Robust Character Based Tagging with Domain Lexical Features for Chinese Spoken Language Understanding

Changchun Bao, Yali Li, Ta Li, Jielin Pan and Yonghong Yan
ThinkIT Laboratory
Institute of Acoustics
Chinese Academy of Sciences
Beijing, China, 100190
Email: baochangchun@hccl.ioa.ac.cn

***Abstract*—Word information is useful in natural language understanding. But in Chinese language processing, word information is not given natural. While word-segmentation works well for text in NLU, it deteriorates Chinese SLU because of the flexibility and distortion of spoken utterance plus ASR errors. This paper propose a novel approach, sub-word features, to take use word information and help understanding spoken utterance while retain the robustness of character-wise processing. By means of this approach, we can also effectively use named entity list to improve SLU performance. Experiments show that the sub-word features give an average of 0.7 improvement for ASR, and the usage of named list given an average of 4.7 improvement.**

I. Introduction

With the rapid progress of speech technology, several speech enabled interactive applications have emerged, like spoken dialogue systems (SDS). Spoken dialogue systems have evolved from early command and control systems, to call routing and form filling systems, and to the latest voice search systems, which promises to provide local information through natural spoken interaction. With the demand for natural interaction increases, the requirement for deeper 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 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 [15].

Although sharing a similar goal to natural (or written) language understanding (NLU), the understanding of spoken languages faces some more challenging difficulties. 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. What's more, current spontaneous speech recognizers inevitably bring a lot of 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, single robustness 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 deal 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.

If we opt for character, we are loosing word level information since a word usually means much more than or even quite different from its 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 and highly desired, we may need to pay some price, not only for segmentation, but also for some side effect from noisy input.

The rest of this paper is organized as follows. First we introduce the task, the framewrork and the data for our CSLU system in section II. Then in section III we take a look at the processing unit issue and carry out character-based and word-based experiments. In section IV we further describe how to exploit domain lexical features for CSLU. Before closing, we discuss some related works in section V.

II. Chinese Spoken Language Understanding

A. The Application

The target application behind our work 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), i.e., bank, cinema, hotel, hospital, restaurant, gas station, and sport facility and provides information of contact telephone number, address, price, hotel star grade, and so on. Route planning assistance is also provided, which can help users to find how to get to the interested place by public transportation or self driving. The service area covers Zhongguancun, Haidian District, Beijing.

978-1-4244-5961-2/10/$26.00 ©2010 IEEE

B. CSLU and Named Entity Recognition

Usually for the meaning of a spoken utterance there are two relatively independent aspects: one is semantic, i.e., about entities and their relations, and the other is pragmatic, i.e., about the speaker's intention. Therefore the task of SLU can be decomposed into two subtasks. One is semantic understanding and the other is pragmatic understanding. Semantic understanding ideally should address the recognition of both entities and their relations. But so far only entity recognition is highly focused on for SLU due to some reasons. On the one hand it is difficult to analyze spoken utterances in a full and deep way due to noises like disfluencies and ASR errors. 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. In this paper we mainly 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. And named entity recognition is also one of the major tasks of SLU in the AT&T call routing system [16] and VoiceTone service [3].

In our application there are two classes of named entities (NEs) according to whether an NE is specific to a POI. The common NEs consists of toponym, phone number, traffic means, bus number, and price. The specific NEs are the entities whose meaning is rather related to the task domain. For service type, there are seven classes, including bank, cinema, restaurant, service station, hospital, hotel, sport utility. There are more than 5000 entries in the directory list about toponym and the seven classes of service facility entity.

C. Named Entity Recognition as Sequence Tagging

The task of named entity recognition (NER) is commonly formulated as sequence tagging, e.g., [13]. In our work we take the same approach and build on [1]. In the previous work we dealt with a simpler task (route planning) and domain (only three types of named entities were concerned) and used only transcripts. And a Maximum Entropy model based tagger is used. In current work we choose a more challenging task (voice search for local information) and domain (13 types of named entities). Experiments are carried out on both manual transcripts and speech recognition output with a conditional random field (CRF) based tagger. A brief description of CRF model is given below.

CRF is a statistical sequence modeling framework introduced by Lafferty et al. [8]. In the model the probability assigned to a label sequence $\vec{y}$ for a given input sequence $\vec{x}$ is given as:

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

where f_k is a feature function and λ_k is the corresponding parameters. $Z(\vec{x})$ is a normalization factor.

The CRF model we use is a linear chain one. The classifier used in this paper is an open source implentation – the toolkit of CRF++[1]. Some changes were made to allow more flexible use of feature functions.

D. Data

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. The human has access to the information source from the internet. And a user interacted with the agent on the telephone. Half of these dialogues are used for training and one third for test (test-H). The other is human-computer dialogues. They were collected through human-computer interactions after the voice search system was built. They 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

III. THE UNIT FOR TAGGING

As we mentioned in the introduction, 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 pros and cons for each option and then make a decision based on some experiments.

A. Word vs. Character as 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. (The relation between Chinese words and characters is in some say similar to that between English words and morphemes.) Therefore in Chinese SLU, one has to make a choice between word or character as the basic processing unit. This is an issue similar to 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 processing 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 [17], [14]. 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

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

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 collect enough dialogue data, let alone have them further annotated. This leaves us two options: either we segment with some off-the-shelf toolkit,[2] usually trained on available training data, or we discard word but 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 right away, we are loosing word level information. But a word usually means much more than or even quite different from its component characters. For example, the meaning of "地铁" (subway) is far from the meaning of "地" (field) and "铁" (iron). This is very much so for many proper names or named entities. Will the loss of missing word information outweigh the grain from being away from segmentation noise?

We will answer these questions empirically below.

B. Experiment

In order to compare word-based NER and character-base NER, we carry out a set of experiments. In these experiments, we use the same human-human training set to train CRF models. Then we test with this model on human-human test set and human-machine test sets (test-A and test-B) using both transcripts and ASR outputs.

The baseline feature templates 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 kinds of features, x is the lexical unit and can be word or character, depending on the choice, i is the relative position to current unit. For example, $x[0]$ denotes current unit and $x[-1]$ denotes previous one. The word-wise experiment data is gained from original character-wise data using the ICTCLAS toolkit.

Using the above features, we trained two models, one for word and the other for character. Then we test both models on the three 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, 13.2% for test-C1 and 8.8% for test-C2. (The details of the speech recognizer can be found in [10].) The NER performance in terms of F_1 measure (a harmonic mean of precision and recall) is given in II

As shown in experiment result, we think that as robustness is one of key issues for SLU, word segmentation for SLU as preprocess maybe is not suitable for flexible natural spoken utterance.

From the result, we can see the overall performance of word-segmented-based semantic parsing is not good as character-based semantic parsing. The reason I think is as follows: first, word segmentation itself is not 100% perfect, wrongly segmented words influence semantic parsing, the lead to a decline of parsing performance on the transcriped data, from character-based strategy to word-segmentation-based. Second, as stated before, character-based strategy, by using character-n-gram features, can make up word information to some degree, and obtain pretty good performance. Third, when input is noisy asr recognized user natural utterance, segmentation lose because of its weak robustness, this break down the parsing performance sharply, on the contrast, character-based strategy is influenced by this noise much more slightly.

Therefore, we choose character as the basic unit for the sake of robustness.

[2]We used ICTCLAS, available at http://www.ictclas.org/.

TABLE II
PERFORMANCES OF CHARACTER-WISE AND WORD-WISE PARSING

input	character-wise	word-wise
test-H/ref	91.20	26.80
test-H/asr	72.15	21.97
test-C/ref	80.59	10.91
test-C/asr	75.27	10.19

IV. EXPLOITING DOMAIN LEXICAL FEATURES

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

A. Featuer Extraction with NEs

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=1-4) beginning with the current character are in a lexicon. (The construction of the lexicon will be explained below.) For each ngram in the lexicon (i.e., it is a word), we add *beginning of word ngram* to the feature list of current character and *inside word ngram* to the feature list of every following character. This can be illustrated in Figure 1.

1) Domain lexicon: We crawled some dedicated websites for the seven types of POIs in the target service area of Zhongguancun and extracted relevant information. As a result, we get a list of more than 5,000 named entities (only about 10% 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 and they are very likely to be out of 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 we decomposed those named entities into their components. 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 put into the lexicon. With such a lexicon we improved not only the robustness against variations but also the processing efficiency since the size of the lexicon is significantly reduced.

B. Bare use of domain information

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 the items are names of POIs (e.g., hotels, restaurants, banks, etc.), addresses, contact numbers, etc.. We need 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.

C. Experiments

We carried out three series of experiments to see if/how exploiting domain lexical information improves NER performance. In the first series, we enhanced features by incorporating word/subword information. In the second, we augmented training data by taking all POI information as utterances. In the third, we combined the above two. The results are given in columns +feature, +training and +both in Table III.

In order to check the influence of the proposed sub-word approach, we have carried out a set of experiments. Similar to previous introduced experiments, this set of experiments are carried out on three kinds of test set, the human-human test set, human-machine test set 1 and human-machine test set 2, each test set has both transcription and ASR recognized hypothese for test.

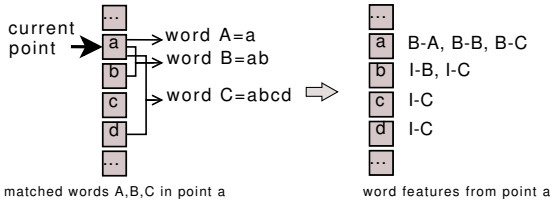

Fig. 1. Feature of beginning of word ngram

TABLE III
PERFORMANCE OF USING DOMAIN LEXICON

input	baseline	+feature	+training	+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

V. RELATED WORK

There are many research groups devoting for the development of SLU. Some have already deployed for application, for example, AT&T has deployed telephone dialogue service systems like How May I Help You(HMIHY) and VoiceTone[3], LUNA project has implemented in the FT3000 telephone service application[2], et al. And there are also many SLU technologies in rapid research, like Microsoft's HMM/CFG SLU approach[7], IBM's fertility methods, Cambridge's HVS models[4], and so on.

Among these approaches, the understanding strategy of our system is similar to AT&T's. In its SLU systems, AT&T uses shallow parsing as slu strategy. SLU is devided into two subtasks, named entity recognition and utterance classification. The named entity recognition is composed of series processing, named entity detection and named entity value extraction. In the named entity extraction process, HMM is used as a robust statistical classifier; and in the named entity value extraction process, word confusion network and grammars are used to extract the values represented by named entities.

For Chinese spoken language understanding, [5] propose a Chinese SLU approach using weakly supervised learning. The SLU framework mainly consists of two kinds of classifiers: topic classifier and slot classifier. Besides the two key components, the system also contains preprocessor and a slot-merger.The semantic parsing is character-wise. The function of preprocessor is to search for sub-strings matching corresponding to certain semantic classes, which resembles named entity recognition. In order to search such sub-strings, it uses a local chart parser which is said to has low level of robustness. The task of topic classifier is similar to utterance classification. When utterance entities and topic are made certain, the slot classifier is to recognize the relation between each entity with utterance topic. And the slot-merger carry out some post-process to amend understanding.

[6] propose a chunk-based Chinese parsing approach for spoken language translation. In this system, user utterance is processed in chunk-wise mannar. It assumed that although Chinese utterances have very flexible order in the sentence level, there is fixed order between several words, and these words can be chunked together to represent a single integrate meaning. So the slu system extend granularity from word-wise to chunk-wise. This approach, in our opion, maybe has a degree of robustness to flexible order of Chinese utterance.

VI. Conclusions

In natural spoken Chinese language processing, word information is not naturally given, and word segmentation doesn't work perfectly in SLU application. So how to use word information to help understanding becomes a problem. In this paper, we propose a novel strategy, sub-word features, to take use of word information while retain the character-wise parsing. The character-wise parsing strategy remain the robustness against ungrammatical speech phenomena and ASR errors. Experiments show that the proposed strategy can improve NER performance for both transcription text and ASR hypotheses. Furthermore, by means of sub-word features strategy, we propose using application knowledge–NE list–to improve NER performance. Experiments show an evident improvement. Further work will focus on searching for suitable approach to combine non-NE sub-word lexicon with NE sub-word lexicon to obtain more improvement, and investigating the way to combine SLU with ASR more actively, in order to deal with ASR errors more powerfully.

Acknowledgment

This work is partially supported by The National High Technology Research and Development Program of China (863 program, 2006AA010102), National Science & Technology Pillar Program (2008BAI50B00), MOST (973 program, 2004CB318106), National Natural Science Foundation of China (10874203, 60875014, 60535030).

References

[1] Bao, Changchun and Xu, Weiqun and Yan, Yonghong, Recognizing named entities in spoken Chinese dialogues with a character-level maximum entropy tagger, INTERSPEECH 2008, pp.1145-1148.

[2] Geraldine Damnati, Frederic Bechet, and Renato De Mori, First implementation of the LUNA spoken luanguage understanding strategy on a telephone service application, Intelligent Information System 2008, pp. 499 - 505, 2008.

[3] Gupta, N. and Tur, Gokhan and Hakkani-Tür, Dilek and Bangalore, Srinivas and Riccardi, Giuseppe and Gilbert, M., The AT&T spoken language understanding system, IEEE Transactions on Audio, Speech & Language Processing, vol.14, no.1, pp.213–222, 2006

[4] He, Yulan and Young, Steve, Semantic processing using the Hidden Vector State model, Computer Speech & Language, vol.19, no.1, pp.85–106, 2005.

[5] Wu, Wei-Lin and Lu, Ru-Zhan and Duan, Jian-Yong and Liu, Hui and Gao, Feng and Chen, Yu-Quan, Spoken language understanding using weakly supervised learning, Computer Speech & Language, vol.24, no.2, pp. 358-382, 2010

[6] Xie, Guodong and Zong, Chengqing and Xu, Bo, Approach to Robust Spoken Chinese Language Parsing, Journal of Chinese Language and Computing, vol.14, no.1, pp.5-19, 2004

[7] Wang, Ye-Yi and Acero, Alex, Spoken Language Understanding - an introduction to the statistcal framework, IEEE Signal Processing Magzine, no. 9, 2005

[8] Lafferty, John D. and McCallum, Andrew Kachites and Pereira, Fernando C. N., Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, In Proc. ICML 2001, pp.282–289, 2001

[9] Brodley, Carla E. and Danyluk, Andrea Pohoreckyj, Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001

[10] Li, Ta and Bao, Changchun and Xu, Weiqun and Pan, Jielin and Yan, Yonghong, Improving Voice Search Using Forward-Backward LVCSR System Combination, Proc. of The Sixth International Symposium on Neural Networks (ISNN 2009), pp. 769–777, 2009.

[11] Chang, Pi-Chuan and Galley, Michel and Manning, Christopher D., Optimizing Chinese Word Segmentation for Machine Translation Performance, Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, pp.224–232, 2008

[12] Hajič, Jan and Ciaramita, Massimiliano and Johansson, Richard and Kawahara, Daisuke and Martí, Maria Antònia and Màrquez, Lluís and Meyers, Adam and Nivre, Joakim and Padó, Sebastian and Štěpánek, Jan and Straňák, Pavel and Surdeanu, Mihai and Xue, Nianwen and Zhang, Yi, The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages, Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task, Boulder, Colorado, 2009.

[13] Tjong Kim Sang, Erik F. and De Meulder, Fien, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, Proceedings of CoNLL-2003, Edmonton, Canada, 2003.

[14] Jin, Guangjin and Chen, Xiao, The Fourth International Chinese Language Processing Bakeoff: Chinese Word Segmentation, Named Entity Recognition and Chinese POS Tagging, Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing, 2008

[15] De Mori, Renato and Bechet, Frédéric and Hakkani-Tur, Dilek and McTear, Michael F. and Riccardi, Giuseppe and Tur, Gokhan, Spoken language understanding, IEEE Signal Processing Magazine, vol.25, no. 3, 2008.

[16] Bechet, Frédéric and Gorin, Allen L. and Wright, Jeremy H. and Hakkani-Tur, Dilek, Detecting and extracting named entities from spontaneous speech in a mixed-initiative spoken dialogue context: How May I Help You?^{sm,tm}, Speech Communication, vol.42, no.2, pp. 207–225, 2004

[17] Huang, Changning and Zhao, Hai, Chinese Word Segmentation: A Decade Review, Journal of Chinese Information Processing, vol.21, no.3, 2007.