Style-Specific Language Model Adaptation using TF*IDF Similarity for Korean Conversational Speech Recognition

Young-Hee Park*, Minhwa Chung*
* Department of Computer Science, Sogang University.

(Received March 11 2004; Revised July 16 2004; accepted July 29 2004)

Abstract

In this paper, we propose a style-specific language model adaptation scheme using n-gram based tf*idf similarity for Korean spontaneous speech recognition. Korean spontaneous speech shows especially different style-specific characteristics such as filled pauses, word omission, and contraction, which are related to function words and depend on preceding or following words. To reflect these style-specific characteristics and overcome insufficient data for training language model, we estimate in-domain dependent n-gram model by relevance weighting of out-of-domain text data according to their n-gram based tf*idf similarity, in which in-domain language model include disfluency model. Recognition results show that n-gram based tf*idf similarity weighting effectively reflects style difference.

Keywords: *Korean conversational speech recognition, Language model adaptation, Disfluencies, Filled pauses.*

I. Introduction

For conversational speech recognition, language model adaptation with large out-of-domain data is useful, since obtaining sufficient language model training data is often difficult in a conversational domain. However conversational speech has different style and content in comparison with the written text corpora[1]. To improve n-gram language models by reflecting style and content specific characteristics, Iyer[1] used relevance weighting using only unigram statistics for estimating similarity, which is good for content but poorly supports the style specific characteristics.

We focus on reflecting style specific characteristics, because Korean conversational speech shows a lot of style specific characteristics as well as disfluencies. Since these phenomena are related to function words and depend on preceding or following words, word sequences are more important than words themselves.

To deal with style specific characteristics, we estimate in-domain dependent n-gram model by relevance weighting of out-of-domain text data according to style and content similarity, where style is represented by n-gram based tf*idf similarity. In addition, we have tested the ability of disfluencies to predict the neighboring words.

II. Korean Conversational Speech Corpus

2.1. Corpus Description

We collected conversational speech corpus by simulating conversations between travel agents and their customers making travel plans. Each conversation consists of three or four detail topics such as hotel reservation and inquiries on how to reach the hotel, so that is complex enough.

The corpus has been transcribed at a spacing unit level. Disfluencies such as filled pauses and word fragment are annotated. Repeat and repair are also annotated. From the analysis of the corpus, the rate of disfluencies is 11.2%,

Corresponding author: Young-Hee Park (Younghp@sogang.ac.kr)
Department of Computer Science, Sogang University Sinsu-Dong, Mapo-Ku 121-742, Korea

Table 1. Corpus summarization and distribution of disfluencies.

No. of conversations	100
No. of utterances	6,006
No. of words	103,406
No. of unique words	2,292
Disfluencies	11.2 %
Filled pauses	9.9 %
Top 3 filled pauses	ye (33.1%) uh (29.7%) ah (10.3%)

which supports that our corpus is useful for research on spontaneous speech. For our language model experiments, since most of disfluencies are filled pauses, we have restricted disfluencies to filled pauses. Top five filled pauses cover 80.9% of filled pauses. Some examples are shown in Table 1[2].

Then we have performed morphological analysis because morphemes are generally considered as basic recognition units for Korean[2]. As a result, we have obtained about 103,406 morphemes (after this, we use word instead of morpheme) corpus. Table 1 shows the size of our corpus. We have segmented each conversation using a turn corresponding to an utterance. Since an utterance consists of several sentences, filled pauses help predict the beginning word of sentences.

2.2. Style Specific Characteristics

In addition to disfluencies, Korean conversational speech has several different characteristics of style in comparison with the written text corpora.

- Final auxiliary particle "*yo*" (a respectful word) appears at the end of verb phrase frequently and is observed 4.4 % in conversational speech against 0.2 % in broadcast news corpus.
  I.e., "*chulbalhagoyo*" (depart).
- Besides, auxiliary particle "*eun/neun*" often follow conjunctive ending "*myeon.*"
  I.e., "*geureomyeoneun*" (then).
- Predicative case particle "*i*" is frequently omitted. Result from force alignment of our corpus shows that 22% of them are omitted.
  I.e., "*yeohaengsaimnida*" (This is travel agency) changes to "*yeohaengsa*(i)*mnida*".
- Conjunctive particle "*hago*" (and) is more often used than "*wa/gwa.*"
  I.e., "*ireumhago beonho*" (name and number).
- Contractions of ending words or case particles; for example, the objective case particle "*eul/reul*" changes to "*l*".
  I.e., "*pyoreul*" (ticket) changes to "*pyol.*"

All of these observed style specific characteristics are function words and their characteristics are dependent upon preceding and following words. However, a function word is not important in information retrieval scheme. In these style specific characteristics, it is not a function word itself, but a word sequence including function words that is important. These characteristics appear rarely in broadcast news corpus and show that simple POS grammar cannot adapt style specific domain language model well.

III. Style-Specific Language Model Adaptation

We aim at style adaptation of in-domain language model by combining in-domain n-gram estimates with the weighted out-of-domain n-gram estimates. The weight comes from n-gram based tf*idf similarity. In-domain language model reflects the prediction ability of disfluencies.

We use 16M words broadcast news articles as large out-of-domain corpus. Interview articles of broadcast news corpus include a few disfluencies which, however, are not labeled. Since our corpus consists of single topic and has style specific words, lexicon includes all the vocabularies from our conversational speech corpus except fragment words.

3.1. N-gram based Document Selection

Similarity between domains can be captured using an information retrieval framework[1,3]. This approach is based on weight matrix processing, which needs to create a keyword matrix that shows the relative importance of the keywords. Since this kind of document classification is mainly based on distribution of keyword frequency (content), it is weak in applying word associations and syntactic information to the similarity measuring. In our approach for measuring similarity between in-domain and all of the out-of-domain documents, we propose a new measure using word n-gram based tf*idf similarity in order to reflect content and style to a language model.

Tf*idf measure is based upon vector space model of document representation. Proposed measure uses bigram words sequences (w_{n-1}, w_n) as keywords for the vector representation, instead of word itself, in order to apply word associations. Keywords are selected by weighting all bigram word sequences with their inverse document frequency defined as follows;

$$idf(t_k) = \log\left(\frac{N}{n_{t_k}}\right) \tag{1}$$

where term t_k represents the word sequences (w_{n-1}, w_n), n_{t_k} represents the number of documents, and N represents the total number of documents. Words that occur in all documents are given zero weight. In our work, keyword candidates are restricted to bigram word sequences of our in-domain vocabulary except disfluencies. In Korean conversational speech, this selection is significant, since style specific characteristics of Korean conversational speech is closely connected to the word sequences involving function words.

Each document d_i is indexed using a vector of terms t_k weighted by

$$wgt(t_k, d_i) = tf_{ik} \cdot idf(t_k), \quad k = 1, \ldots, L \tag{2}$$

where term frequency tf_{ik} is the unigram count of term t_k in document d_i.

When the entire in-domain corpus is document I, the cosine similarity between the two documents is defined as follows:

$$Sim(d_i, I) = \frac{\sum_{j=1}^{L} wgt(t_j, d_i) wgt(t_j, I)}{\sqrt{\left(\sum_{l=1}^{L} wgt(t_l, d_i)^2\right)\left(\sum_{l=1}^{L} wgt(t_l, I)^2\right)}} \tag{3}$$

All the documents in the broadcast news corpus are ranked by the decreasing similarity coefficient $Sim(d_i, I)$, which lies between 0 and 1, and indicates greater similarity with values closer to 1. The similarity weight for a particular document is then $v(d_i) = Sim(d_i, I)$.

In order to make the out-of-domain language model adapt in-domain style, we combine n-gram counts of out-of-domain documents with an adaptation weight $v(d_i)$. When h is word history for word w and $C_O^i(h,w)$ is the n-gram count in the ith document of the out-of-domain, the combined count $C_O(h,w)$ is as follows:

$$C_O(h,w) = \sum_i v(d_i) \times C_O^i(h,w) \tag{4}$$

3.2. Disfluency Model

The optimal handling of disfluencies in the language model is not defined, yet. However, for filled pauses only, many recognition systems have modeled filled pauses in the language models such as unigram model or treating them as normal words 000. In order to confirm the effect of disfluency model, we assume that they have ability to predict the neighboring words before and after them. We treat filled pauses as normal words in the context for language modeling.

In our experiments, since filled pauses are observed in in-domain data only, we model them in in-domain language model. In order to test the predicting ability of filled pauses, we model them with three ways as follows.

In-Dis1: disfluencies are predicted by unigram probabilities. w_d and w_n are disfluency and normal word, respectively. h_s denotes the word history obtained by skipping disfluencies in training text.

$$\begin{cases} C_{I_{new}}(h, w_d) = C_I(w_d) \\ C_{I_{new}}(h, w_n) = C_I(h_s, w_n) \end{cases} \tag{5}$$

In-Dis2: disfluencies are predicted by normal words; however, they are not used for predicting normal words.

$$C_{I_{new}}(h,w) = C_I(h_s, w) \qquad w = w_d \text{ or } w_n \tag{6}$$

In-Dis3: disfluencies are predicted by normal words and they are also used for predicting normal words. $h_{n,d}$ is the word history containing disfluencies.

$$C_{I_{new}}(h,w) = C_I(h_{s,d}, w) \qquad w = w_d \text{ or } w_n \tag{7}$$

Table 2. Perplexity of the adapted out-of-domain LMs (uni-TFIDF: conventional unigram based tf*idf weighting, bi-TFIDF: proposed bigram based tf*idf weighting).

Weight	Perplexity
No weight	203.6
uni-TFIDF	191.2
bi-TFIDF	186.7

3.3. Language Model Adaptation

We combine in-domain estimates with out-of-domain estimates with two ways as follow.

First, we have directly combined n-gram counts of in-domain, $C_{I_{new}}(h,w)$ with n-gram counts of out-of-domain, $C_O(h,w)$, in which they are combined without weight.

Second, since the size of out-of-domain is bigger than those of in-domain, we have linearly interpolated between in-domain language model and out-of-domain language model to reduce the influence of the size of out-of-domain data.

IV. Experiments

The recognition system used in this experiment is based on the 1-pass semi-dynamic trigram network decoder, which was originally developed for Korean read speech dictation task 0. For training of acoustic model, 84 conversation speech and about 20 hours of read speech are used. Acoustic model is a set of continuous HMMs and each state has 6 Gaussian mixtures. Previous experiment 0, in which filled pauses are modeled by separate HMMs, shows that using just one HMM corresponding to a filled pause with the longest duration is effective for improving the recognition performance. Therefore, one noise HMM and one filled pause HMM are used.

From the conversational speech corpus, 84 conversations are used for training, 8 conversations are used for estimating interpolation coefficients, and the remaining 8 conversations are used for test. We have used the same test set for language modeling and speech recognition experiments. Lexicon includes all normal words and three filled pauses with the highest frequencies shown in Table 1.

Table 2 shows the perplexity of out-of-domain language models adapted by the style characteristics of in-domain. Our proposed LM with bi-TFIDF weighting shows the best result and reduces the perplexity by 8.3% over no-weighting LM and by 2.4% over uni-TFIDF weighting[1]. This means that the relevance weighting scheme using an information retrieval framework works well in language model adaptation, and the proposed style adaptation language model reflects effectively the style characteristics of Korean conversational speech.

Table 3. Perplexities and recognition results of disfluency models of in-domain LMs.

Model	Perplexity	WER (%)
In-Dis1	39.9	28.9
In-Dis2	31.9	27.9
In-Dis3	27.8	27.2

Table 3 shows the recognition and perplexity results with three disfluency trigram language models of in-domain. While in the model In-Dis1 disfluencies are predicted by unigram, In-Dis2 and In-Dis3 use trigram for predicting disfluencies. Comparing the results of In-Dis1 and In-Dis2, we see that disfluencies are context dependent, which reduce the perplexity by 8% and the WER by 1%. We also observe that disfluencies in the history (In-Dis3) further reduce the perplexity by 4.1% and the WER by 0.7% due to their predicting ability. In our experiments, we have performed the recognition at the utterance level in one pass and an utterance consists of several sentences. Therefore we believe that the predicting ability of filled pauses between sentences makes our disfluency model more effective.

A comparison of the perplexities in Table 2 with those of Table 3 shows that the in-domain data are quite different from the out-of-domain data.

Table 4 shows the effect of combining In-Dis3 model with out-of-domain models according to the combining method. The performance of weighting schemes is similar to the results of Table 2. Note that adding out-of-domain data improves performance on both perplexity and WER over the in-domain data only (In-Dis3). No-weight addition of data degrades performance over all other models due to the difference between domains. From the linear interpolation, we could reduce the effect of the size of out-of-domain data and reduced over 1% WER.

Table 4. Perplexities and recognition results of various adopted language models according to the way combining In-Dis3 with out-of-domain.

	Count combine		Interpolation	
	PP	WER	PP	WER
In-Dis3	27.8	27.2	27.8	27.2
+No weight	52.0	29.2	27.1	26.3
+uni-TFIDF	31.9	26.8	25.8	25.8
+bi-TFIDF	29.9	26.1	24.6	24.9

Finally, by using the proposed style-specific adapted language model, we have obtained 24.9% WER as the best result.

V. Conclusions

In this paper, we have analyzed the style characteristics of Korean conversational speech such as the frequent usage of a specific auxiliary particle "*yo*", omission of the predicative case particle "*i*", and contractions in comparison with the written text data. Then we have described our style-specific language model adaptation for Korean conversational speech recognition. For style-specific language model adaptation, we have proposed a relevance weighting of out-of-domain text data with n-gram based tf*idf similarity.

The n-gram based IR weighting approach captures both content and style similarity, while the IR weighting approaches capture content similarity only. We have obtained absolute reductions of WER 1.7% by disfluency model (In-Dis3) and 2.3% by the n-gram based tf*idf similarity weighting, respectively. In sum, our model reduces 4.0% WER absolutely.

For further researches, we need to investigate more detailed annotations about disfluencies, discourse markers, and other style specific features. We plan to use the data from web, since more out-of-domain data improve the performance of language model for conversational speech recognition.

Acknowledgements

This work has been supported by Ministry of Science & Technologies, Brain Neuro Informatics Research Program (Project No. M1-0107-01-0003). We would like thank for ETRI for allowing us to use the conversational speech database.

References

1. R. Iyer and M. Ostendorf, "Relevance weighting for combining multi-domain data for N-gram language modeling," Computer Speech and Language, Vol. 13, pp. 267-282, 1999.
2. Y.-H. Park and M. Chung, "Analysis of Korean spontaneous speech characteristics for spoken dialogue recognition," Journal of The Acoustical Society of Korea, vol 21, no 3, pp.330-338, 2002.
3. M. Mahajan, D. Beeferman and X. D. Huang, "Improved topic-dependent language modeling using information retrieval techniques," Proc. ICASSP, vol. 1, pp.541-544, 1999.
4. A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," Proc. ICASSP, vol. 1, pp.405-408, 1996.
5. M. Siu and M. Ostendorf, "Modeling disfluencies in conversational speech," Proc. ICSLP, vol. 1, pp.621-625, 1996.
6. D.-H. Ahn and M. Chung, "Compact subnetwork-based large vocabulary continuous speech recognition," Proc. ICSLP, vol. 1, pp.725-728, 2002.

[Profile]

- **Young-Hee Park**

The Journal of Acoustic Society of Korea, Vol.21, No.3E, 2002.

- **Minhwa Chung**

The Journal of Acoustic Society of Korea, Vol.21, No.4E, 2002.