언어모델 인터뷰 영향 평가를 통한 텍스트 코퍼스 균형 및 사이즈간의 통계 분석

정의정, 이영직

한국전자통신연구원 음성언어팀

Statistical Analysis Between Size and Balance of Text Corpus by Evaluation of the effect of Interview Sentence in Language Modeling

Eui_Jung, Jung, Youngjik Lee

Spoken Language Processing Team, ETRI

{euijoung, ylee}@etri.re.kr

Abstract

This paper analyzes statistically the relationship between size and balance of text corpus by evaluation of the effect of interview sentences in language model for Korean broadcast news transcription system. Our Korean broadcast news transcription system's ultimate purpose is to recognize not interview speech, but the anchor's and reporter's speech in broadcast news show. But the gathered text corpus for constructing language model consists of interview sentences a portion of the whole, 15% approximately. The characteristic of Interview sentence is different from the anchor's and the reporter's in one thing or another. Therefore it disturbs the anchor and reporter oriented language modeling. In this paper, we evaluate the effect of interview sentences in language model for Korean broadcast news transcription system and analyze statistically the relationship between size and balance of text corpus by making an experiment as the same procedure according to varying the size of corpus.

1. Introduction

Language model is very effective in large vocabulary continuous speech recognition (LVCSR), for example, broadcast news recognition. However, construction of statistical language model needs supporting of large text corpora. That is, if the text corpora are not sufficient, it is impossible to predict the relationship among words exactly. If there is an empirical research on large text corpora as the reference of design or selection of the training set for language modeling, the possible waste of human and monetary resource will be reduced. We evaluate the effect of interview sentences in language model for Korean broadcast news transcription system, and then analyze statistically the relationship between size and balance of text corpora by making an experiment as the same procedure according to varying the size of corpus. This paper is organized as follows. In section 2, we describe speech data and text corpora. In section 3, we present language model performance. Finally conclusions are drawn in Section 4.

2. Speech data and Text Corpora

The ETRI speech database is obtained by extracting audio signals from KBS9 News video tapes. The database consists of 54 day broadcast news from November 1, 1999 to February 1, 2000 including anchor and reporter speech and is used all as training set excepting

test set with 2 days of December 17, 1999 and January 17, 2000. To extract vocabulary and language models, we gather through broadcast news company(KBS, MBC)'s Internet website and the corpora's size is 14 millions totally. Detailly speaking, KBS 9 News transcriptions are 4 years from Sept. 1996 to Nov. 2000 and MBC News Desk transcriptions 4.5 years from Feb. 1996 to Sept. 2000. The gathered text corpora consist of interview sentences a portion of the whole, approximately 15%. The characteristic of Interview sentence is different from the anchor's and the reporter's in one thing or another.

Table 1. Text Corpora

Corpus	Size	Periods		
		Company	From	To
TD1	8 Millions	KBS	Sept. 1996	Feb. 1999
		MBC	Feb. 1996	Feb. 1999
TD2	14 Millions	KBS	Sept. 1996	Nov. 2000
		MBC	Feb. 1996	Sept. 2000

3. Language model performance

3.1 Experimental setup

As shown in Figure 1, we use the JRTk recognition toolkit to train acoustic model and SLM toolkit[3] to get trigram language model. We use a part-of-speech (POS) tagger[1] to segment eojeols into morphemes. Pronunciation dictionary is automatically obtained by using a morpheme-based grapheme-to-phoneme converter [2]. Speech signals are sampled at 16kHz to produce 16 bit data. The window size is 16ms and the frame shift is 10ms. We also use melcepstrum and its differentia coefficients as feature. Then a linear discriminant analysis (LDA) is performed to produce 24 dimensional features. We use 40 basic phonemes. For each phoneme, we use a 3-state hidden Markov model without skip transition. For observation probability, we use senone-based acoustic modeling with inter-morpheme coarticulation considered. The maximum context width is 2 for both left and right directions for intraword contexts and 1 for interword contexts. We use 3,000 senones and each senone has its own codebook with 16 Gaussian mixtures. For context clustering we use 47 detailed phoneme categories (e.g., vowel, consonants, fricative, and so on) as context questions in the decision tree. The vocabulary size is 64,014 including human and nonhuman noise. we use the backoff smoothing method to estimate probabilities with small data[4].

3.2 The Effect of Interview Sentences in LM

Our Korean broadcast news transcription system's ultimate purpose is to recognize not interview speech, but the anchor's and reporter's speech in broadcast news show. To evaluate the effect of interview sentences in text corpora, we construct the language models with text corpora TD1 and TD2 in two cases: one is " remove interview sentences" and the other is " does not remove interview sentences" .

Table 2. Recognition of Anchor and Reporter speech

In constructing LM	1999-12-17	2000-01-17
Does not Remove Interview sentences	80.2%	76.9%
Remove Interview sentences	81.7%	77.8%
ERR	7.58	3.90

Table 3. Recognition of Interview speech

In constructing LM	1999-12-17	2000-01-17
Does not Remove Interview sentences	36.6%	42.9%
Remove Interview sentences	35.2%	40.5%
ERR	-2.21	-4.21

In Table 2 and 3, TD1 is used in constructing LM, and in Table 4 and 5, TD2 is used.

Table 4. Recognition of Anchor and Reporter speech

In constructing LM	1999-12-17	2000-01-17
Does not Remove Interview sentences	83.9%	79.4%
Remove Interview sentences	84.0%	79.6%
ERR	0.62	0.97

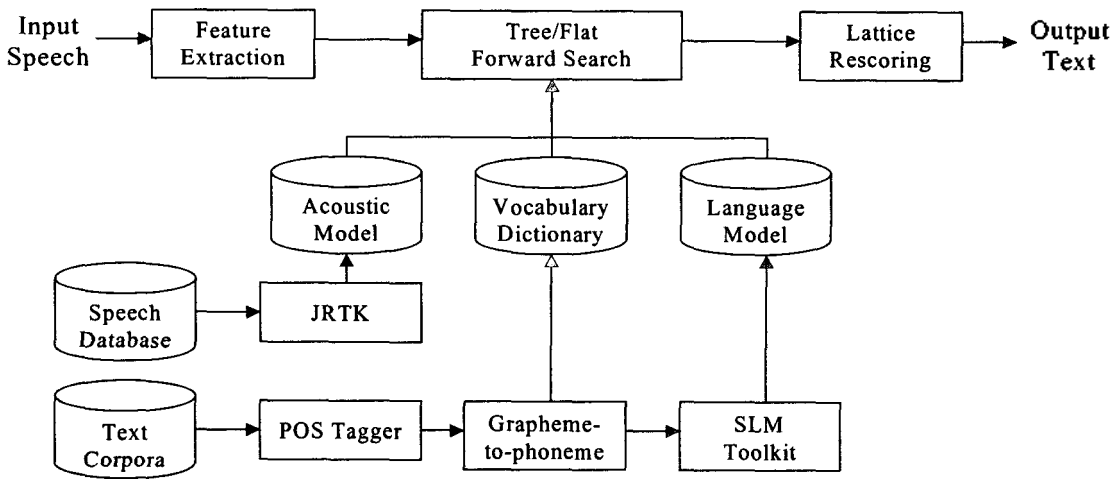

Figure 1. Experimental Setup of Broadcast News Recognition System

Table 5. Recognition of Interview speech

In constructing LM	1999-12-17	2000-01-17
Does not Remove Interview sentences	39.6%	44.2%
Remove Interview sentences	35.4%	41.5%
ERR	-6.95	-4.84

The characteristic of Interview sentence is different from the anchor's and the reporter's in one thing or another and disturbs the anchor and reporter oriented language modeling. Therefore removing interview sentences in constructing language model is effective for our system. In Table 2 and 4, when text corpora size is bigger, the effect of removing interview sentences is less remarkable. That is, if text corpora size is sufficiently big, we have no regard for the balance of corpora. But when text corpora are not enough, we consider the balance of corpora to improve the performance of LM.

3.3 The Effect of Corpus size in LM

To evaluate the effect of text corpus size, we construct the language model with text corpora TD1(8millions) and TD2(14millions). In Table 6 and 7, interview sentences are not removed in text corpora in constructing language model, and in Table 8 and 9, they are removed.

Table 6. Recognition of Anchor and Reporter speech

Text Corpora	1999-12-17	2000-01-17
TD1	80.2%	76.9%
TD2	83.9%	79.4%
ERR	18.69	10.82

Table 7. Recognition of Interview speech

Text Corpora	1999-12-17	2000-01-17
TD1	36.6%	42.9%
TD2	39.6%	44.2%
ERR	4.73	2.28

Table 8. Recognition of Anchor and Reporter speech

Text Corpora	1999-12-17	2000-01-17
TD1	81.7%	77.8%
TD2	84.0%	79.6%
ERR	12.57	8.11

Table 9. Recognition of Interview speech

Text Corpora	1999-12-17	2000-01-17
TD1	35.2%	40.5%
TD2	35.4%	41.5%
ERR	0.31	1.68

When text corpus size is increased, the performance of language model is improved remarkable as a matter of course. The effect of text corpus size is more effective than removing the interview sentences in text corpora.

3.4 Comparison of Perplexity

Perplexity is a widely used measure of language model quality.

Table 10. Comparison of Perplexity

Removing Interview	TD1 (8 millions)		TD2 (14 millions)	
	Do	Do not	Do	Do not
PP	288.62	268.09	220.02	220.50
Entropy	8.17	8.07	7.78	7.78
3-gram hit rate	45.89%	47.15%	52.59%	51.71%
2-gram hit rate	35.83%	35.67%	32.88%	33.59%
1-gram hit rate	18.27%	17.19%	14.53%	14.70%
OOV rate	1.78%	1.62%	1.38%	1.32%

In Table 10, the performance of the TD1 (8 millions) text corpora based LM is less remarkable than one of the TD2 (14 millions) text corpora based LM. That is, Text corpora size is critical for the quality of the LM. Whereas, we can' t find that removing the interview sentences is more effective in unsufficient text corpora. This is because the correlation between a language model' s perplexity and its effect on a speech recognition system is not as strong as was once thought. Perplexity is based solely on the probabilities of the words which actually occur in the test text. It does not consider the alternative words which may be competing with the correct word in the decorder of a speech recognizer. It' s time to start the exploration of new language model evaluation measure([5],[6]).

4. Conclusions

In this paper, we remove interview sentences in text corpora to construct language model for Korean broadcast news transcription system whose ultimate purpose is to recognize not interview speech, but the anchor's and reporter's speech in broadcast news show and test its performance. When text corpora size is bigger, the effect of removing interview sentences is less remarkable. That is, If text corpora size is sufficiently big, we have no regard for the balance of corpora. But when text corpora are not enough, we consider the balance of corpora to improve the performance of LM. We also use perplexity as a measure of language model performance and we confirm that text corpora size is critical for the quality of the LM. But we can' t find that removing the interview sentences is more effective in unsufficient text corpora. Perplexity can' t describe the balance of the text corpora.

ACKNOWLEDGEMENTS

This research was funded by the Korea Ministry of Information and Communication.

REFERENCE

[1] J.-H. Kim, Lexical disambiguation with error-Driven Learning, Ph.D. dissert. Dept. Computer Science, Korea Advanced Institue of Science and Technology, 1996.

[2] J. Jeon, S. Cha, M. Chung, J. Park, and K. Hwang, " Automatic generation of Korean pronunciation variants by multistage applications of phonological rules," ICSLP ' 98, Sydney, Austrailia, Dec. 1998.

[3] P. Clarkson and R. Rosenfeld, " Statistical language modeling using the CMU-Cambridge toolkit," EUROSPEECH ' 97, pp. 2707-2710, 1997.

[4] S. M. Kalt, " Estimation of probabilities from sparse data for the language model component of a speech recognizer," IEEE Trans. ASSP, Vol. 35, pp. 400-401, 1987.

[5] Chen,S.,Beeferman, D., and Rosenfeld, T." Evaluation Metrics for Language Models" In Proceedings of the DARPA Broadcast Transcription and Understanding Workshop, 1998.

[6] Clarkson, P. and Rosenfeld, T." Towards Improved Language Model Evaluation Measures" , In Proceedings of EUROSPEECH' 99, Sep. 5-9, 1999 Budapest, Hungary.