ＴＲ－ＳＬＴ－００４９

Comparison of Chinese, Japanese and English: Applying Multi-class N-gram Language Model

Rui Yang, Hirofumi Yamamoto, Yoshinori Sagisaka

２００３年９月１８日

概要

This document presents my research work in ATR for one and a half months. In Multi-class Composite N-gram language model (Yamamoto, 2001), we use different weight to merge backward information and forward information. It shows the improvement of perplexity, both on Japanese and English for small corpus size (Fadi Badra, 2003, TR-SLT-0046). This experiment applied the Multi-class bigram language model to Chinese, to determine which weight of the backward and forward information is better for Chinese. Based on corpus of BTEC (Chinese version, rev.1), the result of the experiment shows that Multi-class Ngram model is also effective for Chinese language. The best weight of backward information is 1.0, while that of forward information is 0.0.

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目２番地２　TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

Introduction

Word N-grams language model has been widely used as a statistical language model for language processing. Word N-grams language model is model that gives the transition probability of the next word from the previous N-1 word sequence based on a statistical analysis of the huge text corpus. In this case, the accuracy of the word prediction capability will increase according to the number of the order N, but also the number of word transition combinations will exponentially increase. The size of training data for reliable transition probability values will also dramatically increase. This is a critical problem for spoken language in that it is difficult to collect training data sufficient enough for a reliable model.

As a solution to this problem, class N-grams language model is proposed. In class N-grams, multiple words are mapped to one word class and the transition probabilities form word to word are approximated to the probabilities from word class to word class.

Multi-class N-gram language model has been proposed and been extended to Multi-class Composite N-gram model. In the Multi-class method of word clustering, multiple classes are assigned to each word by clustering the following and preceding word characteristics separately. This word clustering is performed based on the word connectivity in the corpus.

In this report, we applied the Multi-class bigram language model to Chinese, to determine which weight of the backward (preceding words) and forward (following words) information is better for Chinese. Based on BTEC corpus (Chinese version, rev.1), the result of the experiment shows that Multi-class Ngram model is also effective for Chinese language. The best weight of backward information is 1.0, while that of forward information is 0.0.

Background

The Multi-class language model has already been tested on Japanese. It has had satisfying result, achieving 9.5% perplexity reduction compared to word N-grams and 16% word error rate reduction. It also reduced the parameter size of 40%.

For English, we can achieve some effective result with only small training corpora .

And after compared the weight of connectivity information, we can find the best value for English is around 0.8, while that for Japanese is 1.0.

Purpose

- To apply *Multi-class bigram language model* to Chinese.
- To find a good weight of connectivity information.
- To make the comparison among Chinese, Japanese and English, try to point out the resemblance.

Language modeling

N-gram model

One of the most common approximations for a language model is the N-gram model. The idea is to restrict the history of a word to its N-1 preceding ones when estimating its probability. The transition probability is given in the formula,

$$P = P(Wi \mid Wi\text{-}n\text{+}1, \cdots, Wi\text{-}2, Wi\text{-}1)$$

For example, in a bigram (N=2) model, each word will be estimated using only the very word preceding. Such as,

P (实验 | 进行 了 语音 识别) ≈ P(实验 | 识别)

Class N-grams

Class N-grams were proposed to resolve the problem that a huge number of parameters are required in word N-grams. In class N-grams, the transition probability of the next word from the previous N-1 word sequence is given in the formula,

$$P(Wi \mid Wi\text{-}n\text{+}1, \cdots, Wi\text{-}2, Wi\text{-}1) \approx P(Ci \mid Ci\text{-}n\text{+}1, \cdots, Ci\text{-}2, Ci\text{-}1) \cdot P(Wi \mid Ci)$$

Where, *Ci* represents the word class to which the word *Wi* belongs.

Problems in the Definition of Word Classes

In Class N-grams, word classes are used to represent the connectivity between words. In the conventional word class definition, word connectivity for which words follow and that for which word precedes are treated as the same neighboring characteristics without distinction. Therefore, only the words that have the same word connectivity for the following words and the preceding word belong to the same word class, and this word class definition cannot represent the word connectivity attribute efficiently.

For example, English words “a” and “an”, both will be classified by POS as an indefinite article, and be assigned to the same word class. In this case, information about the difference with the following word connectivity will be lost. On the other hand, a different class assignment for both words will cause the information about the community in the preceding word connectivity to be lost. This directional distinction is quite crucial for languages with reflection such as French and Japanese.

Multi-class

According to the preceding and following information, we can have examples of English and Chinese (Fig.1, Fig.2), where "to-class" means the set of preceding information, "from-class" means the set of following information.

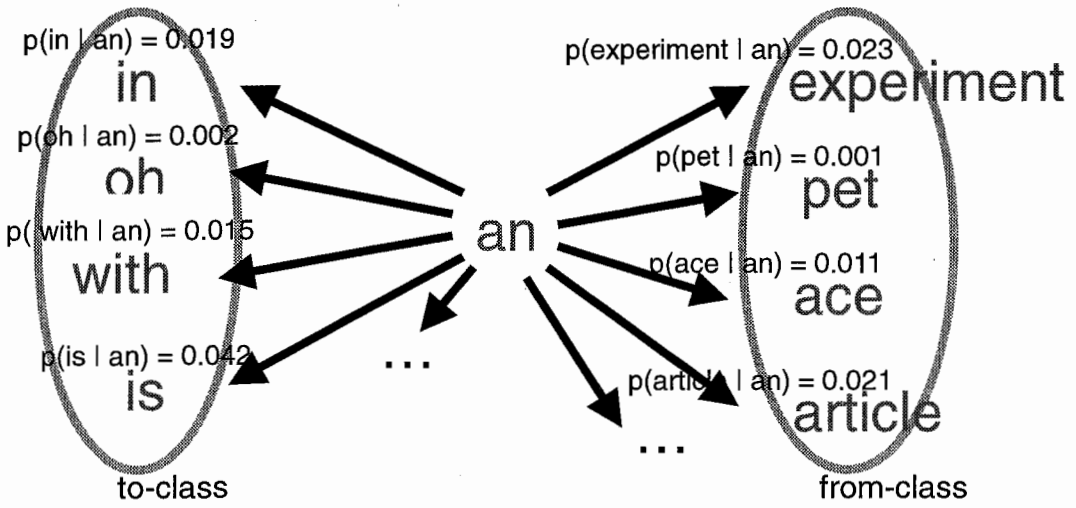

Fig.1 to-class and from-class -- English

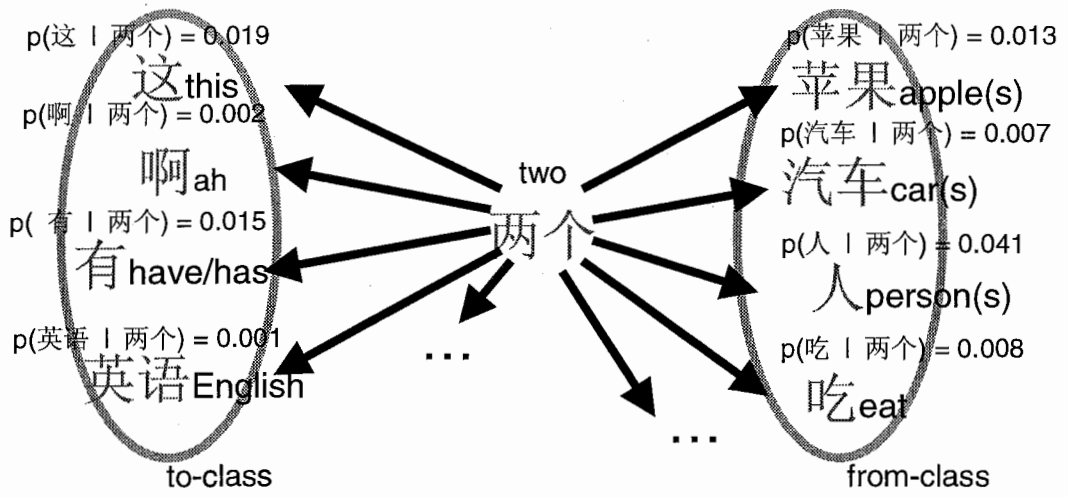

Fig.2 to-class and from-class -- Chinese

The vectors are used to represent word neighboring characteristics. The elements of the vectors are forward or backward word bigram probabilities to the clustering target word after being smoothed. And we consider that word pairs that have a small distance between vectors also have similar word neighboring characteristics (Brown et al., 1992) (Bai et al., 1998).

In this method, the same vector is assigned to words that do not appear in the corpus, and the same word cluster will be assigned to these words. To avoid excessively rough clustering over difficult POS, we cluster the words under the condition that only words with the same POS can belong to the same cluster. POS that have the same connectivity in each Multi-class are merged. Word clustering is thus performed in the following manner.

1. Assign one unique class per word.s.

2. Assign a vector to each class or to each word X. This represents the word connectivity attribute.

   $v^t(x) = [p^t(w_1 \mid x), p^t(w_2 \mid x), ..., p^t(w_N \mid x)]$

   $v^f(x) = [p^f(w_1 \mid x), p^f(w_2 \mid x), ..., p^f(w_N \mid x)]$

   Where, $v^t(x)$ represents the preceding word connectivity, $v^f(x)$ represents the following word connectivity, and p^t is the value of the probability of the succeeding class-word bigram or word bigram, while p^f is the same for the preceding one.

3. Merge the two classes. We choose classes whose dispersion weighted with the unigram probability results in the lowest rise, and merge these two classes:

   $$U_{new} = \sum_w (p(w) D(v(c_{new}(w)), v(w)))$$

   $$U_{old} = \sum_w (p(w) D(v(c_{old}(w)), v(w)))$$

   where we merge the classes whose merge cost $U_{new} - U_{old}$ is the lowest. $D(v_c, v_w)$ represents the square of the Euclidean distance between vector v_c and v_w, c_{old} represents the classes before merging, and c_{new} represents the classes after merging.

4. Repeat step 2 until the number of classes is reduced to the desired number.

Environment Settings

- Corpus:

  BTEC (Chinese Version revise 1), with 1,070,580 words

- Lexicon:

  made of corpus words.

- Test set:

  81k words (including BoS and EoS)

- N-grams number:

  bigram

- Smoothing method:

  “Good-Turing”

Procedure of Experiment

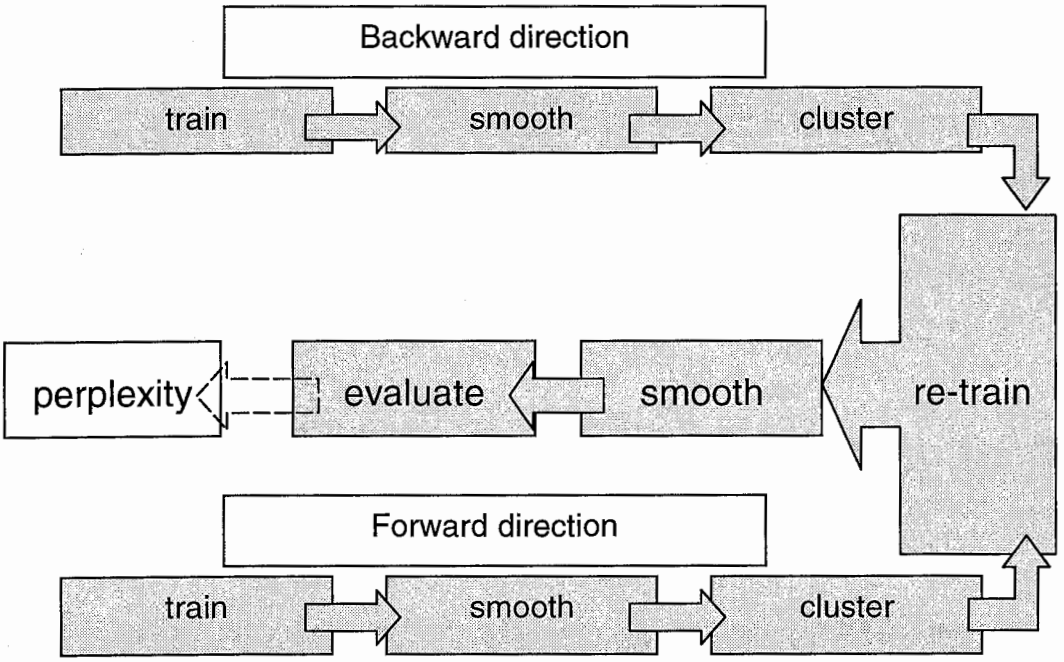

Fig.3 Procedure of Experiment

Result

The curves were drawn for different connectivity information each time.

Explanation of Graph

- Left to "/" = "set of classes for *Ct*(*w*i)"
- Right to "/" = "set of classes for *Cf*(*w*i-1)"
- b = backward (preceding information)
- f = forward (following information)
- First number of each line is the weight λ for different connectivity information

For example, when λ =0.8, "0.8b+0.2f / 0.2b+0.8f" means 80% of preceding information and 20% of following information were merged for to-class *Ct*(*w*i); and 20% of preceding information and 80% of following information were merged for from-class *Cf*(*w*i-1).

The zoomed-in graph were shown as following, x-axis is number of classes, y-axis is perplexity,

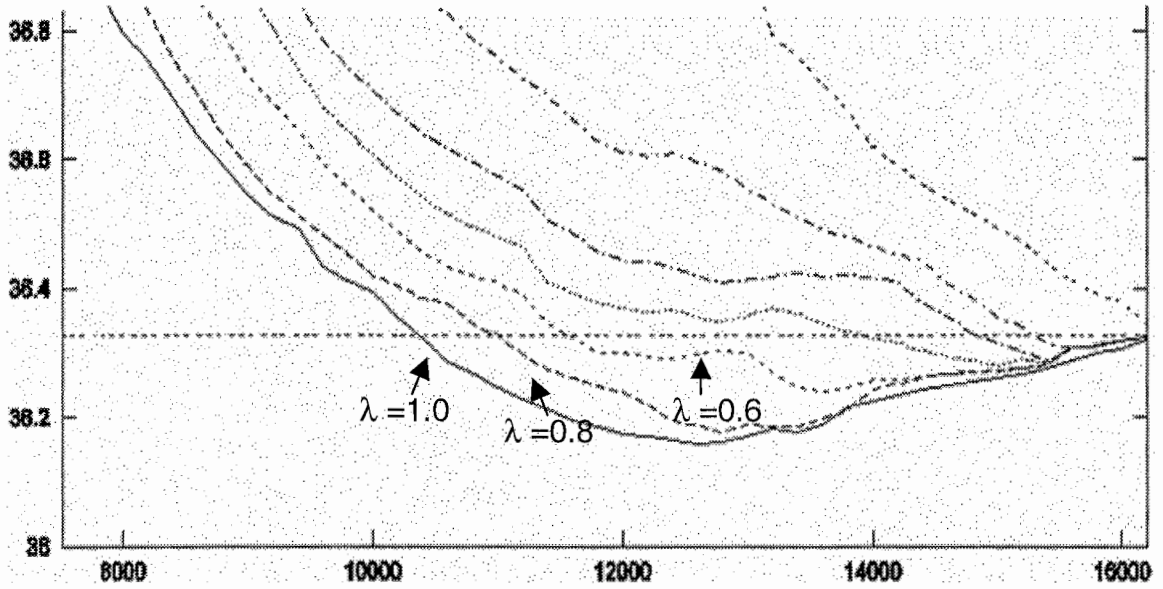

Fig.4 Result (zoomed in)

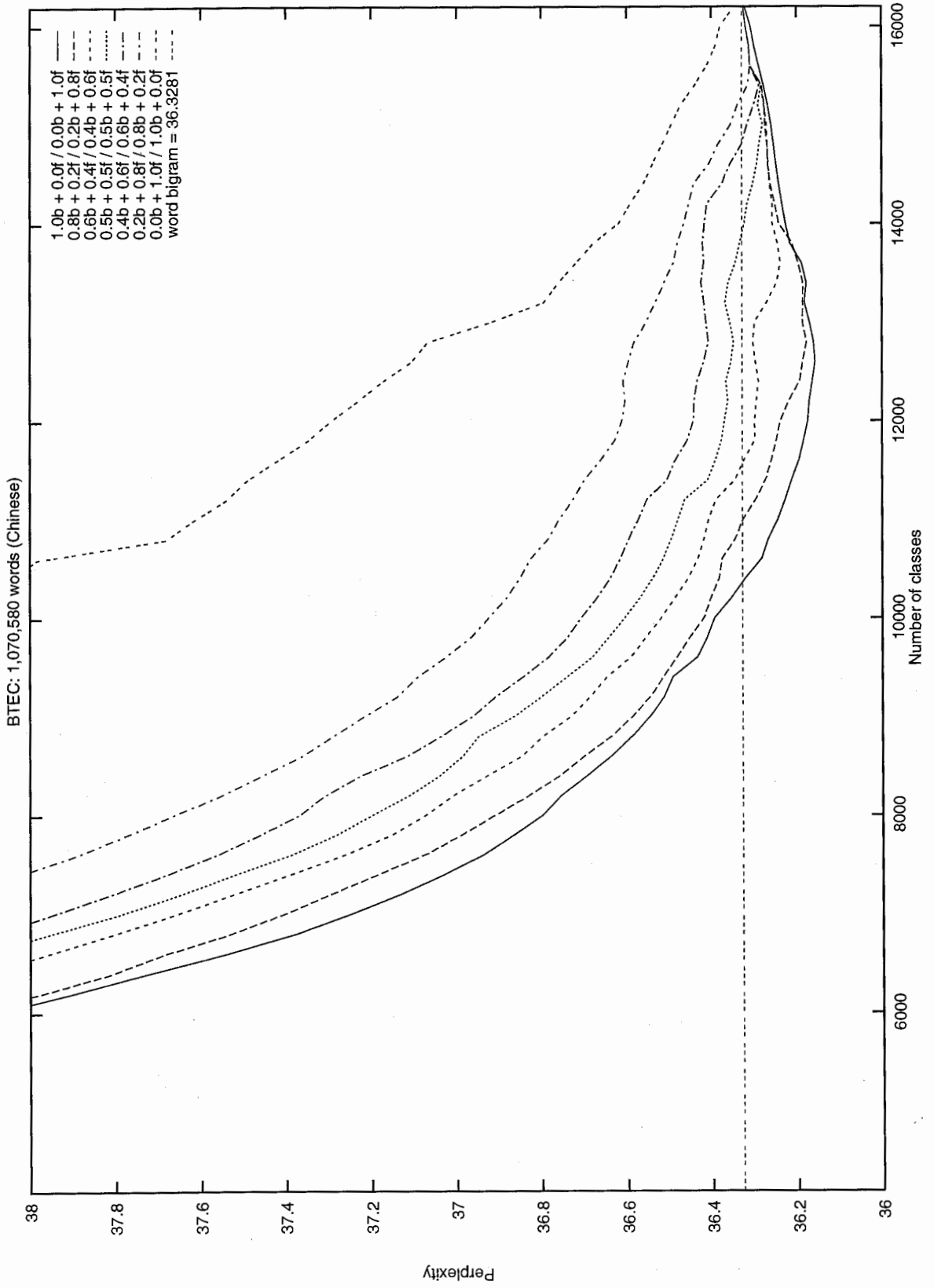

Fig.5 Graph of Result Curves

Comparison

	Chinese	Japanese**	English**
Training Corpus	BTEC(C)*rev*1	BTEC(J)	BTEC(E)
LM	Multi-class	Multi-class	Multi-class
Best λ	1.0	1.0	0.8
*Perplexity**	yes	yes	Only for small corpus size
*Number of Parameters**	not distinct	yes	not distinct

** Improved or not? compare with 2gram*

*** The data of Japanese and English is from Fadi Badra's technical report (TR-SLT-0046).*

Tab.1 Comparison

Conclusions

1. Multi-class bi-gram language model is also effective on Chinese.

2. Compared within different value of connectivity information, the best result is λ =1.0

3. Compared to the other two languages, Japanese and English, the result of Chinese is similar to that of Japanese, but different in the performance of parameter reduction.

Acknowledgments

I would like to thank *Dr. Yamamoto* and *Dr. Zhang* for all the directions they have given to me.

And also, I would like to thank all the members of *Dept.2* for their important help.

References

Books :

Manning and Shutze, 1999, Foundations of Statistical Natural Language Processing, *MIT press* .

Articles :

P. F. Brown, P. V. de Souza, R. L. Mercer, V. J. della Pietra, and J. C. Lai. 1992. Class-based N-gram Models of Natural Language. Computation Linguistics, 18(4):467-479.

Shuanghu Bai, Haizhou Li, and Baosheng Yuan. 1998. Building classes-based language models with contextual statistics. *Proc. ICASSP*, pages 173-176

H. Yamamoto, S. Isogai, and Y. Sagisaka, 2001, Multi-class Composite N-gram Language Model for Spoken Language Processing using Multiple Word Clusters.

H. Yamamoto and Y. Sagisaka, 1999, Multi-class Composite N-gram Based on Connection Direction. *Proc. ICASSP*, pages 533-536.

Jianfeng Gao, Joshua T. Goodman, and Jiangbo Miao, 2001, The Use Of Clustering Techniques for Language Modeling – Application to Asian Languages. Computational Linguistics and Chinese Language Processing, Vol. 6, No.1, pages 1-34.

Jianfeng Gao, Joshua T. Goodman, Mingjing Li, and Kai-Fu Lee, Toward a Unified Approach to Statistical Language Modeling for Chinese, *ACM Transactions on Asian Language Processing*

Fadi Badra and Hirofumi Yamamoto, 2003, Comparative Study on Multi-class Composite N-grams Applied to English and Japanese, ATR technical report, TR-SLT-0046