2008－SLP－72 (6)
2008／7／18

Spoken Document Retrieval by Using a Hierarchical Language Model

Xinhui HU† Youzheng WU† and Hideki KASHIOKA†

† NiCT/ATR 2-2-2 Hikaridai, Seikacho, Souraku-gun, Kyoto, 619-0223 Japan
E-mail: † {xinhui.hu, youzheng.wu, hideki.kashioka}@nict.go.jp

Abstract We propose a new scheme for searching keywords in a speech document. A hierarchical language model is introduced for the recognition and indexing phases. This language model comprises two independently trained layers, an upper layer comprising a conventional class-based n-gram for recognizing in-vocabulary (IV) words, and a lower layer comprising a sub-word n-gram for recognizing two types of out-of-vocabulary words (OOV), Japanese personal names and locations. The recognized sub-word sequences in a lattice are pruned with a finite state automata (FSA). By using the recognized subwords and their position information within the OOV words, the subwords are linked to form potential OOV words. A confusion network is adopted for the indexing phase; the network is built by using the IV words and the pruned OOV words in the lattices. Evaluations on the keyword search, including IV words and OOV words (both personal and location names), are conducted. The experimental results show that the hierarchical language model has a considerable high ability to identify the above proper names that are not registered in recognition lexicon, whereas that for IV words is not significantly improved.

Keyword Out-of-Vocabulary Search, Spoken Document, Hierarchical Language Model

階層化言語モデルによる音声ドキュメントの検索

胡 新輝† 呉 友政† 柏岡 秀紀†

† NiCT/ATR 〒619-0223 京都府相楽郡精華町光台 2-2-2
E-mail: † {xinhui.hu, youzheng.wu, hideki.kashioka@nict.go.jp

あらまし 本稿では、音声ドキュメントに対するキーワード検索について、階層的言語モデルを用いたラティス構造をインデックスに利用する検索手法を提案する。階層的言語モデルは、音声ドキュメントの認識、および検索のインデクシングにおいて利用し、階層は、二階層から構成されている。上位の階層は、既知の単語からなる従来のクラス n-gram の階層であり、下位の階層は、日本人名および地名の、未知語を処理するためのサブワード n-gram の階層である。認識結果として扱うラティスには、未知語のタイプやその候補の位置情報を利用し、FSA により枝狩りされたサブワード列から構成される未知語が含まれる。このようなラティスを簡潔に表現する一つである confusion network（CN）を用いたインデクシング処理を行うことにより、既知語、未知語を同時に処理できる手法となっている。検索の評価は、既知語および未知語を含むものとして評価した。実験結果として、この階層的モデルが認識辞書に未登録の日本人の人名や地名を効果的に検出した。また、既知語の効率は、従来のクラス n-gram 手法に比較して、ほとんど変化しておらず、悪影響はみられなかった。

キーワード 未知語検索、音声ドキュメント, 階層言語モデル

1. Introduction

In content-based speech retrievals, large vocabulary continuous speech recognition (LVCSR) is generally necessary to build speech transcription which is used for indexing and retrieving. Although significant improvement has being achieved in LVCSR in recent decades, two factors continue to limit its application in spoken document retrieval (SDR). One is the problem of out-of-vocabulary (OOV) words, even for clean speech, and another is the problem of speech recognition errors during the transcription of words, particularly in the case of spontaneous speech. Almost all automatic speech recognition (ASR) systems have a closed vocabulary system. Words that are not included in the closed vocabulary system will not be recognized by the ASR system, thereby causing recognition errors. Personal names and locations are the two main resources of out-of-vocabulary (OOV) words, since we cannot add all such names into the recognition lexicon. On the other hand, however, for information retrieval, personal names and locations are often the most interesting targets of retrieval tasks. Speech recognition error is a fatal weakness in spoken document retrieval. In contrast to text retrieval, the

occurrence of errors in spoken document retrieval can be regarded as a peculiar characteristic with it. When a word is misrecognized as another word (referring to substitution) or when it disappears from the recognized text (referring to deletion), it cannot be retrieved.

To resolve the problem of OOV words, the representative method is to introduce smaller processing unit than word into the indexing and retrieval phases, such as phoneme, syllable, or morpheme-like unit [1, 2, 3, 4]. To overcome the problem of recognition error, there are also several methods being proposed. In [5], multiple LVCSR models are employed to improve the recognition and retrieval performance. However, at the present stage, the lattice-based approaches may be the most attractive ones for avoiding speech recognition errors [6], such as the position specific posterior lattices (PSPL) [7], and confusion network [8, 9].

This study aims at applying the keyword search paradigm of text documents to speech documents, particularly with focus on spontaneous speech. The process of keyword search is an indispensible procedure for any text oriented information retrieval (IR) system. For example, in Questions and Answers (QA), a query is generally analyzed and keywords are extracted from it. Then, on the basis of the extracted keywords, the final answer is retrieved. Hence, it is important to obtain stable and reliable keyword searches.

Here, we will be engaged in studying keyword search for in-vocabulary (IV) words and OOV words. Two types of proper nouns including Japanese personal names and location names (JPL) are regarded as OOV words. Furthermore, the personal names are divided into Japanese family name (JSE), Japanese given name (JNA).

The rest of this paper is organized as follows. In section 2, we introduce the hierarchical language model that is used for speech recognition, including training and evaluating speech recognition performance. Further, we present methods for pruning lattices and combining subword strings within lattices to form logical units of names; these logical units are formed by using the syntactical structure of the three OOV classes — JSE, JNA, and JPL. In section 3, we explain the method of applying the vector space model to the confusion network that is obtained from the recognition lattice where both IV words and logical OOV units are contained. In section 4, we will provide a description on the selection of keywords that are used for the experiments. The experimental setup and results are provided in section 5. Finally, we present our conclusions in section 6.

2. Hierarchical Language Model for Lattice Building and Alignment Processes

2.1 Hierarchical language model for constructing the recognition lattice

The hierarchical language model used for speech recognition is a class-based n-gram model [10], as shown in Figure 1; the upper layer generates IV words and classes of OOV words. The lower layer generates the *mora*, a morpheme-like sequence of subwords of the corresponding classes of OOV words.

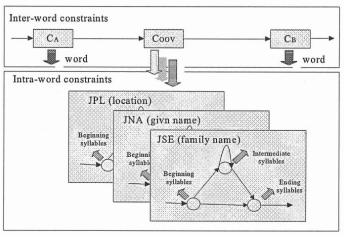

Figure.1 Configuration of the Hierarchical Language Model.

The upper and lower layers are trained independently by using different language corpora. The upper layer of the model comprises an interclass n-gram, it is trained in the same manner as that used in a conventional class-based model; however, classes for the target proper names should be defined in this step. For comparison with the proposed hierarchical language model, we take the upper layer model alone as a ***baseline model***, in which no subword model is combined, and all proper nouns in lexicon are remained as they were.

The lower layer consists of a set of intra-class models. All these subword models are trained independently by using their corresponding name lists. For each class, "B," "I," "E," and "BE," which denote the beginning syllables, intermediate syllables, ending syllables, and a single word,

respectively, are used to tag the position of the subword inside its name class. When all subword's model are trained, hey are combined to the upper layer where corresponding OOV classes are found. During the combination, all personal and location names are removed from the original upper layer model.

The output of the speech recognition carried out by using this hierarchical language model is of the form "w1 w2 p1_JSE_B p2_JSE_I p3_JSE_E w3 w4..." Here, w1, w2, etc., refer to the recognized IV words, and p1, p2, p3, etc., refer to the *mora* (phonemes)of the recognized subwords; JSE_B, JSE_I, and JSE_E refer to the beginning, intermediate, and ending syllables, respectively, of a Japanese family name.

For the training data, the baseline language model is trained by using the manual transcript of the corpus of spontaneous Japanese (CSJ) [11]. It contains approximately seven million words. During language model training and lexicon construction, the fillers and hesitations among the transcripts are not taken into account. Finally, a lexicon with a vocabulary of 6.2K words is obtained. The subword models are bigram models, whose family name, given name, and location are trained by using corpora of 300K, 295K, and 80K words, respectively.

30 test sets of speech data provided by the CSJ corpus are evaluated for the recognition performance with the above language models. The average word accuracy rate using the hierarchical language model is 60.37%, meanwhile, the average accuracy using the baseline model is 61.90%. Here, the acoustic model is in the travel domain.

2.2 Pruning the lattice and combining the subwords

Since a common word is generally composed of several subwords, the length of subwords is smaller than that of IV words. So the number of nodes in the lattice will increase when the subwords are output. Therefore, the total number of paths of possible recognized candidates will expand rapidly. We will employ the confusion network for retrieving information. The confusion network is a compact lattice that aligns confusion sets (words or subwords) at a given time. If the confusion network containing both words and subwords is simply converted from the original lattice of the recognizer, the alignment of the confusion set will be disturbed, a long IV word will be aligned with an individual subword, and many subwords will be aligned by an interval of subwords; this means that subwords are aligned with other subwords and even a long IV word will be aligned to a short subword. Figure 2 shows an example of a confusion network that is obtained from the recognition lattice for the word "Hiroshima." The network shown in Figure 2(a) is obtained directly from the original lattice of the recognizer. In such structures, a considerable amount of noise is added to the post-processing stage.

In order to construct a confusion network in which confusion sets having similar pronunciations are aligned together, the lattice of the recognizer is pruned and subwords are combined before constructing the confusion network. The steps involved in this procedure are as follows:

(1) Use finite state automata (FSA) to prune unreasonable subword's candidate transitions in the result lattice.

(2) Combine subwords having the same class (JFN, JNA, and JPL). Since the positions of the subwords (corresponding to "B", "I", "E", or "BE" tags mention before) of these classes are tagged inside the classes, it is reasonable to combine them into a pseudo-word.

Finally, a pruned and combined graph like in Figure 2 (b) is obtained for the Figure 2(a).

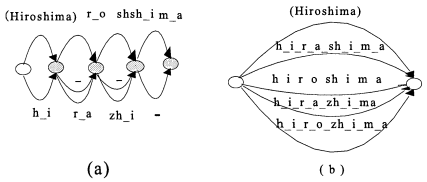

Figure. 2 Examples of confusion networks with different units of a confusion set for the recognition of the word "Hiroshima."

3. Appling the Vector Space Model to the Confusion Network

The retrieving process is based on confusion network, to which the vector space model is applied. Follows brief the calculations of parameters of the retrieval model.

3.1 Indexing and ranking of confusion networks

Let D represent a document modeled by a confusion network. We define $Pr(w|o,D)$ as the posterior probability of a term w at position o in D in order to refer to the occurrence of w in the network. The term frequency tf is the number of times a term occurs in a document. In the proposed method, tf is evaluated by summing the posterior probabilities of all the occurrences of the term in the confusion network. It is calculated by the following equation:

$$tf(w,D) = \sum_{i=1}^{|occ(w,D)|} P(w \mid o_i, D) \qquad (1)$$

Here, the frequency of a term w is the number of times w occurs in D. $|occ(w,D)|$ denotes the segment number of the confusion network in D, which contains w.

The inverse document frequency idf indicates the relative importance of a term in the corpus. We also calculate it on the basis of the posterior probability of w, as shown in the following equation:

$$idf(w) = \log(N / \sum_{D \in C} O(w,D)) \qquad (2)$$

Here,

$$O(w,D) = \begin{cases} 1 & if \sum_{i=1}^{occ(w,D)} P(w \mid o,D) \geq 0.5 \\ 0 & otherwise \end{cases} \qquad (3)$$

implies that if the summation of w's posterior probability is greater than a threshold (0.5), it is considered to have occurred in this document. N denotes the total number of documents contained in all the corpora. C denotes the number of documents in a single corpus.

4. Keyword Selection

2 types of test sets are used for the evaluation of the retrieval. The first set is the IV word set (denoted as IV), which is extracted from the manual transcript of target speech documents. Because the transcripts are used to construct the language model of the recognizer, all the words are contained in the recognition lexicon when the baseline model is used. The second set (denoted as OOV) is a name list containing personal names and locations, which is used for evaluating the OOV word retrieval. The names are extracted from the same documents, but the hierarchical language model is built after all personal names and location name are removed; hence, these names do not appear in the speech recognition lexicon when the hierarchical language model is used. Following semi-automatic procedure is used for selecting these keywords.

4.1 TF/IDF-based method for selecting keywords

The keywords for retrieval are selected on the basis of the term frequency (TF) and inverted document frequency (IDF).

(1) For each word within a document, compute its TF.
(2) Compute IDF of a word among the whole corpus.
(3) Compute TF × IDF of a word within a document, select the highest 100 words as the keywords of the document.
(4) Obtain the keyword collection of the whole corpus, and define it as the keyword database.

4.2 IV Queries and OOV Queries

After the keyword database is obtained, we then build queries of IV keywords and OOV keywords. For building IV queries, keyword groups containing 1, 2, 3, 4 words are picked up. The keyword groups that appear in less than 4 documents are removed from the query candidates. Then, 40 queries containing different keyword counts are manually selected as the IV queries.

The building of the OOV queries is similar to that of the IV queries, but only the proper nouns are extracted from the keyword database. 40 OOV queries, including 18 single names (family name, given name, location), 17 full personal names (family name + given name), and 5 other combinations, are finally built. Examples of these 2 types of keyword queries are shown in appendix A.1 and A.2.

5. Experiments and Result

5.1 Experimental setups

For the retrieval experiments, 3244 files of the CSJ speech data [11] are used. Each file contains speech data worth approximately 10–15 min. Three types of speeches—academic presentations, simulated public speeches, and read-aloud

speeches—are presented in these data sets. The manual transcripts of the documents are also provided with the corpus.

We use standard information retrieval (IR) measures provided by the trec_eval [12] program: mean average precision (MAP) and precision after R (R-P), where R denotes the number of relevant documents.

5.2 Results

There are two objectives of the experiments on keyword retrievals.

The first objective is to investigate the effectiveness of the hierarchical model. It is desirable to improve the searching of OOV words while ensuring that the IV words are not significantly influenced. These two desirable outcomes often conflict with each other.

The second objective is to investigate the effectiveness of the confusion network. For this purpose, the retrieval effectiveness of the following indexes are compared: (1) manual transcript text (Trans.), (2) 1-best recognition result, and (3) confusion network (CN) output.

The statistics of the retrieval performance of the above indexes are shown in Table 1 for the IV queries, in Table 2 for the OOV queries.

Table 1 Mean average precision (MAP) and recall precision (R-P) for IV Queries.

LM		1-best	CN	Trans.
Base.	MAP	0.6181	0.6552	0.7775
	R-P	0.5648	0.5800	0.7121
Hierar chical.	MAP	0.6498	0.6524	-
	R-P	0.5814	0.5863	-

Table 2 Mean average precision (MAP) and recall precision (R-P) for the OOV Queries.

LM		1-best	CN	Trans.
Base.	MAP	0	0	-
	R-P	0	0	-
Hierar chical.	MAP	0.2803	0.3455	-
	R-P	0.3069	0.3345	-

6. Conclusion and Discussion

We have presented a method for searching OOV words in SDR by using a hierarchical language model. The subword models corresponding to OOV words are trained independent of the baseline model, and they can be trained by using a group of corpora when necessary. For example, if we require to process organization names, we can train a new subword model when an organization name list is prepared, and combine it with the baseline model. Therefore, this implementation method is convenient for expanding new type of OOV words.

In this study, we have successfully retrieved 2 types of proper names—Japanese names (including family and given names), and locations. The conclusions of this study can be summarized as follows:

(1) As shown in table 1, for IV keywords retrieval, there is no big difference between the baseline model and the proposed hierarchical model. That means the performance for IV words is not influenced by this hierarchical model so much.

(2) The use of a confusion network for searching OOV words yields a better performance as compared to the 1-best method. As shown in Table 2, the improvements in the MAP and R-P due to the use of the confusion network as compared to 1-best are 23.26% and 8.90%, respectively.

The above results reveal that the hierarchical model is effective for searching OOV words, and lattice-based processing such as that using a confusion network is useful for searching subwords.

However, the overall performance of the SDR remains low. One reason for the low performance is the poor speech recognition. Especially for subword words, their recognition accuracy is much more influenced by the performance of recognizer than common words. We have conducted recognition experiments on different test sets, and found large difference exists among these data. The OOV word recognition accuracy of a test set belonging to travel domain is as high as 80%, but it is only 40% for the test set which is extracted from the CSJ corpus, although the IV word's accuracy is in the same scale. So it is important to improve recognition accuracy for subword's retrieval. Another reason is that the OOV keyword character sequences, which are used as the query keywords, are only selected by the first candidate of a pronunciation lexicon. In fact, many OOV words have several pronunciations or have different behaves in different context. For example, a given name "Souseki" can be expressed as either " s_o_u_s_e_k_i", or "s_o_o_s_e_k_i."

In the future, in addition to the personal name and location, the organization name will also be added to the group of OOV words for retrieving, and multiple pronunciation of a subword's *mora* will be taken into consideration to expand OOV's queries.

7. Acknowledgments

This study was partly supported by a Grant-in-Aid for Scientific Research on Priority Areas in Japan as a part of Cyber Infrastructure for the Information Explosion Era, under Grant No. 19024074.

References

[1] Kenney, Ng, "Subword-based Approaches for Spoken Document Retrieval," Ph.D. thesis, Massachusetts Institute of Technology.

[2] Ville, T. and Mikko, K., "Using Latent Semantic Indexing for Morph-based Spoken Document Retrieval," ICSLP Proc., pp.341–344 (2006).

[3] Logan, B., Moreno, P. and Deshmukh, O., "Word and Subword Indexing approaches for Reducing the Effects of OOV Queries on Spoken Audio," in Proc. HLT (2002).

[4] Yu, P. and Seide, F., "A Hybrid Word /Phoneme-based Approach for Improved Vocabulary-independent Search in Spontaneous Speech," ICSLP Proc. (2004).

[5] Nishizaki, S.N., "Robust Spoken Document Retrieval Methods for Misrecognition and Out of Vocabulary Keywords," IEICE Transactions, Vol.J86-D-II, No.10, pp.1369–1381 (2003).

[6] Saraclar, M. and Sproat, R., "Lattice-based Search for Spoken Utterance Retrieval," in Proc. of HLT-NAACL, pp.129–136 (2004).

[7] Chelba, C. and Acero A., "Position Specific Posterior Lattices for Indexing Speech," in Proc. Of ACL, (2005).

[8] Turunen, V. and Kurimo, M., "Indexing Confusion Networks for Morph-based Spoken Document Retrieval," ACM SIGIR Proc. (2007).

[9] Hori, T., Hetheringtong, I..L., Hazen, T. J. and Glass, J. R., "Open-vocabulary Spoken Utterance Retrieval Using Confusion Networks," ICASSP Proc. (2007).

[10] Tanigaki, K., Yamamoto, H. and Sagisaka, Y.,"A Hierarchical Language Model Incorporating Class-dependent Word Model for OOV Words Recognition," ICSLP Proc., 123–126 (2000).

[11] Maekawa, K., Koiso, H., Furui, H. and Isahara, H., "Spontaneous Speech Corpus of Japanese," LREC Proc., pp.947–952 (2000)

[12] Garofolo, J. S. , Auzanne, C. G. P. and Voorhees, E. M., "The TREC Spoken Document Retrieval Track: A Success Story," TREC-9 Proc (2000).

Appendix

A.1 Examples of Query Keywords (IV)

#Query ID	Keywords	#Relevant Document
IVQ01	誤答 アイエヌ 単音 未修 エーエヌ	4
IVQ06	コウモリ シーエフテラ 超音波	2
IVQ12	音像 左耳 受聴	7
IVQ13	新婦 媒酌人 披露宴	4
IVQ20	自動詞 他動詞 連用形	3
IVQ26	女声 男声	6
IVQ27	石器 縄文	6
IVQ35	考古学	10
IVQ36	最年少	3
IVQ39	社会民主党	1

A.2 Examples of Query Keywords (OOV)

#Query ID	Keywords	#Relevant Document
OOVQ01	夏目 漱石	11
OOVQ02	徳川 家康	10
OOVQ03	長嶋 茂雄	8
OOVQ09	福沢 諭吉	6
OOVQ19	三軒茶屋 二子玉川	2
OOVQ20	岐阜 佐倉	2
OOVQ30	橋本	20
OOVQ31	江の島	8
OOVQ38	石井	13
OOVQ40	藤原	12