to utilize this large data set to improve the language modeling.

The most common way to utilize the variable-quality data is to use Brute-Force approach. It generates a single model based on counts from both the original training data (seed set) and the variable-quality data (optional set) as follows:

$$C_{bf}(w_{i-2},w_{i-1},w_i) = C_{SS}(w_{i-2},w_{i-1},w_i) + C_{OS}(w_{i-2},w_{i-1},w_i)$$

where $Css(w_{i-2},w_{i-1},w_i)$ is the count of word sequence $w_{i-2}w_{i-1}w_i$ in seed set; while $Cos(w_{i-2},w_{i-1},w_i)$ is the count of $w_{i-2}w_{i-1}w_i$ in the optional set.

The Brute-Force approach is simple and efficient. It equals putting all data sets together and then training an LM. But this method can not guarantee the performance improvement, especially when the combined data sets are of variable-quality.

Instead of combining counts, one might also directly combine two language models trained on the two data sets respectively using a single interpolation as follows:

$$P_{si}(w_i \mid w_{i-2},w_{i-1}) = \lambda \times P_{SS}(w_i \mid w_{i-2},w_{i-1}) + (1-\lambda)\times P_{OS}(w_i \mid w_{i-2},w_{i-1})$$

where $Pss(w_i|w_{i-2},w_{i-1})$ is the trigram probability of the LM generated on the seed set, $Pos(w_i|w_{i-2},w_{i-1})$ is trigram probability of the LM trained on the optional set. The interpolation weight λ is tuned by EM algorithm using a cross-validation-like approach. Although the Single Interpolation Scheme is much more effective than Brute-Force Technique, it is still sub-optimal because perplexity is the tuning metric and its correlation with error rate is not as strong as once thought.

Recently, several improvements over traditional methods of count combination and model combination have been proposed. For example, the n-gram distribution based method [8] combines counts from different data sets by n-gram distribution adaptation. However, to our knowledge, almost all previous algorithms are based on an assumption: the more data the better. It is true because having more data can cover more language features. But blindly adding new corpus will sometimes hurt the model because the quality of the adding corpus may be poor (unsuitable to the application domain, or containing errors). Therefore, the negative effect of introducing noise will probably outweigh the positive effect of having more language features.

3. CONPRESSED CONTEXT-DEPENDENT INTERPOLATION SCHEME

As described in the previous section, most previous LM combination methods assume that the more data the better. But for many real tasks, such as large vocabulary speech recognition, a large portion of available data (e.g. web data) is of variable quality, and may contain errors (e.g. incorrect language usages). If these corpora are directly added into training data, it is highly possible that the performance of the resulting LM will be hurt.

Even for these “bad” corpora, it is still possible to utilize a portion of them to improve the performance of the LM. We propose a new approach to combining data. Our methods consists of three steps: (1) we first filter good data from the variable-quality corpora; (2) we then use the Witten-Bell smoothing technique [9] to estimate the context-dependent interpolation weights, and (3) we finally combine the seed data set with the variable-quality data sets at the level of n-gram counts.

We denote the good-quality training data set as *seed set*, and denote the other training sets of variable-quality as *optional sets*. Our approach combines trigram counts using the formula shown below:

$$C(w_{i-2}w_{i-1}w_i) = \alpha \times C_{ss}(w_{i-2}w_{i-1}w_i) + (1-\alpha)\frac{N_{ss}}{N_{os}}C_{os}(w_{i-2}w_{i-1}w_i) \quad (3.1)$$

$$\text{where } \alpha = \frac{C_{ss}(w_{i-1}w_i)}{C_{ss}(w_{i-1}w_i) + R_{ss}(w_{i-1}w_i)} \quad (3.2)$$

Here $C_{ss}(w_{i-2}w_{i-1}w_i)$ is the counts that $w_{i-2}w_{i-1}w_i$ appears in the seed set, $C_{os}(w_{i-2}w_{i-1}w_i)$ is the counts that $w_{i-2}w_{i-1}w_i$ appears in the optional set, N_{ss} and N_{os} are respectively the total counts of trigrams in the seed set and optional set. $R_{ss}(w_{i-1}w_i)$ is the number of unique words following the bigram $w_{i-1}w_i$ in the seed set.

We discount the n-gram counts of the optional set by $\frac{N_{ss}}{N_{os}}$ for two reasons. First, we want to compress the optional set into the same size with the seed set. Usually, the optional set is much larger than the seed set, directly interpolating the two count models will be meaningless. Second, this discount is also used to filter good data from the optional set. We only keep those n-grams with counts larger than a threshold, because highly-frequent n-grams are very likely to be correct. On the contrary, it is difficult to decide whether a low-frequency n-gram is right or not. So we simply do not take them into consideration.

Notice that the interpolation weight α is no less than 0.5. Thus we always give more emphasis on the seed set. The larger the α is, the more possible that the seed set itself is enough for LM training. On the contrary, if α is relatively small, the seed set will not be considered containing sufficient information for LM training, and we will try to improve the training by picking up supplemental information from the optional set.

Since Equations (3.1) and (3.2) use discount and context-dependent data-driven interpolation weights, this algorithm will be referred to as the compressed context-dependent interpolation (CCDI) approach afterwards.

4. EXPERIMENTS

We test our approach in the application of large vocabulary traditional Chinese speech recognition. In order to simplify the decoder, we do not take into account the acoustic ambiguity. We collected about 2.1 gigabytes of traditional Chinese text data, which contain several Taiwan newspapers and web news. Among them, only about 420 megabytes corpus is of good quality and is used as the seed set (SS), while the remaining 1.7 gigabytes of data which contains many errors (e.g. wrong usages) is used as the optional set (OS). We also collected an 1 megabyte balanced test data, along with 1 megabyte balanced developing data used to estimate the interpolation weight for single interpolation scheme. The lexicon we used contains 43,818 entries.

In our experiments, all language models were pruned to the proper size using a entropy-based cutoff method [10]. We compared different LM combination approaches in terms of both perplexity and character error rate (CER) of phone-character conversion. While perplexity is the most common metric for evaluating a LM, CER is a measure that we are particularly interested in because it is strongly correlated to the performance of a speech recognizer. The experiment results are shown in table 1.

We first combined data sets by Brute-Forth (BF) approach, which is equal to putting all corpora together. The results are even worse than the baseline system, which used the SS as training data only. It turns out that more data does not always lead to better performance in our experiments. Unsurprisingly, we found that most other traditional approaches to combining counts or combining models, such as the single interpolation scheme, do not work in our situation as shown in Table 1.

Data	Approach	Perplexity	CER
SS	Baseline	181.07	7.72%
OS		1199.75	15.22%
SS + OS	BF	317.35	8.21%
SS + OS	SI	193.26	8.38%
SS + OS	CCDI	184.80	6.44%

Table 1: Perplexity and CER results using the brute-force (BF) approach, the single interpolation (SI) scheme, and the compressed context-dependent interpolation (CCDI) approach, compared to the baseline LM trained on the seed set.

We then performed experiments of combining *n*-gram counts by CCDI. The results are very encouraging. We achieved 16.5% error rate reduction over the baseline system. The results show the effectiveness of our method.

We also found that while we obtained a reduction in CER by the CCDI approach, a slight increase in perplexity is observed. The fact that perplexity increases correspond to the better recognition results is also reported in [5,8]. It shows that perplexity might not be a good measure to evaluate the performance of an LM. Recent research in LM evaluation measures [11] may help to find a better criterion for estimating language model interpolation weight.

5. CONCLUSION

In this paper, we present a new algorithm, the compressed context-dependent interpolation approach, to combine heterogeneous corpora at the level of n-gram counts. We first pick up correct information from variable-quality corpus, and then interpolate it with good-quality corpus using context-dependent interpolation weight. Experimental results show that the CCDI approach achieve a 16.5% reduction in CER of phone-character conversion, while most previous approaches do not work in our experiments because the assumption they are based on, “ the more data the better”, did not satisfied. The basic idea behind our method is that we not only want to filter good data, but also want to balance it among all the training data in order to give greater emphasis to data that better matches real usage scenarios or better balances our overall training set.

6. ACKNOWLEDGES

We would like to thank Mingjing Li, Michael Yang, Terry Wang, Frank Seide, Ashley Chang, and many other researchers from Microsoft Research Asia. Thanks also to Guihong Cao from Nankai University, Ming Zhang from Tsinghua University, for their helps on data collection and model training.

7. REFERENCES

[1] Jelinek F. “Self-organized language modeling for speech recognition”, in *Readings in Speech Recognition*, Morgan-Kaufmann, San Mateo, CA, pp.450-506, 1990.

[2] Miller D., Leek R., Schwartz R. M. “A hidden Markov model information retrieval system”, in *Proc. 22nd International Conference on Research and Development in Information Retrieval*, Berkeley, CA, pp. 214-221, 1999.

[3] Zue V. W. “Navigating the information superhighway using spoken language interfaces”, *IEEE Expert*, vol.10, no. 5, pp.39-43, October, 1995

[4] Gao J., Goodman J., Li M., and Lee K. F. “Toward a unified approach to Chinese SLM”, *ACM Transactions on*

Asian Language Information Processing, Vol. 1. No. 1, pp 3-33, March 2002.

[5] Iyer, R., Ostendorf, M., and Gish, H. "Using out-of-domain data to improve in-domain language models", *IEEE Signal Process Letter*. 4:8, Aug., 1997.

[6] Clarkson, P. and Robinson, A. "Language model adaptation using mixtures and an exponentially decaying cache", In *Proceedings of the ICASSP-97*, 1997

[7] Seymore, K., and Rosenfeld, R. "Using story topics for language model adaptation", In *ICASSP-97*, 1997.

[8] Gao, J., Li, M., and Lee, K. F. "N-gram distribution based language model adaptation", In *ICSLP-2000,* Beijing, Oct. 16-20, 2000

[9] Witten, I. H., and Bell, T. C. "The zero frequency problem: Estimating the probabilities of novel events in adaptative text compression", *IEEE Transactions on Information Theory*, 37(4): pp.1085-1094, July, 1991

[10] Stolcke, A., "Entropy-based Pruning of Backoff Language Models", In *Proc. DRAPA News Transcription and Understanding Workshop*, Lansdowne, VA. pp.270-274, 1998

[11] Clarkson, P., and Robinson, T., "Improved language modeling through better language model evaluation measures", *Computer Speech and Language* 15. 39-53, 2001