Robust Understanding of Spoken Chinese through Character-based Tagging and Prior Knowledge Exploitation

Weiqun Xu, Changchun Bao, Yali Li, Jielin Pan and Yonghong Yan

The Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences
21 Beisihuan West Road, Beijing, China, 100190
{xuweiqun,baochangchun,liyali,panjielin,yanyonghong}@hccl.ioa.ac.cn

***Abstract*—Robustness is one of the most challenging issues for spoken language understanding (SLU). In this paper we studied the semantic understanding of Chinese spoken language for a voice search dialogue system. We first simplified the problem of semantic understanding into a named entity recognition (NER) task, which was further formulated as sequential tagging. We carried out experiments to opt for character over word as the tagging unit. Then two approaches were proposed to exploit prior knowledge – in the form of a domain lexicon – into the character-based tagging framework. One enriched tagger features by incorporating more formal lexical features with a domain lexicon. The other made plain use of domain entities by simply adding them to the training data. Experiment results show that both approaches are effective. The best performance is achieved by combining the above two complimentary approaches. By exploiting prior knowledge we improved the NER performance from 75.27 to 90.24 in F_1 score on a field test set using speech recognizer output.**

I. Introduction

Spoken dialogue systems [1], powered by the ever progressive speech technologies, have evolved from early command and control systems, to call routing and form filling systems, and to the latest voice search systems [2]. With the ever increasing demand for natural interaction, the requirement for robust understanding of spoken input becomes more and more urgent, since the success of those interactive applications relies not only on what is said but also more on what is meant. This has fostered the study of spoken language understanding (SLU), which aims to interpret the signs given by a speech signal in terms of some meaning representation [3].

Although sharing a similar goal to natural (or written) language understanding (NLU), the understanding of spoken languages faces some more challenges. One is the spontaneous speech phenomena or disfluencies abundant in natural interactions, like false starts, hesitations, self-corrections, and filled pauses, etc. This renders a lot of utterances ungrammatical as compared with written sentences for NLU. Furthermore, current spontaneous speech recognizers inevitably bring many errors at a rate much higher than that for read speech or broadcast news recognizers. This makes the recognizer output even worse. So many noises, either from spontaneous phenomena or brought by imperfect recognizers, make robustness standing out as one of the most important and challenging issues for SLU.

In addition, we need to decide if the basic processing unit is chosen to be word or character since we are faced with Chinese SLU (CSLU) . For English it is straightforward to choose word as the basic processing unit because there are natural and agreed boundaries between words in a sentence. But for Chinese it is quite different. There are no natural boundaries between words in a sentence, and worse still, there is no universal agreement on the definition of word boundaries or on word segmentation criteria. If we opt for character, we are losing word level information since the meanings of some words differ a lot from those of their component characters. This is very much so for many proper names or named entities. If we opt for word, a more natural meaning-bearing unit, we may need to pay some price, not only for segmentation, but also for possible noises from segmentation.

The rest of this paper is organized as follows. First we give a brief historical sketch of SLU and highlight the currently dominant statistical approach in section II. Then we introduce the statistical tagging framework for our CSLU system in section III and the target application domain and data in section IV. In section V we take a look at the tagging unit and carry out character-based and word-based tagging experiments. In section VI we further describe how to exploit domain information for CSLU. Before closing, we discuss some related works in section VII.

II. Spoken Language Understanding

The study of SLU began about half a century ago and has always been an active research area. In the 1970's there was the ARPA speech understanding project [4]. In the early 1990's there was the DARPA spoken language system program in spoken language understanding focusing on the Air Travel Information System (ATIS) domain in the 1990's [5]. At the same period there was another European ESPRIT SUNDIAL project [6]. The LUNA project[1] is a recent three-year (2006-2009) EU FP6 IST project dedicated to real-time SLU of spontaneous speech in telephone applications. Over the years

[1]See http://www.ist-luna.eu/.

SLU studies have achieved enormous progresses both in depth and breadth (for a recent comprehensive overview, see [7]).

In the past ten years or so, the statistical framework has become the dominant paradigm in SLU [8], [3]. Various models and approaches have been tried to address the problems of robustness and limited annotated data in SLU. Wang and Acero introduced an HMM/CFG composite model [9] to integrate easy-to-obtain domain knowledge into a data-driven statistical learning framework. A discriminative model based on the model of conditional random fields is further investigated for the same purpose [10]. In the the AT&T HMIHY call routing system, a two-step process is built to detect and extract named entities from spoken utterances through tightly coupling speech recognition and understanding and combining knowledge-based methods and data-driven approaches [11]. Tur et al. further investigated how to use active and semi-supervised learning to reduce the number of labeled training examples by selectively sampling a subset of the unlabeled data and exploiting the unselected ones [12]. In the AT&T VoiceTone dialogue system, the understanding of user intent is carried out by extending the boosting algorithm to incorporating prior knowledge [13]. He and Young proposed a hidden vector state (HVS) model [14] to extend the basic discrete Markov model by expanding each state to encode the stack of a push-down automaton so as to efficiently encode hierarchical context. Jeong and Lee studied a transfer learning approach to multi-domain SLU using a triangular-chain structured model [15]. Dinarelli et al. tried to combine generative and discriminative models for SLU through discriminatively re-ranking a list of ranked hypotheses produced by a generative model [16].

III. The Statistical Tagging Framework

Usually for the meaning of a spoken utterance there are two relatively independent aspects: one is semantic, i.e., entities and their relations, and the other is pragmatic, i.e., speaker's intention. Therefore the task of SLU can be decomposed into two sub-tasks. One is semantic understanding and the other is pragmatic understanding. Here we are concerned with semantic understanding.

Semantic understanding ideally should address the recognition of both entities and their relations. But so far some representative SLU systems mainly addressed entity recognition, e.g., the AT&T call routing system [11] and Voice Tone system [13]. This is mainly due to the following reasons. On the one hand it is difficult to analyze spoken utterances in a full and deep way due to noises like disfluencies and errors from automatic speech recognizer (ASR). On the other hand spoken utterances are relatively simpler than written sentences and entity recognition can satisfy the basic needs for some applications. For pragmatic understanding or intention recognition, dialogue act recognition is widely studied, e.g. [17], [18]. In this paper we work on the recognition of named entities (salient domain entities, a subset of entities) for semantic understanding of Chinese spoken utterances. This is a shallow and partial approach but works well in terms of robustness, as will be seen below.

The task of named entity recognition (NER), a well-studied key task in information extraction, is commonly formulated as sequence tagging [19]. In our work we take the same approach and build on our previous work [20]. The tagging is carried out with a conditional random field (CRF) based tagger.

CRF is a statistical sequence modeling framework introduced in [21], [22]. In the model the probability assigned to a label sequence $\vec{y}$ for a given input sequence $\vec{x}$ is given as:

$$p(\vec{y}|\vec{x}) = \frac{1}{Z(\vec{x})} \exp \sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(\vec{x}, y_{t-1}, y_t)$$

where f_k is a feature function and λ_k is the corresponding parameters, k is the feature index, t is the position index inside a sequence, and $Z(\vec{x})$ is a normalization factor.

The CRF model we use is a linear chain one. The CRF toolkit used is an open source implementation – CRF++[2]. Some changes were made to allow for more flexible use of features.

IV. Application Domain and Data

The target application is a spoken dialogue system for local search, with which users can search for information through natural speech. Currently the system covers seven types of points of interest (POI), including bank, cinema, hotel, hospital, restaurant, gas station and gym, and provides information of contact telephone number, address, price, hotel star grade, restaurant cuisine style, and so on. The service area covers Zhongguancun, Haidian District, Beijing.

In our work two types of voice search dialogues are used. One is human-human dialogues, collected via the Wizard-of-Oz (WOZ) setup, where a human acted as an information-providing agent, who has access to the information source from the Internet. And a user interacted with the agent over the telephone. About half of these dialogues are used for training and one sixth for test (test-H). The other is human-computer dialogues. They were collected through human-computer interactions in a field test after the voice search system was built and are used as another test set (test-C). All the data were manually transcribed and annotated with named entities. Statistics about the data is given in Table I in terms of the number of utterances and characters.

TABLE I
STATISTICS ABOUT THE DATA

set	no. utterances	no. characters
training	5,258	52,884
test-H	1,512	17,501
test-C	1,411	13,531

V. The Unit for Tagging

As we mentioned earlier, for Chinese spoken language understanding, we need to decide which unit to choose for tagging, character or word. We will first have a look at the

[2]Available at http://crfpp.sf.net.

pros and cons for each option and then make a decision based on experiment results.

A. Word vs. Character as Tagging Unit

Unlike several Indo-European languages, e.g., English, there are no natural boundaries between words in Chinese sentences. Chinese sentences look like characters concatenated together, without any gap between characters. Therefore for Chinese SLU, one has to make a choice between word or character as the basic tagging unit. This is a similar issue for several other Chinese language processing tasks, like syntactic chunking and parsing.

Chinese words, like English ones, are natural and independent meaning bearing unit. Ideally they are very desirable to be chosen as the basic tagging unit. But it is not always easy to obtain such boundary information. The task of Chinese word segmentation (CWS) is designed to address this problem. It is well-studied and great progress has been achieved in the past few years [23], [24]. With plenty of manually annotated training data (usually news text and around a million characters or more), CWS can be taken as a resolved issue. But for situations where there is no matching training data, it is still an open issue.

Unfortunately this is our situation. Almost all currently available annotated data are written texts. But the data we need to deal with are spontaneous spoken dialogues. It is less likely we could easily collect enough dialogue data, let alone have them further annotated. This leaves us two options: either we segment with some ready toolkit (usually trained on available training data) or we give up word and use character as the basic unit. If we use word from an automatic segmenter, there will be some noise brought over from segmentation errors. In addition speech recognition errors may make it worse. Will the gain from word information outweigh the loss due to noises? If we use character, we are losing word level information. But a word usually means much more than or even quite different from its component characters. For example, the meaning of "中关村" (Zhaoguancun, a technology hub in Beijing) is far from the meaning of "中" (middle), "关" (pass) and "村" (village). This is very much so for many proper names or named entities. Will the loss of missing word information outweigh the gain from being free from segmentation noise?

B. Experiment

In order to compare word-based and character-based NER, we carry out a set of experiments. In these experiments, we use the same human-human data set to train CRF models.

The baseline features for CRF include: (1) unigram features $(x[i], i = -2, \ldots, 2)$, i.e., current lexical unit, two before and two after, and (2) bigram features $(x[i]/x[i+1], i = -2, \ldots, 1)$, i.e., a concatenation of adjacent units. For both types of features, x is the lexical unit and can be word or character, depending on the choice, i is the position index relative to the current unit. For example, $x[0]$ denotes current unit and $x[-1]$ denotes the unit before. The word segmentation was done with the ICTCLAS toolkit.[3]

Using the above features, we trained two models, one for word and the other for character. Then we tested both models on the two test sets, using both manual transcripts and ASR outputs. The ASR performance in terms of character error rate (CER) is 23.7% for test-H and 8.8% for test-C. (The details of the speech recognizer can be found in [25].) The NER performance in terms of F_1 measure (a harmonic mean of precision and recall) is given in Table II, where *ref* is manual transcript and *asr* is ASR output.

TABLE II
RESULTS OF CHARACTER-BASED AND WORD-BASED NER (IN F_1)

input	character-based	word-based
test-H/ref	91.20	81.65
test-H/asr	72.15	64.67
test-C/ref	80.59	68.73
test-C/asr	75.27	65.01

From the results we can see the significant performance degradation for word-based NER. This is partly due to that the word sequences are very noisy due to many segmentation errors and that the word-based NER is very sensitive and therefore vulnerable to noises (either from segmentation or from ASR). The latter is more damaging since even if we had perfect word segmenter, ASR noises are still hard to avoid. Therefore, we have to choose character as the basic unit for the sake of robustness.

Between the two test sets, the performance on the human-computer test set is lower, though human-computer dialogues are simpler than human-human dialogues. This is due to the mismatch between the training and test sets. The training data are human-human dialogues and test-H matches training data better than test-C. We can see the mismatch from the out-of-vocabulary rate of named entities (NEs) for the two test sets. For test-H, it is 21.4%. But for test-C, it is 52.5%.

VI. EXPLOITING PRIOR KNOWLEDGE

As robustness is our major concern for CSLU, we have to opt for character as the basic tagging unit. This way we treat sentences as concatenated strings of characters, without explicitly taking into account word level information. But we believe that there is some prior knowledge, i.e., lexical information above characters that we could make good use of, esp. for the domain lexicon, mainly the NEs from the POI information. Therefore we try two approaches to incorporating domain lexical features into the character-based NER. One utilizes lexical features of NEs during feature extraction, and the other makes plain use of NEs by adding them to the training data as if they were sentence fragments.

A. Domain Information and Lexicon as Prior Knowledge

To build a domain lexicon, we crawled some dedicated websites for the seven types of POIs in the target service

[3]Available at http://www.ictclas.org/.

area of Zhongguancun and extracted relevant information. As a result, we get a list of 4,866 named entities (only about 25% of them occurred in collected dialogue data for training). We could use the list directly as the domain lexicon. But it is very ineffective for some NEs, esp. for those POI names and addresses. For POI names there may be some variations which are very likely to be out of the list. For example, 郭林家常菜 (GUOLIN homely dish) is a restaurant name. Some may refer to it by its full name, some by its variant "郭林餐馆" (GUOLIN restaurant), other by its shortened name "郭林" (GUOLIN). For addresses there may be different shortened forms. Therefore they were decomposed into components or sub-words (some may be words in different context). Some examples are given below:

```
POI names:
 - 郭林 (GUOLIN)  / 家常菜 (homely dish)
 - 翠宫 (Jade Palace) / 饭店 (Hotel)
 - 肯德基 (KFC) / 保福寺 (BAOFUSI) / 店 (subbranch)
Address:
 - 北京市 (Beijing) / 海淀区 (Haidian District) /
   北          四          环 (North Fourth Ring) / 西
路 (West Road) / 21 (21)/号 (number)
```

In the end those components were added to the lexicon. In the lexicon there are 938 entries. With such a lexicon we improved not only the coverage and therefore the robustness against variations but also the processing efficiency since the size of the lexicon is significantly reduced.

B. Enriching Features with the Domain Lexicon

The idea of incorporating word information into character-based tagging is simple. For a character in a sentence, we check if the character ngrams ($n = 1, 2, 3, 4$) beginning with the current character are in a lexicon. For each ngram in the lexicon, we add *beginning of ngram* to the feature list of current character and *inside ngram* to the feature list of every following character. This can be illustrated in Figure 1. In the example, for current character a in a character sequence (i.e., a sentence), we look up ngrams a, ab, abc and $abcd$ in the lexicon. Assume a, ab and $abcd$ are A, B, C in the lexicon, then we add B-A, B-B, B-C to the feature list of a, I-B, I-C to that of b, and I-C to the that of c and d. Here B-X indicates the beginning character of ngram X and I-X any character of ngram X other than the beginning one. This is the notation used in [19].

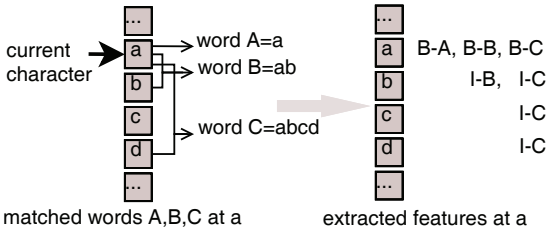

Fig. 1. Feature extraction with domain lexicon

C. Augmenting Training Data with the Domain Lexicon

The NEs that occurred in the collected dialogues are only a small part (about 10%) of the POI information we collected. If we use the tagger trained with the dialogue data for practical use, it is very likely to meet NEs that are out of the set in the training data. Recall that we had the semantics of those items from the collected POI information, i.e., we know that the items are names of POIs (e.g., hotels, restaurants, banks, etc.), addresses, contact numbers, etc. We need to find some way to further harness that. One seemingly naive but practically highly effective approach is to treat those items as utterances and add them to the training data. This is partly inspired by the fact that some NEs in the collected dialogues did appear in the form of elliptical utterances, where there are no other characters or words but NEs.

D. Experiments

We carried out three series of experiments to see if and how exploiting domain lexical information improves character-based NER performance. In the first series, we enhanced features by incorporating formal word/sub-word information. In the second, we augmented training data with domain knowledge by taking all POIs as utterances and their semantic tags as annotations. In the third, we combined the above two. The results are given in Table III, in columns *+feature*, *+training* and *+both* respectively.

TABLE III
PERFORMANCE OF NER USING DOMAIN INFORMATION (IN F_1)

input	baseline	+feature	+data	+both
test-H/ref	91.20	93.51	93.66	95.29
test-H/asr	72.15	73.56	74.74	75.08
test-C/ref	80.59	89.10	94.73	97.14
test-C/asr	75.27	83.64	87.37	90.24

From the table we can see that all approaches achieved notable improvement. By using domain lexicon for feature extraction, we are able to capture rich lexical constraints from words and sub-words, which are helpful for the NER robustness. By simply adding the list of domain entities to the training data, the coverage for the NE tagger is significantly improved. Both approaches are effective in exploiting domain information for character-based NER, but from different angles. This motivated us to combine both. From the results (the last column *+both*), we can see that the improvement is remarkable. For all types of input, we consistently achieved best performance. This may indicate that the contributions from both approaches are complementary to some degree.

It seems very hard to improve the NER results on the input test-H/asr (ASR output of human-human dialogues). But if we recall that the ASR performance is as high as 23.7% in CER, it is easy to understand why. It is the noisier ASR output that is to blame for the lower NER performance, since there is no out-of-vocabulary NE.

In addition, the best NER results on test-C are higher than that on test-H, even test-H matches the training data better.

This is because human-computer dialogues are simpler than human-human dialogues.

VII. Related Work

In this paper we built on previous work [20] and addressed the robust (semantic) understanding of spoken Chinese language for a more complex task and domain. The most similar works in English SLU are carried out in the AT&T call routing system [11] and VoiceTone system [13]. In [11], a hybrid method is used to recognize four types of NEs: two generic ones (date and phone number) and two task specific ones (Item amount and Which Bill). For the generic ones, they achieved the best F-scores of 74.5 for date and 93.9 for phone number. In [13], NER is done with a rule-based approach and they achieved F-scores of 68.6 for 13-digit account number, 66.5 for date, 86.8 for place, 65.1 for phone number, etc. (no overall F-score is given). In our work we formulated the SLU task as character-based sequence tagging problem and used CRF-based tagger. By exploiting domain information we improved the recognition performance of 13 types of NEs from 75.27 to 90.24 (F_1) on a field test set using ASR output. (NB: since we do not work on the same data, the figures are only indicative and should not be directly compared.)

There are also two major related works in CSLU [26], [27]. In [26] the approach is based on words (or chunks) instead of characters. The chunks are segmented with some hand-crafted rules. The understanding is carried out on the chunks with an HMM-based tagger. In [27] CSLU is taken as a two-stage classification task: first topic classification and then slot classification. The slot classifier, roughly equivalent to our NER, is also based on words. They mainly focused on how to employ weakly supervised learning to reduce manual labeling effort. But we focused on how to improve the robustness through character-based statistical tagging with domain information exploitation. It is worth pointing out that neither of them carried out experiments on the understanding of ASR output. But this is what a real-world SLU system has to face. We showed the effectiveness of our approach on both manual transcript and ASR output.

VIII. Conclusions

In this paper we addressed the robust semantic understanding of spoken Chinese for a voice search dialogue system. The robustness is achieved via three means:

1) We formulated the problem of semantic understanding as an NER task, which is conveniently solved through statistical sequential tagging.
2) Under the statistical framework, we chose character instead of word as the tagging unit. Results showed that the character-based tagging is much more robust than the word-based one against the data sparseness and ASR noisiness problems despite the loss of word-level information.
3) we further exploited prior knowledge in the character-based statistical tagging framework in two ways. One is to enrich tagger features by incorporating more lexical (word and sub-word) features with a domain lexicon. The other is to make plain use of domain entities by simply adding them to the training data. Experiment results show that both approaches are effective. The best performance is achieved by combining the above two complimentary approaches.

The statistical sequential tagging approach to NER is not new. But to the best of our knowledge, this is the first piece of work that employs the character-based tagging framework to solve the problem of shallow semantic understanding of spoken Chinese. What's more, this apporach is shown to be amenable to prior knowledge exploitation.

There are several directions for future work. But the very next we would like to work on is to extend the input and output of SLU from single item to rich representations, like n-best, confusion networks, etc., to further improve robustness. Re-ranking would be another further direction.

Acknowledgment

This work is partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 10874203, 60875014, 61072124, 11074275, 11161140319).

References

[1] M. F. McTear, "Spoken dialogue technology: enabling the conversational user interface," *ACM Computing Surveys*, vol. 34, no. 1, pp. 90–169, 2002.
[2] Y. Wang, D. Yu, Y.-C. Ju, and A. Acero, "An introduction to voice search," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 28–38, May 2008.
[3] R. De Mori, F. Bechet, D. Hakkani-Tur, M. F. McTear, G. Riccardi, and G. Tur, "Spoken language understanding," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 50–58, May 2008.
[4] D. H. Klatt, "Review of the ARPA speech understanding project," *The Journal of the Acoustical Society of America*, vol. 60, no. S1, pp. S10–S10, 1976.
[5] P. Price, "Evaluation of spoken language systems: the atis domain," in *HLT '90: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA: Association for Computational Linguistics, 1990, pp. 91–95.
[6] J. Peckham, "Speech understanding and dialouge over the telephone: an overview of the ESPRIT SUNDIAL," in *HLT '91: Proceedings of the workshop on Speech and Natural Language*. Morristown, NJ, USA: Association for Computational Linguistics, 1991, pp. 14–27.
[7] G. Tur and R. De Mori, Eds., *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. Wiley, 2011.
[8] Y.-Y. Wang, L. Deng, and A. Acero, "Spoken language understanding," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 16–31, Sep. 2005.
[9] Y. Wang and A. Acero, "Combination of cfg and n-gram modeling in semantic grammar learning," in *Proceedings of the Eurospeech Conference*, 2003, pp. 2809–2812.
[10] Y. Wang, A. Acero, M. Mahajan, and J. Lee, "Combining statistical and knowledge-based spoken language understanding in conditional models," in *ACL*. The Association for Computer Linguistics, 2006, pp. 882 – 889.
[11] F. Bechet, A. L. Gorin, J. H. Wright, and D. Hakkani-Tur, "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, Feb. 2004.
[12] G. Tur, D. Hakkani-Tur, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding." *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.
[13] N. Gupta, G. Tur, D. Hakkani-Tür, S. Bangalore, G. Riccardi, and M. Gilbert, "The at&t spoken language understanding system," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 1, pp. 213–222, 2006.

[14] Y. He and S. Young, "Semantic processing using the hidden vector state model," *Computer Speech & Language*, vol. 19, no. 1, pp. 85–106, 2005.
[15] M. Jeong and G. G. Lee, "Multi-domain spoken language understanding with transfer learning," *Speech Communication*, vol. 51, no. 5, pp. 412 – 424, 2009.
[16] M. Dinarelli, A. Moschitti, and G. Riccardi, "Re-ranking models for spoken language understanding," in *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2009, pp. 202–210.
[17] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.
[18] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching dbn," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 7, pp. 1303–1314, 2008.
[19] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the conll-2003 shared task: Language-independent named entity recognition," in *Proceedings of CoNLL-2003*, W. Daelemans and M. Osborne, Eds. Edmonton, Canada: Association for Computational Linguistics, 2003, pp. 142–147.
[20] C. Bao, W. Xu, and Y. Yan, "Recognizing named entities in spoken chinese dialogues with a character-level maximum entropy tagger," in *INTERSPEECH 2008*. Brisbane, Australia: ISCA, 2008, pp. 1145–1148.
[21] J. D. Lafferty, A. K. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, C. E. Brodley and A. P. Danyluk, Eds. Williamstown, MA, USA: Morgan Kaufmann, 2001, pp. 282–289.
[22] C. A. Sutton and A. K. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to Statistical Relational Learning*, ser. Adaptive Computation and Machine Learning, L. Getoor and B. Taskar, Eds. The MIT Press, 2007, ch. 4, pp. 93–127.
[23] C. Huang and H. Zhao, "Chinese word segmentation: A decade review," *Journal of Chinese Information Processing*, vol. 21, no. 3, pp. 8–20, 2007, in Chinese.
[24] G. Jin and X. Chen, "The fourth international chinese language processing bakeoff: Chinese word segmentation, named entity recognition and chinese pos tagging," in *Proceedings of the Sixth SIGHAN Workshop*. Hyderabad, India: Asian Federation of Natural Language Processing, 2008, pp. 69–81.
[25] T. Li, C. Bao, W. Xu, J. Pan, and Y. Yan, "Improving voice search using forward-backward lvcsr system combination," in *Proc. of ISNN 2009*. Wuhan, China: Springer, 2009, pp. 769–777.
[26] G. Xie, C. Zong, and B. Xu, "Approach to robust spoken chinese language parsing," *Journal of Chinese Language and Computing*, vol. 14, no. 1, pp. 5–19, 2004, in Chinese.
[27] W.-L. Wu, R.-Z. Lu, J.-Y. Duan, H. Liu, F. Gao, and Y.-Q. Chen, "Spoken language understanding using weakly supervised learning," *Computer Speech & Language*, vol. 24, no. 2, pp. 358 – 382, 2010.