Efficient language model development for spoken dialogue recognition and its evaluation on operator's speech at call centers

Kiyokazu MIKI, Kaichiro HATAZAKI, and Hiroaki HATTORI

Media and Information Research Laboratories, NEC Corporation
1753 Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa 211-8666, JAPAN
k-miki@bq.jp.nec.com, k-hatazaki@ah.jp.nec.com, h-hattori@bp.jp.nec.com

Abstract. While a language model for recognition of spoken dialogue is ideally built from a very large, specific-task-oriented corpus, a great amount of time and effort is required to develop such a corpus, and this involves both the audio recording and written transcription of large amounts of speech data. Training data for a language model should match the target task in both topic and style. What is needed, then, is a method to utilize previously existing spoken dialogue corpora that are not necessarily related to the specific target-task. Such corpora would be combined with documents related to the topic of the target-task to develop a language model for the target spoken-dialogue. In this paper, we propose a method for combining previously existing corpora with key phrases (i.e. phrases that contain keywords) extracted from task related documents. Even though the added data is from documents related to the target dialogue, since it consists of key phrases, stylistic differences (between document data and the actual dialogue to which the model will be applied) are not a problem. We have produced a model using this method and have evaluated it in use on actual spoken dialogue collected at call centers. Experimental results show that a relative 13% reduction in word error rate could be achieved with the addition of key phrases. This performance is nearly as good as that which would be achieved on the basis of a large, expensive transcript-corpus, and the cost of producing the key phrase data is essentially negligible. Such cost reduction achieved by our method will enable speech recognition applications to be more widely used.

Keywords: Spoken dialogue recognition, Language model adaptation, Call center.

1 Introduction

Call centers are particularly appropriate for use as speech recognition application domains because almost all of their operations are based on spoken dialogue. The many attempts made to date to apply speech recognition technology to call center applications have mostly been based on IVR (Interactive Voice Response) including automatic call-routing, which is designed for taking over human operator work in recognizing and responding to customer speech. On the other hand, we are developing another speech recognition application that recognizes operator speech and helps human operator work. For example, product names are extracted from speech recognition results of the operator speech and the information about them is displayed to the operator. An application focusing on operator speech would seem more promising because of the advantage it has for ease of speech recognition. While customer speech is telephone-transmitted, operator speech can be recorded directly from a headset microphone, and the sound quality is much higher. Further, the enunciation of professionally trained operators can be expected to be clearer from the beginning.

Even for the recognition of operator speech at call centers, however, a large training corpus is needed to build a language model. The development of such a corpus requires the recording and transcription of a very large number of utterances. The cost of such work is high because fully-trained transcribers are needed and, even for them, the transcription will require more than 4 times the amount of time that was needed for the original recording. Moreover, because spoken dialogue in business situations is often highly confidential, collecting such dialogue can itself present a problem, and strict management will be required even when such collection has been possible. In order to reduce these costs and increase the possibilities for applying speech recognition technology, a method is needed for using inexpensively

produced, target-specific data-sets to adapt previously existing language models for use on specific target tasks. Such method is one of the methods known as "language model adaptation" and reviewed in [1]. Especially when combining conversational-style data with document-style data, adding unigram probabilities from document-style data to an existing class bigram learned from conversational-style data is proposed in [2]. In [3] specific-task-oriented data are collected from WWW. Search queries which are comprised of the phrases from existing conversational-style corpora and the words about the topic of the specific-task are used to get the data that match the style and topic of the target recognition task. In [4], if a small amount of specific-task-oriented data is available, similarity measure is used for selecting the training data from the data collected from WWW. In [5] verbs and noun phrases of the target-task are extracted from the data collected from WWW by syntactic analysis, and the artificial sentences that include the verbs and noun phrases are generated with templates. In [6] task related predicates and arguments are extracted from task related documents by semantic analysis, and conversational-style data are generated with templates. Statistical transformation model is estimated from parallel aligned corpus of the conversational-style transcripts and their document-style texts with statistical machine translation technique, and the task related documents are transformed into conversational-style data by the model [7].

In this paper, we propose a method for combining previously existing spoken dialogue corpora with key phrases (i.e. phrases that contain keywords) extracted from task related documents. Even though the added data is from documents related to the target dialogue, since it consists of key phrases, stylistic differences (between document data and the actual dialogue to which the model will be applied) are not a problem. We have evaluated our method in recognition tests on actual spoken dialogue collected at call centers.

2 Data Combining

In the following discussion in this paper, we determined to use only the following two kinds of language resources for combination with previously existing corpora:

1. A small amount of transcripts for the target task
2. Documents related to the target task

These two resources seemed to be the most feasible for producing a model at low cost. Documents related to the target task include, for example, dialogue summaries. For call centers, they often keep summarized reports of phone calls in storage already. We set as our goal the achievement of adequate recognition performance on the basis of the addition of these resources. We introduce three language model data-combination methods here:

1. Corpus mixture
2. Keyword addition based on class membership mixture
3. Key phrase addition (addition of phrases (N-gram) containing keywords)

2.1 Corpus Mixture

Corpus mixture is the simplest method of the three. It simply mixes multiple corpora for the production of a single language model. With corpus mixture, the word-occurrence probability in an N-gram model $p(w_n \mid w_1, \cdots, w_{n-1})$ may be expressed as:

$$p(w_n \mid w_1, \cdots, w_{n-1}) = \frac{\sum_k \lambda_k C_k(w_1, \cdots, w_n)}{\sum_k \lambda_k C_k(w_1, \cdots, w_{n-1})} \quad . \tag{1}$$

$C_k(w_1, \cdots, w_m)$ refers to the number of appearances of word sequence $w_1, \cdots, w_m$ in corpus *k*. λ is a weight value. Because the data employed here in the adaptation is characterized by actual phrasings, rather than by use of keywords, this method is only appropriate when mixing corpora that are of same

styles, for example, when adding transcripts of the target spoken dialogue to other spoken dialogue corpora.

2.2 Keyword Addition Based on Class Membership Mixture

Keyword addition is an adaptation method that imports keywords for the target task into a base language-model. Because this method employs keywords that are common among corpora related to a given topic, it is suitable for use even when mixing corpora that are of different styles. While documents, for example, are of a written language style, transcripts are of a spoken language style, and keyword addition makes it possible to reduce the problems created by this difference. Combining a class N-gram language model with this keyword adding method is proposed in [2]. The approximation of a word bigram probability in a class bigram language model may be expressed as:

$$p(w_2 \mid w_1) \approx p(c_2 \mid c_1) p(w_2 \mid c_2) \ , \tag{2}$$

where c is the class of w. We assume that each word in a vocabulary belongs to only one class. $p(c_2 \mid c_1)$ is a class bigram probability, and $p(w \mid c)$ is a class membership probability. The former represents the style of the corpus from which the class bigram probability is to be estimated. The latter represents the vocabulary distribution of the training corpus. For keywords, class membership probability $p(w \mid c)$ may be expressed as:

$$p(w \mid c) = \frac{\sum_k \lambda_k C_k(w)}{\sum_k \lambda_k C_k(c)} \ , \tag{3}$$

For words other than keywords, the base language-model can be used as is. With this method, the definition of each class and the assignment of each word to a class must be the same for all corpora. It does not suffer from the mixing of different-style corpora, for example, adding documents to spoken dialogue corpora, but it is unable to represent the contexts surrounding keywords.

2.3 Key Phrase Addition

In order to combine robustness with regard to differences in style with an ability to include word contexts, we have developed a language model adaptation method that adds key phrases (phrases that contain keywords) to a base language-model. In our past work with text retrieval systems, this method helped improve speech recognition performance when question-style texts and document style manuals were mixed [8]. Key phrase (N-gram) probability may be expressed as:

$$p(w_n \mid w_1, \cdots, w_{n-1}) = \frac{\sum_k \lambda_k C_k(w_1, \cdots, w_n)}{\sum_k \lambda_k C_k(w_1, \cdots, w_{n-1})} \ , \tag{4}$$

if the word sequence $w_1, \cdots, w_n$ contains keywords. For non-key N-grams the original base language-model can be used as is. With this method, it is possible to represent the contexts surrounding keywords. Here, “keyword context” refers to words that appear in the general vicinity of a given keyword. These may include function words (e.g. “with” following a keyword “trouble”) or sequential keywords (e.g. a keyword “desktop” followed by another keyword “PC”). Fig. 1 illustrates what kind of data in task related documents are added to existing conversational-style corpora by this method. Keyword context is independent of differences in style and is an important factor in speech recognition.

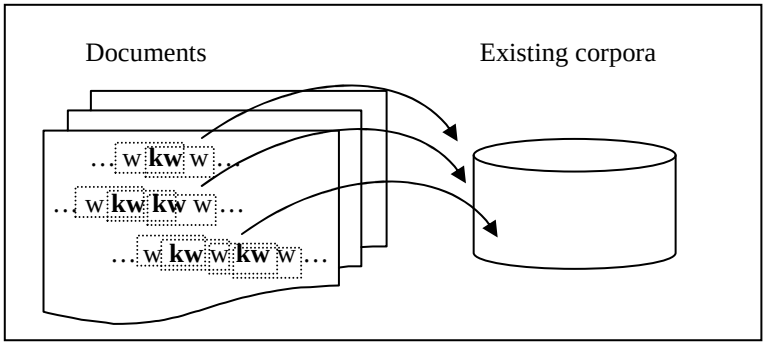

Fig. 1. Key Phrase Addition

3 Experiments and Discussion

3.1 Conditions

We used a call center as the domain for our target task and conducted speech recognition experiments on the actual Japanese-language speech data of operators in three call centers. The first two, call centers A and B, were product customer-service desks, while the third, call center C, was a services customer-service desk. Our target call center was A. Language resources used in the experiments are listed on Table 1.

Table 1. Language Resources

Name	CC	Type	Words	Sentences
Trans-A(S)	A	Transcripts	46K	5.5K
Doc-A	A	Documents	1.6M	85K
Trans-B(L)	B	Transcripts	692K	86K
Trans-C(L)	C	Transcripts	499K	83K
Trans-A(L)	A	Transcripts	444K	61K

Trans-A(L) [L=large] includes Trans-A(S) [S=small] and was collected at CCA (call center A) to verify the performance of the best condition in the experiment. Trans-A(S) was prepared to verify the performance of the situation that a small amount of transcripts is available. Doc-A were summarized call-reports provided by CCA. Trans-B(L) and Trans-C(L) were previously existing transcripts also made available to us. In previous work we have found that at least 100 hours of recorded data are needed to achieve good recognition performance for spoken dialogue at call centers. All large transcripts (labeled (L) in Table 1) were transcribed from over 100 hours of recordings. Small transcripts (labeled (S) in Table 1) represent roughly 10 hours of recordings. We tested our method on the Japanese-language speech of 6 male operators (the total volume of speech was roughly 64K words).

To capture the style and topic of the target-task, keyword set is obtained as follows:

1. Firstly, calculate tf-idf value for the vocabulary of Doc-A by using Doc-A and additional 10K newspaper articles.
2. Rank the vocabulary according to the score (tf-idf value).
3. Extract top ranked 10K words.
4. Furthermore, extract “content words” (e.g. nouns, verbs, adjectives, etc.; as opposed to such “function words” as conjunctions, post-position particles, etc.) from the extracted 10K words as the keywords.

As an acoustic model, we used a speaker-independent HMM representing context-dependent triphones. We used a statistical word trigram model as a language model and employed a back-off

smoothing method which uses class bigrams rather than word unigrams when no word bigrams are available. Word classes are based upon their syntactic part of speech. When a single vocabulary item functioned in one place as a part of speech different from the part of speech that it functioned as in another place, we treated the two instances as different words. The recognition process was composed of 2 stages. In the first stage, a beam search with a word bigram language model produces a word lattice, and in the second stage, that word lattice is rescored using a word trigram language model.

Table 2. Language Model Properties and Recognition Rates

Language Resources	Adaptation	Word		Characters		OOV	PP
		Cor.	Acc.	Cor.	Acc.		
1. Doc-A		54.5%	45.1%	63.2%	59.0%	6.7%	49972
2. Trans-C(L)		63.3%	50.6%	66.6%	60.8%	11.5%	1856.5
3. Trans-B(L)		70.2%	60.3%	75.2%	70.3%	4.3%	318.4
4. Trans-A(S)		70.4%	59.3%	75.7%	70.4%	6.8%	442.3
5. Trans-B(L)+Trans-A(S)	CM	72.6%	63.9%	77.9%	73.3%	2.4%	225.1
6. Trans-B(L)+Doc-A	KW	72.1%	64.2%	78.0%	73.5%	1.8%	247.5
7. Trans-B(L)+Doc-A	KP	73.1%	65.6%	78.8%	74.7%	1.5%	259.7
8. Trans-A(L)		75.2%	67.3%	80.4%	75.9%	1.5%	148.6

3.2 Results

Experimental results are shown in Table 2. The Language Resources column indicates what language resources were used. The Adaptation column shows the adaptation method used. CM indicates corpus mixture, KW indicates keyword addition, and KP indicates key phrase addition. Word Cor. is "percent correct" and Word Acc. is "word accuracy" with respect to words as units. The character columns indicate the same values with respect to characters as units. In the Japanese language, the divisions between individual words are often ambiguous, and the recognition rates with respect to words as units will differ depending on the type of segmentation that has been conducted. That is why we employ the additional character-based measures. OOV refers to words not found in the vocabulary that we used. No filled-pauses (e.g. "… ah …," "… er …," etc.) taken as an OOV item will be included in the final OOV percentage because of the great variation in the ways that they are notated. PP represents the test set perplexity of the evaluation data. When calculating test set perplexity, we assign 10^{-7} (=1/10M) probability to words which do not exist in the training corpus. For all adaptation, we use λ=1.

Our experiments showed that the language model trained from a large amount of transcripts for the target call center (8.Trans-A(L)) gave the best performance. All adaptation methods (5~7) improved recognition performance. We can see a correlation between recognition performance and such language model properties as OOV and PP.

3.3 Discussion

A language model built only from the documents related to the target-task CCA (1.Doc-A) indicates poor performance even though the training data Doc-A represent the topic of CCA. This degradation was caused from the stylistic differences between document data and the actual dialogue to which the model was applied. For example, filled-pauses are never seen in document data but they appear very frequently in spoken dialogue. (OOV rate of this condition seems relatively small, the reason is that filled-pauses are excluded from OOV calculation as mentioned above.)

A comparison of 2.Trans-C(L) with 3.Trans-B(L) reveals that speech recognition performance was significantly affected by dissimilarity in call center type, but greater than 50% word accuracy was achieved even when the call center type was different. This means that a large part of the words in dialogues were independent of the difference in topics. Not indicated in the Table data is the fact that 16.6% of the words in the target data were either "*hai*," meaning "yes," or filled-pauses, and that function words comprised 39.1% of the total.

A language model using a small amount of transcripts for CCA (4.Trans-A(S)) performed at the same level as one using a large amount of transcripts for a similar call center (3.Trans-B(L)). Although their

recognition rates were roughly the same, their errors had different causes: with 4.Trans-A(S) the problem was general data sparseness, while with 3.Trans-B(L) it was a lack of topic words for CCA. That is why a mixing of the two (5.Trans-B(L)+Trans-A(S)) resulted in improved performance. The improvement was a relative 9% reduction in word error rate over that achieved with 3.Trans-B(L) alone.

Language models built from the data produced by combining CCB transcripts with CCA documents (6, 7.Trans-B(L)+Doc-A) also showed improvement. The language model obtained by keyword addition achieved a relative 10% reduction over that with 3.Trans-B(L) alone. The key phrase addition (7) outperformed the keyword-addition-based model (6). This shows that keyword contexts from written documents can help improve speech recognition performance, and the degradation in performance that is created by stylistic differences is little with the key phrase addition method.

Table 3. Recognition Rates for each Part of Speech

Part of Speech	Number	Word Cor.			
		Trans-B(L)	KW	KP	Trans-A(L)
Noun	7744	53.6%	66.5%	71.1%	70.9%
Verbal noun	2251	61.2%	68.3%	75.0%	75.8%
(Japanese NA-type) adjective	817	58.0%	63.3%	67.8%	69.6%
Numeral	2155	76.8%	75.6%	85.7%	86.8%
Verb	5039	70.6%	70.7%	70.3%	74.3%
Adjective	741	72.2%	72.2%	73.7%	76.4%
Hai	3815	92.2%	92.1%	89.9%	91.2%
Filled pause	7747	54.9%	54.8%	52.9%	57.2%

Recognition rates for each part of speech are indicated in Table 3. The Number column indicates the number of appearances of words for respective parts of speech. These parts of speech may be categorized into three groups on the basis of improvement obtained with adaptation. The first, the top four items in the Table, yielded the most improvement. They are content words which have no conjugation, i.e., they are fixed in forms that are common to both transcripts and documents. Numerals are also distinctive in that, while no improvement was seen when using keyword addition, improvement was seen when using key phrase addition. Numerals are meaningful when they appear in a significant sequence, as, for example, when they appear as a product code number, but the addition of a single numeral without a significant context is meaningless with respect to speech recognition performance. It is notable that, for this first group, recognition performance is equivalent to that of a model created from a large amount of transcripts, and that these are words which are generally meaningful ones for recognition results. With the second group, little or no improvement was achieved by adaptation. These are conjugated words, and they differ between spoken language and written language. At call centers operators use polite forms of speech, and in the Japanese language, this involves marked changes in form and, at times, even replacement with honorific terms. With the last group, adaptation produced negative effects. Since these are only found in spoken dialogue, the addition of document data weakens the effectiveness of their inclusion. Results here indicate the advantage of using keyword addition to reduce the negative effects caused by differences in style.

4 Summary and Future Work

We have introduced here an efficient language model adaptation method and have described an evaluation of it on actual spoken dialogue collected at call centers. Task related documents can be used effectively as adaptation data. We use N-grams containing keywords in order to reduce the negative effects of differences in style with spoken dialogue. Experimental results show that adaptation with documents improves recognition performance. In a comparison of methods for adaptation with documents, our proposed method achieved the best performance. In the future, we intend to conduct more extensive field tests, and we will focus further on keyword selection methods and techniques for revising documents stylistically. It appears that criteria representing semantic features should be imported into keyword selection. Paraphrasing technology may be of help in revising documents into a spoken dialogue style.

References

1. Bellegarda, J.R.: Statistical Language Model Adaptation: Review and Perspectives. Speech Communication, Vol. 42 (2004) 93-108
2. Witschel, P., Höge, H.: Experiments in Adaptation of Language Models for Commercial Applications. Proceedings of Eurospeech'97, Vol. 4 (1997) 1967-1970
3. Bulyko, I., Ostendorf, M., Stolcke, A.: Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures. Proceedings of HLT-NAACL (2003) 7-9
4. Sarikaya, R., Gravano, A., Gao, Y.: Rapid Language Model Development using External Resources for New Spoken Dialogue Domains. Proceedings of ICASSP 2005, Vol. I (2005) I-573-576
5. Akbacak, M., Gao, Y., Gu, L., Kuo, H-K.J.: Rapid Transition to New Spoken Dialogue Domains: Language Model Training Using Knowledge from Previous Domain Applications and Web Text Resources. Proceedings of INTERSPEECH 2005 (2005) 1873-1876
6. Hakkani-Tür, D., Rahim, M.: Bootstrapping Language Models for Spoken Dialogue Systems from the World Wide Web. Proceedings of ICASSP 2006, Vol. I (2006) 1065-1068
7. Akita, Y., Kawahara, T.: Efficient Estimation of Language Model Statistics of Spontaneous Speech via Statistical Transformation Model. Proceedings of ICASSP 2006, Vol. I (2006) I-1049-1052
8. Ishikawa, S., Ikeda, T., Miki, K., Adachi, F., Isotani, R., Iso, K., Okumura, A.: Speech-activated Text Retrieval System for Multimodal Cellular Phones. Proceedings of ICASSP 2004, Vol. I (2004) I-453-456