Language Model for the Web Search Task in a Spoken Dialogue System for Children

Jumpei Miyake
junpei-m@is.naist.jp

Shota Takeuchi
shota-t@is.naist.jp

Hiromichi Kawanami
kawanami@is.naist.jp

Hiroshi Saruwatari
sawatari@is.naist.jp

Kiyohiro Shikano
shikano@is.naist.jp

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan

ABSTRACT

In this paper, we propose a method to improve the speech recognition accuracy for web search utterances to a spoken dialogue system. Speech data with a dialogue system are obtained by our speech-oriented information guidance system, "Takemaru-kun" [1], which has been in operation at a public community center since November 2002. From the results of manual labeling of the utterances, child utterances account for about 80%. Most of the web search utterances are out-of-domain words, i.e. trendy words or proper nouns. In order to adapt it to a wider domain, we propose to expand the language model and the vocabulary by collecting from various web resources such as weblogs and open dictionaries. First, we analyze the characteristics of the adult and child web search utterances separately. Then, we make a comparative study of a variety of learning corpora for language model construction. Finally, comparison of the performance of the language models is conducted.

1. INTRODUCTION

Database search using speech recognition has been focused on recently. Additionally, the Internet resources have turned out to be useful those not only for people who use computers regularly, but also for those who are not familiar with typing, the children and the elderly for example. We have continued to develop a speech-oriented dialogue system, which is also a kind of database search.

A variety of dialogue-processing methods are applied to achieve the human-machine communication. However, most of the dialogue systems lack of versatility because of the narrow task domain of the dialogue. One web application called "w3voice"[1] was developed, which enables the speech recognition function to extend on a general web browser. And a question-answering system using web text mining has also been studied actively. Therefore, the dialogue system needs to adapt to a wider domain in the future. For an example of advanced speech recognition to adapting to a wide domain, there is a public web service called "PodCastle," which provides full-text searching of Japanese podcasts on the basis of automatic speech recognition [2]. We aim at recognizing only the speech input for web search in the dialogue system.

In this paper, we focus on the web search task in the dialogue system. Therefore, we first analyze the utterance data for both children and adults. Next, we perform an evaluation of the various language models for web search.

[1]http://w3voice.jp/skeleton/

Figure 1: Speech-oriented information guidance system, "Takemaru-kun."

2. OVERVIEW OF THE TAKEMARU-KUN SYSTEM

The speech-oriented dialogue system called "Takemaru-kun" has been placed at the North Community Center in Ikoma City, Nara Prefecture (Fig.1). The dialogue system has been operated in a real-environment for five years. Its spoken interface is designed as a simple one-question-one-response strategy, aiming at talk without delay. The task domains of the dialogue system are the facility information, sightseeing information, greetings, self-introduction, chat, weather, news and web search.

The architecture of the speech-oriented dialogue system consists of three modules: input speech discrimination, speech recognition and multimodal response. The architecture is shown in Fig.2. User input is recognized in parallel with age group-dependent acoustic and language models. Voice activity detection and GMM-based rejection of noisy or nonverbal inputs are integrated in the open-source Large Vocabulary Continuous Speech Recognition engine, "Julius" [5]. The age group of the speaker is determined the based on the acoustic likelihood. The response is selected by using a question-answer database [3] prepared beforehand for each age group. The response selection method is based on the similarity measure between the n-best recognition hypotheses and the example question in the QA database. The responses consist of a synthesized voice, web pages and animation. The num-

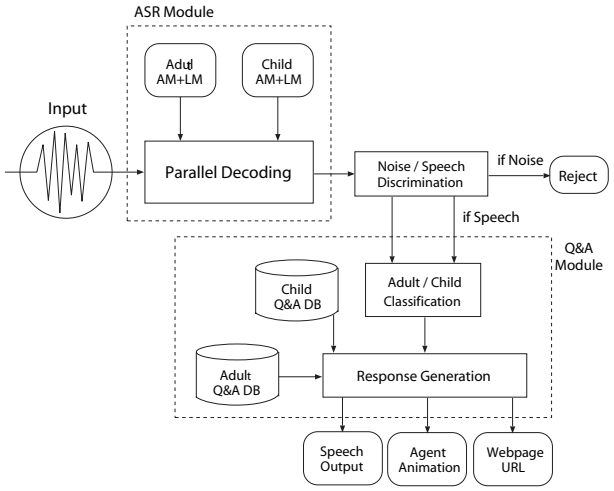

Figure 2: A dialogue system's architecture.

ber of responses is about 360 for both children and adults, although the surface sentences differ for each.

3. UTTERANCE ANALYSIS

For building a corpus suitable for web search, we analyzed the child and adult utterances regarding form and content. The utterance data of the web search task were collected over for about two years from November 2005 to October 2007. The amount of data for adults is 190 and for children is 555. Results of the analysis of utterance form are shown in Fig.3 and utterance content is shown in Fig.4.

In terms of utterance form, one word utterance accounts for about 80% for both adults and children. This analysis shows what words used to perform the keyword search by typing for web search. Therefore, when we construct a language model suitable for web search, we had better collect a corpus including mainly compound words. The sentence utterance includes both questions for conventional Takemaru tasks and questions with a high intention of web search, such as "Where is the garbage can ?" and "Please access the Yahoo! Japan site." A query of web search is used without modifying the result of speech recognition in the our system. Therefore, it is necessary to discriminate the question with a high intention of web search in the sentence utterance in future works. "Other" has utterances that have no meaning, for example, fillers.

In terms of utterance content, there are a lot of utterances concerning proper nouns, including trendy words, in both children and adults. Since their domains are very wide, it is difficult to specify the topic to construct the language model. With children, there are a lot of utterances concerned with TV cartoons, TV games and general words, for example, "baseball" and "soccer." On the other hand, with adults, there are a lot of utterances concerned with local information, the facility and sightseeing.

Therefore, it is necessary to collect a corpus including wide domain contents and to adapt the model to both trendy information and local information. Additionally, as most of utterances are in word form, it is also necessary to remove

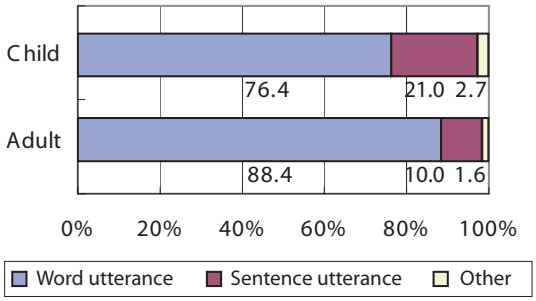

Figure 3: Utterance forms in web search.

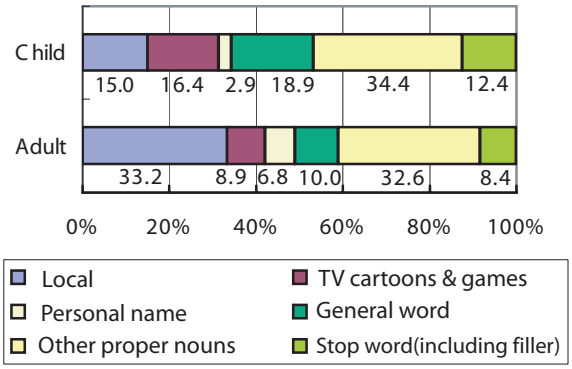

Figure 4: Utterance contents in web search.

useless words, for example, conjunctions and prepositions.

4. CONSTRUCTING LANGUAGE MODEL

In constructing a language model for web search, the training corpus must include a very wide range of topics and a very large vocabulary, for example, the topics of general, trendy information, yesterday's and today's events. Additionally, it must adapt to a domain of local topics for the dialogue system.

For the training corpora of trendy information, we collected texts from the Web that are extensively added to and updated every day. In the web text, we collected a user participation type keyword registration site, keyword ranking sites of web search and weblogs because they are the corpora in which various events in the world and user's interests were reflected. The web sites from which we collected texts are as follows:

- **A user participation type keyword registration site**
  - Wikipedia(Japanese)[2]
  - Hatena Keyword[3]
- **keyword ranking sites of web search and weblogs**
  - Yahoo! Web search keyword ranking(Japan)[4]
  - goo Web search keyword ranking[5]

[2]Wikipedia - http://ja.wikipedia.org
[3]Hatena - http://d.hatena.ne.jp/keyword
[4]Yahoo! Japan - http://searchranking.yahoo.co.jp/
[5]goo - http://ranking.goo.ne.jp/keyword

Table 1: Language model before and after domain adaptation

Language model	Word counts	Vocabulary counts	
		Before	After
Takemaru (conventional)	396741	10407	-
Hatena	408170	92165	100719
Wikipedia	1057803	79292	89031
Keyword ranking	434213	23457	30212

– Keyword of topic by @Wiki[6]
– BLOG360[7]

Both Wikipedia and Hatena keywords are called collective intelligence and many users register various keywords and detail descriptions on the sites. The number of keywords registered in them is more than 20,000. They include wide domain contents. Hatena includes many more trendy keywords than Wikipedia, since its means of registering is simple. Moreover, Hatena generates a synergistic effect in that users want to register the interesting words, since each user can share the interesting words through weblogs.

The method of collecting sentences from the web for a training corpus has been the most widely used [4]. In our case, we don't use sentences but keywords from the web site, because of consutructing a language model including a mainly compound words.

For the local information, we used the language models that are employed in the Takemaru-kun system. The language model is robust for local information and the utterances including fillers, since it was made by the transcriptions of the spontaneous-utterance data. The corpus for this model has 106,325 sentences and the vocabulary size is 10,407.

Texts of the corpus are segmented by the Japanese language morphologic analysis tool, "ChaSen ver2.4.2."[8] The pronunciations of words are given by the dictionary, "ipadic-2.7.0."[9] This dictionally has 239,631 words. Language models are the 3-gram model and the back-off smoothing method is the Witten-Bell discounting method [6]. Finally, we merged the trendy word model from the web with the local model to adapt the domain. Language model merging is performed with the SRILM tool [7] and weighting coefficient to merge is 0.5. In the preliminary experiment, we compared the corpora merging with the language model merging. It is supposed that the characteristic of "Takemaru" is weakened since there is a significant difference in the size of vocabulary between "Takemaru" and the other models i.e. "Wikipedia" and "Hatena". In this paper, only the results of the language model merging for which the performance is better are described. Each of the language models before and after merging are shown in Table 1. The corpora of web search and weblog keyword ranking sites are treated as one corpus, "Keyword ranking," since their word counts are small as compared with both "Wikipedia" and "Hatena". We collected the corpora on May 7, 2008 and all keywords are registered at this time.

[6]@Wiki - http://blog.with2.net/trend_words.php
[7]BLOG360 - http://blog360.jp(service terminated)
[8]ChaSen - http://chasen.naist.jp/hiki/ChaSen/
[9]ipadic-2.7.0 - http://chasen-legacy.sourceforge.jp/

Table 2: Experimental conditions

Test data	Period	Total data
Adult	2005.0108-2007.1006	163
Child	2005.0108-2007.1006	423
Student	2008.0111-2008.0116	157
Acoustic model	2002.10-2004.10	
Speech Recognizer	Julius Ver. 4.0.1	

5. EVALUATION

5.1 Experimental Conditions

We evaluated the performance of the language models that were constructed by web text corpora and that have been used in Takemaru. The measure of the evaluation is word correct rate. None of the test data in this experiment include training data, thus evaluation is conducted by open test. The test data are the child and adult utterance data collected from the web search task of Takemaru operated in the North Community Center, i.e. "Child" data and "Adult" data. In addition, we collected new web search task data by more than ten students in a student room at our college for this experiment, i.e. "Student" data. Therefore, "Student" data is different from both "Child" data and "Adult" data in terms of environment. However, this is not a crucial difference because the microphone has a high directional characteristic and a user speaks into the microphone at a close distance. By comparison with the "Child" data and "Adult" data, most topics of "Student" data are trendy information and it contain little local information.

The acoustic model is set as the initial model to the model which is trained with the JNAS database [8] and trained with each adult or child utterance data collected in the Takemaru system. Ths context-dependent phonetic-tied mixture model [9] for real-time decoding is employed. The speech recognizer is "Julius Ver.4.0.1". Other experimental conditions are shown in Table 2. OOV (Out-Of-Vocabulary) and Perplexity of test data for the language models are shown in Table 3.

5.2 Experimental Result

We evaluated the performance of the language models by word correct rate. The test data is the utterance data in the web search task and is not included in the training data. Results before and after merging with "Takemaru" are shown in Figs.5 to 7.

In terms of the results before merging, the best result is the "Hatena" model, whose vocablary size is largest and OOV is low. The "Hatena" model can recognize a wide domain and trendy information. The "Takemaru" model is not proper for "Student" data. Both "Adult" data and "Child" data include much local information. on the other hand, "Student" data include little local information and much trendy information. The "Takemaru" model, which is the conventional model is not suitable for trendy information. Interestingly, although the "Keyword ranking" model is 50,000 vocabulary

Table 3: OOV and Perplexity of test data before and after model merging

Adult	Before		After	
	OOVs	ppl	OOVs	ppl
Takemaru	16.0%	21.9	-	-
Wikipedia	5.1%	127.5	2.6%	63.5
Hatena	5.6%	78.4	4.1%	44.7
Keyword ranking	12.9%	47.9	7.2%	36.4
Child	Before		After	
	OOVs	ppl	OOVs	ppl
Takemaru	10.5%	32.5	-	-
Wikipedia	4.0%	92.0	2.8%	44.4
Hatena	2.1%	68.9	1.5%	39.9
Keyword ranking	7.3%	36.7	3.6%	31.4
Student	Before		After	
	OOVs	ppl	OOVs	ppl
Takemaru	21.8%	49.9	-	-
Wikipedia	7.6%	122.3	5.5%	114.8
Hatena	3.4%	78.9	2.3%	84.6
Keyword ranking	10.9%	41.2	6.7%	51.7

words smaller than "Wikipedia" data and "Hatena" data, the performance of "Keyword ranking" data was almost equal to both "Wikipedia" data and "Hatena" data. This suggests that the domains of utterances in web search by the spoken dialogue system are similar to the domains of web search by keyboard input.

In terms of the results after merging, all models increased the word correct rate of the test data about 15% or more than the conventional language model since the models after merging can be adapted to both local information and trendy information.

Namely, it is important for the speech recognition in web search of the dialogue system to collect the corpora of a wide domain, and adapt with the domain matched to a topic. However, it does not only depend on vocabulary size. It also suggests it is necessary to collect a training corpus suitable for web search efficiently.

6. CONCLUSIONS

We constructed a language model considering both trendy and local information. It improved the speech recognition accuracy about 15% over the conventional model. In this paper, pronunciation of words in corpora is given from a general dictionary. Therefore, it often contains trendy words or new words for which the correct pronunciation is not given. We will examine a method to automatically give the correct pronunciation to the trendy words in future works.

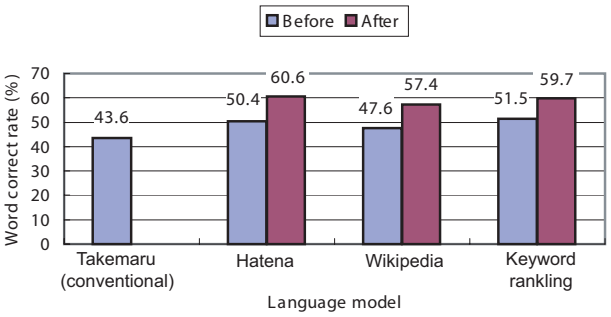

Figure 5: Results of word correct rate for "Child."

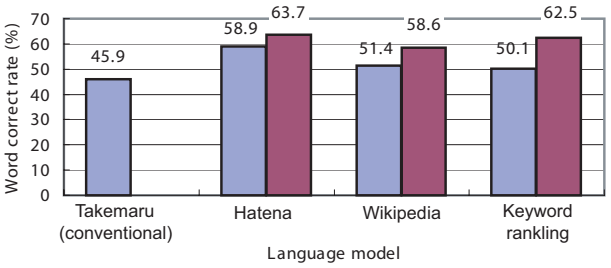

Figure 6: Results of word correct rate for "Adult."

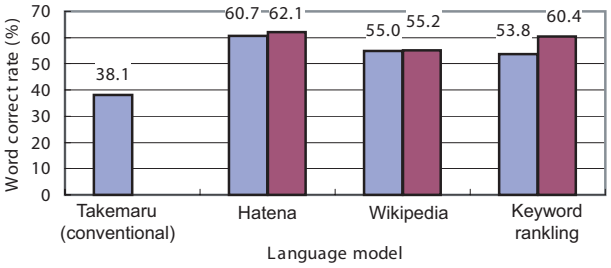

Figure 7: Results of word correct rate for "Student."

7. REFERENCES

[1] Ryuichi Nishimura, Akinobu Lee, Hiroshi Saruwatari and Kiyohiro Shikano, *"Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability,"* in Proc. ICASSP 2004, vol.1, pp.433–436, 2004.

[2] Jun Ogata, Masataka Goto, and Kouichirou Eto, *"Automatic Transcription for a Web 2.0 Service to Search Podcasts,"* in Proc. INTERSPEECH 2007, pp.2617–2620, 2007.

[3] Shota Takeuchi et al., *"Construction and Optimization of a Question and Answer Database for a Real-environment Speech-oriented Guidance System,"* in Proc. Oriental COCOSDA 2007, pp.149–154, December 2007.

[4] Motoyuki Suzuki et al., *"Unsupervised language model adaptation based on automatic text collection from WWW,"* in Proc. INTERSPEECH 2006, pp. 2202–2205 2006.

[5] "*Julius, an Open-Source Large Vocabulary CSR Engine* – http://julius.sourceforge.jp/,"

[6] I.H. Witten and T.C. Bell, *"The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression,"* IEEE Transactions on Information Theory, Volume 37, Issue 4, pp.1085–1094,Jul,1991.

[7] Andreas Stolcke, *"SRILM - an Extensible Language Modeling Toolkit,"* in Proc. ICSLP 2002, pp.901–904, 2002.

[8] Itou Katunobu, Yamamoto Mikio, Takeda Kazuya, Takezawa Toshiyuki, Matsuoka Tatsuo, Kobayashi Tetsunori, Shikano Kiyohiro and Itahashi Shuichi, *"JNAS : Japanese speech corpus for large vocabulary continuous speech recognition research,"* The Journal of th Acoustical Society of Japan, vol.20, pp.199–206, 1999.

[9] Akinobu Lee, Tatsuya Kawahata, Kazuya Takeda, and Kiyohiro Shikano, *"A New Phonetic Tied-Mixture Model for Efficient Decoding,"* in Proc. ICASP 2002, pp.1269–1272, 2000.