Chinese Word Segmentation with Multiple Postprocessors in HIT-IRLab

Huipeng Zhang Ting Liu Jinshan Ma Xiantao Liao
Information Retrieval Lab, Harbin Institute of Technology, Harbin, 150001 CHINA
{zhp,tliu,mjs,taozi}@ir.hit.edu.cn

Abstract

This paper presents the results of the system IRLAS[1] from HIT-IRLab in the Second International Chinese Word Segmentation Bakeoff. IRLAS consists of several basic components and multiple postprocessors. The basic components include basic segmentation, factoid recognition, and named entity recognition. These components maintain a segment graph together. The postprocessors include merging of adjoining words, morphologically derived word recognition, and new word identification. These postprocessors do some modifications on the best word sequence which is selected from the segment graph. Our system participated in the open and closed tracks of PK corpus and ranked #4 and #3 respectively. Our scores were very close to the highest level. It proves that our system has reached the current state of the art.

1 Introduction

IRLAS participated in both the open and closed tracks of PK corpus. The sections below descript in detail the components of our system and the tracks we participated in.

The structure of this paper is as follows. Section 2 presents the system description. Section 3 describes in detail the tracks we participated in. Section 4 gives some experiments and discussions. Section 5 enumerates some external factors that affect our performance. Section 6 gives our conclusion.

2 System Description

2.1 Basic Segmentation

When a line is input into the system, it is first split into sentences separated by period. The reason to split a line into sentences is that in named entity recognition, the processing of several shorter sentences can reach a higher named entity recall rate than that of a long sentence. The reason to split the line only by period is for the simplicity for programming, and the sentences separated by period are short enough to process.

Then every sentence is segmented into single atoms. For example, a sentence like "HIT-IRLab 参加了第二届 SIGHAN 分词评测。" will be segmented as "HIT-IRLab/参/加/了/第/二/届/SIGHAN/分/词/评/测/。".

After atom segmentation, a segment graph is created. The number of nodes in the graph is the number of atoms plus 1, and every atom corresponds to an arc in the graph.

Then all the words in the dictionary[2] that appear in the sentence will be added to the segment graph. The graph contains various information such as the bigram possibility of every word. Figure 1 shows the segment graph of the above sentence after basic segmentation.

2.2 Factoid Recognition

After basic segmentation, a graph with all the atoms and all the words in the dictionary is set up. On this basis, we find out all the factoids

[1] IRLAS is the abbreviation for "Information Retrieval Lab Lexical Analysis System".

[2] The dictionary is trained with training corpus.

Figure 1: The segment graph
Note: the probability of each word is not shown in the graph.

such as numbers, times and e-mails with a set of rules. Then, we also add all these factoids to the segment graph.

2.3 Named Entity Recognition

Then we will recognize the named entities such as persons and locations. First, we select N^3 best paths from the segment graph with Dijkstra algorithm. Then for every path of the N+1 paths[4] (N best paths and the atom path), we perform a process of Roles Tagging with HMM model (Zhang et al. 2003). The process of it is much like that of Part of Speech Tagging. Then with the best role sequence of every path, we can find out all the named entities and add them to the segment graph as usual. Take the sentence "张会鹏是个好学生。" for example. After basic segmentation and factoid recognition, the N+1 paths are as follows:

张/会/鹏/是/个/好/学生/。
张/会/鹏/是/个/好/学/生/。

Then for each path, the process of Roles Tagging is performed and the following role sequences are generated:

X/S/W/N/O/O/O/O[5]
X/S/W/N/O/O/O/O/O

From these role sequences, we can find out that "XSW" is a 3-character Chinese name. So the word "张会鹏" is recognized as a person name and be added to the segment graph.

2.4 Merging of Adjoining Words

After the steps above, the segment graph is completed and a best word sequence is generated with Dijkstra algorithm. This merging operation and all the following operations are done to the best word sequence.

There are many inconsistencies in the PK corpus. For example, in PK training corpus, the word "就是" sometimes is considered as one word, but sometimes is considered as two separate words as "就 是". The inconsistencies lower the system's performance to some extent.

To solve this problem, we first train from the training corpus the probability of a word to be one word and the probability to be two separate words. Then we perform a process of merging: if two adjoining words in the best word sequence are more likely to be one word, then we just merge them together.

2.5 Morphologically Derived Word Recognition

To deal with the words with the postfix like "性", "者", "率" and so on, we perform the process to merge the preceding word and the postfix into one word. We train a list of postfixes from the training corpus. Then we scan the best word sequence, if there is a single character word that appears in the postfix list, we merge the preceding word and this postfix into one word. For example, a best word sequence like "长跑 者 身 披 彩带" will be converted to "长跑者 身 披 彩带" after this operation.

2.6 New Word Identification

As for the words that are not in the dictionary and cannot be identified with the steps above, we perform a process of New Word Identification (NWI). We train from the training corpus the probability of a word to be independent and the probability to be a special part of another word. In our system, we only consider the words that have one or two characters. Then we scan

[3] N is a constant which is 8 in our system.
[4] It may be smaller than N+1 if the sentence is short enough; exactly, N+1 is the upper bound of the path number.
[5] X, S, W, N and O are all roles for person name recognition, X is surname, S is the first character of given name, W is the second character of given name, N is the word following a person name, and O is other remote context. We defined 17 roles for person name recognition and 10 roles for location name recognition.

the best word sequence, if the product of the probabilities of two adjoining words exceed a threshold, then we merge the two words into one word.

Take the word "甘薯" for example. It is segmented as "甘 薯" after all the above steps since this word is not in the dictionary. We find that the word "甘" has a probability of 0.83 to be the first character of a two character word, and the word "薯" has a probability of 0.94 to be the last character of a two character word. The product of them is 0.78 which is larger than 0.65, which is the threshold in our system. So the word "甘薯" is recognized as a single word.

3 Tracks

3.1 Closed Track

As for the PK closed track, we first extract all the common words and tokens from the training corpus and set up a dictionary of 55,335 entries. Then we extract every kind of named entity respectively. With these named entities, we train parameters for Roles Tagging. We also train all the other parameters mentioned in Section 2 with the training corpus.

3.2 Open Track

The PK open track is similar to closed one. In open track, we use all the 6 months corpus of *People's Daily* and set up a dictionary of 107,749 entries. Additionally, we find 101 new words from the Web and add them to the dictionary. We train the parameters of named entity recognition with a person list and a location list in our laboratory. The training of other parameters is the same with closed track.

4 Experiments and Discussions

We do several experiments on PK test corpus to see the contribution of each postprocessor. We cut off one postprocessor at a time from the complete system and record its F-score. The evaluation results are shown in Table 1. In the table, MDW represents Morphologically Derived Word Recognition, and NWI represents New Word Identification.

	PK open	PK closed
Complete System	96.5%	94.9%
Without Merging	96.3%	94.7%
Without MDW	96.6%	94.4%
Without NWI	96.5%	94.9%

Table 1: Evaluation results of IRLAS with each postprocessor cut off at a time

From Table 1, we can come to some interesting facts:

- The Merging of Adjoining Words has good effect on both open and closed tracks. So we can conclude that this module can solve the problem of inconsistent training corpus to some extent.
- Morphologically Derived Word Recognition does some harm in open track, but it has a very good effect in closed track. Maybe it is because that in open track, we can make a comparatively larger dictionary since we can use any resource we have. So most MDWs[6] are in the dictionary and the MDWs that are not in the dictionary are mostly difficult to recognize. So it does more harm than good in many cases. But in closed track, we have a small dictionary and many common MDWs are not in the dictionary. So it does much more good in closed track.
- New Word Identification is minimal in both open and closed tracks. Maybe it is because that the above steps have recognized the most OOV words and it is hard to recognize any more new words.

5 External Factors That Affect Our Performance

The difference on the definition of words is the main factor that affects our performance. In many cases such as "异彩纷呈", "极大", "世纪颂" are all considered as one word in our system but not so in the PK gold standard corpus. Another factor is the inconsistencies in training corpus, although this problem has been solved to some extent with the module of merging. But

[6] It refers to Morphologically Derived Words.

because the inconsistencies also exist in test corpus and there are some instances that a word is more likely to be a single word in training corpus but more likely to be separated into two words in test corpus. For example, the word "紧跟" is more likely to be a single word in training corpus but is more likely to be separated into two words in test corpus. There is another factor that affects MDW, many postfixes in our system are not considered as postfixes in PK gold standard corpus. For example, the word "太空港" is recognized as a MDW in our system since "港" is a postfix, however, it is segmented into two separate words as "太空 港" in PK gold standard corpus.

6 Conclusion

Through the second SIGHAN bakeoff, we find the segmentation model and the algorithm in our system is effective and the multiple postprocessors we use can also enhance the performance of our system. At the same time, we also find some problems of us. It also has potential for us to improve our system. Take MDW for example, we can make use of more features such as the POS and the length of the preceding word to enhance the recall and precision rate. The bakeoff points out the direction for us to improve our system.

References

Huaping Zhang, Qun Liu, Hongkui Yu, Xueqi Cheng, Shuo Bai, *Chinese Named Entity Recognition Using Role Model*. International Journal of Computational Linguistics and Chinese Language Processing, 2003, Vol.8(2)

Andi Wu, Zixin Jiang, 2000. *Statistically-Enhanced New Word Identification in a Rule-Based Chinese System*. In Proceedings of the Second Chinese Language Processing Workshop, pp. 46-51, HKUST, Hong Kong.

Huaping Zhang, Hongkui Yu, Deyi Xiong, Qun Liu, *HHMM-based Chinese Lexical Analyzer ICTCLAS*. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 11-12, 2003, Sapporo, Japan.

Aitao Chen, *Chinese Word Segmentation Using Minimal Linguistics Knowledge*. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 11-12, 2003, Sapporo, Japan.

Andi Wu, *Chinese Word Segmentation in MSR-NLP*. In Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, July 11-12, 2003, Sapporo, Japan.