Web データから抽出したスタイル依存テキストを用いた中国語自由発話言語モデルの構築

胡新輝[†]　松田繁樹[†]　柏岡秀紀[†]

我々は、中国語自由発話音声認識用言語モデルの学習データを収集するために、TFIDF を用いた文クラスタリング方法と Perplexity に基いたスコアリング方法を統合して Web からテキスト文を選択する方法を提案する。Sogou 中国語インターネットコーパスに対してフィルタリング処理と形態素解析処理をした後、すべての文をスタイルによってクラスタリングし、あらかじめ準備した自由発話シード文に近いクラスタを決定する。それから、これらのクラスタに含まれるテキスト文毎に、更にシードモデルに対する Perplexity を評価し文選択を行う。
この提案法を用いて得られた４百万のテキスト文を用いて言語モデルの推定を行った実験として、オープンドメインの自由発話テストセットに対して、これらのテキストで学習された言語モデルを既存の旅行会話用言語モデルに線形結合したモデルを評価した。評価結果から、ベースモデル（旅行会話用）より文字エラー率が 6.2 ポイント,ランタイムの選択方法より４．９ポイント減少した。また、Perplexity ベースのみの方法に比べて 1.90 ポイントの改善が確認された。

Data Collection From Web Resources For Constructing Chinese Spontaneous Language Model Using Sentence Style Clustering

Xinhui HU[†]　Shigeki MATSUDA[†] and Hideki KASHIOKA[†]

In this paper, we present our work on data collection by combining sentence clustering and perplexity-based scoring approach for constructing a Chinese spontaneous language model of a speech recognition system. We used the Sogou Chinese Internet corpus [1], a Chinese web archive, as the data resource. After filtering processing and word segmentation, the web texts were clustered based on their writing styles, and optimal clusters were chosen by referring to a set of spontaneous seed sentences. Then, each sentence of the optimal clusters was further evaluated by a perplexity-based scoring approach to decide if it would be selected.

With the proposed method, we selected over 4 M sentences. Using the language model interpolated with the one trained by these selected sentences and the baseline (seed) model, speech recognition evaluations were conducted on an open domain spontaneous test set. We reduced the character error rate an average of 6.2 points over the baseline model, and 4.9 points over the model constructed from randomly sampled web sentences. We verified that the proposed method is superior to only using the perplexity-based scoring approach, with 1.90 points of reduction in the character error rate. It is shown to be efficient for selecting spontaneous sentences from web data.

1. Introduction

A language model (LM), which is an important component of an automatic speech recognition (ASR) system, tries to capture the properties of a language and to predict the next word in a speech sequence. It is generally constructed using a textual corpus. Its performance depends heavily on the size and quality of the corpus. Here, quality refers to the matching extent between the corpus content and the recognition task, and to the similarity between the style of the texts and the speech indicating whether it is read or spontaneous.

To build an LM of a spontaneous open domain, the textual style of the training corpus should contain more “colloquial” than “read” and “written” expressions and should cover as many topics as possible.

However, the manual construction cost of such a spontaneous text corpus is very high with low efficiency. Thus, automatic approaches to collecting such data are required. With the rapid development of the Internet, the web is becoming a great data treasure trove and is receiving more and more attention from researchers. Web data are usually retrieved by keywords for download using a web search engine [2, 3, 4]. Misu et al. used word perplexity as the similarity criterion, chose queries from seed utterances, and retrieved effectively relevant utterances of these queries for a speech dialogue system [5]. These methods are valid for selection of topic adaptation sentences, but have difficulties improving the selection of spontaneous sentences because selecting keywords that characterize spontaneous speech is complicated. Gathering keywords is also difficult for such various topics in the case of open domains.

The style of spontaneous speech is different from read speech. For example, spontaneous speech has many fillers and pauses. Two methods are mainly used to improve a spontaneous LM. One is transformation from a written format LM to a spoken format. Hori et al.

[†] 情報通信研究機構
National Institute of Information and Communications Technology

composed a weighted finite-state transducer (WFST) that translates sentence styles to integrate LMs of different styles of speaking or dialect and different vocabularies [6]. Akita et al. significantly reduced the perplexity and the word error rate (WER) by transforming a document-style model into a spoken style based on a statistical machine translation framework [7]. Another method generates disfluencies that simulate spontaneous sentences by predicting fillers and short pauses in document sentences [8]. Masumura et al. used a naive Bayes classifier to select speech-like texts from downloaded web data and then used the same method as [8] to convert the texts into a spontaneous format. An LM trained by the generated data gave as high performance as the large-scale spontaneous speech corpus [9].

Compared with a written corpus, the Chinese spontaneous corpora are considerably insufficient. Our work aims to improve the selection efficiency of Chinese spontaneous sentences while covering wide domains.

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in others. It is a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, information retrieval, and bioinformatics. Spontaneous texts have their own particular characteristics of word usage, for example, the distribution of parts-of-speech (POS) and word order in sentences are different from read speech and written texts. Therefore, we hope to utilize the clustering approach to group texts from the viewpoint of text style, so that the selection of sentences having spontaneous style can be improved.

In this study, we adopt a combination of a clustering-based approach and perplexity-based approach for data selection. This paper is organized as follows. Section 2 briefly describes the entire system configuration and the adopted word segmentation method. Section 3 introduces the concept of style clustering and presents the sentence selection method. Section 4 reports the experiment results, and Section 5 concludes the paper.

2. Web Data Preparation

2.1 System configuration

Figure 1 shows the system configuration of our study. The original web data are filtered to remove the HTML tags, the Java script codes, etc. and to normalize them. Then, these data are segmented into word texts by a word segmentation system. After that, the segmented sentences are clustered, and optimized clusters are chosen from them. In this stage, a seed sentence set, which is extracted from the speech transcript of several field experiments, is used. These seed sentences are assumed to be characterized by the style of conversational and spontaneous speech. The phenomenon appearing in spontaneous speech are reflected in the transcripts, such as filler and repeated words. After clustering, the perplexity of each sentence in the optimized clusters is computed, and based on this, the decision is made whether to select the sentence. Here, a seed LM trained by a corpus of the travel domain and a set of spontaneous speech transcripts is used for computing the perplexity.

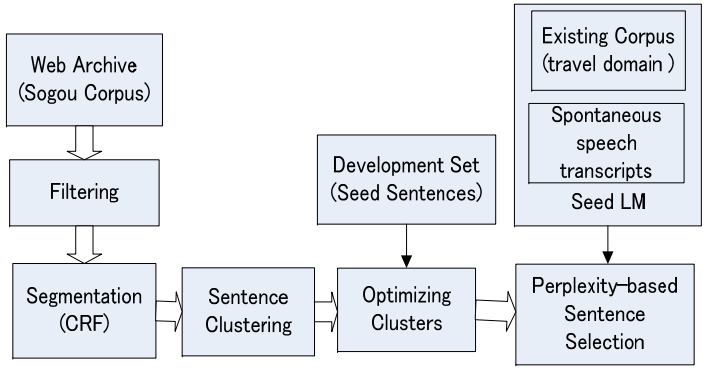

Figure 1. System configuration

2.2 Chinese word segmentation

Word segmentation is generally indispensable for most Chinese language processing schemes since there are no natural word delimiters in Chinese, such as spaces in English.

2.2.1 *CRF-based segmentation*

Recently, the character-based tagging method, such as maximum entropy [10] and the conditional random field (CRF) [11] model, has become the dominant technique for Chinese word segmentation due to its global optimization and its ability to detect new words. In this study, we rebuilt an existing CRF-based system [12] for word segmentation using new training data.

2.2.2 *Training data for word segmenter*

The SINICA balanced corpus (659.1 K sentences) [13] and two files of the LDC2007T03 Tagged Chinese Gigaword corpus [14] (cna_cmn_200401, 33.9 K sentences, xin_cmn_200401, 36.4 K sentences) were used as training data for the CRF segmenter. These

tagged data are based on identical specifications and cover a wide range of modern Chinese fields. Therefore, these data are helpful for segmenting web texts. The segmenter is used to segment the filtered sentences. 3.09 billion words containing a 2.5 M vocabulary (unique words) were estimated in the final segmented data.

3. Data Selections

3.1 Open domain and spontaneous speech styles

Content words, especially nouns, play the most important roles in determining the topics of a document. Masumura et al. retrieved web pages using *all* known nouns as search queries to acquire a comprehensive range of text topics [9]. In contrast, we hypothesize that if a query keyword does not contain any nouns, then the retrieved results will have no predefined topics. The texts (or transcripts) of spontaneous speech are greatly different from read and written texts. They are particularly characterized by such fillers 呃, 啊 (um, eh), 这, 这个 (it, this), 那个 (that), 那么 (then), filled pauses, repetitions, and ellipses. From our investigations on an existing textual corpus, which is for constructing LMs of conversational ASR in the travel domain, we found many words whose combinations appear more frequently in colloquial conversations than in written texts. Examples include 我 想 (I want), 我 要 (I'm going), 给 我 (give me), 请问 这个 (please tell me about this), and 能不能 (could you). All of these phenomena show that spontaneous styles are mainly characterized by non-nouns.

3.2 Sentence style clustering

Motivated by the fact that topic clustering is mainly based on noun distribution in documents, we propose style clustering based on the distributions of parts-of-speech (POS) excepting noun, which is concretely achieved by removing nouns from the clustering vocabulary.

3.2.1 *Clustering algorithm*

We use the CLUTO toolkit [15] to perform the clustering. It finds a predefined number of clusters based on a specific criterion. We chose the following function to maximize the within-class similarity:

$$(S_1 S_2 \ldots S_K)^* = \arg\max \sum_{i=1}^{K} \sqrt{\sum_{v,u \in S_i} sim\,(v,u)}\,, \qquad (1)$$

where K is the desired number of clusters, S_i is the set of documents belonging to the i^{th} cluster, $\boldsymbol{v}$ and $\boldsymbol{u}$ represent two documents, and $sim(\boldsymbol{v},\boldsymbol{u})$ is their similarity. We use the cosine distance to measure the similarity between two documents:

$$sim\,(u,v) = \frac{\vec{v} \cdot \vec{u}}{\| \vec{v} \| \ \| \vec{u} \|}\,, \qquad (2)$$

where $\vec{v}$ and $\vec{u}$ are the feature vectors representing the two documents based on the style hypotheses. The elements in every feature vector are scaled based on their term and inverse document frequencies (TF and IDF). These terms are limited by the style clustering vocabulary words.

3.2.2 *Implementation*

The original Sogou corpus consisted of 128 individual compressed files, each of which contains millions of sentences. Based on implementation considerations, we divided the clustering into two steps.

1) We regarded each sentence as a document and clustered each file into 100 clusters. 12,800 clusters/documents were obtained.

2) We further clustered the above 12,800 documents into 100 clusters.

3.2.3 *Optimized style clusters*

After the above clustering, we built a 3-gram word LM with each cluster. Then, based on the perplexity of a development set (seed sentences) from an existing training corpus and field experiments, all the clusters were ranked by their perplexities.

Optimized clusters were obtained by accumulating clusters beginning from the minimum perplexity and observing the perplexity of the seed sentence set to the LM trained by the accumulated clusters. Figure 2 shows the perplexity changes with the accumulations of clusters. We found that at the point of +newC3, the perplexity is minimum. These accumulated clusters up to this point are regarded as the optimized clusters, where 14 M sentences are contained.

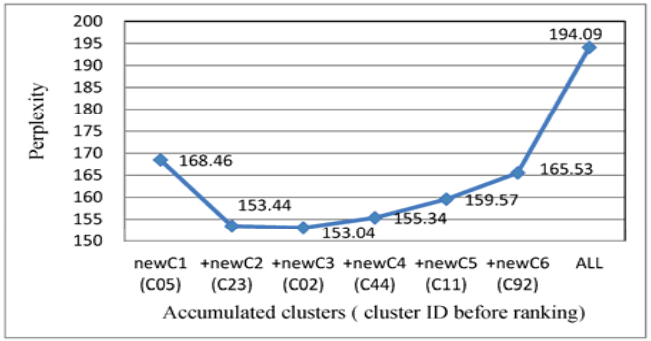

Figure. 2 Perplexities with accumulation of clusters

3.3 Perplexity-based scoring

While the previous clustering is focused on grouping a large quantity of texts by their POS distribution, the word content is not emphasized in the clustered results. However, sentence style is also captured frequently by word sequence. For example, as mentioned before, spontaneous utterances have many filled words, hesitations and repeated words. Because the sentence perplexity is a metric to evaluate the probability of a word sequence of a sentence, it is useful for capturing the characteristics of word sequences [17]. Here, we use the sentence perplexity as a measure for further selecting spontaneous like sentences from the clustered data. To achieve this, a seed language model is trained with an existing training corpus in the travel domain and several sets of speech transcripts collected from field experiments. The data for training the seed LM are shown in Table 1. VoiceTra is the public mobile phone S2ST service [16] mentioned above. With HS, they are real data with spontaneous styles.

The sentence selection is conducted empirically. For each sentence in the optimized clusters, the sentence perplexity is calculated. Sentences whose perplexity is smaller than a threshold are selected for the final data.

Table 1 Data used for seed model

Name	Size	Domain	Collected from:
VoiceTra (Seeds)	1.4 K	Open	Mobile phone　S2ST service
HS	28 K	Travel	Field experiments
BTEC	500 K	Travel	Texts, translation

3.4 Language Model with selected sentences

Linear interpolated LM is adopted for building the final LM. It is formulated as follows:

$$LM = \lambda \cdot LM_{base} + (1-\lambda) \cdot LM_{Selected} \qquad (3)$$

Here, LM_{base} is the baseline LM trained by the data in Table 1, and $LM_{Selected}$ is the LM trained by the selected sentences after the above perplexity-based selection.

λ is the weighting factor tuned by a development set from the VoiceTra data.

4. Experiments

4.1 Experimental settings

4.1.1 *Data set for evaluation and development*

To evaluate the quality of the selected sentences, we built an LM as shown in Eq.(3) using these data and used it for speech recognition. We selected 606 utterances (EVA01) from the VoiceTra for recognition and another 606 utterances (DEV01) for development.

4.1.2 *Perplexity threshold and selected data size*

Figure 3 shows the selected sentences in different perplexity thresholds and the perplexity of DEV01 relative to the LM trained by the corresponding selected sentences. The best threshold is found at 900, where the perplexity is minimum. At this point, 4.1 M sentences were selected.

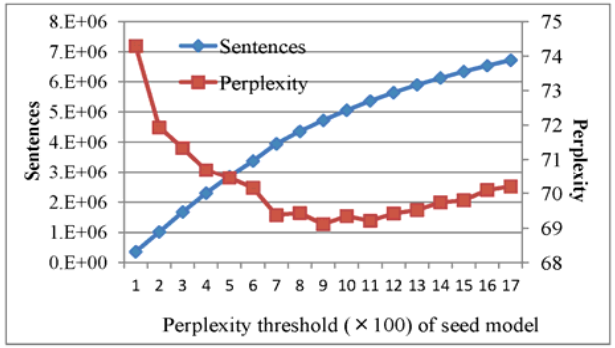

Figure.3 Perplexities of DEV01 and collected utterances in different perplexity thresholds

4.1.3 *Other selections for comparisons*

For comparisons with the proposed selection method (PROP), we conducted the following selections from the same web data.

(1) BALN: Baseline LM trained by the existing training corpus (500K) in the travel domain.
(2) RAND: Approximately the same scale of sentences (4.1 M) as in PROP were randomly selected from all the web data.
(3) OPTI: All sentences in the optimized clusters were selected, so no further perplexity-based selection was needed.
(4) PPLX: Selection was conducted only by sentence perplexity relative to the seed model without clustering before it.
(5) TOPI: The clustering is based on the topics; this means that nouns are used for clustering.

4.2 Evaluation results

Figure 4 shows the recognition results (character error rate: CER) of test set EVA01 using the LMs trained by different data selections with corresponding selected sentence counts and word vocabulary sizes. In the case of BALN (baseline LM), the CER is 40.65%.
From it, we can see that all purposeful selections more effectively decreased CER than BALN and RAND. The PROP is the combination of OPIT and PPLX, it outperforms the BALN with 6.2 points, and the RAND with 4.90 points; the CER of PROP is 1.90 points lower than the PPLX. By comparison between two different clustering methods, the CER of PROP is 0.50 points lower than TOPI, which demonstrates that our proposed style-based clustering effectively models the spontaneous open domain speech, although the improvement is not so great.

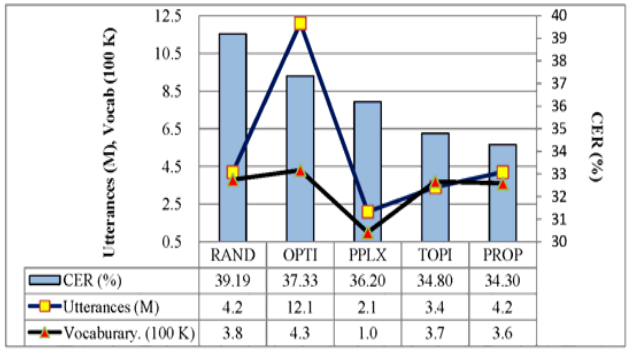

Figure. 4 Recognition performance (CER), training data size, and vocabulary size in different data selections

5. Conclusions

In this study, we proposed a method that combined sentence style clustering and perplexity-based scoring approach to select sentences from a Chinese web archive for training a Chinese LM of an ASR system. The purpose of this work is to collect as many spontaneous like sentences from the web data as possible. By clustering texts based on non-noun POS words and measuring perplexity relative to LM trained by the corpus in which the spontaneous transcripts are added, the spontaneous sentence selection is enhanced. As a result, we selected over 4 M sentences covering 350 K vocabulary words from the web data. Using the LM interpolated by these selected data with the existing corpus, which are mainly in the travel domain, we achieved an average of 6.2 points in CER reduction over the baseline LM which is constructed by the travel domain corpus, and 4.9 points over a model constructed by random selection from the web data. Compared with solely using the perplexity-based similarity method, we verified that 1.90 points of reduction in CER was obtained by combination with the sentence clustering. We also verified that the style-based clustering is more effective than the topic-based one, although the difference is not big. The reason for the small difference is supposed to be that spontaneous texts are insufficient in number on the Web, and the clustering performance is influenced by it.

For future work, we will improve the selection of clustering vocabulary so that the spontaneous texts are better characterized. Meanwhile, we will also study simulations that add filler and short pauses to the texts that are to be selected.

References

1) http://www.sogou.com/labs/dl/t.html
2) X. Zhu and R. Rosenfield, "Improving Trigram Language Modeling with the World-Wide-Web," Proc. ICASSP, pp. 533-536, 2001.
3) I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling Using Class-dependent Mixtures," Proc HLT, vol. 2, pp. 7-9, 2003.
4) C. Munteanu, G. Penn, and R. Baecker, "Web-based Language modeling for Automatic Lecture Transcription," Proc. Interspeech, pp. 2353-2356, 2007.
5) T. Misu and T. Kawahara, "A Bootstrapping Approach for Developing Language Model of New

Spoken Dialogue Systems by Selecting Web Text," Proc. Interspeech, pp. 9-13, 2006.
6) T. Hori, D. Willett, and Y. Minami, "Language Model Adaptation Using WFST-Based Speaking Style Translation," Proc. ICASSP, vol. 1, pp. 228-231, 2003.
7) Y. Akita and T. Kawahara, "Topic-Independent Speaking-Style Transformation of Language Model for Spontaneous Speech Recognition," Proc. ICASSP, pp. IV33-36, 2007.
8) K. Ohta, M. Tsuchiya, and S. Nakagawa, "Effective Use of Pause Information in Language Modeling for Speech Recognition," Proc. Interspeech, pp. 2691-2694, 2009.
9) R. Masumura, S. Hahm, and A. Ito, "Training a Language Model Using Web Data for Large Vocabulary Japanese Spontaneous Speech Recognition," Proc. Interspeech, pp. 1465-1468, 2011.
10) N. W. Xue and L. Shen, "Chinese Word Segmentation as LMR Tagging," Proc 2nd SIGHAN, pp. 176-179, 2003.
11) F. C. Peng, F. F. Feng, and A. McCallum, "Chinese Segmentation and New Word Detection Using Conditional Random Fields," Proc. COLING, pp. 562-568, 2004.
12) X. H. Hu and H. Kashioka, "Chinese Character-based Segmentation & POS-tagging and Named Entity Identification with a CRF Chunker," Proc. 5th International Symposium on Chinese Spoken Language Processing, pp. 693-702, 2006.
13) http://www.sinica.edu.tw/SinicaCorpus/
14) http://www.ldc.upenn.edu/
15) http://glaros.dtc.umn.edu/gkhome/views/cluto
16) http://mastar.jp/translation/voicetra-en.html
17) R. Iyer and M. Ostendorf, “Relevance Weighting for Combining Multi-domain Data for n-gram Language Modeling,” Computer Speech and Language, I3, pp267-282, 1999.