VOCABULARY OPTIMIZATION BASED ON PERPLEXITY

Kyuwoong Hwang

Spoken Language Section, Electronics and Telecommunications Research Institute
161 Kusung-Dong Yusong-Gu, Taejon, Korea
Interactive systems laboratories, Carnegie Mellon Univ., Pittsburgh USA
hkw@cs.cmu.edu

ABSTRACT

In this paper, we suggest a method to optimize the vocabulary for a given task using the perplexity criterion. The optimization allows us to reduce the size of the vocabulary at the same perplexity of the original word based vocabulary or to reduce perplexity at the same vocabulary size. This new approach is an alternative to phoneme n-gram language model in the speech recognition search stage. We show the convergence of our approach on the Korean training corpus. This method may provide an optimized speech recognizer for a given task. We used phonemes, syllables, morphemes as the basic units for the optimization and reduced the size of the vocabulary to the half of the original word vocabulary size for the morpheme case.

1. INTRODUCTION

In general, a word is the basic unit in language modeling and in the search stage of speech recognition. But, in terms of automatic speech recognition performance this may not be the most efficient unit. In this paper, we will redesign the vocabulary of a given task using the perplexity criterion. Recently, there have been some efforts to use other units than word for language modeling and speech recognition. One approach is to separate a word into its more basic units(morphemes)[1, 2]. In languages such as Korean, German and Japanese, e.g. a verb, may be conjugated depending on its syntactic role. Another approach is to concatenate a sequence of words which occurs frequently into an ensemble of words[5].

In this paper, we suggest a statistical method to optimize the vocabulary based on perplexity. In a diagram of the perplexity as a function of the vocabulary size, we can choose an operation point by either reducing the perplexity of the given task or reducing the size of the vocabulary based on a perplexity threshold. The perplexity of a task can be reduced by choosing a smaller perplexity operation point and the size of a vocabulary can be reduced in the search stage by choosing a larger perplexity operation point. This will increase the recognition rate of a speech recognition system and in case of smaller vocabulary size reduce the computational load due to the reduced search tree. This method may be an alternative to using phoneme n-gram in an usual search stage.

2. ALGORITHM

There are many perplexity measures proposed[8, 6]. Among them, we are using a phoneme based perplexity since the word perplexity can be only applied to the same vocabulary. If the vocabulary itself is changed, the word perplexity loses its meaning. Phoneme perplexity is a more appropriate measure of task difficulty.

$$PP = (word\,perplexity)^{\frac{1}{average\,length\,of\,words}}$$

If we constraint more on the possibility of a phoneme sequence, it will reduce the phoneme perplexity. Based on this intuition, we first reduce the constraints on phoneme sequences by splitting a word into its building blocks and then increase constraints by concatenating the blocks based on their frequency of occurrence as in Figure 1. In order to reduce the phoneme perplexity, we ignore the concept of words and convert the transcription of a speech database into its basic unit sequence. Then, by concatenating frequently occurred unit sequence into a word-like unit, we make a vocabulary which is optimized on a specific task.

The summary of our algorithm is as follows.

1. Build a unit sequence of a transcription by converting the words according to their sub-word unit sequences. This sequence will be its pronunciation when a phone is used as a basic unit, syllables when a syllable is used as a basic unit and morpheme analysis result when a morpheme is used as a basic unit.
2. Find a pair which has the highest frequency. When more than one pair have the highest frequency, choose the longest one in terms of its basic unit.
3. If the maximum frequency is less than a given threshold, stop. Otherwise continue.
4. Concatenate the resulting pair into a word-like unit in the transcription.
5. Go to 2.

After building a vocabulary from the training corpus, we concatenate the phoneme sequence of the test corpus into word-like units using the vocabulary generated from the training corpus. This is performed by concatenating longer units first.

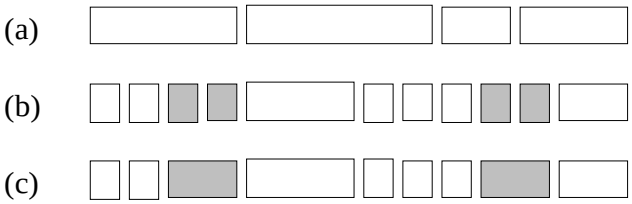

Figure 1. Merging basic units based on its frequency. (a) Transcriptions in words. (b) Transcriptions in basic units. (c) The most frequent sequence is merged and becomes a new unit.

3. EXPERIMENTAL RESULTS

The Korean Spontaneous Scheduling Database(KSST)[4] was used for this experiment. The database is the recording of spontaneous dialogs between two people scheduling a meeting. 300 dialogs were used for the training and the remaining 38 dialogs were used for testing. The training corpus contains 3190 sentences and 8680 unique words. This relatively large vocabulary is due to the Korean dialog structure which covers a broad range of topics even in a simple scheduling dialog and is due to heavy verb conjugating and agglutinative characteristics of the Korean language. Therefore, these facts leave space for optimization.

	training	test
# dialog	300	38
# utterance	3190	477
# word	35595	6487
# syllable	102641	18324

Table 1. Text corpus

To measure perplexity, we use the CMU Statistical Language Modeling Toolkit[7] and to perform recognition experiments, the JANUS speech recognition system[3]. In Experiment I, we use a phoneme as a basic unit and the perplexity is decreased for training data but does not decrease as much for test data. The main reasons are due to the fact small data size as well as the fact that converting a word into phonemes in Korean may result into very small units without linguistic information that would exist in the form of words. Therefore, in the next experiments II and III we use bigger segments of a word such as syllable and morpheme as basic units to preserve some linguistic knowledge which may reside in a word form.

3.1. Using phonemes as basic units

In Figure 2, we show the perplexity versus vocabulary size for the KSST task. The perplexity for the training corpus converges with the increasing size of the vocabulary. For the training data we achieved a similar perplexity at half the size of the vocabulary. Most of the resulting word-like units in the new vocabulary are morphemes and regular words. However, for the KSST database the test set perplexity does not converge as well as training set due to the small size of the training set.

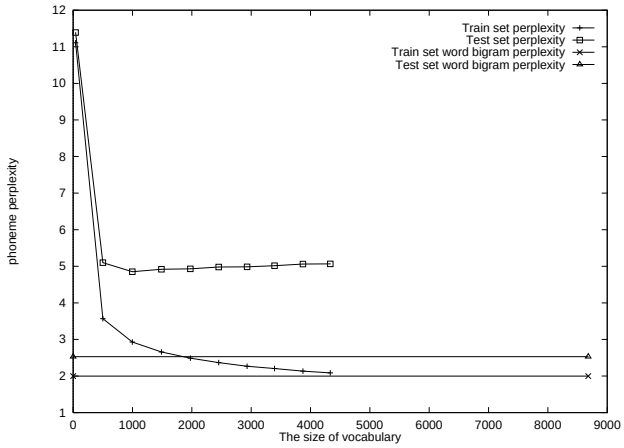

Figure 2. Perplexities (The size of vocabulary for word bigram perplexity is 8680.)

3.2. Using syllables and morphemes as basic units

In experiment II, we use Korean syllables as basic units. First, by converting words into their syllable sequences we make a transcription represented in syllables. Then, we merge the sequence of syllables which occurs frequently in training corpus into a new pseudo word.
In experiment III, we used Korean morphemes as basic units. First, by converting words into their morpheme sequences we make a transcription represented in morphemes. Then, we merge the sequence of morphemes which occurs frequently in the training corpus into a new pseudo word.

In Figure 3, we show the bigram and trigram phoneme perplexities of syllable-based optimization and morpheme-based optimization with those of the normal word case. Perplexities converge in all cases. The reason that the graph ends near 6000 is that we can't find any sequence which has higher frequency than the threshold at that point due to the small size of corpus.

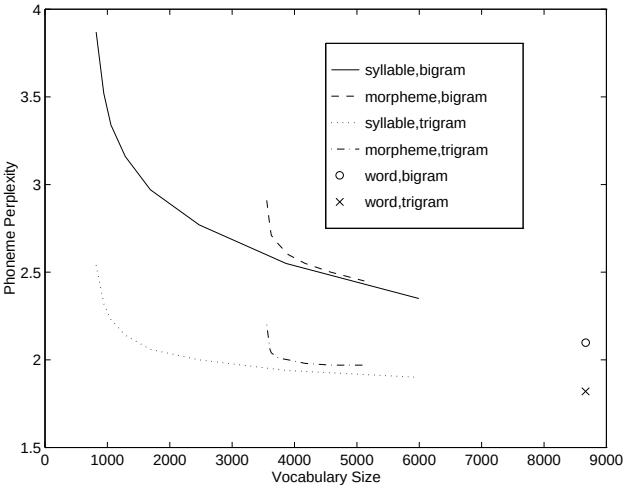

Figure 3. Training set perplexities

In Figure 4, the perplexities tend to increase as the vocabulary size increases even though in Figure 7 the performance increases as the vocabulary size increases before over-optimization of the vocabulary. This perplexity increasement can be viewed as the result of over-training of the vocabulary and the indirect relation between the perplexity and the performance of a recognition system.

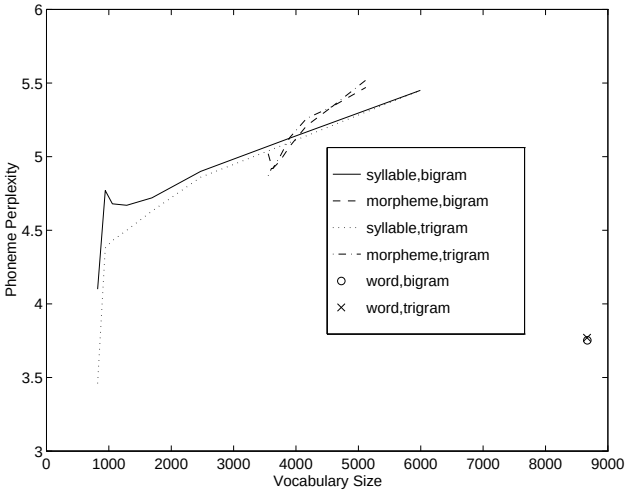

Figure 4. Test set perplexities

In Figure 5, as the merging of basic units in the training transcription progresses, the training corpus vocabulary covers less test corpus. But, in experiment II and III these coverages are much higher than that of a normal word vocabulary.

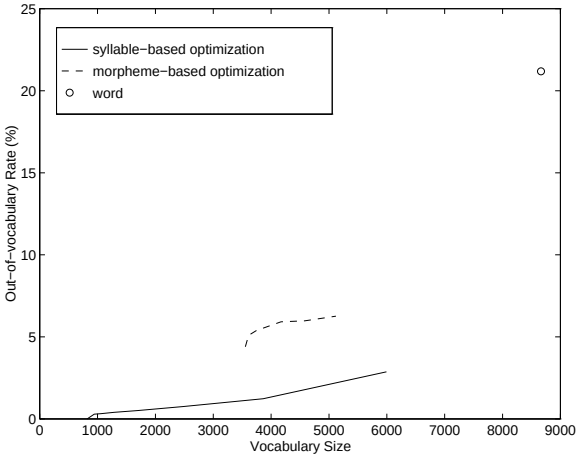

Figure 5. Out of vocabulary rates

In Figure 6, the average length of the word-like units are increased as we continue merging basic unit sequences based on their frequencies. The point where the length starts to saturate is the point where the perplexity start to saturate in Figure 3 and the performance curve shows the best result in Figure 7.

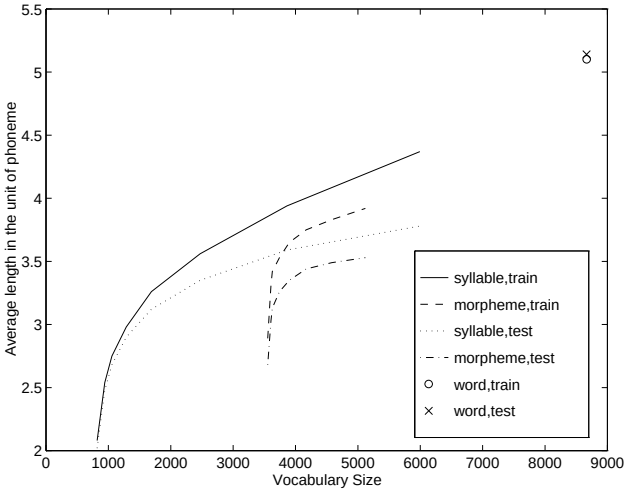

Figure 6. Average length of units in terms of phone

Recognition experiments were done for the new optimized vocabularies and language models derived from the corpuses represented by that optimized word-like units. The recognition system is trained by the transcriptions represented by words. This may give preference to the normal word case in recognition experiments. To see the tendency of the proposed optimization algorithm, 24 utterances are used for test experiments.

As you can see from Figure 7, the performance was best when we used the conventional words as the units in the search vocabulary. For the selection of basic units, morpheme units are better than syllable units. The best performance among morpheme derived vocabularies is 77.4% which is similar to the performance of the word vocabulary with only almost half of the size of the word vocabulary.

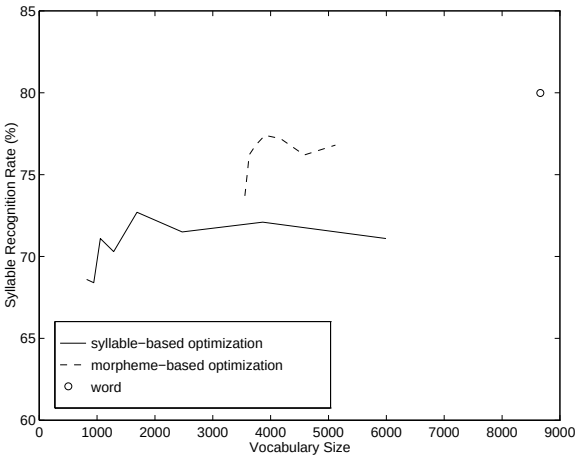

Figure 7. Syllable recognition rates

Because the recognition results of a syllable based system and a morpheme based system are syllables and morphemes, we use syllable recognition rate as the measure

of the performance of each system to make a comparison among systems possible. The syllable recognition rate is calculated by a normal recognition result aligning procedure after converting the reference transcriptions and the recognition results into syllable sequences.

4. CONCLUSION

In this paper, we proposed a method to redesign the vocabulary of a task to be optimized for speech recognition. In general, a word is accepted as a basic unit of a speech recognition system without suspicion. However, we show that a word may not be the best solution for speech recognition and a better word-like unit can be obtained by optimizing the vocabulary using the perplexity criterion. Even though the test corpus perplexity did not show as good result as training corpus perplexity due to the small size of training corpus, the training corpus perplexity showed that this approach may be an alternative to optimize the vocabulary for a given task and in recognition experiments, it showed similar performance with half size vocabulary than word case. This approach will show good results in tasks where the training data covers the domain well such as medium sized human-computer interaction task with pre-defined grammars.

5. ACKNOWLEDGEMENT

The author would like to thank members of the Interactive Systems Laboratories for many fruitful discussions. The author acknowledges the Interactive Systems Laboratories for the provision of JANUS speech recognition system and the Spoken Language Section of the Electronics and Telecommunications Research Institute for the provision of Korean spoken language database.

REFERENCES

[1] P. Geunter. Using morphology towards better large-vocabulary speech recognition systems. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, 1995.

[2] A. Ito and M. Kohda. Language modeling by string pattern N-gram for Japanese speech recognition. In *Proc. of Int. Conf. of Spoken Language Processing*, pages 169–172, Philadelphia, PA, USA, 1996.

[3] A. Lavie, A. Waibel, L. Levin, D. Gates, M. Gavalda, T. Zeppenfeld, P. Zhan, and O. Glickman. Translation of conversational speech with JANUS-II. In *Proc. of Int. Conf. of Spoken Language Processing*, Philadelphia, PA, USA, October 1996.

[4] H. S. Lee, J. Park, and H. R. Kim. An implementation of Korean spontaneous speech recognition system. In *Proc. of Int. Conf. on Signal Processing Applications and Technology*, pages 1801–1805, Boston, MA, USA, October 1996.

[5] H. Masataki and Y. Sagisaka. Variable-order N-gram generation by word-class splitting and consecutive word grouping. In *Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 188–191, 1996.

[6] H. Ney, U. Essen, and R. Kneser. On structuring probabilistic dependences in stochastic language modelling. *Computer Speech and Language*, 8:1–38, 1994.

[7] R. Rosenfeld. The CMU statistical language modeling toolkit and its use in the 1994 ARPA CSR evaluation. In *Proc. of the Spoken Language Technology Workshop*, 1995.

[8] J. Ueberla. Analysing a simple language model - some general conclusions for language models for speech recognition. *Computer Speech and Language*, 8:153–176, 1994.