Journal of Information Processing Vol. 17 72–81 (Feb. 2009)

Regular Paper

Named Entity Recognition from Speech Using Discriminative Models and Speech Recognition Confidence

Katsuhito Sudoh,[†1] Hajime Tsukada[†1] and Hideki Isozaki[†1]

This paper proposes a discriminative named entity recognition (NER) method from automatic speech recognition (ASR) results. The proposed method uses the confidence of the ASR result as a feature that represents whether each word has been correctly recognized. Consequently, it provides robust NER for the noisy input caused by ASR errors. The NER model is trained using ASR results and reference transcriptions with named entity (NE) annotation. Experimental results using support vector machines (SVMs) and speech data from Japanese newspaper articles show that the proposed method outperformed a simple application of text-based NER to the ASR results, especially in terms of improving precision.

1. Introduction

Nowadays, information is accessible worldwide over the internet. Most information on the internet is written in text, so such text data are used as information sources by natural language processing (NLP) applications like information retrieval, information extraction, and summarization. In addition to text data, more and more audio/video data are becoming available as network bandwidth becomes wider. Audio/video data usually contain spoken language information, and they are also important information sources for NLP applications, which have been attracting much interest. Typical examples of this approach include DARPA's global autonomous language exploitation (GALE) program.

Among the types of language information, named entities (NEs), which include such expressions as people's names and temporal entities (date and time), hold key information in documents and play an important role in NLP applications. For extracting spoken language information from audio/video data, we focus on named entity recognition (NER) from automatic speech recognition (ASR) results. However, the approach of NER from ASR results involves ASR errors, which are caused by out-of-vocabulary (OOV) words and mismatches of acoustic/language models, even with state-of-the-art technologies. Although continuous efforts to improve ASR itself are needed, developing a robust NER for noisy word sequences containing ASR errors is also important.

Most conventional NER methods from speech [1)–5)] use generative models similar to hidden Markov models (HMMs) to work with name categories as states and words as observations. Since such generative models assume that observations are independent of each other, non-independent features are difficult to use. However, in NER tasks, the rich representation of observations using various non-independent features, such as part-of-speech and capitalization, is effective. For this reason, recent studies on text-based NER use discriminative models including maximum entropy (ME) models [6),7)], support vector machines (SVMs) [8)], and conditional random fields (CRFs) [9)] with such non-independent features. Zhai, et al. [10)] applied such a text-based discriminative NER method to ASR results. A problem with applying text-based NER is that ASR errors cause NER errors. Palmer and Ostendorf [2)] addressed this problem by rejecting erroneous ASR word hypotheses based on ASR confidence. However, their NER model is based on a generative model and holds a feature-independence constraint.

In this paper, we extend their approach to discriminative models and propose an NER method that uses ASR confidence as a feature representing whether each word hypothesis is correct. Training data for the NER model are ASR results and reference transcriptions with NE annotation. Experimental results using support vector machines (SVMs) and speech data from Japanese newspaper articles show that the proposed method outperformed a simple application of text-based NER to ASR results.

2. NER with Discriminative Models

2.1 The NER Problem

NER is a task that identifies NEs and their name categories (`PERSON`,

†1 NTT Communication Science Laboratories

LOCATION, TIME...). For example, in the passage "The prime minister of Japan, Yasuo Fukuda...," the word *Japan* is identified as LOCATION NE and the compound word *Yasuo Fukuda* is identified as PERSON NE. This is a kind of chunking problem and can be solved by classifying words into NE classes. Each NE class represents a name category and a chunking state, or the non-NE category (OTHER). In the case of the Start/End method [11], four chunking states are defined: BEGIN (beginning of an NE), MIDDLE (middle of an NE), END (end of an NE), and SINGLE (a single-word NE). In the previous example, the NE class of the word *Yasuo* is PERSON-BEGIN and that of *Fukuda* is PERSON-END.

2.2 SVM-based Method

This paper's research is founded on the SVM-based NER method [8], which shows good performance in Japanese. There are two problems of NER using SVMs: (1) SVMs can only solve binary classification problems; (2) A sequence of most likely NE classes may not be consistent (for example, PERSON-END may follow LOCATION-BEGIN). For problem (1), the method reduces multi-class problems of NER to a group of binary classification problems distinguishing members of a class from members of other classes using N (N = number of NE classes) SVMs. The answer class for a word is defined as the class of the SVM that returns the highest score among all others, because there is not always only one SVM that classifies the word to its "positive" class. Moreover, to overcome problem (2), the method finds the best consistent answer class sequence by a Viterbi search over all answer class sequences. The search is based on probability-like values derived from sigmoid function $s_n(x) = 1/(1 + \exp(-\beta_n x))$ with an SVM output score x.

3. Proposed Method

3.1 ASR Error Problems on NER

Even for text data, NER is not so easy because the NE class of a word differs according to its context, and we have to determine it using limited training data. The NER of ASR results is more difficult because ASR errors occur with both NE constituent words and non-NE words. ASR error problems on the NER of ASR results may involve these cases:

(a) If some of the NE constituent words are misrecognized, the rest may be identified as an incorrect NE.
(b) Falsely inserted words may be identified as incorrect NEs.
(c) Even if all constituent words of an NE have been correctly recognized, ASR errors on its context words may cause NER errors.

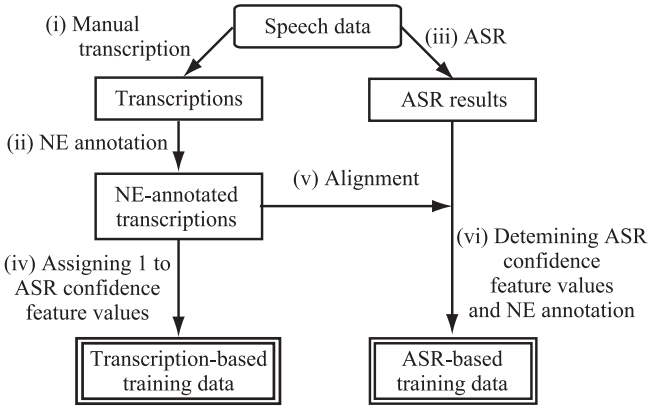

Fig. 1 Procedure for preparing training data.

For robust NER that overcomes these problems, we focus on precisely identifying correct NEs in which all constituent words are correctly recognized by ASR.

Here, we do not aim to correct ASR errors to identify *misrecognized* NEs in which one or more constituent words are misrecognized. Identifying such misrecognized NEs is a different and more difficult problem beyond the scope of this paper, because such a task requires error correction or information recovery from misrecognized words.

3.2 Discriminative NER with ASR Confidence Feature

Identifying misrecognized words helps to solve the ASR error problems described above. In this paper, we incorporate *ASR confidence feature* into a discriminative NER method. The ASR confidence feature is a binary feature of a word, which indicates whether the word has been correctly recognized.

3.2.1 Training

The training data are prepared from speech data in the following procedure illustrated in **Fig. 1**. First, (i) we manually transcribe the speech data and (ii) annotate the transcriptions with NE labels. (iii) The speech data are also automat-

Table 1 Example of transcription-based training data.

Word	Confidence	NE label
Murayama	1	PERSON-BEGIN
Tomiichi	1	PERSON-END
shusho	1	OTHER
wa	1	OTHER
nento	1	DATE-SINGLE

Table 2 Example of ASR-based training data.

Word	Confidence	NE label
Murayama	1	OTHER
shi	0	OTHER
ni	0	OTHER
ichi	0	OTHER
shiyo	0	OTHER
wa	1	OTHER
nento	1	DATE-SINGLE

ically recognized to obtain ASR results. NE-annotated transcriptions, which are used for training the NER model of the proposed method, can be regarded as correct ASR results, so (iv) we make *transcription-based training data* in which the ASR confidence feature value of every transcribed word is 1 (correct). Next, (v) we align the ASR results to the NE-annotated transcription at the word level. (vi) We determine the ASR confidence feature values (i.e., ASR correctness) of ASR word hypotheses and annotate ASR results with NE labels, using the word alignment. We call the annotated ASR results *ASR-based training data*. Note that we ignore the misrecognized NEs in the annotation, based on the focus of this paper described in Section 3.1. In other words, if some of the constituent words of an NE are misrecognized, all of the words constituting the NE, including correctly recognized ones, are labeled as non-NE words (OTHER). Using these transcription- and ASR-based training data, we train SVMs for NER.

Examples of transcription- and ASR-based training data are shown in **Table 1** and **Table 2**. Since the word *Tomiichi* in the person's name *Murayama Tomiichi* is misrecognized by ASR, the correctly recognized word *Murayama* is also labeled OTHER in Table 2. Using these training data, we train the SVMs as described in Section 2.

3.2.2 Testing

In testing, the ASR confidence feature value of each ASR word hypothesis cannot be determined because reference transcriptions are not available. In the proposed method, we determine the feature value based on whether the ASR confidence score of the word hypothesis (described in detail in Section 3.3) exceeds threshold t_w. That procedure works for the rejection of NEs whose constituent words have low confidence in the ASR level. Since our NER model is trained to classify misrecognized word hypotheses into the non-NE class, more word hypotheses are regarded as non-NE words with larger threshold t_w, and therefore more NEs are rejected. t_w controls a trade-off between precision and recall in NER, which differs by application requirements.

3.3 ASR Confidence Scoring

ASR confidence scoring is an important technique in many ASR applications, and various methods have been proposed. Note that the proposed NER method does not depend on a specific ASR confidence scoring method and that any ASR confidence scoring methods are applicable.

A commonly used and effective confidence measure is word posterior probability over a word graph [12]. Word posterior probability $p([w;\tau,t]|\boldsymbol{X})$ of word w at time interval $[\tau,t]$ for speech signal $\boldsymbol{X}$ is calculated as follows [12]:

$$p([w;\tau,t]|\boldsymbol{X}) = \sum_{\boldsymbol{W}\in\boldsymbol{W}[w;\tau,t]} \frac{\left\{p(\boldsymbol{X}|\boldsymbol{W})\,(p(\boldsymbol{W}))^{\beta}\right\}^{\alpha}}{p(\boldsymbol{X})}, \quad (1)$$

where $\boldsymbol{W}$ is a sentence hypothesis, $\boldsymbol{W}[w;\tau,t]$ is the set of sentence hypotheses that include w in $[\tau,t]$, $p(\boldsymbol{X}|\boldsymbol{W})$ is an acoustic model score, $p(\boldsymbol{W})$ is a language model score, α is a scaling parameter ($\alpha<1$), and β is a language model weight. α is used for scaling the large dynamic range of $p(\boldsymbol{X}|\boldsymbol{W})(p(\boldsymbol{W}))^{\beta}$ to avoid a few of the top hypotheses dominating posterior probabilities. $p(\boldsymbol{X})$, which is approximated by the sum over all sentence hypotheses, is denoted as

$$p(\boldsymbol{X}) = \sum_{\boldsymbol{W}} \left\{p(\boldsymbol{X}|\boldsymbol{W})\,(p(\boldsymbol{W}))^{\beta}\right\}^{\alpha}. \quad (2)$$

$p([w;\tau,t]|\boldsymbol{X})$ can be efficiently calculated using a forward-backward algorithm.

Here, in word graphs, two or more word hypotheses of an identical word may appear at overlapping time intervals. The posterior probability of "word ap-

pearance" is distributed over these hypotheses and may become relatively small compared to other competing word hypotheses. Soong, et al.[13] used the sum of word posterior probabilities of those overlapping word hypotheses, called generalized word posterior probability (GWPP). The GWPP of a word hypothesis $[w;\tau,t]$ is denoted as follows:

$$GWPP([w;\tau,t]|\boldsymbol{X}) = \sum_{\substack{[w;\tau',t'] \\ \text{s.t. } [\tau',t']\cap[\tau,t]\neq\emptyset}} p([w;\tau',t']|\boldsymbol{X}). \quad (3)$$

Word posterior probability is a useful confidence measure, and many studies have been conducted to further improve ASR confidence scoring by integrating other features using neural networks[14], linear discriminant analysis[15], and SVMs[16]. In this paper, we use another SVM for ASR confidence scoring to distinguish correctly recognized words from misrecognized words, and we use output scores of the SVM as ASR confidence scores. The SVM is also trained using the ASR-based training data, where the ASR correctness of the word hypotheses is given.

4. Experiments

We conducted the following NER experiments from speech data to investigate the performance of the proposed method compared to the simple application of a text-based NER to ASR results.

4.1 Setup

The experiments were conducted in the following condition, also illustrated in **Fig. 2**.

4.1.1 Speech Data

We simulated the procedure shown in Fig. 1 by using an NE-annotated text corpus and read speech of the text. The NE-annotated text corpus, which is used as training data in the Information Retrieval and Extraction Exercise (IREX) workshop[17], consisted of 1,174 Japanese newspaper articles (10,718 sentences) and 18,610 NEs in eight categories (artifact, organization, location, person, date, time, money, and percent). The number of speakers of the read speech was 106 (about 100 sentences per speaker). The sentences were tokenized into words

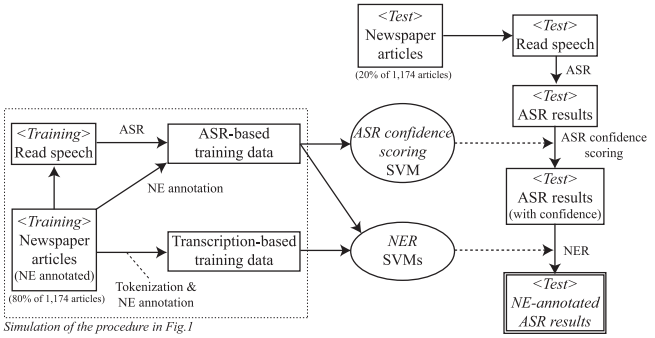

Fig. 2 Experiment condition for the proposed method.

with part-of-speech by the Japanese morphological analyzer ChaSen ⋆1. In tokenization, unreadable tokens such as parentheses and punctuation marks were removed for consistency with speech data. As a result of tokenization, the text corpus had 264,388 words of 60 part-of-speech types.

The experiments were conducted with 5-fold cross validation, using 80% of the 1,174 articles, the ASR results of the corresponding speech data for training SVMs (both for ASR confidence scoring and for NER). The rest of these data were used for the test.

4.1.2 ASR

We used an ASR engine[18] with a speaker-independent acoustic model for read speech. The language model was a word 3-gram model of 426,015 vocabulary words, which was trained using other Japanese newspaper articles (about 340 M words) tokenized by ChaSen. The test-set perplexity over the text corpus was 79.05. The text corpus had 235 (1.26%) NEs that contained OOV words. Scaling parameter α was set to 0.01, which showed the best ASR correct/incorrect classification results using word posterior probabilities in the training set in terms of receiver operator characteristic (ROC) curves. Language model weight β was set to 15, which is a commonly used value in the ASR engine. As a result, word accuracy over the overall speech data was 79.67% (word correct: 85.17%). In the ASR results, 82.00% of the NEs in the text corpus remained; in other words, all

⋆1 http://chasen.naist.jp/hiki/ChaSen/ (in Japanese)

of the constituent words of those NEs were correctly recognized. The rest were lost in the ASR process and could not be recognized by NER.

4.1.3 ASR Confidence Scoring Model

The ASR confidence scoring model (SVM) was trained with soft margin parameter $C = 0.01$ (empirically chosen), using a quadratic kernel $(1 + \vec{x} \cdot \vec{y})^2$ and the following features:

- word surface
- part-of-speech
- GWPP

These features of the two preceeding and two succeeding words were also used. Here, the word surface feature was represented by such binary features as "whether the word surface is *Murayama*," "whether the word surface is *shusho*," and so on. The part-of-speech feature was also represented by binary features, including "whether the part-of-speech is a noun," "whether the part-of-speech is a verb," and so on. In terms of the GWPP feature, we reduced real-valued GWPP p to ten binary features (empirically chosen) as: "whether p was larger than 0 and less than or equal to 0.1," "whether p was larger than 0.1 and less than or equal to 0.2," ... "whether p was larger than 0.9 and less than or equal to 1" for compatibility with other binary features.

Figure 3 shows the ROC curves of ASR correct/incorrect classification using SVM-based ASR confidence scoring, word posterior probabilities, and GWPPs, where

$$\text{Correct acceptance rate} = \frac{\#\text{ correctly recognized words estimated as correct}}{\#\text{ correctly recognized words}},$$

$$\text{False acceptance rate} = \frac{\#\text{ misrecognized words estimated as correct}}{\#\text{ misrecognized words}}.$$

These rates are averaged over five cross-validation test sets. The SVM-based ASR confidence scoring showed better performance in ASR error estimation than GWPPs by integrating multiple features. The features used in the experiment were chosen without a careful feature selection and may not be optimal. However, this paper does not aim to pursue the best ASR confidence scoring performance itself, but to improve NER by the proposed method with the instance of ASR confidence scoring methods. We therefore used these features in the following experiments as an instance of ASR confidence scoring methods.

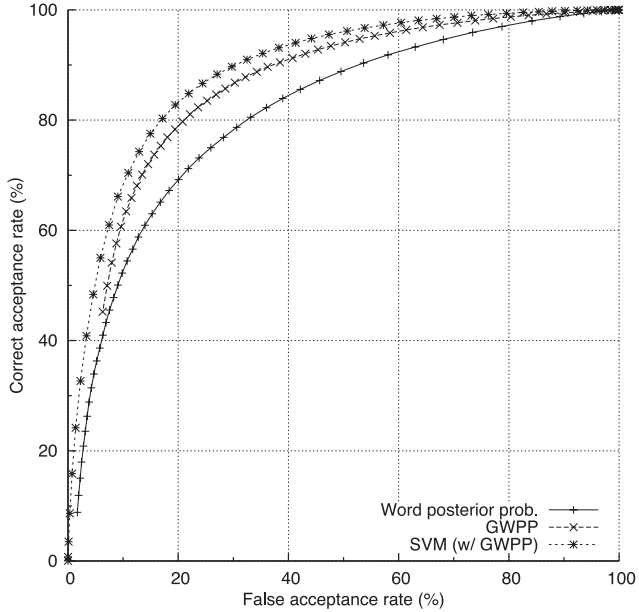

Fig. 3 ROC curves of ASR error estimation with word posterior probabilities, GWPPs, and SVMs.

4.1.4 NER Model

For the SVMs for NER, we followed the settings in Ref. 8) and used soft margin parameter $C = 0.1$, the parameter of sigmoid function $\beta_n = 1.0$, a quadratic kernel $(1 + \vec{x} \cdot \vec{y})^2$, and the following features:

- word surface
- part-of-speech
- character type
- ASR confidence

These features of the two preceeding and two succeeding words were also used. The word surface and part-of-speech features were represented by binary features

identical to those used in the SVM for ASR confidence scoring. The character-type feature reflected the three different kinds of characters used in Japanese. In this paper we define ten character types:

`single-kanji`: written in a single Chinese character
`all-kanji`: written in Chinese characters
`hiragana`: written in *hiragana* Japanese phonograms
`katakana`: written in *katakana* Japanese phonograms (often used for imported words)
`number`: representing a number
`single-capital`: written in a single capitalized roman letter
`all-capital`: written in capitalized roman letters
`capitalized`: only first letter is capitalized
`roman`: other roman letter words
`others`: all other words.

The character-type feature was represented by ten binary features corresponding to the above character types. For the chunking state definition we employed the Start/End method, and thus the number of NE classes, which equaled the number of SVMs, was 33 (eight NE categories * four chunking states + non-NE class).

4.2 Evaluation Metrics

We evaluated the NER methods based on precision, recall, and F-measure denoted as follows.

$$\text{Precision} = \frac{\#\text{ correctly recognized NEs}}{\#\text{ recognized NEs}}$$

$$\text{Recall} = \frac{\#\text{ correctly recognized NEs}}{\#\text{ NEs in text corpus}}$$

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

In evaluations, a recognized NE was accepted as correct when its surface and NE category were identical to those of its reference NE. Here, differences in word segmentation did not matter, due to segmentation ambiguities in Japanese.

4.3 Compared Methods

The difference between the baseline, which is the simple application of a text-based NER to the ASR results, and the proposed method is the use of the ASR confidence feature and the ASR-based training data. Accordingly, we compared several methods with and without these attributes to investigate their effects. We also compared two ASR confidence scoring methods (SVM-based and GWPP) with respect to the ASR confidence feature, to investigate the difference in NER performance with varying ASR confidence scoring performance. The list of compared methods is shown in **Table 3**.

Table 3 Compared methods.

Method	Training data	Confidence scoring
NoConf-T (Baseline)	Trans.	N/A
NoConf-A	ASR	
NoConf-TA	Trans.+ASR	
Reject-T$_{GWPP}$	Trans.	GWPP
Reject-A$_{GWPP}$	ASR	
Reject-TA$_{GWPP}$	Trans.+ASR	
Reject-T	Trans.	SVM
Reject-A	ASR	
Reject-TA	Trans.+ASR	
Conf-A$_{GWPP}$	ASR	GWPP
Conf-TA$_{GWPP}$	Trans.+ASR	
Conf-A	ASR	SVM
Conf-TA (Proposed)	Trans.+ASR	
GM-NoConf-T	Trans.	N/A
GM-Conf-TA	Trans.+ASR	SVM
Conf-A$_{UB}$	ASR	Oracle
Conf-TA$_{UB}$	Trans.+ASR	
(Test data are reference transriptions)		
Trans-T	Trans.	N/A
GM-Trans-T		

4.3.1 NoConf Group

First, we considered methods without the ASR confidence feature. A baseline method, called *NoConf-T*, only used the transcription-based training data. As variations of the baseline method, *NoConf-A* used the ASR-based training data, and *NoConf-TA* used both the transcription and ASR-based training data.

4.3.2 Conf Group

Second, we introduced the ASR confidence feature. The proposed method called *Conf-TA* used both transcription-based and ASR-based training data with the ASR confidence feature. As variations of the proposed method, *Conf-A* only used the ASR-based training data. We also tested the models of Conf-TA and Conf-A using test data with another ASR confidence scoring method that used GWPPs as confidence scores, as *Conf-TA*$_{GWPP}$ and *Conf-A*$_{GWPP}$, respectively. Furthermore, we tested these models using test data with *perfect* ASR confidence scoring where the ASR confidence feature values were oracle, as *Conf-TA*$_{UB}$ and *Conf-A*$_{UB}$. These were for investigating upper-bound performance of the proposed method.

4.3.3 Reject Group

Next, we considered methods with an ASR-level rejection strategy. In the methods of the Reject group, word hypotheses that had ASR confidence scores lower than threshold t_w were rejected and replaced with OOV symbols before the NER process. This "word rejection" strategy explicitly rejected low-confidence word hypotheses, while the proposed method assigned *zero* as their ASR confidence feature values instead. *Reject-T* used the same model as the baseline method (NoConf-T) for the ASR results to which rejection applied. As its variations, *Reject-A* used the ASR-based training data and *Reject-TA* used both transcription- and ASR-based training data. In the ASR-based training data used for Reject-A and Reject-TA, misrecognized words were also rejected and replaced with OOV symbols. We also tested these models using test data with the GWPP confidence scores as *Reject-T*$_{GWPP}$, *Reject-A*$_{GWPP}$, and *Reject-TA*$_{GWPP}$.

4.3.4 GM Group

For comparison with a generative-model-based method, we implemented an NER method with confidence-based ASR error modeling [2] as *GM-Conf-TA*, trained using both the transcription- and ASR-based training data. We conducted experiments with it using our SVM-based confidence scoring. Furthermore, we also applied the generative NER method, trained using only the transcription-based training data, to the ASR results as *GM-NoConf-T*.

Note that our generative model implementation included word surface and part-of-speech information, in the way described in Ref. 19).

Table 4 NER results in F-measure, precision, and recall.

Method	F-measure (%)	Precision (%)	Recall (%)
NoConf-T (Baseline)	67.22	70.80	63.98
NoConf-A	65.65	78.84	56.25
NoConf-TA	67.20	77.74	59.17
Reject-T$_{GWPP}$	67.93	74.73	62.27
Reject-A$_{GWPP}$	67.47	75.62	60.91
Reject-TA$_{GWPP}$	68.61	76.16	62.43
Reject-T	68.06	75.54	61.93
Reject-A	67.98	76.93	60.89
Reject-TA	69.10	77.99	62.03
Conf-A$_{GWPP}$	67.63	75.93	60.97
Conf-TA$_{GWPP}$	68.70	76.60	62.28
Conf-A	67.92	77.62	60.37
Conf-TA (Proposed)	**69.28**	78.63	61.89
GM-NoConf-T	57.08	53.99	60.55
GM-Conf-TA	60.13	63.36	57.21
(Results with perfect ASR confidence scoring)			
Conf-A$_{UB}$	72.07	86.30	61.88
Conf-TA$_{UB}$	73.31	87.97	63.00
(Results with reference transcriptions)			
Trans-T	84.23	86.42	82.16
GM-Trans-T	71.07	67.27	75.34

4.3.5 Trans Group

We applied the baseline model to the transcriptions for reference as *Trans-T*. The generative NER model was also applied to the transcriptions as *GM-Trans-T*.

4.4 NER Results

Table 4 shows the NER results. Here, ASR confidence threshold t_w used in the methods of the Reject and Conf groups was optimized to maximize the NER F-measure by a 2-fold cross validation. In terms of the effect of t_w, **Figure 4** shows the precisions and recalls by those methods with varying t_w.

As shown in Table 4, the proposed method achieved the best F-measure, 69.28%, among the compared methods. It was 2.0% better than the baseline

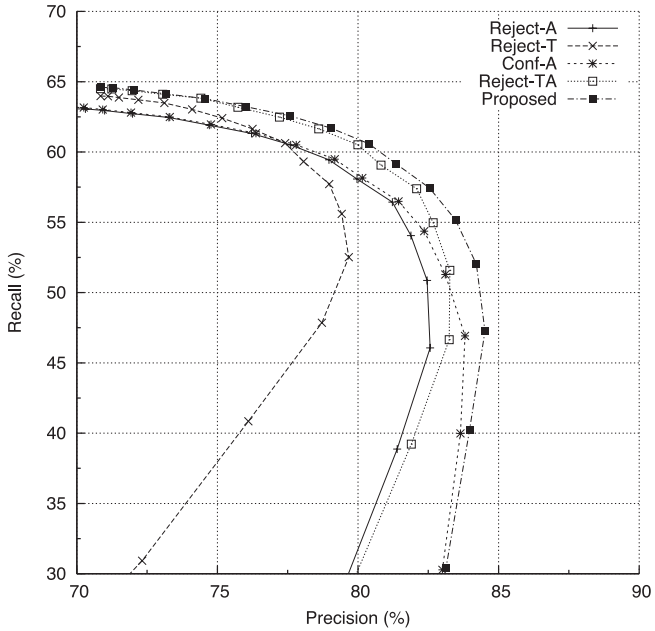

Fig. 4 NER precision and recall with varying ASR confidence threshold t_w.

result (67.22%), with an 7.8% increase in precision (70.80% to 78.63%) and a 2.1% decrease in recall (63.98% to 61.89%). The advantage of the proposed method over the closest result (69.10% by Reject-TA) is shown more clearly in Fig. 4. The proposed method consistently outperformed Reject-TA, especially in higher precision ranges. Therefore, the proposed method efficiently rejects erroneous NEs with a small loss and enables more accurate information extraction from speech.

The results in Table 4 also show:

- Using both training data was better than using either one, while using only ASR-based training data resulted in a large decrease in recall with an increase in precision.
- From a comparison among oracle, SVM- and GWPP-based ASR confidence scoring, better NER results were obtained through better ASR confidence scoring.
- GM group showed about 15% worse results in F-measure than their counterparts of SVM-based NER methods.

5. Discussion

5.1 ASR-based Training Data

The large decrease in recall with the use of the ASR-based training data is considered to be from our NE annotation rule. In the ASR-based training data, all constituent words of misrecognized NEs, even correctly recognized ones, are labeled as non-NE words. This causes NER models to be too *strict*, and therefore not only incomplete NEs but possible NEs are rejected excessively. Using both training data, the decrease in recall becomes smaller and the F-measure improves. This suggests that the use of both training data balances recognition of correct NEs and rejection of incorrect or incomplete NEs.

5.2 ASR Confidence Feature

The advantage of the proposed method over Reject-TA in high precision ranges suggests that the use of the ASR confidence feature is more effective than explicit word rejection. The difference between the two methods appears in how they deal with low-confidence words that are determined as incorrect.

For low-confidence NE constituent words, the words should be labeled as non-NE words by both the proposed method and Reject-TA. This is because misrecognized words are labeled as non-NE words in the ASR-based training data, and there is no difference between the two methods at that point. On the other hand, for low-confidence context words, Reject-TA rejects them and identifies NEs without their features. In contrast, the proposed method was able to utilize their features such as "the next word was *foo*, but it was incorrect." The availability of misrecognized word features is a key advantage of the proposed method, which becomes more effective with larger ASR confidence threshold t_w (i.e., in higher precision ranges).

5.3 NER Based on Generative Models

In general, a generative model requires fine tuning of model structure and parameters to incorporate various features due to the feature-independence constraints. Our implementation of the generative NER method in the experiment

may not be sufficiently tuned; however, SVM-based methods achieved sufficiently higher F-measure results than the generative method with much less tuning effort.

6. Related Work

Some recent studies on NER from speech [3)–5),10),20)] have considered more than 1-best ASR results in the form of N-best lists and word lattices. Using many ASR hypotheses helps recover 1-best ASR errors on NEs and may improve NER accuracy.

Generative NER models were used for multi-pass ASR and NER searches using word lattices [3)–5)]. These studies showed the advantage of using multiple ASR hypotheses, but they still have difficulty in using non-independent features as described in Section 1. Discriminative training of an HMM-like NER model [20)] was also proposed, but that approach did not utilize non-independent effective features.

On the other hand, discriminative NER models were also applied to multiple ASR hypotheses. Zhai, et al. [10)] used ME-based NER with character-related features. The NER method was applied to N-best ASR results and the N-best NER results were merged by weighted voting based on several sentence-level scores such as ASR and NER scores. Their method does not use ASR confidence as a ME model feature, but our ASR confidence feature can be used with theirs and other discriminative models for further improvement.

7. Conclusion

We proposed an NER method from ASR results that incorporates ASR confidence as a feature of discriminative models. The NER model is trained using both transcription- and ASR-based training data. In experiments using SVMs, the proposed method showed a higher NER F-measure, especially in terms of improving precision, than a simple application of text-based NER to ASR results. The method effectively rejects erroneous NEs due to ASR errors with a small drop in recall, due to both the ASR confidence feature and ASR-based training.

Our approach can be applied to other tasks of spoken language processing. It can also be extended to other noisy inputs such as those from optical character recognition (OCR), since confidence itself is not limited to speech. For further improvement, we will consider N-best ASR results or word lattices as inputs and introduce more speech-specific features such as word durations and prosodic features.

Acknowledgments We would like to thank Dr. Takaaki Hori for his help in using SOLON. We also thank the members of Knowledge Processing Research Group for valuable discussion.

References

1) Miller, D., Schwartz, R., Weischedel, R. and Stone, R.: Named Entity Extraction from Broadcast News, *Proc. DARPA Broadcast News Workshop*, pp.37–40 (1999).
2) Palmer, D.D. and Ostendorf, M.: Improving Information Extraction by Modeling Errors in Speech Recognizer Output, *Proc. HLT* (2001).
3) Horlock, J. and King, S.: Named Entity Extraction from Word Lattices, *Proc. EUROSPEECH*, pp.1265–1268 (2003).
4) Béchet, F., Gorin, A.L., Wright, J.H. and Hakkani-Tür, D.: Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How May I Help You?, *Speech Communication*, Vol.42, No.2, pp.207–225 (2004).
5) Favre, B., Béchet, F. and Nocéra, P.: Robust Named Entity extraction from large spoken archives, *Proc. HLT-EMNLP*, pp.491–498 (2005).
6) Borthwick, A.: A Maximum Entropy Approach to Named Entity Recognition, PhD Thesis, New York University (1999).
7) Chieu, H.L. and Ng, H.T.: Named Entity Recognition with a Maximum Entropy Approach, *Proc. CoNLL*, pp.160–163 (2003).
8) Isozaki, H. and Kazawa, H.: Efficient Support Vector Classifiers for Named Entity Recognition, *Proc. COLING*, pp.390–396 (2002).
9) McCallum, A. and Li, W.: Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-enhanced Lexicons, *Proc. CoNLL* (2003).
10) Zhai, L., Fung, P., Schwartz, R., Carpuat, M. and Wu, D.: Using N-best Lists for Named Entity Recognition from Chinese Speech, *Proc. HLT-NAACL*, pp.37–40 (2004).
11) Sekine, S., Grishman, R. and Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proc. 6th Workshop on Very Large Corpora*, pp.171–178 (1998).
12) Wessel, F., Schlüter, R., Macherey, K. and Ney, H.: Confidence Measures for Large Vocabulary Continuous Speech Recognition, *IEEE Trans. Speech and Audio Processing*, Vol.9, No.3, pp.288–298 (2001).

13) Soong, F.K., Lo, W.-K. and Nakamura, S.: Generalized Word Posterior Probability (GWPP) for Measuring Reliability of Recognized Words, *Proc. SWIM* (2004).
14) Schaaf, T. and Kemp, T.: Confidence Measures for Spontaneous Speech Recognition, *Proc. ICASSP*, Vol.II, pp.875–878 (1997).
15) Kamppari, S.O. and Hazen, T.J.: Word and Phone Level Acoustic Confidence Scoring, *Proc. ICASSP* (2000).
16) Zhang, R. and Rudnicky, A.I.: Word Level Confidence Annotation using Combinations of Features, *Proc. EUROSPEECH*, pp.2105–2108 (2001).
17) Sekine, S. and Eriguchi, Y.: Japanese Named Entity Extraction Evaluation—Analysis of Results, *Proc. COLING*, pp.25–30 (2000).
18) Hori, T., Hori, C. and Minami, Y.: Fast on-the-fly composition for weighted finite-state transducers in 1.8 million-word vocabulary continuous-speech recognition, *Proc. ICSLP*, Vol.1, pp.289–292 (2004).
19) Palmer, D.D., Ostendorf, M. and Burger, J.D.: Robust information extraction from automatically generated speech transcriptions, *Speech Communication*, Vol.32, No.1, pp.95–109 (2000).
20) Horlock, J. and King, S.: Discriminative Methods for Improving Named Entity Extraction on Speech Data, *Proc. EUROSPEECH*, pp.2765–2768 (2003).

(Received June 4, 2008)
(Accepted November 5, 2008)
(Released February 4, 2009)

Katsuhito Sudoh is a research scientist of NTT Communication Science Laboratories. His research interests include spoken language processing and machine translation. He is a member of ACL, ASJ, and NLP. He received B.E. in 2000 and M.I. (Master of Informatics degree) in 2002, from Kyoto University.

Hajime Tsukada is a senior research scientist of NTT Communication Science Laboratories. His research interests include statistical machine translation as well as speech and language processing. He is a member of ACL, ASJ, IEICE, and etc. He received B.S. in 1987 and M.S. in 1989, both from Tokyo Institute of Technology in Information Science.

Hideki Isozaki is a senior research scientist, supervisor, and group leader of Communication Science Laboratories, NTT Corporation. His research interests include question answering, information extraction, and information retrieval. He is a member of ACL, IEICE, JSAI, and NLP. He received B.E. (1983), M.E. (1986), and Ph.D. (1998) from the University of Tokyo. He joined NTT in 1986.