AN EXTREMELY LARGE VOCABULARY APPROACH TO NAMED ENTITY EXTRACTION FROM SPEECH

Takaaki Hori and Atsushi Nakamura

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{hori,ats}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper describes an approach to Named Entity (NE) extraction from speech data, in which an extremely large vocabulary lexicon including all NEs occurring in a large text corpus is used for Automatic Speech Recognition (ASR). Accordingly, NEs appear in the recognition results just as they are. Our approach is implemented by the following steps: (1) run an NE-tagger for a whole text corpus and make an NE-tagged corpus in which each NE is padded with its category, (2) construct a lexicon and a language model for ASR using the tagged corpus where each NE is considered as a regular word, and (3) run the speech recognizer in one pass. Although a very large vocabulary is necessary to ensure a high coverage of NEs, that is no longer a major problem since we recently achieved real-time extremely large vocabulary ASR using a WFST framework. In experiments on NE extraction from spoken queries for an open-domain question-answering system, our approach yielded higher F-measure values than a conventional approach.

1. INTRODUCTION

Named Entity (NE) extraction is a basic technology for Natural Language Processing (NLP) applications such as information retrieval, question answering, summarization, etc. NEs are, for example, proper names (of persons, locations, organizations, titles, etc.), temporal entities (of date, time of day), and numerical expressions (of length, weight, amount, etc). The purpose of NE extraction is to identify NE expressions (word or word sequence) and their categories (person, location, etc.) included in text documents. The extracted NEs are used for indexing, classifying, and understanding the documents.

In the past decade, speech applications have been evolving by combining Automatic Speech Recognition (ASR) with NLP. Some NLP technologies are extended to deal with speech inputs, and used in speech applications such as spoken document retrieval, spoken interactive question answering [1], speech summarization [2], and so on. NE extraction can also be a key component of such applications.

In a typical approach to the NE extraction from speech, we firstly convert speech into text by ASR, and then extract NEs by analyzing the text with an NE-tagger. This cascade approach, however, has a problem due to speech recognition errors which are unavoidable even though we employ current state-of-the-art technology. Namely, common NE extraction techniques that only assume text inputs do not take into account the recognition errors, and the analysis fails at the spot of errors. To mitigate the impact of errors, there are some approaches which attempt to recompose the lost NEs from multiple ASR hypotheses represented in an N-best list or a word lattice [8][10]. It is, however, not easy to accurately compose an NE from parts scattered in separate hypotheses, and, in the worst case, additional errors might be produced in the process of recomposition. Furthermore, there is concern that the cascade of an ASR system and an NE-tagger, each of which is not error-free, might include a structural defect in the sense that the NE-tagger may expand errors propagated from the ASR system.

We take a different approach from the above ones, in which the NE-tagger is applied to analysis of a reliable text corpus containing various NEs. And, we utilize the NE-tagged corpus for lexicon and language model construction. One advantage in this approach is that the NE-tagger can perform and contribute at its full potential since the inputs from the text corpus are almost error-free. It rarely or never encounters an impracticable input differently from the case that it is placed on the back-end of ASR. Another advantage is that the statistics concerning NEs is directly reflected on lexicon and language model construction. An NE can be composed by concatenation of already-lexicalized words. Explicit lexical entries for such NEs and N-gram probabilities considering NEs or their classes impose strong constraints on a word sequence conforming to an NE, and this can improve the recognition accuracy of NEs. Note that it is infeasible to acquire such a knowledge only through counting word N-gram frequencies from a non-NE-tagged corpus. A remaining problem in this approach is that it inevitably enlarges vocabulary size. In our investigation, the total number of different words and NEs appearing in newspaper articles of recent 12 years

amounted to about 1.8 million. It has been infeasible to deal with such a huge vocabulary in conventional ASR systems because of enormous processing time and memory usage. Our recent work has, however, achieved a real-time ASR that can deal with an extremely large vocabulary of nearly 2 million words in the framework of Weighted Finite-State Transducers (WFSTs) [4]. This eliminates the problem of huge vocabulary, and makes our approach a practical solution for spoken NE extraction.

Our approach is implemented by the following steps: (1) run an NE-tagger for a whole text corpus and make an NE-tagged corpus in which each NE is padded with its category, (2) construct a lexicon and a language model for ASR using the tagged corpus where each NE is considered as a regular word, and (3) run the speech recognizer in one pass.

In experiments on NE extraction from spoken queries for the open-domain question-answering system, the results demonstrate that our system can recognize NEs more accurately than the ASR system with a standard vocabulary set, and it improves the F-measure compared to NE extraction from 1-best ASR hypotheses.

2. NAMED ENTITY EXTRACTION

The purpose of NE extraction is to automatically identify NEs and their categories in a sentence. Recently, classifier-based NE extraction has become one of the most effective approaches in the field of NLP. Among the classifier-based methods, a classifier that discriminates between NEs and other words is constructed and used to detect NEs. The classifier is previously trained using a corpus manually tagged with NEs. Generally, Hidden Markov Models (HMMs) [5], Maximum Entropy (ME) Models [6], and Support Vector Machines (SVMs) [7], etc. are used as classifiers, and they can statistically be trained using labeled data.

In [7], Japanese NE extraction using SVM-based classifiers is investigated. In this method, for example, the words in "George Herbert Bush said Clinton is ..." are classified as follows:
"President"=OTHER, "George"=PERSON-BEGIN,
"Herbert"=PERSON-MIDDLE, "BUSH"=PERSON-END,
"said"=OTHER, "Clinton"=PERSON-SINGLE,
"is"=OTHER.
In this way, the first word of a person's name is labeled as PERSON-BEGIN, while the last word is labeled as PERSON-END. Other words in the name are PERSON-MIDDLE. If a person's name is expressed by a single word, it is labeled as PERSON-SINGLE, and if a word does not belong to any category, it is labeled as OTHER.

For each word, 15 features were derived by picking up three features (part-of-speech, character type, and the word itself) from each of the current word, the preceding two words, and the succeeding two words. The features are transformed to a feature vector $\boldsymbol{x} = (x[1], \ldots, x[D])$ described as:

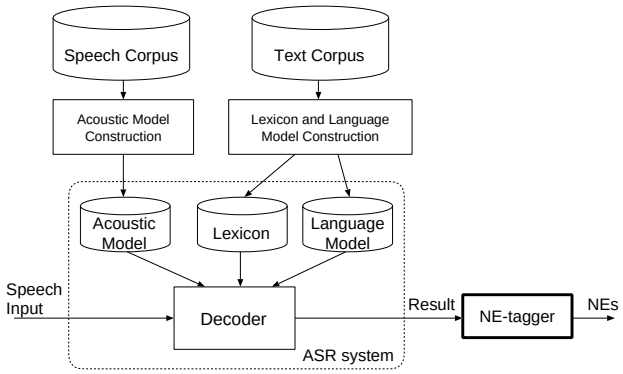

Fig. 1. Baseline spoken NE extraction system

$x[1] = 0$ // Current word is not 'Alice'
$x[2] = 1$ // Current word is 'Bush'
$x[3] = 0$ // Current word is not 'Charlie'
:
$x[15029] = 1$ // Current POS is a proper noun
$x[15030] = 0$ // Current POS is not a verb
:
$x[39181] = 0$ // Previous word is not 'Henry'
$x[39182] = 1$ // Previous word is 'Herbert'

In each vector, only 15 elements are 1, and the other elements are 0.

The SVM classifier can be trained using a sequence of such vectors with labels. Although an SVM can solve only two-class problems, it can be extended to cover n-class problems. In [7], an SVM for each label (PERSON-BEGIN, PERSON-MIDDLE, OTHER, etc.) is prepared, and each SVM decides whether an input vector belongs to the corresponding label or not. Thus, SVMs can be applied to NE extraction by performing one-versus-all-others classification. However, it is necessary to maintain consistency between labels, i.e. for example, an NE of person's name has to start with PERSON-BEGIN or PERSON-SINGLE and end with PERSON-END or PERSON-SINGLE itself. The Viterbi search can be used to ensure totally consistent NE extraction.

3. EXTREMELY LARGE VOCABULARY ASR FOR SPOKEN NE EXTRACTION

A simple implementation of NE extraction from speech is to just combine an ASR system and an NE extraction system in series as shown in Fig. 1. The ASR system features a conventional architecture consisting of an acoustic model, a lexicon, a language model, and a decoder. The models are trained using speech and text data. The NE extraction system receives the word sequence from the ASR system, and extracts NEs from the sequence.

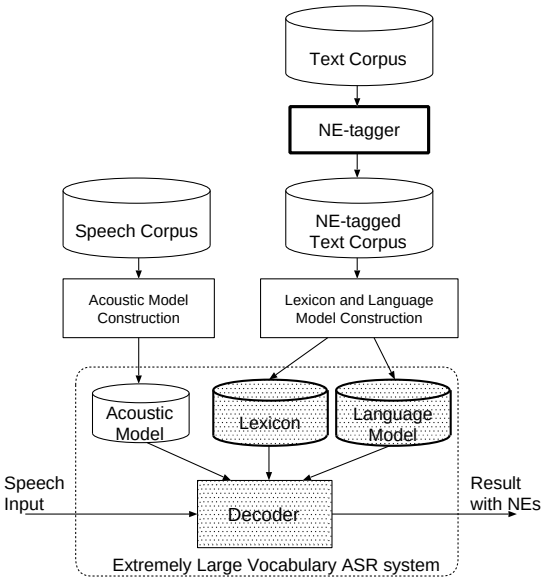

Fig. 2. Extremely large vocabulary spoken NE extraction system

On the other hand, in our proposed method, first NE extraction is applied to all text data for estimating a lexicon and a language model as shown in Fig. 2. In this way, NEs are included in the lexicon and the language model. The extracted NEs appear in recognition results just as they are. In the ASR system, since each NE is considered as a compound word, constraints over several words can effectively work to improve the recognition accuracy of NEs. However, an enormous number of NEs have to be included in the lexicon and the language model to ensure a high coverage. Although it is difficult to execute conventional ASR systems with such a very large vocabulary, our WFST-based ASR system can work with such an extremely large vocabulary [3].

Our ASR system is based on Weighted Finite-State Transducers (WFSTs) [4], utilizing the fast on-the-fly composition algorithm for decoding [3]. This algorithm is different from the original WFST approach in which a simple Viterbi search in a fully-composed single network represented in the WFST form. In the algorithm, two WFSTs are composed during decoding as necessary. One of the two is a WFST to translate speech to a word sequence with unigram probabilities, and the other is a WFST of a trigram model that gives trigram probabilities to the input word sequence. In this algorithm, a Viterbi search is performed based on the former WFST, while the latter WFST is only used to re-score the hypotheses generated during decoding. Since this re-scoring is very efficient, the total amount of computation is almost the same as when using only the former WFST. As a result, the speed of our decoder is faster than that of decoding with the fully-composed WFST. Furthermore, since our algorithm does not construct a fully-composed network, it is also memory-efficient. Consequently, we had achieved one-pass real-time speech recognition in an extremely large vocabulary of 1.8 million words.

4. EXPERIMENTS

We conducted experiments on the task for a Japanese spoken interactive open-domain question-answering (ODQA) system developed at NTT Communication Science Labs [1]. In this task, a user asks the system a question where domain is not restricted, after which the system finds the answer from a large corpus of news texts covering the last 12 years. Since the system cannot know what the user will ask in advance, the speech recognizer has to cover an extensive vocabulary.

A lexicon and a trigram language model for ASR were estimated using newspaper articles of the last 12 years and about 14,000 interrogative sentences. The interrogative sentences were multiplied by 120 in counting the frequencies to make language models. SVM-based named entity extraction, described in Section 2, was performed for all the training data, and the extracted named entities were dealt with as regular words. Each NE is labeled with its category. Thus, even if the word sequences of two NEs are identical, they are different NEs when their categories do not correspond each other. The category set comprises of 45 elements including proper names, date and time expressions, numerical expressions, etc. Consequently, the total number of different words in the data increased from 500,000 to 1.8 million.

Tied-state triphone HMMs with 3,000 states and 16 Gaussians per state were trained by using read speech data uttered by about 300 speakers (approximately 50 hours). The speeches were digitized with 16-kHz sampling and 16-bit quantization. Feature vectors contained 39 elements consisting of 12 MFCCs and a log-energy, their delta and delta-delta components.

The evaluation data consisted of 500 spoken queries for the QA system, which were composed by 25 male and 25 female subjects freely and read by the same subjects. Each subject composed and uttered 100 sentences. The ratio of Out-Of-Vocabulary (OOV) words including OOV NEs was 0.4%, and the test-set perplexity by the 1.8 million-word vocabulary trigram was 98.1. The extracted NEs are supposed to be used as keywords to retrieve the relevant news text.

First, we investigated the effect of incorporating NEs directly into the lexicon for ASR. Table 1 shows recognition accuracy in cases of a normal lexicon (vocabulary size: 500K) and an extremely large lexicon with NEs (vocabulary size: 1.8M). Since the two ASR systems are based on different definitions of word units, it is difficult to compare the performance of the two systems by word accuracy. Thus, we used Japanese character accuracy instead of that. The character accuracy is computed as well as word accuracy by counting errors character by character. By incorporating NEs in the lexicon, 14% of the relative errors were removed.

Table 1. Speech recognition accuracy[%]

	Baseline 500K	Word+NE 1.8M
Character Accuracy	89.0	90.5
(Word Accuracy)	(86.4)	(86.5)

Table 2. F-measure of NE extraction[%]

	baseline	proposed	manual transcription
NE surface	77.9	79.9	92.5
NE+Category	75.4	77.8	89.2

Next, we evaluated the performance of NE extraction. Table 2 shows F-measure values in cases of the baseline system in Fig. 1 and the proposed system in Fig. 2. The 500K-word lexicon was used for ASR in the baseline system, while the 1.8M-word lexicon was used in the proposed system. Here, "NE surface" indicates the case we assume to be correct extraction when a named entity was correctly detected, and "NE+Category" represents the case we assume to be correct extraction only when both a named entity and its category were correctly recognized. In every case shown in the table, the proposed method outperformed the baseline method. These results demonstrate the effectiveness of our approach using an extremely large vocabulary including NEs.

5. CONCLUSIONS

In this paper, we proposed an approach to Named Entity (NE) extraction from speech data, in which an extremely large vocabulary lexicon including all NEs occurring in a large text corpus is used for Automatic Speech Recognition (ASR). Although a very large vocabulary is necessary to obtain a high coverage of NEs, our ASR system can work fast enough despite the large vocabulary of 1.8 million words. In experiments on NE extraction from spoken queries for the spoken interactive open-domain question-answering system, the results demonstrated that our proposed method significantly improved recognition accuracy for NEs compared to the baseline ASR system. Our method also improved the F-measure. Although this method is basically unable to extract Out-Of-Vocabulary (OOV) NEs, the ratio of OOV NEs can be made small enough by using an extremely large vocabulary. In the future, we need to compare our method to other state-of-the-art spoken NE extraction methods using multiple hypotheses such as word lattices from an ASR system.

6. ACKNOWLEDGEMENT

We thank Dr. Hideki Isozaki at NTT Communication Science Labs. for providing an SVM-based named entity tagger, and also thank Dr. Chiori Hori at Carnegie Mellon University for suggesting the application of our ASR system to named entity extraction from speech.

7. REFERENCES

[1] C. Hori, T. Hori, H. Isozaki, E. Maeda, S. Katagiri, S. Furui, "Deriving disambiguous queries in a spoken Interactive ODQA system," in *Proc. of ICASSP2003*, Vol. I, pp. 624–627, 2003.

[2] C. Hori, "A study on statistical methods for automatic speech summarization," Doctoral dissertation, Tokyo Institute of Technology, 2002.

[3] T. Hori, C. Hori, Y. Minami, "Fast on-the-fly composition for weighted finite-state transducers in 1.8 million-word vocabulary continuous-speech recognition," in *Proc. of Interspeech 2004–ICSLP*, Vol. I, pp. 289–292, 2004.

[4] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, Vol. 16, pp. 69–88, 2002.

[5] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, "Named entity extraction from speech," in *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[6] A. Kobayashi, F. J. Och, and H. Ney, "Named entity extraction from Japanese broadcast news," in *Proc. of Interspeech 2003–Eurospeech*, pp. 1125–1128, 2003.

[7] H. Isozaki and H. Kazawa, "Efficient Support Vector Classifiers for Named Entity Recognition," in *Proc. of COLING-2002*, pp. 390–396, 2002.

[8] J. Horlock and S. King, "Named entity extraction from word lattices," in *Proc. of Interspeech 2003–Eurospeech*, pp. 1265–1268, 2003.

[9] M. Surdeanu, J. Turmo, and E. Comelles, "Named entity recognition from spontaneous open-domain speech," in *Proc. of Interspeech 2005–Eurospeech*, pp. 3433–3436, 2005.

[10] B. Favre, F. Bechet, and P. Nocera, "Mining broadcast news data: robust information extraction from word lattices," in *Proc. of Interspeech 2005–Eurospeech*, pp. 601–604, 2005.