Journal of Information Processing Vol.21 No.2 168–175 (Apr. 2013)

[DOI: 10.2197/ipsjjip.21.168]

Regular Paper

Collecting Colloquial and Spontaneous-like Sentences from Web Resources for Constructing Chinese Language Models of Speech Recognition

Xinhui Hu[1,a)] Shigeki Matsuda[1] Chori Hori[1] Hideki Kashioka[1]

Received: June 1, 2012, Accepted: November 2, 2012

Abstract: In this paper, we present our work on collecting training texts from the Web for constructing language models in colloquial and spontaneous Chinese automatic speech recognition systems. The selection involves two steps: first, web texts are selected using a perplexity-based approach in which the style-related words are strengthened by omitting infrequent topic words. Second, the selected texts are then clustered based on non-noun part-of-speech words and optimal clusters are chosen by referring to a set of spontaneous seed sentences. With the proposed method, we selected over 3.80 M sentences. By qualitative analysis on the selected results, the colloquial and spontaneous-speech like texts are effectively selected. The effectiveness of the selection is also quantitatively verified by the speech recognition experiments. Using the language model interpolated with the one trained by these selected sentences and a baseline model, speech recognition evaluations were conducted on an open domain colloquial and spontaneous test set. We effectively reduced the character error rate 4.0% over the baseline model meanwhile the word coverage was also greatly increased. We also verified that the proposed method is superior to a conventional perplexity-based approach with a difference of 1.57% in character error rate.

Keywords: spontaneous text collection, the Web data, Chinese language model, automatic speech recognition

1. Introduction

In state-of-the-art large vocabulary automatic speech recognition (ASR) systems, a large statistical language model (LM) is used, typically an n-gram. The n-gram LM is generally constructed by using a textual corpus. Its performance depends heavily on the size and quality of the corpus. Here, the quality refers to the matching extent between the corpus content and the recognition task, and to the similarity between the style of the text and the target speech. The ideal training text is the transcript of target speech because it truthfully reflects the speech content and style. However, the manual construction cost of such a text corpus is very high, and the efficiency is also very low, particulary for transcribing colloquial and spontaneous speech. Thus, automatic approaches to collect such text data are required.

From the viewpoints of LM training of ASR system, texts are categorized into "read," "colloquial," and "spontaneous" styles. The "read" text refers to formal text mainly found in usual text like newspapers and published books. The read texts are generally formal, with correct grammar; The "colloquial" texts are mainly chat-like text such as Twitter, micro-blogs, etc. Compared with the "read" texts, the colloquial texts are generally informal, possibly containing incorrect grammar sentences. Furthermore, the "spontaneous" or "spoken" refers to texts corresponding to conversational speech, there are disfluencies such as fillers, false starts, and hesitations etc in it. Different ASR tasks ask for different training data. For example, a LM of a broadcast news ASR system can be effectively trained by using collected newspaper data [1]. Compared with the "read" data, it is difficult to collect the "colloquial" and "spontaneous" texts for constructing LM. With the increased internet services such as Twitter, many texts of free-writing style appear on the web, these texts are very close to "colloquial" and "spoken." Thus, it is possible to obtain colloquial and spontaneous texts from the web for training LM.

There have been studies on collecting web data for constructing LMs [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 [3]. Moore et al. adopted an approach of selecting training data based on comparing the entropy according to domain-specific and non-domain-specific LMs [4]. They showed that this produced a better LM than either random data selection or method based on perplexity according to a domain-specific LM.

These methods are valid for topic adaptation, but have difficulty in improving the selection of colloquial and spontaneous sentences because selecting keywords that characterize colloquial and spontaneous speech is complicated. Gathering keywords is also difficult for various topics in the case of open domains.

The main evidences characterizing spontaneous speech are disfluencies such as filled pause, repetition, repair and false start, many efforts have been made to cope with the detection and correction of these disfluencies [5], [6], [7]. For example, Duchateau et al. proposed an approach to improve on the robustness of

[1] National Institute of Information and Communications Technology (NICT), Souraku, Kyoto 619–0289, Japan
[a)] xinhui.hu@nict.go.jp

a plain trigram LM by manipulating predication contexts containing repetitions, hesitations or restarts [7]. Besides studies focusing on detecting disfluencies in speech, dealing with spontaneous style of training texts are also studied. Two methods are conventionally used to improve modeling spontaneous languages from the aspect of training texts. One is transformation from a written format model to a spoken format. Hori et al. composed a weighted finite-state transducer (WFST) that translates sentence styles to integrate LMs of different styles of speaking or dialect and different vocabularies [8]. Akita et al. significantly reduced the perplexity and word error rate (WER) by transforming a document-style model into a spoken style based on a statistical machine translation framework [9]. Another method generates disfluencies as done by Ohta et al. that simulate spontaneous sentences by predicting fillers and short pauses in document sentences [10]. Masumura et al. used a naive Bayes classifier to select speech-like texts from downloaded web data and then used the same method as Ohta et al. did to convert the texts into a spontaneous format. Experiments on spontaneous speech recognition showed that an LM trained by the generated data performed as well as the large-scale spontaneous speech corpus [11].

The perplexity-based approach is generally realized by using a n-gram LM, it is easily influenced by the domain of the LM. Those texts that are related to the domain of the LM are easily selected, otherwise, they will be rejected by this approach. For example, a sentence containing out-of-vocabulary (OOV) words generally has a high perplexity, so it may not be selected even it is an adequate one. Another problem of the n-gram-based LM is its difficulty to predict words in a long distance context. Those sentences containing long distance context are more difficult to be picked up than short distance context by using n-gram LM. The n-gram based approach tends to collect sentences containing short context sentences. Therefore, in this way, many sentences having similar construction are collected, but sentences containing word pairs which are separated by other words are not easily selected. The utilization of n-gram LM will result in small variations of structures of collected sentences and cause an over-fitting problem in training process.

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that objects in the same cluster are more similar to each other than to those in others. Generally, correct grammars are used and regular word orders are formed in written texts, however, the word order and grammar of the spontaneous are irregular, even incorrect. It is more difficult to trace spontaneous style than written style by using word order. Different from the perplexity-based approach, the similarity criterion for clustering generally does not ask for word sequence order, and has no limitation on word distance of context. Therefore, we hope to utilize the clustering approach to group texts from the viewpoint of text style to overcome the shortages of the other approaches such as perplexity-based approach, and enhance the selection of colloquial and spontaneous sentences.

In this study, we proposed a method to integrate a perplexity-based approach and a clustering-based approach for data selection so that both approaches complement each other. This paper is organized as follows: Section 2 briefly introduce characteristics of Chinese texts related to spoken language, including colloquial speech and spontaneous speech, by using corresponding corpora. Section 3 describes the system configuration of this study. In Section 4, after introducing the construction of a conditional random field (CRF) based word segmentation and POS-tagging system, we will present our proposed selection procedures in detail. Section 5 reports the experiment results, and Section 6 sets forth our conclusions.

2. Characteristics of Colloquial and Spontaneous Chinese Texts and Seed Data

2.1 Distinctions between Different Speech Styles

In the same way as English and Japanese, Chinese texts related to spoken speech can be categorized into three styles:

(1) Read - Speech based on prepared text in advance, such as TV programs. The text is generally formal, with correct grammar.

(2) Colloquial - Speech occurred in daily life. It is generally informal, possibly containing incorrect grammar sentences. It generally means “spoken language” that includes slang and idiomatic phrases used in everyday speech.

(3) Spontaneous - Colloquial expressions with disfluencies such as fillers, false starts, and hesitations etc.

We categorize all speech-related texts into the above three types so that the data collection can be conducted based on the degree of spontaneity. Among these three categories, the read speech has the least spontaneity, and the spontaneous speech has the highest spontaneity.

2.2 Corpora for Analysis of Colloquial and Spontaneous Speech

We have prepared several corpora corresponding to different styles. The description about these data are shown in the **Table 1**. We use these data to build the baseline or seed LM to select necessary sentences since they cover these styles.

The BTEC [12] (Basic Travel Expression Corpus) is a manually corrected textual corpus for building LMs of Chinese ASR systems, in the travel domain. The texts in it are colloquial expressions mainly inspired by phrase books for tourists and translation from other languages like English and Japanese, there is no disfluency in it. The VoiceTra [13] [*1] corpus contains manual transcripts of speech collected from a speech-to-speech translation (S2ST) service. It is also our target system and it ASR performance we are working on improving. By detailed checks on the speech, its content is found open domain, and its speaking style is found to be colloquial and spontaneous, main part of this

Table 1 Chinese corpora with different styles.

Name	Size (Sentences)	Style	Domain
BTEC	528 K	Colloquial	Travel
VoiceTra	74.7 K (4.7 K) [*2]	Colloquial and Spontaneous	Open
CTS	170 K	Spontaneous	Open

[*1] Although its content is described as travel-related conversation in its web site, it is confirmed to be open domain by native speaker’s manual checks, over 20% of them are non-travel related conversations.

[*2] Sentences used as seed data.

data is colloquial expressions, and the particular characteristics of spontaneous speech such as fillers, hesitations are frequently found existing in it.

The CTS (Chinese Transcripts of Spontaneous speech) contains transcripts of a Chinese (Mandarin) telephone conversational speech corpus. This corpus is collected in mainland China, 500 speakers (composed 250 conversations), mainly university students, are contained in it. Each conversation lasts about 30 minutes, no predefined topics are given to the speakers. So the domain of the corpus is open, and style is spontaneous. The disfluences like repetition, correction, filler pauses, are also transcribed.

Some examples excerpted from the above corpora are shown as follows:

(1) 好的 我 的 卡 是 万事达信用卡。(All right, my card is the Master credit card.) [BTEC]
(2) 半价 是 多少钱 啊? (How much is the half-price?) [BTEC]
(3) 我 的 公司 呃 在 呃 光台 二丁目。(My company is um located at um Hikaridai 2-chome. [VoiceTra]
(4) 请问 哪 哪里 有 厕所? (Please tell me where (where) is the toilet?) [VoiceTra]
(5) 明天 到 了 日本 到 了 就 给 我 打电话。(Tomorrow, when you arrive at Japan, (after arriving) please phone me at once.) [VoiceTra]
(6) 跟 我 寝室 啊 差不多 了 大家 马马虎虎 都 会 有 联系 然后。(It seems the same as my dormitory, um, almost, everyone is just so so, we have connections with each other, and then.. [CTS]
(7) 我 跟 我 跟 我 跟 那个 钟楠 的 没 啥 有 过 联系。(I, I, I, am seldom in touch with that guy, Zhong-Nan.) [CTS]

It is difficult to give a clear definition of each style to distinguish them, because there are overlapping areas among them. However, although still remaining ambiguities, we assumed a sentence is spontaneous when disfluencies are found in it, such as the (5) and (6) of the above examples. The other 4 sentences are regarded as common colloquial ones.

2.3 Seed Data for Constraining Spontaneous Style

In this study, we use a seed data to catch necessary data from the data resource. The seed data should be the same as the target task in both content and style. It is generally collected from the real environment of the ASR applications.

The VoiceTra service system [13] is the target system to improve in this study. We have collected 74.7 K utterances (as shown in the Table 1) from this system. Here, we use 4.7 K of them as seed data. However, all of the transcripts are used to investigate the possible maximum achievement by using a large quantity of adaptation data from real environment, and is used for comparisons with the other selections.

3. System and Data Descriptions

3.1 System Configuration

Figure 1 shows the system configuration of this study. The Sogou corpus [14], a Chinese web archive, is used as the data resource. This archive contains 135.4 million Web pages from 5.3 million Chinese Web sites collected by Sogou.com from June, 2006 to January, 2007. Sogou.com is one of the largest commercial search engines in Chinese Web environment. The original web data are filtered to remove HTML tags, Java script codes, etc., and to normalize them. Then, these data are segmented and part-of-speech (POS) tagged into word texts by a conditional random field (CRF) based word segmentation and POS-tagging system [17]. After that, these sentences are processed by two steps: first, they are selected by using perplexity-based selection. A seed LM is used here for evaluating sentence perplexity. Second, the selected sentences are clustered based on their style and the optimal clusters are chosen as the final data. In this step, a seed data is used to detect the optimization. After the above selection, an adaptation LM is built by these selected sentences, and it is then used to adapt the baseline (seed) LM to build a final LM for an ASR system.

3.2 Seed LM

The seed LM in the system is trained by using three sets of training data: BTEC, CTS, VoiceTra (4.7 K) as shown in the Table 1, they correspond to colloquial, spontaneous styles, and the target task (mix of colloquial and spontaneous), respectively.

The BTEC is a Chinese corpus in the travel domain, with about a 46.5 K word vocabulary, and covers a wide range of travel expressions. It is basically in a colloquial style. Here are some examples excerpted from this corpus. "请问 乌 龙茶 多少钱? (How much does the Oolong tea cost?)," "到 京都站 要 多长时间? (How long will it take to go to the Kyoto station?)" Since it's large and has been manually checked, it has a wide coverage of colloquial expressions, and a high guarantee in quality.

The CTS corpus is created by performing automatic word segmentation and POS-tagging to a manual transcripts of a Chinese spontaneous speech corpus. With it, the insufficiency of the real

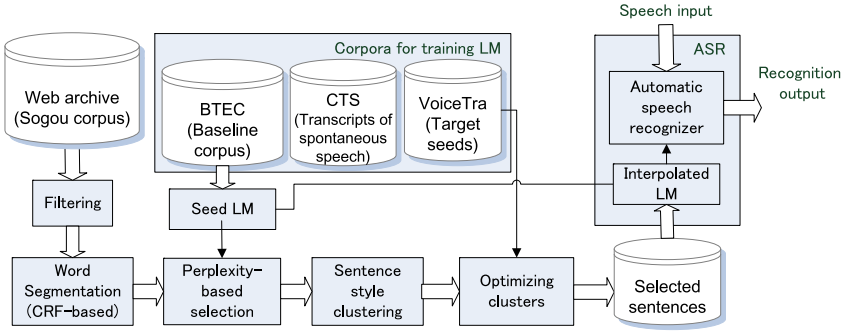


Fig. 1 System configuration.

spontaneous samples can be compensated.

The seed LM is built by linear interpolating a n-gram LM (with an interpolation weight of 0.46) trained by the VoiceTra seed data with a n-gram LM trained by the BTEC and CTS corpus. The vocabulary of the seed LM is 67.2 K words.

4. Data Selection Procedures

4.1 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.

4.1.1 CRF-based Word Segmentation and POS-tagging

Recently, the character-based tagging method, such as CRF-based model [15] and maximum entropy [16], has become the dominant technique for Chinese word segmentation and POS-tagging due to its global optimization and its ability to detect new words. So it is particulary helpful for processing Web data due to the data hugeness and existence of many new words. In this study, we built an CRF-based segmenter [17] for word segmentation and POS-tagging.

4.1.2 Training Data for the CRF-based Word Segmenter

The SINICA balanced corpus [18] (659.1 K sentences) and two files of the LDC2007T03 Tagged Chinese Gigaword corpus [19] [*3] were used as training data for the segmenter. These annotated data are based on identical specifications and cover a wide range of modern Chinese fields. So, it is suitable to process web text. We also used features of the BTEC lexicon in which a lot of colloquial conversation related words are included. The features that are used for training segmentation model are instantiations of the features shown as follows. They are chosen in a context window with a five-character length when the training data is scanned.

(a) $C_{-2}, C_{-1}, C_0, C_1, C_2$
(b) $C_{-2}C_{-1}, C_{-1}C_0, C_0C_1, C_1C_2, C_{-1}C_1$
(c) $C_{-1}C_0C_1$
(d) T_{-2}, T_{-1}
(e) D_{-2}, D_{-1}, D_0

Here, C_i represents the character at the position i, (a) stands for the current character and previous and next two characters, (b) stands for connection between the first character and the second character, (c) means the connection of 3 characters among the previous, present, and the next character. T_i of the (d) means the POS tag set of the character at the position i. In (e), $D_i = 1$ when the character C_i belongs to a vocabulary which is composed of high-frequency words (top 10 K) of the BTEC lexicon, otherwise, it is 0.

With this segmenter, the word segmentation and POS-tagging was conducted to the filtered sentences of the Web data. In all, 130 billion words containing a 2.5 M word vocabulary (with cut-off 25) were estimated in the final data.

We conducted evaluations on the segmenter using the "as_testing," a test set (14 K sentences) of the 2005 Sighan workshop on Chinese segmentation [20]. The F-score of word segmentation and POS-tagging is 0.932, while the F-score of pure word segmentation is 0.940.

[*3] these two files were cna_cmn_200401, containing 33.9 K sentences, and xin_cmn_200401, containing 36.4 K sentences, respectively.

4.2 Selection by Perplexity-based Approach

In the first selection step, we adopted a criterion of sentence perplexity; or equivalently empirical cross-entropy with respect to a seed n-gram LM q as a measure of the distance between the LM and the sentence. The cross-entropy $H(p, q)$ of a sentence with empirical word distribution p given the LM q is:

$$H(p,q) = -\sum_{w_1,\ldots,w_n} p(w_1,\ldots,w_n)\log q(w_n|w_1,\ldots,w_{n-1}) \quad (1)$$

where the sum ranges over words of the sentence, $p(w_1, \ldots, w_n)$ gives the relative observation frequency of words in the sentence, and $q(w|h)$ returns the probability that history h is followed by word w according to the language model q. The perplexity of the sentence is therefore $2^{H(p,q)}$. So, selecting the sentences with the lowest perplexity is equivalent to choosing the sentences with the lowest cross-entropy according to the language model. Motivated by the indexing keywords in information retrieval processing where the infrequent words are referred as topic-related, and the frequent words are referred as the non-topic or style related, we move the topic words out of the perplexity measuring. Concretely, we omit the n-gram items containing nouns of low frequency from the sum in the above equation.

The sentence selection is conducted by comparing its perplexity with a threshold. When it is smaller than the threshold, the sentence is selected. The threshold is decided by analyzing the perplexities of a development set that contains 1,800 utterances of VoiceTra transcripts; The threshold was intentionally set at a high level, so that sentences containing n-gram terms different from the seed data and containing OOV words can be easily picked up. Using this development set, the threshold was selected at 200. Only 50 of the 1,800 utterances were found over this threshold. It was verified that almost all of the 50 utterances contain only OOV words such as proper nouns. Some examples survived sentences at this step are shown as follows:

(1) 目前在岛上的游客安排的最后一班船返程时间是那个7点钟。(Now, those tourists left in the island are assigned to the last ship which will be scheduled at (that) 7 o'clock.)
(2) 伊犁都不敢喝了, 各么苏州本地双喜啊行的? (I don't dare drink Yi-Li milk, how about Suzhou's local brand, the Double-Happiness?)
(3) 我困得了困得了呃困得了 !!! (I am sleeping, sleeping, um, sleeping!!!)
(4) 你还在这里啊！呃，别忘了带钥匙。(You are still in here, aren't you!, um, don't forget your key.)

4.3 Selection by Sentence-style Clustering

4.3.1 Definition of Sentence-style Clustering

Motivated by the fact that topic clustering is mainly based on noun distribution in document clustering, we propose style clustering based on the distributions of POS other than nouns, which is concretely achieved by removing nouns from the clustering vocabulary.

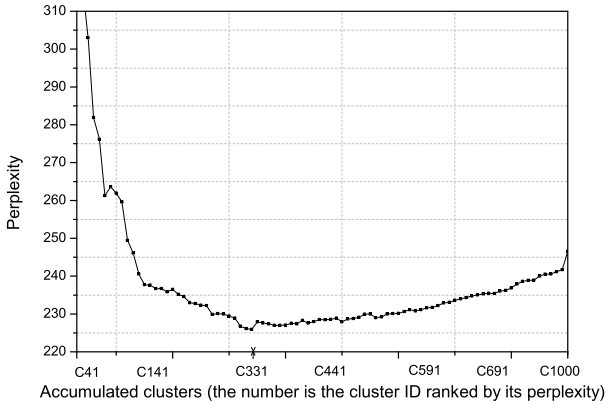


Fig. 2 Perplexities with accumulation of clusters.

4.3.2 Clustering Algorithm

The clustering process finds a predefined number of clusters based on a specific criterion. We chose the following function to maximize the within-class similarity:

$$(S_1 S_2 \ldots S_K)^* = maximize \sum_{i=1}^{K} \sqrt{\sum_{v,u \in S_i} sim(v,u)} \qquad (2)$$

where K (= 1,000) is the desired number of clusters which is obtained empirically by taking both performance and computation into consideration, S_i is the set of sentences belonging to the i^{th} cluster, and v and u are the feature vectors representing the two sentences. The elements in each feature vector are scaled based on term frequency (*TF*), a fundamental parameter in information retrieval processing, of the sentence. The terms are limited by the clustering vocabulary. Another important parameter *IDF* (inverted document frequency) in the information retrieval is not used here since it mainly focus on finding infrequent terms. The $sim(v,u)$ is their similarity, which is computed by the measure of the cosine distance.

The method of Repeated Bisection [21] is adopted. We used the bayon toolkit [22] to realize the above clustering.

4.3.3 Optimized Style Clusters

After the clustering, we built a word 3-gram LM for each cluster. All the clusters were evaluated by the perplexities of the seed data with respect to their corresponding LMs. Here, the perplexity is calculated in the same way as described in the previous Section 4.2 that it is obtained by excluding infrequent words. Then, the optimized clusters were obtained as follows: The clusters were accumulated beginning with the minimum perplexity; meanwhile, the perplexity of the seed set to the LM trained by the accumulated clusters was observed. Here, the unigram probability of the OOV word class is assigned a very small value (here, a negative value −7 of its logarithmic probability is used) for the LM. **Figure 2** shows the perplexity changes with the accumulations of clusters. As shown in the figure, at the point of C331, the perplexity is the minimum. These accumulated clusters, containing totally 3.80 M sentences with a vocabulary of 345.6 K words, are regarded as the optimized clusters.

After all, the selected sentences after different steps are shown in **Table 2**.

Some examples survived sentences at this step are shown as follows:

Table 2 Selected sentences after each steps.

Step	Descriptions	Sentences
0	All Web	11 billion
1	Perplexity-based selection	13.13 M
2	Sentence-style clustering	3.80 M

(1) 这才是现代生活呃…有多少这样的童鞋[*4] 请举手！(This is what a modern life should be, um, How many such classmates are there in here, please raise your hand!)

(2) 有人领到那个什么麦旋风了么. 多收了三五 (Does anyone get that (what) Mcflurry, I paid more 3 or 5 . . .)

(3) 前面有一个…呃……是呃……乡村耶！(There is, um, yes, a village ahead.)

(4) 不是风筝是什么? 还断定……意思就是 UFO? 扯淡！(Isn't that a kite? Why do you conclude . . . you mean it is an UFO? Nonsense!)

4.4 Language Model with Selected Sentences

A 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 existing data - BTEC, CTS, and VoiceTra seed sentences, and $LM_{selected}$ is the LM trained by the selected sentences.

λ (= 0.28) is the weighting factor for tuning the final model. It is obtained using a development set from the VoiceTra data.

5. Experiments

5.1 Experimental Settings

5.1.1 Data Set for Development and Evaluation

To evaluate the quality of the selected sentences, we built an LM as described in Eq. (3) and used it for speech recognition experiments. We selected 606 utterances (EVA01) from the VoiceTra as the evaluation set and another 606 utterances (DEV01) as the development set for tuning LM.

5.1.2 Other Selections for Comparisons

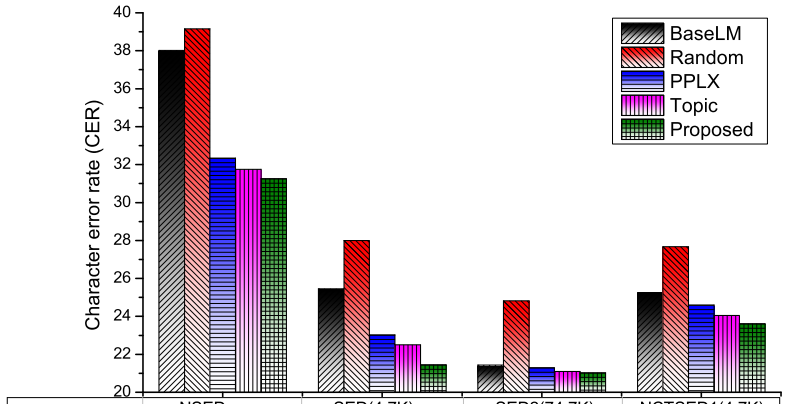
For comparisons with the proposed selection method (Proposed), we made the following selections from the same web data. Except for particular notes, for comparability, the number of sentences in all cases are approximately set at the same scale as the Proposed (approximately 3.80 M sentences).

(1) BaseLM: Baseline LM trained by the BTEC, CTS, and seed sentences (4.7 K) of the VoiceTra. Totally 532.7 K sentences are contained in the training data.

(2) Random: Sentences randomly selected from all the web data.

(3) PPLX: Sentences selected only by the perplexity-based method. The second step, the clustering-based selection, is not used. The selection is performed in the order of sentence perplexity. By constructing an LM using the selected sentences, the perplexity of DEV01 to this LM is investigated, and its change is found to be flat after the sentences are accumulated to a certain amount. We selected a point as the threshold point, at this point, the sentence count is the

[*4] 童鞋 (baby shoes) is a harmonic glossary of 同学 (classmate) frequently appeared in the Web.



	NSED	SED(4.7K)	SED2(74.7K)	NCTSED1(4.7K)
BaseLM	38.02	25.45	21.44	25.25
Random	39.15	28.00	24.82	27.68
PPLX	32.34	23.02	21.30	24.60
Topic	31.75	22.50	21.10	24.05
Proposed	31.25	21.45	21.03	23.61

Training seed data

Fig. 3 Recognition performances (CER) using different sentence selections.

same as the Proposed, the sentence perplexity at this point is regarded as the perplexity threshold. All sentences with smaller perplexities than the threshold are selected for this method. We also verified that the threshold point is in the flat range of perplexity with the sentence accumulation.

(4) Topic: The clustering in the second step is based on the topics; this means that only nouns are used for clustering. In this case, the optimized clusters are determined in a same manner, and its own point of minimum perplexity is found.

(5) Proposed: Sentences selected by the proposed approach in which style-based clustering and perplexity-based selection are integrated.

5.1.3 Investigations on Seed Sentence Sizes in the Baseline LM

To investigate the influence of the seed sentence size on the performance of the LM, three different sizes of seed sentences (NSED (0 sentences), SED1 (4.7 K sentences) and SED2 (74.7 K sentences)) are compared when they are used in the training seed LM. Another case (NCTSED1) is also investigated when the CTS is removed from the SED1.

5.2 Evaluation Results and Analysis

5.2.1 Recognition Performance

Figure 3 shows the recognition results (character error rate: CER) of test set EVA01 using different data selections, in different seed sizes.

From this, we can see that all purposeful selections (PPLX, Topic and Proposed) more effectively decreased the CER than the baseline LM (BaseLM) and random selection (Random). It is obvious that the random selection will deteriorate the performance of LM. Compared with the BaseLM, when no seed sentences were used in it, the improvement with the Proposed was 6.77% (from 38.02% to 31.25%). However, with the increase of seed sentences used in the baseline LM, the improvements decreased. For example, in the case of SED1 where 4.7 K seed sentences are used, the improvement was 4.00% (from 25.45% to 21.45%), and when the seed sentences were increased to 74.7 K (SED2), the improvement became 0.71% (from 21.74% to 21.03%). Compared with only using the perplexity-based approach (PPLX), the addition of clustering processing (both Topic and proposed) further decreased the CER. These facts verified that the collected sentences are refined in content and in style by these clustering approaches. Among the two clustering approaches, the style clustering (Proposed) outperformed the topic clustering (Topic), with difference of 1.05%.

When the CTS is not used for training the seed LM, the recognition performance is worsen than it is used. This verified that the usage of CTS effectively contributed the improvement of the selection, particulary for the spontaneous style.

There is a phenomena to be noted. Though the words used in Topic and Proposed are quite contrary, the noun plays main role in the Topic while non-noun plays main role in the Proposed, both contributed to improving the CER. This can be explained that the target speech (VoiceTRA) has a strong bias towards some topics, such as travel. Having such characteristics in target data, it is believed that these two approaches can be used complementary.

5.2.2 Qualitative Analysis on the Selected Texts

Qualitative analysis also shows the proposed method improved the selection of spontaneous-style. To check the selection effectiveness of the proposed method, we analyzed the selected sentences.

By checking the selected texts, we found that the colloquial texts are effectively collected by using the proposed method. Many new conversational texts are found in the collected data. These are largely owned to the usage of a colloquial-based corpora, especially the usage of the BTEC corpus. Differences were found between the conventional perplexity-based approach and the proposed approach.

(1) Effectiveness of the style clustering

The following sentences have been shown as the results of the Section 4.3. They are selected by the Proposed method, but disappeared in the PPLX.

- 前面 有 一个 … 呃 …… 是 呃 …… 乡村 耶！(There is, um, yes, a village ahead.)

- 不是 风筝 是 什么? 还 断定 …… 意思 就是 UFO? 扯淡！(Isn't that a kite? Why do you conclude . . . you mean it is an UFO? Nonsense!)

It can be explained by the fact that the sentence perplexities of these sentences are high due to the proper noun "乡村 (village)," or "风筝 (kite)," but these words are not found in the infrequent word list, so these words are contained in the perplexity calculation of the Eq. (1). The perplexities of these sentences are higher than the threshold, so they are not selected in the PPLX method. However, in the clustering-based approach, such nouns are removed from similarity measuring, so they can easily be grouped into their correct clusters, and are finally selected.

The proposed approach is also verified to be effective when compared with the topic-based approach (Topic). From Fig. 3, the CER of the Proposed is less than the Topic. By checking the sentences selected by the Topic, most of them are on travel domain. This is understandable because the seed data in this study is mainly in this domain. The following sentences are not selected by it, but they are selected by the proposed approach. From their contents, sentences are not belonged to the travel domain. So, the Proposed method shows its ability to select non-topic or open domain data.

- 昨天 的 面试 非常 顺利, 谢谢 您。(Yesterday's interview was very smooth, thank you.)
- 你 的 打印机 怎么 了? (What is wrong with your printer?)

(2) Effectiveness of the seed data

The selection of spontaneous text is improved by the addition of the spontaneous corpus CTS. For examples, after adding it into the training seed LM, the following sentences appeared in the results. Such kinds of sentences are often seen in texts such as Blogs, their style is more free than the other formal texts. The disfluency-like sentences are found in these examples. They showed that the adoption of the spontaneous data CTS in training the seed LM is effective.

- 你 你 什么 意思? (What do you, you mean?)
- 不够 不够 不够 不够。(Not enough, Not enough, Not enough, Not enough)
- 呃。我 看 不 下去 了。(oh, I can't bear to watch it.)
- 不错。好 好 好 好 好 啊。(Very good. OK, OK, OK, OK, OK.)
- 我 来 看看 是 什么 东东。(Let me see what it is.)

6. Conclusions

In this study, we proposed a method of integrating a perplexity-based approach and a sentence-style clustering based approach to select colloquial and spontaneous-like sentences from the web for training LMs of a Chinese ASR system. Two particular techniques were explored to enhance the style selection: 1) perplexity measuring among the frequent words which are related to non-topic or style related; 2) clustering sentences based on non-noun POS words which are also referred to as characterizing style. In the first technique, a seed LM trained by using colloquial and spontaneous textual corpora containing a small vocabulary and adding a small seed data set (4.7 K sentences) is used. In the second step, a development set is used to find the optimized clusters.

As a result, we selected over 3.80 M sentences (with a vocabulary of 345.6 K words). Compared with the baseline model, which vocabulary is 67.2 K words, the word coverage is greatly increased with the selection. For the recognition performance, using a LM interpolated by these selected texts to the baseline LM, we achieved a definite reduction (4.0%) of CER in a Chinese colloquial and spontaneous ASR experiment over the baseline LM. The performance of the proposed method can cope with directly adding 74.7 K sentences of transcripts from the target speech into the baseline LM training corpus.

Compared with the conventional perplexity-based approach (PPLX), our proposed approach achieved a reduction of 1.57% in CER. Our experiments also showed that style-based clustering outperformed topic-based clustering, with a difference of 1.05%. The above facts can be explained by the fact that the style-based method purified the texts by finding the optimized clusters by which the colloquial and spontaneous styles are further characterized, and the over-fitting problem typically observed in the PPLX method is overcome.

By qualitative analysis on the selected texts, we found that colloquial and spontaneous-like style texts have been successfully collected by the proposed method.

In summary, it was verified that the proposed approach efficiently improved the selection of colloquial and spontaneous speech like sentences, and improved the spontaneous speech recognition performance.

For future work, we will study the characteristics of Chinese colloquial and spontaneous speech quantitatively and apply them to the selection procedures. Furthermore, to compensate for the scarcity of real spontaneous texts, we will study transformation and simulation of spontaneous style to predict disfluency for the selected texts.

References

[1] Gauvain, J., Lamel, L., Adda, G. and Jardino, M.: Recent advances in transcribing television and radio broadcasts, *Proc. European Conference on Speech Communication and Technology*, Vol.2, pp.655–658 (1999).

[2] Bulyko, I., Ostendorf, M., Siu, M., Ng, T. and Cetin, O.: Web Resources for Language Modeling in Conversational Speech Recognition, *ACM Trans. Speech and Language Processing* (*TSLP*), Vol.5, No.1, pp.1–25 (Dec. 2007).

[3] Misu, T. and Kawahara, T.: A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Text, *Proc. Interspeech*, pp.9–13 (2006).

[4] Moore, R. and Levis, W.: Intelligent Selection of Language Model Training Data, *Proc. ACL*, pp.220–224 (2010).

[5] Dufour, R., Jousse, V., Esteve, Y., Bechet, F. and Linares, G.: Spontaneous Speech Characterization and Detection in Large Audio Database, *Proc. SPECOM*, pp.21–25 (2009).

[6] Lin, C.K. and Lee, L.S.: Improved Features and Models for Detecting Edit Disfluencies in Transcribing Spontaneous Mandarin Speech, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.17, No.7, pp.1263–1279 (2009).

[7] Duchateau, J., Laureys, T. and Wambacq, P.: Adding Robustness to Language Models for Spontaneous Speech Recognition, *Proc. COST278 and ISCA Tutorial and Research Workshop* (*ITRW*) *on Robustness Issues in Conversational Interaction*, pp.11–13 (2004).

[8] Hori, T., Willett, D. and Minami, Y.: Language Model Adaptation Using WFST-based Speaking Style Translation, *Proc. ICASSP*, Vol.1, pp.228–231 (2003).

[9] Akita, Y. and Kawahara, T., Topic-independent Speaking-style Transformation of Language Model for Spontaneous Speech Recognition, *Proc. ICASSP*, pp.IV33–36 (2007).

[10] Ohta, K., Tsuchiya, M. and Nakagawa, S.: Effective Use of Pause Information in Language Modeling for Speech Recognition, *Proc. Interspeech*, pp.2691–2694 (2009).

[11] Masumura, R., Hahm, S. and Ito, A.: Training a Language Model Using Web Data for Large Vocabulary Japanese Spontaneous Speech Recognition, *Proc. Interspeech*, pp.1465–1468 (2011).
[12] Hu, X.H., Isotani, R., Kawai, H. and Nakamura, S.: Evaluations of an Annotated Chinese Conversational Corpus in Travel Domain for the Language Model of Speech Recognition, *Proc. Interspeech*, pp.1910–1913 (2010).
[13] available from ⟨http://mastar.jp/translation/voicetra-en.html⟩ (accessed 2012-03-10).
[14] Liu, Y.Q., Zhang, M., Cen, R.W., Ru, L.Y. and Ma, S.P.: Data Cleansing for Web Information Retrieval Using Query Independent Features, *Journal of the American Society for Information Science and Technology*, Vol.58, No.12, pp.1–15 (2007).
[15] Peng, F.C., Feng, F.F. and McCallum, A.: Chinese Segmentation and New Word Detection Using Conditional Random Fields, *Proc. COLING*, pp.562–568 (2004).
[16] Xue, N.W. and Shen, L.: Chinese Word Segmentation as LMR Tagging, *Proc. 2nd SIGHAN*, pp.176–179 (2003).
[17] Hu, X.H. and Kashioka, H.: Chinese Character-based Segmentation & POS-tagging and Named Entity Identification with a CRF Chunker, *Proc. 5th International Symposium on Chinese Spoken Language Processing* (*ISCSLP*), pp.693–702 (2006).
[18] Chen, K.J., Huang, C.R., Chang, L.P. and Hsu, H.L.: Sinica Corpus: Design Methodology for Balanced Corpora, *Proc. 11th Pacific Asia Conference on Language, Information and Computation* (*PACLIC*), pp.167–176 (1996).
[19] available from ⟨http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2007T03⟩ (accessed 2012-03-10).
[20] Emerson, T.: The Second International Chinese Word Segmentation Bakeoff, *Proc. 4th SIGHAN Workshop on Chinese Language Processing*, pp.123–133 (2005).
[21] Zhao, Y. and Karypis, G.: Criterion Functions for Document Clustering Experiments and Analysis, *Machine Learning*, Assinippi Park, Norwell, MA (2003).
[22] available from ⟨http://code.google.com/p/bayon⟩ (accessed 2012-05-30).

Xinhui Hu received his B.E. degree in electronic engineering and M.E. degree in communication and system from Harbin Institute of Technology in 1983, and 1988 respectively. He received his Ph.D. degree from the University of Tokyo in 1995. From 1995 to 2000, he joined the Fujisoft Incorporated as a system engineer. From 2001 to 2003, he joined the Research and Development Center of Toshiba as a researcher. Since 2003, he joined ATR Spoken Language Translation Research Laboratories as a researcher. Since 2009, he joined Spoken Language Communication Laboratory of the National Institute of Information and Communications Technology (NICT) as a researcher. Dr. Hu is a member of Information Processing Society of Japan and the Acoustical Society of Japan. His main interests include language model, speech recognition and information retrieval.

Shigeki Matsuda received his Ph.D. degree from Japan Advanced Institute of Science and Technology (JAIST) in 2003, and joined ATR Spoken Language Communication Laboratories as a researcher. From 2009, he joined Spoken Language Communication Laboratory of the National Institute of Information and Communications Technology (NICT) as a researcher. He is engaged in research on speech recognition and speech signal processing, and is a member of the Acoustic Society of Japan, Information Processing Society of Japan, and IEICE.

Chori Hori received her Ph.D. in information science and engineering at Tokyo Institute of Technology (TITECH) in March 2002. She was a researcher in NTT Communication Science Laboratories (CS Labs) at Nippon Telegraph and Telephone Corporation (NTT), Kyoto, Japan from 2002 to 2004. She was a project researcher at InterACT in Carnegie Mellon University (CMU) in Pittsburgh from 2004 to 2006. She is currently a senior researcher at Spoken Language Communication Laboratory at National Institute of Information and Communications Technology (NICT), Kyoto, Japan since 2007. She was the editor of ITU-T recommendations F.745 and H.625 which were approved in 2010. She has received Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2003, 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation (TAF) in 2009, and International Cooperation Award from the ITU Association of Japan (ITU-AJ) in 2011. She is a member of IEEE, ASJ, and IEICE.

Hideki Kashioka received his Ph.D. degrees in computer science from Osaka University, Osaka, Japan in 1993. From 1993 to 2009, he worked for ATR. From 2006, he works for NICT. He is currently the director of Spoken Language Communication Laboratory at Universal Communication Research Institute, NICT. He is also a visiting associate professor at the Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan from 1999.