A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts

Teruhisa Misu, Tatsuya Kawahara

School of informatics, Kyoto University
Sakyo-ku, Kyoto, Japan

misu@ar.media.kyoto-u.ac.jp

Abstract

This paper proposes a bootstrapping method of constructing statistical language models for new spoken dialogue systems by collecting and selecting sentences from the World Wide Web (WWW). To make effective search queries that cover the target domain in full detail, we exploit the document set described about the target domain as seeding data. An important issue is how to filter the retrieved Web pages, since all of the retrieved Web texts are not necessarily suitable as training data. We induct an existing dialogue corpus of different domain to prefer the texts of spoken style. The proposed method was evaluated on two different tasks of software support and sightseeing guidance, and significant reduction of the word error rate was achieved. We show that it is vital to incorporate the dialogue corpus, though not relevant to the target domain, in the text selection phase.

Index Terms: speech recognition, language model, spoken dialogue system, web text selection.

1. Introduction

The quality of the language model directly affects the performance of the spoken dialogue system. It is desirable to use a statistical language model trained with a large amount of data matched to the task domain. When constructing a new spoken dialogue system, however, it is almost impossible to prepare a large amount of user utterances. Thus, an initial language model is made by handcrafting a grammar, or conducting a Wizard-of-Oz data collection. Such an approach is costly and often unreliable, and more automated methodology is preferred.

Recently, there have been several studies for making use of the World Wide Web (WWW), which is the largest text resource, to complement data collection. For example, Zhu et al.[1] used the n-gram count in the Web search result to estimate unreliable trigram probabilities in a spoken document retrieval task. Bulyko et al.[2] used n-gram entries that appeared frequently in the Switchboard corpus as a search query to retrieve relevant texts from the Web. Sarikaya et al.[3] adopted a similar approach for spoken dialogue systems to enhance the training data; search queries were made using a small amount of domain specific data, and retrieved texts were filtered by considering the similarity with the domain specific data using the BLEU score.

However, these studies assume that an initial baseline language model is available by some data collection. (In previous works, thousands of seeding sentences were used.) In this work, we propose a bootstrapping approach which does not require any domain specific data. Instead, we primarily use a document set or knowledge base (KB) about the target domain. This assumption is reasonable for many spoken dialogue systems of the information retrieval type, because they can be realized by retrieving the document set described about the target domain. For example, a restaurant retrieval system can be realized by retrieving a set of Web pages about the restaurants. A document set is also used in question-answering systems as a background knowledge source. In addition, gathering such in-domain documents is much easier than collecting user utterances. By using these texts as seeding data, we can retrieve much more texts of Web pages for the domain.

But the major problem is that these document texts are in written style, thus not necessarily matched to the style of user utterances to a spoken dialogue system. Therefore, we induct a spoken dialogue corpus of different domains to effectively select texts relevant to the target spoken dialogue system. Here, we assume any existing large corpus which is readily available.

In this paper, the proposed scheme is applied to two different tasks of information retrieval and question-answering. We demonstrate that a language model of sufficient quality can be made in absence of domain-specific data.

2. Proposed Scheme and Evaluated Task Domains

In recent years, the target of spoken dialogue systems is being extended from simple databases to general documents including manuals[4] and news articles[5]. In these kinds of systems, replies of the system are generated based on the document set or knowledge base (KB). In this work, we

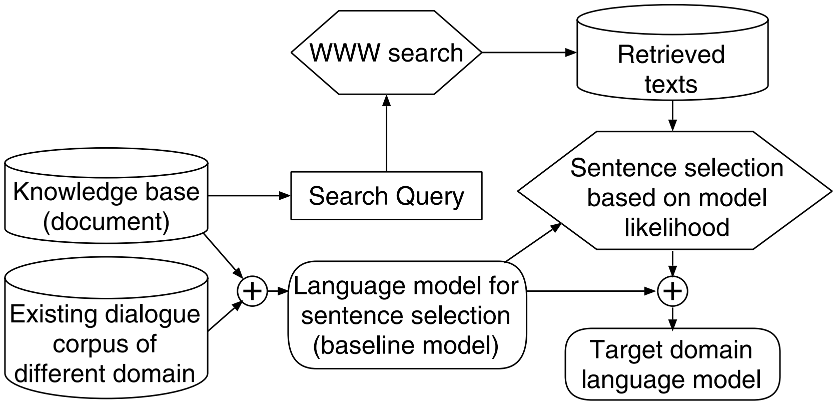
WWW search
Retrieved texts
Knowledge base (document)
Search Query
Sentence selection based on model likelihood
Existing dialogue corpus of different domain
Language model for sentence selection (baseline model)
Target domain language model

Figure 1: Overview of proposed scheme

consider effective use of these kinds of KBs as seeding data for language model construction. By making search queries using the KB to retrieve from the Web, it is expected to collect a large amount of texts to cover the target domain in full detail.

Out of the retrieved texts, we need to select sentences suitable for training data. As a criterion for the sentence selection, we can use the degree of similarity with the seeding data. However, using the KB only as reference data is not enough because the KB is in written style, and thus does not match with user utterances to the dialogue system. We therefore introduce an existing dialogue corpus of different domains to generate a baseline model by interpolating with the KB texts. With this model, we can check if the Web texts are suitable for the training data both in terms of the domain and utterance style.

The flow of the proposed method is illustrated in Figure 1 and summarized as below.

1. Generate a Web search query by extracting keywords out of the KB.
2. Retrieve relevant texts from the WWW.
3. Construct a baseline language model using the KB text and an existing dialogue corpus of different domains.
4. Select training data from the retrieved Web texts based on the model likelihood of the baseline language model.
5. Train the target language model using the KB texts, existing dialogue corpus and selected Web texts.

In this work, we adopt a software support task and a sightseeing guidance task as target domains, and construct language models for speech recognition. As the KB of the software support task, we use the software support document provided by Microsoft Corporation. This KB is used in our document retrieval system "Speech Dialogue Navigator[6]". The KBs of the sightseeing guidance task are the official tourist information of Kyoto city[1] and Wikipedia[2] documents concerning Kyoto. These documents are used as the knowledge source in our upcoming question-answering system. As the existing dialogue corpus of a different domain, we use the CIAIR in-car spoken dialogue corpus which was collected at Nagoya University[7]. This corpus consists of restaurant search dialogues, and the dialogue partners of the users are human operator, WOZ system, and spoken dialogue system, thus the corpus covers various kinds of utterance styles. We select user-side utterances from this corpus and use them as style-matched data. Table 1 shows the size of the above mentioned corpora.

Table 1: Specification of training data

	# sentences	# words
Software support task		
Software support articles	88,440	1.7M
Sightseeing guidance task		
Wikipedia	10,081	0.11M
Tourist information	2,903	0.06M
dialogue corpus of different domain		
CIAIR in-car spoken dialogue corpus	24,701	0.24M

3. Collection of Web Texts

When collecting the training data from the WWW, search query generation is the first important step. In this work, search queries are made by automatically extracting characteristic words of every document in the KB by calculating a TF*IDF score. We select nouns that have large TF*IDF

[1] http://raku.city.kyoto.jp/sight.phtml
[2] http://wikipedia.org/

Table 2: Size of collected Web texts

	# pages	# sentences	# words
Software support	3.4M	88M	1,870M
Sightseeing guidance	0.3M	12M	250M

values for a search query (about 5 words per document), and submit them to the Web search engine.

In the sightseeing guidance domain, however, we use the title of every document (such as a name of a sightseeing spot and a name of a person) in the KB for a search query, because the document size for each entry is much smaller and the keywords in the title well represent the document.

The generated queries were fed into Web search engine Google. The number of the retrieved pages was set to 500 pages in maximum per query. We downloaded only html files from the retrieved results. The retrieved documents are filtered by stripping HTML tags, and split into sentence units.

Table 2 shows the size of the collected Web texts.

4. Sentence Selection from Web Texts

Out of the collected texts, we select "matched" sentences both in terms of the domain and in utterance style, thus appropriate for training data of language model. In many of previous works, all sentences in the retrieved Web pages are used as training data. However, sentences are not necessarily suitable for modeling user utterances in the target spoken dialogue domain.

In order to measure the similarity in the domain and style, we use the baseline language model trained with the KB of the target domain and the existing CIAIR dialogue corpus. Here, incorporation of the dialogue corpus, even though it is of a different domain, is vital to ensure that the selected sentences are appropriate for language model training for the dialogue system. As a similarity measure, we use the average log likelihood or word perplexity calculated by the baseline 3-gram language model. For an unknown word, the minimum probability in the baseline language model is given as a penalty. We then select sentences whose likelihoods are large (perplexities are smaller) than a threshold θ.

Finally, the language model is trained by simply combining all of the KB, the existing CIAIR dialogue corpus and the Web texts selected by the above procedure.

5. Evaluation of Language Model in Speech Recognition Experiment

For the two task domains described in Section 2, the respective language models were generated by the proposed scheme, and evaluated in speech recognition performance. Word error rate (WER) was calculated for evaluation data

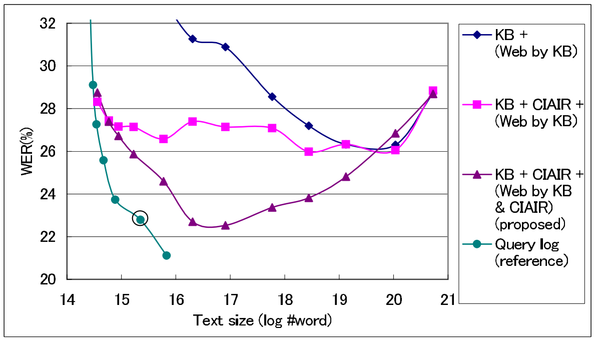

Figure 2: Training text size vs. WER (software support)

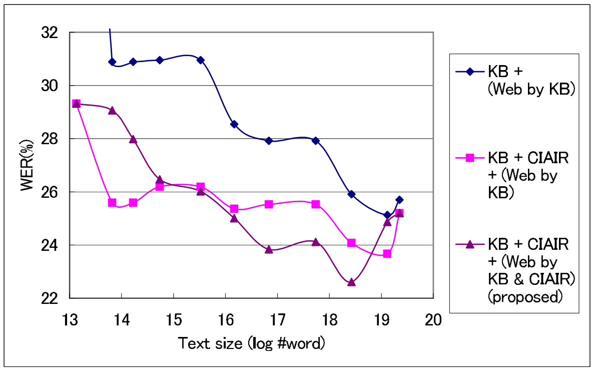

Figure 3: Training text size vs. WER (sightseeing guidance)

of 499 utterances by 30 users in the software support task and 220 utterances by four users in the sightseeing guidance task. For speech recognition, our decoder Julius 3.5[8] and a speaker-independent triphone model were used. The results were shown in Figure 2 and Figure 3. The proposed method was marked as (KB+CIAIR+(Web by KB & CIAIR)) in these figures. In these figures, we plot word error rates against the natural logarithm of the amount of the data used for language model training by changing the threshold θ used for Web text selection. For comparison, we also plot the results when the baseline language model for text selection was made without the existing CIAIR dialogue corpus; that is Web texts are selected using the language model based on the KB only. In this case, we tested two methods: language model is generated by mixing with the CIAIR corpus (KB+CIAIR+(Web by KB)), and without it (KB+(Web by KB)), in order to carefully investigate the effect of the existing CIAIR corpus. The former model shows better performance overall, but there is little difference at the best operating point. In contrast, the proposed method achieved much better performance than those models.

As shown in these figures, using all retrieved Web texts lead to degradation in performance. The result clearly confirms the necessity to select suitable sentences from the retrieved texts. Moreover, in the proposed scheme, the rele-

Table 3: Comparison of language models in ASR performance (WER%)

	Software support	Sightseeing guidance
KB+CIAIR **(baseline)**	28.5	28.4
KB+(Web by KB)	27.0	25.1
KB+CIAIR+(Web by KB)	26.7	24.4
KB+CIAIR+(Web by KB&CIAIR) **(proposed)**	22.8	22.6

vant sentences were more appropriately selected, since the amount of the Web texts at the best operating point (giving minimal WER) was smaller compared with other cases.

We then determined the value of the threshold θ by 2-fold cross validation by splitting the test set into two (set-1 & set-2), that is, set-1 was used as a development set to estimate the threshold θ for evaluation of set-2, and vice versa. We also evaluated the baseline model trained with the KB and CIAIR corpus (without Web texts), by optimizing the interpolation weights of the two. Table 3 shows these results. Approximately, the optimal point was chosen by the cross validation in each case. The difference in WER between (KB+CIAIR+(Web by KB)) and (KB+(Web by KB)) (0.3% in the software support task and 0.7% in the sightseeing guidance task) was due to mixing with the style-matched CIAIR corpus in generating the final language model. The difference between (KB+CIAIR+(Web by KB)) and the proposed method (3.9% in the software support task and 1.8% in the sightseeing guidance task) was due to the use of the CIAIR dialogue corpus for selecting the Web texts. This result demonstrates that it is vitally important to use utterance style-matched data in the text selection stage.

The overall improvement obtained by the proposed method was 5.7% absolute in the software support task, and 5.8% in the sightseeing guidance task. These improvements were statistically significant ($p < 0.01$).

For reference, we also investigate the case where a large number of domain specific data was available in the software support task. We used query sentences collected by the Dialog Navigator[9], which accepts typed-text input. This data is not a transcript of spoken query, but is much similar to the target utterances compared to the KB text and the CIAIR dialogue corpus. The reference language model was made using the KB text and query sentences which were randomly selected from the log data. The result was plotted as (Query log) in Figure 2. It is shown that the proposed method achieved almost comparable performance to the case where three million domain-specific sentences were available (shown by a circle in the figure).

6. Conclusion

We have proposed a bootstrapping method of constructing a statistical language model for a new spoken dialogue system by collecting and selecting sentences from the WWW. The method uses a document set described about the target domain as a seed to generate Web search queries. To select style-matched sentences from the retrieved Web texts, we used an existing spoken dialogue corpus of a different domain. Effectiveness of the proposed method was confirmed by speech recognition in two different task domains. The method does not require in-domain data, except a development set (of hundreds of sentences) for determining the threshold, and shows comparable performance to the case using millions of training sentences.

7. References

[1] X. Zhu and R. Rosenfeld, “Improving trigram language modeling with the world wide web,” in *Proc. ICASSP*, 2001, vol. 1, pp. 533–536.

[2] I. Bulyko, M. Ostendorf, and A. Stolcke, “Getting more mileage fromweb text sources for conversational speech language modeling using class-dependent mixtures,” in *Proc. of Human Language Technology 2003 (HLT2003)*, 2003.

[3] R. Sarikaya, A. Gravano, and Y. Gao, “Rapid Language Model Development Using External Resources for New Spoken Dialog Domains,” in *Proc. ICASSP*, 2005, vol. 1, pp. 573–576.

[4] K. Komatani, T. Kawahara, R. Ito, and H. G. Okuno, “Efficient dialogue strategy to find users’ intended items from information query results,” in *Proc. COLING*, 2002, pp. 481–487.

[5] E. Chang, F. Seide, H. M. Meng, Z. Chen, Y. Shi, and Y. C. Li, “A system for spoken query information retrieval on mobile devices,” *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 8, pp. 531–541, 2002.

[6] T. Misu and T. Kawahara, “Speech-based information retrieval system with clarification dialogue strategy,” in *Proc. Human Language Technology Conf. (HLT/EMNLP)*, 2005.

[7] N. Kawaguchi, S. Matsubara, Y. Yamaguchi, K. Takeda, and F. Itakura, “CIAIR In-Car Speech Database,” in *Proc. ICSLP*, 2004, vol. IV.

[8] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, “Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository,” in *Proc. ICSLP*, 2004, vol. IV.

[9] Y. Kiyota, S. Kurohashi, and F. Kido, “”Dialog Navigator”: A question answering system based on large text knowledge base,” in *Proc. COLING*, 2002, pp. 460–466.