Internal Use Only（非公開）

TR－SLT－0052

Chinese Language Modeling based on BTEC Corpus: V1.0

Shuwu Zhang, Weishan Li

September 24, 2003

Abstract

This technical report describes the details of BTEC-based Chinese language modeling V1.0. The procedure of the modeling mainly consists of two steps. The first step is language preprocessing. In this step, a read-style based word segmentation and POS tagging tool has been applied for initial word segmentation of spoken-style BETC text files. Then, two versions of manual word recombination have been employed based on the initial processing for correcting the errors in segmentation, especially for some proper nouns such as foreign person names, city and hotel names. The perplexity changes with manual word recombination have also been investigated. In the second step, a set of language modeling approaches, including word N-gram, composite N-gram, multi-class N-gram, and multi-class composite N-gram, has been compared according to both measures of perplexity and recognition accuracy. Experimental results showed that multi-class composite N-gram with optimal configuration is a good language modeling approach for Chinese language as already testified in Japanese and English languages.

（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

Content

Abstract

This technical report describes the details of BTEC-based Chinese language modeling V1.0. The procedure of the modeling mainly consists of two steps. The first step is language preprocessing. In this step, a read-style based word segmentation and POS tagging tool has been applied for initial word segmentation of spoken-style BETC text files. Then, two versions of manual word recombination have been employed based on the initial processing for correcting the errors in segmentation, especially for some proper nouns such as foreign person names, city and hotel names. The perplexity changes with manual word recombination have also been investigated. In the second step, a set of language modeling approaches, including word N-gram, composite N-gram, multi-class N-gram, and multi-class composite N-gram, has been compared according to both measures of perplexity and recognition accuracy. Experimental results showed that multi-class composite N-gram with optimal configuration is a good language modeling approach for Chinese language as already testified in Japanese and English languages.

1. Introduction

Language modeling is one of indispensable core part of automatic speech recognition (ASR), It is also be regarded as a bridge spanning automatic speech recognition (ASR) to machine translation (MT) in developing speech-to-speech translation (S2ST) technologies. Aimed at the realization of real-word free communication, ATR-SLT has dedicated to the development of spoken S2ST system for simulating human interpreters for many years. Towards this goal, ATR-SLT is now collecting a set of English-Japanese aligned spoken corpora, called the basic travel expression corpus (BTEC), part of which (BTEC phase 1) has been translated into Chinese language.

Based on the initial Chinese version of BTEC corpus, we conducted a series of experiments and comparison on Chinese language modeling for pursuing a set of language models, which will be imported in Chinese ASR system and Chinese-Japanese speech-to-speech (S2S) translation system. This technical report will describe the details of BTEC-based Chinese language modeling V1.0.

2. BTEC Corpus

The Basic Travel Expression Corpus (BTEC) is a set of English-Japanese aligned corpora covering the utterances for every potential subject in travel conservations, together with their translation. The overall statistics of the first collection, called BTEC1, includes 200241 utterances and has been categorized into 20 topical classes such as Shopping, Transportation, Accommodation, Restaurant, Airport, Exchange, etc. (Kikui, et al, 2003). At ATR-SLT, a Chinese version of BTEC1 has been there which was translated mainly from corresponding English version by the collaboration with NLPR. The Chinese version of BTEC1 is now a plain text files. Based on this Chinese version of BTEC1, we conducted a series of language preprocessing as the preparation of language modeling.

3. Chinese language preprocessing

As described in last section, since Chinese version of BTEC1 is plain text files, we have to take some measures of language preprocessing such as segmentation, POS tagging for the language modeling. In fact, BTEC corpora can be regarded as a type of spoken style languages, it is inevitable to cause a certain amount of errors in word segmentation no matter what type of segmentation methods are used. We, thus, took two steps for word segmentation of the text. 1) to initially pre-segment text into words with

POS tags by a read-style oriented n-gram based segmentation & POS tagging tools; then 2) to manually check possible merging word list and do word recombination based on this merging list. Table 1 shows the information through the revisions (word recombination). In the version 1 of revision (seg0_rev1), a small portion of new words were detected, and the tag of new words were simply connected with combined units (characters or short words) by sign "+". The second version of revision was mainly focused on the manual detection of proper nouns such as English and Japanese person name, city name, hotel name, and so on. The total of 3765 new words have been collected in the merging list, and all possible units under the initial segmented corpus (seg0) were replaced automatically into combined words according to the merging list. The most of tags of new words have also been normalized in terms of original definition of tag set. Only a few of new words were tagged with new flags.

Table 1. BTEC Chinese Language Preprocessing

	Tokens		Tags	Voc	New words
	Train	Test			
Initial Seg.: seg0	1303291	72778	42	15426	--
Revision1: Seg0_rev1	1249206	68721	308	16333	907
Revision 2: Seg0_rev2	1188664	64115	54	19191	3765

Actually, this type of language preprocessing is just a kind of no way solution for spoken language in special domain. A potential solution should be with the bootstrapping processing. Fortunately, we may use this manual work as starting point to further do bootstrapping language processing (segmentation and tagging) iteratively for spoken language in upcoming work.

4. Chinese language modeling

Based on above language preprocessing, we conducted a series of experiments and comparison on Chinese language modeling for pursuing a set of the better language models which will be imported in Chinese ASR system and Chinese-Japanese speech-to-speech (S2S) translation system. This section will describe the details of the experiments and comparison step by step.

4.1 word based N-gram modeling

As we know, word based N-gram is one of the best-selling language modeling. We, thus, firstly trained a set of word N-gram models based on different processed corpora and evaluated the performances of these models by measure of perplexity (PP). Table 2 lists the information of the word N-gram modeling.

Table 2. Comparison of word based N-gram under different language preprocessing

Perplexity	Initial:seg0 Tags:42 Train set: 1.3M Wds Test set: 72778	Seg0_rev1: Tags: 308 Train set: 1.25M Test set: 68721	Seg0_rev2: Tags: 54 Train set: 1.19M Test set: 64115
Word bigram	32.95	36.32	39.33
Word trigram	14.02	14.59	17.23

From above table, we found a strange phenomenon, that is, the perplexity increased with the action of word recombination.

Generally speaking, the perplexity of character-based N-gram should be theoretically higher than that of word-based N-gram under very large unlimited corpora. Similarly, the perplexity of short word based N-gram should also be higher compared to that of long word based N-gram as listed in following simple example. We, thus, have to investigate the possible factors caused above problem.

The possible reasons caused PP increased by the processing of word combination would be from one of following factors or their combination.

✧ With so many utterance separators (UTT_END in answer files)? Generally, training text used for very large corpora based language modeling is treated as a long character/word sequence with just a pair of sequence start and end punctuations. But, for limited spoken corpus, we inserted utterance separators (UTT_START, and UTT_END) in each utterance. This could cause partially inaccurate of estimation in some cases.

✧ Related to lexicon size? Since vocabulary size increased after word combination, some redundant words have included in the lexicon. It maybe arose the perplexity increase sometimes.

✧ Smoothing coefficients? In N-gram language modeling, discounting drown from small frequencies of connections is often used for predicting the unseen events. It may cause the difference of evaluation in some occasions by using different smoothing strategies.

✧ Dispersed distribution of rare word connections after combination under unbalanced corpus? Because of very limited corpus of BTEC, the distribution of some connections has been dispersed after word recombination. It could cause re-estimation of the distribution of connections.

```
=================================================================
A simple example:
 Lexicon: 12 words
10003 [您] N IN {|sil}   10405 [早上] Z AO SH ANG {|sil}
10772 [好] H AO {|sil}
12496 [晚上] UAN SH ANG {|sil}
10276 [晚] UAN {|sil}
12636 [安] AN {|sil}
25425 [您好] N IN H AO {|sil}
25426 [早上好] Z AO SH ANG H AO {|sil}
25427 [晚上好]    UAN SH ANG H AO {|sil}
25428 [晚安] UAN AN {|sil}
5 [UTT-START] sil
6 [UTT-END] sil
ANS1: UTT_START  您好   早上好    晚上好    晚安    UTT_END   (6 wds) PP:   1.001699
ANS2: UTT_START 您   好   早上   好    晚上   好    晚   安    UTT_END   (9 wds) PP:   1.436706
=================================================================
```

It is not easy to say which factor mainly cause the perplexity increased. We have to investigate these factors one by one. Following tables show the investigations based on above analysis.

Table 3. Investigation I: with/without UTT_START/UTT_END for each utterance

PP	NO UTT_START/UTT_END	With UTT_START/UTT_END
Seg0	62.84	32.95
Seg0_rev2	77.33	39.33

Table 3 showed that the perplexity of the model based on word combination (seg0_rev2) is still higher compared to initial one even removing utterance separators. Compared to the model with utterance separators, the PP of the model without utterance separators are relatively higher, this is because that some false connections at the joint points between utterances, which didn't occurred in training set, can not be correctly evaluated. We, therefore, may remove the factor of utterance separator from the list of possible influencing factors.

Table 4. Investigation II: With reduced lexicon

PP	Used lexicon Entries: 21802 Words: 19191	Reduced lexicon Entries: 21369 Words: 18780
W2gram (seg0_rev2)	39.33	39.32

Next, we investigated the influence on lexicon size. We deleted all unused word in training from the lexicon. A moderate reduction of PP can be seen from Table 4. But, it is still higher compared to that of original seg0. So it is not a dominating factor for the increase of PP value.

Table 5: Investigation III: Discounting configuration in smooth?

PP	PP in train set (close test with discounting)	PP in test set (test without discounting)	PP in test set (good turing cut-off =5)
Seg0	35.84	30.53	36.32
Seg0_rev2	40.74	35.44	39.33

For the sake of eliminating the influence with discounting configuration in smooth, we moved the discounting in language modeling. The result listed in Table 5 showed that there is no obvious effect on PP changes with the selection of smoothing techniques.

Till now, only last possible factor, i.e. dispersed distribution of word connections after combination under unbalanced corpus, is remained. But it is hard to simulate unless with large enough corpus.

A presumable explanation would be that because the characters used in foreign name are usually rare words and the joint probabilities of these rare words are approaching to 1.0 in training set, it caused that the local connections of rare words has almost equal joint probabilities even with the connections of single characters in non-processed corpus (seg0). But dispersed distribution with the increment of the words via recombination incurred perplexing to judge possible following words.

We have to forejudge by leave one strategy that dispersed distribution of word connections after combination under unbalanced corpus would be the factor caused the increase of PP value after word recombination.

A warning based on above investigation is, thus, proposed that it is limited to evaluate language model simply by the measure of perplexity only, and it is better to do the evaluation combined with PP and recognition accuracy together.

4.2 Composite N-gram modeling

As many paper reported, variable length N-gram model outperforms conventional N-gram model in a specified domain. Since some consecutive words with high joint frequency has been bound into a new joint word in variable length modeling, it virtually improves the quality of the estimation of local connections by partially enlarging the order of predictable preceding word sequence. It, thus, may get the effect of local long order N-gram. In ATR language modeling tool, this type of variable order N-gram modeling is called Composite N-gram.

Based above processed Chinese BTEC corpus (seg0_rev2), we also trained a set of composite bi-gram models with different threshold of frequency. Table 6 shows a comparison of perplexity between word bi-gram model and composite bi-gram model. From the table, we can see that the perplexity of composite model has a substantial improvement compared to word based bi-gram model within the domain (as listed in column 2 and column 3). But with SLDB corpus, because of a special amount of mismatch in the domain, the PP of composite model is contrarily lower than that of word bi-gram. This is because that the joint probabilities of some connections has been dispersed into that of joint words, and joint words didn't take substantial effect on this relatively open domain. It is similar to above analysis on word N-gram comparison with different language preprocessing.

Table 6: Defined words vs. Variable length

Perplexity	Standard Test	Open 1: with seg0	Open 2: with SLDB
Word Bigran	39.33	117.27	1078.38
Composite Num. of joint words:8244	20.69	90.87	1795.85

* SLDB test set has some OOV words in testing

4.3 Multi-class N-gram Modeling

As a kind of class-based N-gram modeling, multi-class N-gram modeling (H. Yamamoto, et al. 2003) has already been employed in language modeling within ATR-SLT. Multi-class N-gram modeling is to cluster words multi-dimensionally into classes, where both of the left and right context Markov dependencies are employed as two kind of different correlative features for the automatic decision of classification.

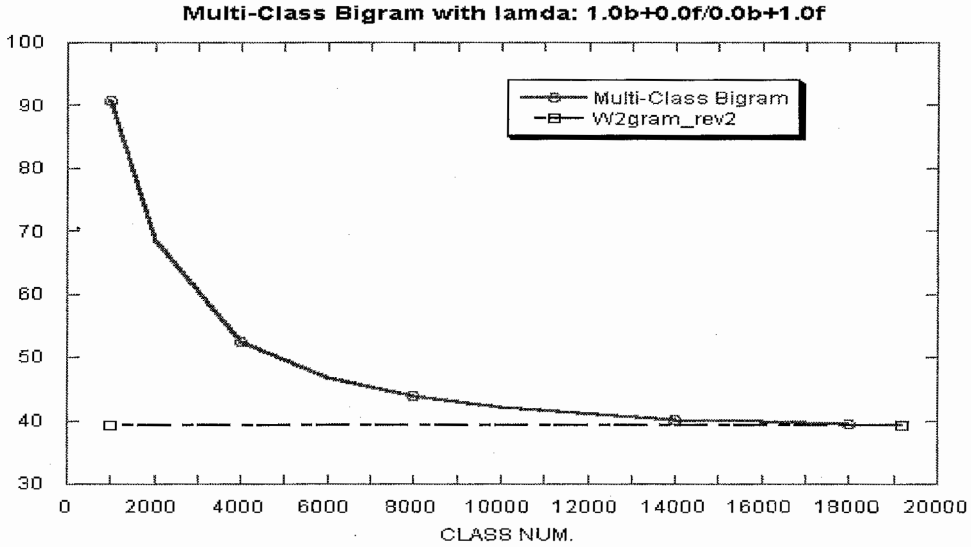

Fig. 1: Multi-class Bi-gram modeling

Based on the corpus seg0_rev2, we experimented the multi-class modeling. Figure 1 shows the perplexity changes in terms of the number of classes in clustering under parameter selection of "1.0b+0.0f/0.0b+1.0f". Compared to conventional word bi-gram, the PP of multi-class bi-gram gradually approaches to that of word bi-gram following the increase of number of classes. Some reports in Japanese, English, Chinese modeling (H. Yamamoto, et al. 2003, F. Badra 2003, R.Yang 2003) were also shown that multi-class bi-gram with special number of classes (say, 12k or 14k) may achieve the lower PP than word bi-gram model in some cases. But this phenomenon did not reflect on this experiment.

4.4 Multi-class Composite N-gram Modeling

Based on above modelings on composite and multi-class, we, further, train multi-class composite modeling by combining multi-class and composite techniques together (H. Yamamoto, et al. 2003). Figure 2 shows the curve of perplexity changes following the num. of classes used in modeling. Among of them, two types of different combination of forward and backward interpolation coefficients are compared. It showed that the models trained with λ =0.5 has a good shape of PP improved compared to the models with λ =1.0.

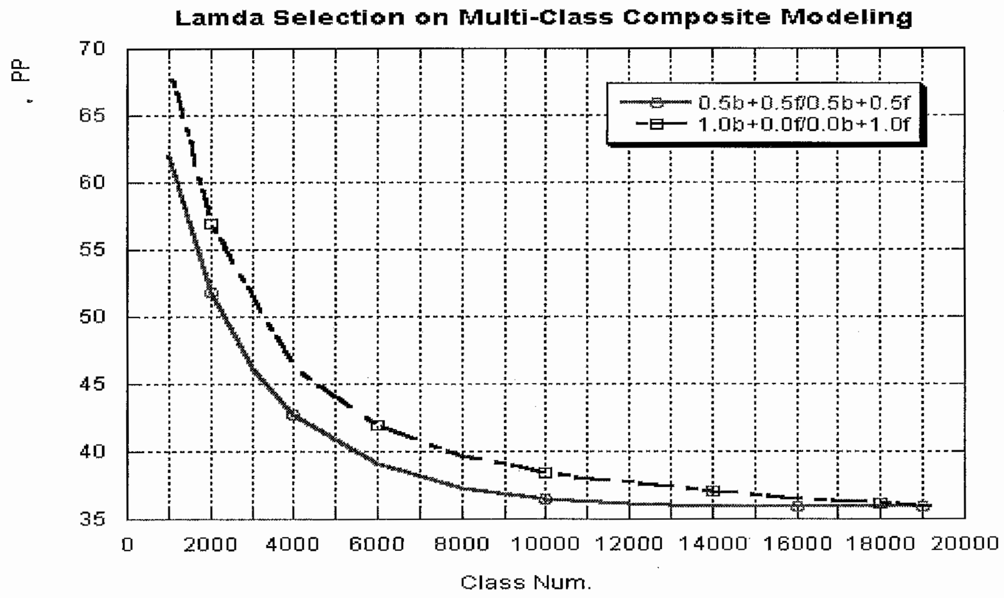

Fig. 2: Multi-class Composite Bi-gram modeling

However, a reversed result was reported in (Yang'2003). The setup difference between current one and Yang'2003 is

✧ Training set: current modeling experiments are based seg0_rev2 corpus, and

Yang'2003 was based seg0_rev1 corpus;

- ✧ Modeling: this is a multi-class composite modeling, and Yang'03 is the multi-class modeling without composite.

Even there is a little difference on setups, we still don't think it is reasonable with completely reversed conclusion. So, it is necessary to further do investigation on this problem in future.

5. Comparison and Evaluation

5.1 Model Comparison

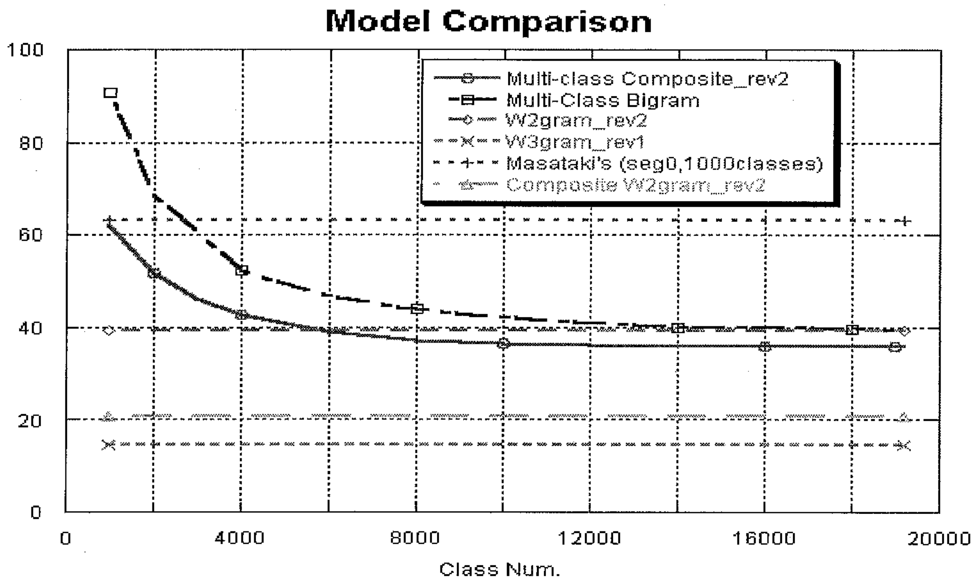

Fig. 3: Model Comparison

Figure 3 gives a comparative analysis on different modeling approaches. From the figure, we may get a priority rank of model selection simply by the measure of perplexity, that is word tri-gram, composite bi-gram, multi-class composite bi-gram with selected number of classes (larger than 6k), multi-class bi-gram, and word bi-gram.

But it is not enough to select model for ASR simply by the value of perplexity. We have to consider more factors, such as word accuracy of recognition, model size for system integration, and son on, to evaluate the real performance of selected language model in ASR. We, thus, conducted a series of experiments for a relatively integrative evaluation of model performance.

5.2 Model Evaluation

Based on above model comparison simply by the measure of perplexity, we further conducted a series of evaluation on language model with more integrative factors, especially with the measure of word accuracy in real ASR.

The acoustic model used in the evaluation was trained based on a combination of 863 Chinese speech database and HKU96 Mandarin speech database in the year of 2000.

It is internal accessible via the public path of /DB/PDB/863/AM/sprec/.

Evaluation based on CAS00 speech test data

Since there is no real test speech data on BTEC up to now, we firstly conducted the ASR experiments based on CAS00 adaptation speech data as tentative test data for the evaluation of language modeling. It consists of five travel-arrangement-based conversations of TAS32006, TAS32015, TAS33012, TAS33014, TAS33016, extracted from SLDB corpus. The speech uttered by szhang, total of 237 utterances.

To do evaluation of LMs, we have to do retrain language models by combing current BTEC corpus with transcriptions of CAS00 Chinese speech DB -- 50 hotel reservation oriented dialogues, which is a subset of SLDB corpus. Table 7 showed a comparative result of evaluation based on different language modeling approaches.

Table 7: Comparison of different modeling approaches based CAS00 test speech

<table>
<tr><td colspan="5">Experimental Setup:
AM: HMnet trained in HKU+863 (2000), LM: BITEC+CAS00 Trans.(HRT) under seg0_rev2
Test: 5 TASXXX sets with 237 utterances in TRA domain; Open for AM & LM</td></tr>
<tr><td>Models</td><td>W2gram</td><td>Multi-class
λ=0. 5</td><td>Multi-class composite:
λ=0. 5</td><td>Composite</td></tr>
<tr><td>Units</td><td>19356 words</td><td>12k classes</td><td>12k classes+
2247 joint. Wds</td><td>19356 wds+2247 joint words</td></tr>
<tr><td>Model Size</td><td>1.5G</td><td>576M</td><td>576.6M</td><td>1.9G</td></tr>
<tr><td>PP</td><td>129.07</td><td>93.27</td><td>83.45</td><td>47.99</td></tr>
<tr><td>Acc./ Corr.</td><td>80.45/86.56</td><td>78.08/86.03</td><td><b>82.89/86.96</b></td><td>84.02/89.58</td></tr>
<tr><td colspan="5">* No optimization for parameters. i.e., all models use the same config & LMscale
Beam=85,85, lmscale=6,10, work_area=1000,60, wdpenalty=20000,20000</td></tr>
</table>

From the table, we can see that multi-class composite model achieves a relatively better performance with regard to integrative factors of word accuracy, perplexity, and model size. Composite model, although, has a good recognition performance in word accuracy, it takes too big of memory cost -- 1.9Gigabyte as showed in the table.

Evaluation based on BTEC speech sample data

Explicitly, it is less evidence for complete evaluation on language modeling if just using out of domain speech data. We, thus, recorded a small set of speech test sample data based on BTEC domain, and conducted a series of explanative LM evaluation. The sample set of speech test data on BTEC consists of 2 male speakers (jzhang & szhang), total of 102 utterances, which has been extracted from BTEC test set: jpn_set01. AM used for the evaluation is still the same AM as above.

The first evaluation is to the test recognition performance of LMs with different language preprocessing as described in section 4.1.

Table 8. Test models with/without manual word recombination

<table>
<tr><td colspan="5">Experimental Setup:
AM: HMnet trained in HKU+863, LM: word bigram under different language preprocessing
Test: 2 males (szhang & jzhang) 102 utterances. <b>Open for AM & LM</b></td></tr>
<tr><td>Models</td><td>Seg0</td><td>Seg0_rev1</td><td>Seg0_rev2</td><td>Rev2+cas00</td></tr>
<tr><td>Units</td><td>15426 wds</td><td>16333 wds</td><td>19191 wds</td><td>19322 wds</td></tr>
<tr><td>Model Size</td><td>952M</td><td>1.07G</td><td>1.47G</td><td>1.5G</td></tr>
<tr><td>PP</td><td>32.95</td><td>36.32</td><td>38.31</td><td>39.68</td></tr>
<tr><td>Acc./ Corr.</td><td>80.19/83.44</td><td>79.97/83.59</td><td><b>82.29/85.79</b></td><td>82.10/85.61</td></tr>
<tr><td colspan="5">*No optimization for parameters. i.e., all models use the same config & LMscale
Beam=85,85, lmscale=6,10, work_area=1000,60, wdpenalty=20000,20000</td></tr>
</table>

Table 8 gives the comparison of word bi-gram based on different pre-processed corpus: seg0, seg0_rev1, and seg0_rev2. The result showed that the model based on seg0_rev2 outperforms the model on initial seg0 in word accuracy although its perplexity is higher than latter one. However, the model trained on the first version of revision (seg0_rev1) is still a little worse in word accuracy compared to both initial seg0 and revised seg0_rev2. It might be explained that the revision on seg0_rev1 would be not revised enough to compensate the dispersed distribution between rare words. Anyway, we reached a remark that the revision on seg0_rev2 is somehow more suitable for advanced language modeling compared to initial seg0 and revised seg0_rev1. We, thus, conducted the evaluation with different language modeling approaches based on seg0_rev2 corpus, hereafter.

Table 9 gives the comparison on different language modeling approaches based on seg0_rev2. A similar remark has also been gotten as the testing on CAS00 speech test data.

Table 9. Test different modelings under the segmentation of seg0_rev2

<table>
<tr><td colspan="5">Experimental Setup:
AM: HMnet trained in HKU+863 (2000), LM: BETC corpus under seg0_rev2
Test: 2 males (szhang & jzhang) 102 utterances. <b>Open for AM &amp; LM</b></td></tr>
<tr><td>Models</td><td>Word bigram</td><td>Multi-class</td><td>Multi-class composite</td><td>Composite</td></tr>
<tr><td>Units</td><td>19191 wds</td><td>12k classes</td><td>12k classes+
2164 joint words</td><td>19191 wds+
2164 joint words</td></tr>
<tr><td>Model Size</td><td>1.5G</td><td>576.26M</td><td>576.40M</td><td>1.75G</td></tr>
<tr><td>PP</td><td>39.33</td><td>39.47</td><td>38.55</td><td>26.07</td></tr>
<tr><td>Acc./ Corr.</td><td>82.29/85.79</td><td>82.10/85.79</td><td><b>82.66/86.10</b></td><td>85.06/87.45</td></tr>
<tr><td colspan="5">*No optimization for parameters. i.e., all models use the same config & LMscale
Beam=85,85, lmscale=6,10, work_area=1000,60, wdpenalty=20000,20000</td></tr>
</table>

Above evaluations have confirmed that multi-class composite modeling under better language preprocessing (seg0_rev2) has an integrative good performance compared to other modeling approaches. Furthermore, we investigate optimal parameter selection within multi-class composite modeling. First, we compared the performance of the models with the changes of interpolation coefficient and the number of classes in clustering. Table 10 gives the message of the comparison.

Table 10. Comparison of multi-class composite models with class changes

<table>
<tr><td colspan="5">Experimental Setup: AM: HMnet trained in HKU+863 (2000),
LM: multi-class composite under seg0_rev2 of BTEC corpus (2164 joint words)
Test: 2 males (szhang & jzhang) 102 utterances. <b>Open for AM &amp; LM</b></td></tr>
<tr><td>Num. of classes</td><td>12k
λ=1. 0</td><td>12k
λ=0. 5</td><td>16k
λ=0. 5</td><td>19191
λ=0. 5</td></tr>
<tr><td>Units</td><td>12k classes</td><td>12k classes</td><td>16k classes</td><td>19191 classes</td></tr>
<tr><td>Model Size</td><td>576.40M</td><td>576.40M</td><td>1.02G</td><td>1.47G</td></tr>
<tr><td>Perplexity</td><td>40.44</td><td>38.55</td><td>38.31</td><td>38.40</td></tr>
<tr><td>Acc./ Corr.</td><td>82.29/85.79</td><td><b>82.66/86.10</b></td><td>82.66/86.53</td><td>81.92/85.98</td></tr>
<tr><td colspan="5">*No optimization for parameters. i.e., all models use the same config & LMscale
Beam=85,85, lmscale=6,10, work_area=1000,60, wdpenalty=20000,20000</td></tr>
</table>

It showed that the selection with interpolation coefficient λ =0.5b+0.5f/0.5f+0.5b and classes num.=12000 is a better combination for multi-class composite modeling under seg0_rev2 corpus.

Table 11. Comparison of multi-class composite with num. changes of joint words

<table>
<tr><td colspan="5">Experimental Setup: AM: HMnet trained in HKU+863 (2000),
LM: multi-class composite under seg0_rev2 of BTEC corpus (λ=0.5, num. of classes=12k)
Test: 2 males (szhang & jzhang) 102 utterances. <b>Open for AM & LM</b></td></tr>
<tr><td>Classes</td><td>Freq=10</td><td>Frq=20</td><td>Frq=30</td><td>Freq=40</td></tr>
<tr><td>Units</td><td>12k classes+
8244 joint words</td><td>12k classes+
3553 joint words</td><td>12k classes+
2164 joint words</td><td>12 classes+
1527 joint words</td></tr>
<tr><td>Model Size</td><td>576.56M</td><td>576.44</td><td>576.4</td><td>576.38</td></tr>
<tr><td>Perplexity</td><td>36.09</td><td>37.98</td><td>38.55</td><td>38.93</td></tr>
<tr><td>Acc./ Corr.</td><td><b>84.69/87.82</b></td><td>82.10/85.61</td><td>82.66/86.10</td><td>82.10/85.24</td></tr>
<tr><td colspan="5">*No optimization for parameters. i.e., all models use the same config & LMscale
Beam=85,85, lmscale=6,10, work_area=1000,60, wdpenalty=20000,20000</td></tr>
</table>

Table 11 further shows the influence on the change of composite connections under fixed interpolation coefficient and class number. It showed that the more composite connections in the model are, the better the performance of model is.

Table 12. Model comparison: model size reduction by setting num. of classes

<table>
<tr><td colspan="5">Experimental Setup: AM: HMnet trained in HKU+863 (2000),
LM: multi-class composite under seg0_rev2 of BTEC corpus, λ=0.5, freq=10
Test: 2 males (szhang & jzhang) 102 utterances. <b>Open for AM & LM</b></td></tr>
<tr><td>Num. of classes</td><td>12k</td><td>10k</td><td>9k</td><td>8k</td></tr>
<tr><td>Units</td><td>12k classes+
8244 joint words</td><td>10k classes+
8244 joint words</td><td>9k classes+
8244 joint words</td><td>8k classes+
8244 joint words</td></tr>
<tr><td>Model Size</td><td>576.56M</td><td>400.57M</td><td>324.57M</td><td>256.57M</td></tr>
<tr><td>Perplexity</td><td>36.09</td><td>36.45</td><td>36.80</td><td>37.24</td></tr>
<tr><td>Acc./ Corr.</td><td><b>84.69/87.82</b></td><td>84.32/87.82</td><td>84.13/87.27</td><td>83.58/87.08</td></tr>
<tr><td colspan="5">*No optimization for parameters. i.e., all models use the same config & LMscale
Beam=85,85, lmscale=6,10, work_area=1000,60, wdpenalty=20000,20000</td></tr>
</table>

Table 12 also gives the message on model size reduction by reducing the number of classes. It has a slight word accuracy reduction by selecting the model with number of classes less than 12000. Considering the acceptance quota in memory size of current ATR recognition engine, we prefer to select the model under 12k classes without the loss of word accuracy.

Table 13. Model comparison: open LM vs. Close LM

Experimental Setup: AM: HMnet trained in HKU+863 (2000),

LM: multi-class composite under seg0_rev2 of BTEC corpus (λ=0. 5, num. of classes=12k)

Test: 2 males (szhang & jzhang) 102 utterances. **Open for AM & LM**

Models	Open LM	Close LM	No optimization for parameters. i.e., all models use the same config & LMscale
			Beam=85,85, lmscale=6,10,
			work_area=1000,60,
			wdpenalty=20000,20000

Units	12k classes+ 2164 joint words	12k classes+ 2164 joint words	
Model Size	576.4M	576.4M	
PP	38.55	33.43	
Acc./ Corr.	82.66/86.10	83.03/86.53	

A comparison between open LM and close LM is also given in Table 13. It showed that current train set has a good coverage to test set since only a slight improvement achieved by using close LM.

Table 14. Comparison of ASR configurations

Experimental Setup: **AM:** HMnet trained in HKU+863 (2000),

LM: multi-class composite under seg0_rev2, λ=0. 5, **freq=10,num. of classes:12k**

Test: 2 males(szhang & jzhang) 102 Utterances.

Configuration				Performance	
Beam	LM scale	Work Area	Word Penalty	Realtime	Acc./Crr.
60,60	4,8	800,60	0,0	4.58	70.85/81.37
70,70	6,10	800,60	20000,20000	1.53	70.66/79.34
	7,12	900,60	20000,20000	1.03	49.45/65.87
	4,8	800,60	0,0	12.28	76.39/84.07
85,85	6,10	1000,60	20000,20000	6.91	84.69/87.82
		900,60	20000,20000	6.88	**84.69/87.82**
	7,12	800,60	20000,20000	2.77	79.15/83.89
90,90	6,10	900,60	20000,20000	9.76	85.95/87.27
	7,12	900,60	20000,20000	3.53	83.21/85.79
95,95	7,12	900,60	20000,20000	5.63	84.13/87.27
100,100	8,14	800,60	0,0	5.96	81.0/84.32

Finally, a set of comparison on ASR configurations is conducted as showed in Table 14. Since current acoustical model and language model are still not yet final version for real Chinese ASR system building. We have to retry the optimal combination of ASR decoding configuration later.

6. Conclusion

Based on translated Chinese version of BTEC corpus, we have conducted a series of experiments on Chinese language modeling. Experimental results showed that multi-class composite N-gram with optimal configuration is also a good language modeling approach for Chinese as already testified in both Japanese and English.

Since current processed version of Chinese BTEC corpus (seg0_rev2) still has a portion of non-normalized proper nouns and segmentation errors while the definition of Part-of-speech (POS) is not closely suitable to BTEC, a quasi spoken-style domain-limited corpus. We should further do more studies and manual works on these points in near future. Meanwhile, some advanced investigations on ASR system building with real BTEC speech data should continually been taken.

Annex

Where is the model?

The model has been put into the directory:

/DB/PDB/BTEC-CHN/LM/V1.0/

Language model: *seg0_rev2.cc2gram.10.12k.bin (λ =0. 5, freq=10,number. of classes:12k)*

Lexicon: *rev2_composite10.lex+*

Acknowledge

I would like to express my highly appreciation to Dr. Seiich Yamamoto (Director of ATR-SLT) and Dr. Kikui Genichiro (Head of Dept. 2 at ATR-SLT) for kindly inviting me back ATR-SLT again, as well Kikui Genechiro-san, Hirofumi Yamamoto-san and professor Yoshinori Sagisaka for the warm supports and helps during the stay.

Also, I would like to thank Jinsong Zhang-san, Weishan Li-san, Xinhui Hu-san, and Rui Yang-san for the happy collaboration on the work

Thanks a lot to all of the dept.2 members and TSG members for their helps.

References

Kikui, G., Sumita, E., Takezawa, T. , Yamamoto, S., (2003) "Creating Corpora for Speech-to-Speech Translation", Eurospeech'2003

Yamamoto, H., Isogai, S., Sagisaka, Y., (2003), "Multi-class composite N-gram language modeling", Speech Communication 41 (2003) 369-379

Fadi Badra, Hirofumi Yamamoto, (2003), "Comparative Study on Multi-Class Composite N-grams Applied to English and Japanese", Technical Report, No: TR-ATR-SLT-0046, August 27, 2003.

Rui Yang, Hirofumi Yamamoto, Yoshinori Sagisaka, (2003), "Comparison of Chinese, Japanese and English: Applying Multi-class N-gram Language Model", Technical Report, No. TR-ATR-SLT-0049, September 17, 2003