Chinese word segmentation and named entity recognition based on a context-dependent Mutual Information Independence Model

Zhang Min Zhou GuoDong Yang LingPeng Ji DongHong

Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore, 119613
Email: (mzhang, zhougd, lpyang, dhji)@i2r.a-star.edu.sg

Abstract

This paper briefly describes our system in the third SIGHAN bakeoff on Chinese word segmentation and named entity recognition. This is done via a word chunking strategy using a context-dependent Mutual Information Independence Model. Evaluation shows that our system performs well on all the word segmentation closed tracks and achieves very good scalability across different corpora. It also shows that the use of the same strategy in named entity recognition shows promising performance given the fact that we only spend less than three days in total on extending the system in word segmentation to incorporate named entity recognition, including training and formal testing.

1 Introduction

Word segmentation and named entity recognition aim at recognizing the implicit word boundaries and proper nouns, such as names of persons, locations and organizations, respectively in plain Chinese text, and are critical in Chinese information processing. However, there exist two problems when developing a practical word segmentation or named entity recognition system for large open applications, i.e. the resolution of ambiguous segmentations and the identification of OOV words or OOV entity names.

In order to resolve above problems, we developed a purely statistical Chinese word segmentation system and a named entity recognition system using a three-stage strategy under an unified framework.

The first stage is called known word segmentation, which aims to segment an input sequence of Chinese characters into a sequence of known words (called word atoms in this paper). In this paper, all Chinese characters are regarded as known words and a word unigram model is applied to perform this task for efficiency. Also, for convenience, all the English characters are transformed into the Chinese counterparts in preprocessing, which will be recovered just before outputting results.

The second stage is the word and/or named entity identification and classification on the sequence of atomic words in the first step. Here, a word chunking strategy is applied to detect words and/or entity names by chunking one or more atomic words together according to the word formation patterns of the word atoms and optional entity name formation patterns for named entity recognition. The problem of word segmentation and/or entity name recognition are re-cast as chunking one or more word atoms together to form a new word and/or entity name, and a discriminative Markov model, named Mutual Information Independence Model (MIIM), is adopted in chunking. Besides, a SVM plus sigmoid model is applied to integrate various types of contexts and implement the discriminative modeling in MIIM.

The third step is post processing, which tries to further resolve ambiguous segmentations and unknown word segmentation. Due to time limit, this is only done in Chinese word segmentation. No post processing is done on Chinese named entity recognition.

The rest of this paper is as follows: Section 2 describes the context-dependent Mutual Information Independence Model in details while purely statistical post-processing in Chinese word segmentation is presented in Section 3. Finally, we report the results of our system in Chinese word segmentation and named entity recognition in Section 4 and conclude our work in Section 5.

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 154–157,
Sydney, July 2006. ©2006 Association for Computational Linguistics

2 Mutual Information Independence Model

In this paper, we use a discriminative Markov model, called Mutual Information Independence Model (MIIM) as proposed by Zhou et al (2002), for Chinese word segmentation and named entity recognition. MIIM is derived from a conditional probability model. Given an observation sequence $O_1^n = o_1 o_2 \cdots o_n$, MIIM finds a stochastic optimal state(tag) sequence $S_1^n = s_1 s_2 \cdots s_n$ that maximizes:

$$\log P(S_1^n \mid O_1^n) = \sum_{i=2}^{n} PMI(s_i, S_1^{i-1}) + \sum_{i=1}^{n} \log P(s_i \mid O_1^n)$$

We call the above model the Mutual Information Independence Model due to its Pair-wise Mutual Information (PMI) assumption (Zhou et al 2002). The above model consists of two sub-models: the state transition model $\sum_{i=2}^{n} PMI(s_i, S_1^{i-1})$, which can be computed by applying ngram modeling, and the output model $\sum_{i=1}^{n} \log P(s_i \mid O_1^n)$, which can be estimated by any probability-based classifier, such as a maximum entropy classifier or a SVM plus sigmoid classifier (Zhou et al 2006). In this competition, the SVM plus sigmoid classifier is used in Chinese word segmentation while a simple backoff approach as described in Zhou et al (2002) is used in named entity recognition.

Here, a variant of the Viterbi algorithm (Viterbi 1967) in decoding the standard Hidden Markov Model (HMM) (Rabiner 1989) is implemented to find the most likely state sequence by replacing the state transition model and the output model of the standard HMM with the state transition model and the output model of the MIIM, respectively. The above MIIM has been successfully applied in many applications, such as text chunking (Zhou 2004), Chinese word segmentation (Zhou 2005), English named entity recognition in the newswire domain (Zhou et al 2002) and the biomedical domain (Zhou et al 2004; Zhou et al 2006).

For Chinese word segmentation and named entity recognition by chunking, a word or a entity name is regarded as a chunk of one or more word atoms and we have:

- $o_i =< p_i, w_i >$; w_i is the $i-th$ word atom in the sequence of word atoms $W_1^n = w_1 w_2 \cdots w_n$; p_i is the word formation pattern of the word atom w_i. Here p_i measures the word formation power of the word atom w_i and consists of:
  - The percentage of w_i occurring as a whole word (round to 10%)
  - The percentage of w_i occurring at the beginning of other words (round to 10%)
  - The percentage of w_i occurring at the end of other words (round to 10%)
  - The length of w_i
  - Especially for named entity recognition, the percentages of a word occurring in different entity types (round to 10%).
- s_i : the states are used to bracket and differentiate various types of words and optional entity types for named entity recognition. In this way, Chinese word segmentation and named entity recognition can be regarded as a bracketing and classification process. s_i is structural and consists of two parts:
  - **Boundary category (B)**: it includes four values: {O, B, M, E}, where O means that current word atom is a whOle word or entity name and B/M/E means that current word atom is at the Beginning/in the Middle/at the End of a word or entity name.
  - **Unit category (W)**: It is used to denote the type of the word or entity name.

Because of the limited number of boundary and unit categories, the current word atom formation pattern p_i described above is added into the state transition model in MIIM. This makes the above MIIM context dependent as follows:

$$\log P(S_1^n \mid O_1^n)$$
$$= \sum_{i=2}^{n} PMI(s_i, S_1^{i-1} \mid p_{i-1} p_i) + \sum_{i=1}^{n} \log P(s_i \mid O_1^n)$$

3 Post Processing in Word Segmentation

The third step is post processing, which tries to resolve ambiguous segmentations and false unknown word generation raised in the second step. Due to time limit, this is only done in Chinese word segmentation, i.e. no post processing is done on Chinese named entity recognition.

A simple pattern-based method is employed to capture context information to correct the segmentation errors generated in the second steps. The pattern is designed as follows:

<Ambiguous Entry (*AE*)> | <Left Context, Right Context> => <Proper Segmentation>

The ambiguity entry (*AE*) means ambiguous segmentations or forced-generated unknown words. We use the 1st and 2nd words before *AE* as the left context and the 1st and 2nd words after *AE* as the right context. To reduce sparseness, we also only use the 1st left and right words as context. This means that there are two patterns generated for the same context. All the patterns are automatically learned from training corpus using the following algorithm.

```
LearningPatterns()
// Input: training corpus
// Output: patterns
BEGIN
    (1) Training a MIIM model using training corpus
    (2) Using the MIIM model to segment training corpus
    (3) Aligning the training corpus with the segmented training corpus
    (4) Extracting error segmentations
    (5) Generating disambiguation patterns using the left and right context
    (6) Removing the conflicting entries if two patterns have the same left hand side but different right hand side.
END
```

4 Evaluation

We first develop our system using the PKU data released in the Second SIGHAN Bakeoff last year. Then, we train and evaluate it on the Third SIGHAN Bakeoff corpora without any fine-tuning. We only carry out our evaluation on the closed tracks. It means that we do not use any additional knowledge beyond the training corpus. Precision (**P**), Recall (**R**), F-measure (**F**), OOV Recall and IV Recall are adopted to measure the performance of word segmentation. Accuracy (**A**), Precision (**P**), Recall (**R**) and F-measure (**F**) are adopted to measure the performance of NER. Tables 1, 2 and 3 in the next page report the performance of our algorithm on different corpus in the SIGHAN Bakeoff 02 and Bakeoff 03, respectively. For the performance of other systems, please refer to http://sighan.cs.uchicago.edu/bakeoff2005/data/results.php.htm for the Chinese bakeoff 2005 and http://sighan.cs.uchicago.edu/bakeoff2006/longstats.html for the Chinese bakeoff 2006.

Comparison against other systems shows that our system achieves the state-of-the-art performance on all Chinese word segmentation closed tracks and shows good scalability across different corpora. The small performance gap should be able to overcome by replacing the word unigram model with the more powerful word bigram model. Due to very limited time of less than three days, although our NER system under the unified framework as Chinese word segmentation does not achieve the state-of-the-art, its performance in NER is quite promising and provides a good platform for further improvement. Error analysis reveals that OOV is still an open problem that is far from to resolve. In addition, different corpus defines different segmentation principles. This will stress OOV handling in the extreme. Therefore a system trained on one genre usually performances worse when faced with text from a different register.

5 Conclusion

This paper proposes a purely unified statistical three-stage strategy in Chinese word segmentation and named entity recognition, which are based on a context-dependent Mutual Information Independence Model. Evaluation shows that our system achieves the states-of-the-art segmentation performance and provides a good platform for further performance improvement of Chinese NER.

References

Rabiner L. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *IEEE 77(*2), pages257-285.

Viterbi A.J. 1967. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, IT 13(2), 260-269.

Zhou GuoDong and Su Jain. 2002. Named Entity Recognition Using a HMM-based Chunk Tagger, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'2002)*. Philadelphia. July 2002. pp473-480.

Zhou GuoDong, Zhang Jie, Su Jian, Shen Dan and Tan ChewLim. 2004. Recognizing Names in Biomedical Texts: a Machine Learning Approach. *Bioinformatics*. 20(7): 1178-1190. DOI: 10.1093/bioinformatics/bth060. 2004. ISSN: 1460-2059

Zhou GuoDong. 2004. Discriminative hidden Markov modeling with long state dependence using a kNN ensemble. *Proceedings of 20th International Conference on Computational Linguistics (COLING'2004).* 23-27 Aug, 2004, Geneva, Switzerland.

Zhou GuoDong. 2005. A chunking strategy towards unknown word detection in Chinese word segmentation. *Proceedings of 2nd International Joint Conference on Natural Language Processing (IJCNLP'2005), Lecture Notes in Computer Science (LNCS 3651)*

Zhou GuoDong. 2006. Recognizing names in biomedical texts using Mutual Information Independence Model and SVM plus Sigmod. *International Journal of Medical Informatics* (Article in Press). ISSN 1386-5056

Tables

Task	P	R	F	OOV Recall	IV Recall
CityU	0.9 38	0.952	94.5	0.578	0.967
MSRA	0.952	0.962	95.7	0.51	0.98
CKIP	0.94	0.957	94.8	0.502	0.976
PKU	0.952	0.952	95.2	0.71	0.967

Table 1: Performance of Word Segmentation on Closed Tracks in the SIGHAN Bakeoff 02

Task	P	R	F	OOV Recall	IV Recall
CityU	0.968	0.961	96.5	0.633	0.983
MSRA	0.961	0.953	95.7	0.499	0.977
CKIP	0.958	0.941	94.9	0.554	0.976
UPUC	0.936	0.917	92.6	0.617	0.966

Table 2: Performance of Word Segmentation on Closed Tracks in the SIGHAN Bakeoff 03

Task	A	P	R	F
MSRA	0.9743	0.8150	0.7882	79.92
CityU	0.9725	0.8466	0.8061	82.59

Table 3: Performance of NER on Closed Tracks in the SIGHAN Bakeoff 03