CORRECTING WORD SEGMENTATION AND PART-OF-SPEECH TAGGING ERRORS FOR CHINESE NAMED ENTITY RECOGNITION

Tianfang Yao Wei Ding Gregor Erbach

Computational Linguistics Department, Saarland University

D-66041Saarbrücken, Germany

yao@coli.uni-sb.de wding@dfki.de gor@acm.org

Abstract: In the exploration of Chinese named entity recognition for a specific domain, the authors found that the errors caused during word segmentation and part-of-speech (POS) tagging have obstructed the improvement of the recognition performance. In order to further enhance recognition recall and precision, the authors propose an error correction approach for Chinese named entity recognition. In the error correction component, transformation-based machine learning is adopted because it is suitable to fix Chinese word segmentation and POS tagging errors and produce effective correcting rules automatically. The Chinese named entity recognition component utilizes Finite-State Cascades which are automatically constructed by POS rules with semantic constraints. A prototype system, CNERS (Chinese Named Entity Recognition System), has been implemented. The experimental result shows that the recognition performance of most named entities have significantly been improved. On the other hand, the system is also fast and reliable.

Key words: information extraction, named entity recognition, machine learning, finite-state cascades

1. INTRODUCTION

Information Extraction (IE) is a key language technology that aims to extract facts from documents. Since the early 90's IE technology has taken a rapid development, driven by the series of Message Understanding Conferences (MUC's) in the government-sponsored TIPSTER program [6]. It is now coming on to the market and is of great significance for information end-user industries of all kinds, especially finance companies, banks, publishers and governments [10]. As we know, named entities (NEs) are an important constituent in natural language sentences. Therefore, NE recognition (NER) is also a fundamental task of IE. In general, Chinese named entities include person name, person title, location name, organization name, product name, time, date, monetary, percentage and so on [3].

Chinese is not a segmented language, so that the words in a sentence must be segmented before they are processed by IE component. Although most papers related with Chinese IE did not deal with the relationship between word segmentation or part-of-speech (POS) tagging and the performance of IE [1], we notice that these errors have obstructed the improvement of NER performance. In order to change this situation, we propose an error correction approach for Chinese NER in this paper. Transformation-based machine learning [2, 5] is adopted in our model because it is suitable to fix Chinese word segmentation and POS tagging errors and produce effective correcting rules automatically. After using this approach, the recognition recall and precision of most named entities have apparently been enhanced.

Figure 1 is the model of our NER. The dotted line shows the flow process for the training texts; while the solid line is one for the testing texts. When training, the texts are segmented and tagged, then the error correction rules are produced and some of them are selected as the regular rules under the appropriate conditions. Thereafter, the errors caused during word segmentation and POS tagging in testing texts can automatically be corrected through utilizing such error correction rules. Among the six NEs, personal name (PN), time word (TW) and location name (LN) are recognized immediately after the error correction; while team name (TN), competition title (CT) and personal identity (PI) will be recognized by NER component.

This paper is organized as follows. Section 2 illustrates the error correction approach for word segmentation and POS tagging. Section 3 briefly gives the outline for NER component. Section 4 introduces the

[1] In [9] the authors have investigated the relationship between word segmentation and information retrieval.

prototype system and shows the experimental results and the appropriate analysis. Finally, section 5 draws the conclusions for this approach.

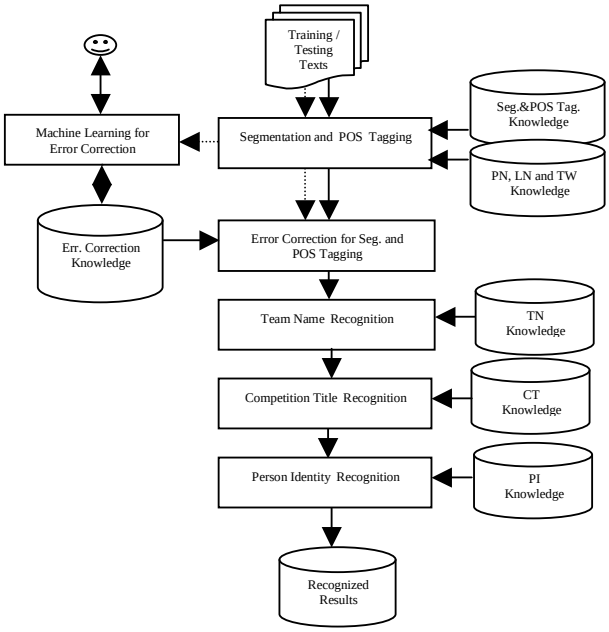

Figure -1. Named Entity Recognition Model

2. CORRECTION FOR WORD SEGMENTATION AND POS TAGGING ERRORS

For the purpose of ensuring good quality in segmenting word and tagging POS, we compared different existing Chinese word segmentation and POS tagging systems and introduced the Modern Chinese Automatic Word Segmentation and POS Tagging System [8] as the first processing component in our model. In this system, the word segmentation unit is based on the Association-Backtracking algorithm and mainly depends on Chinese language knowledge, such as word-building, form-building and syntax; while the POS tagging unit utilizes the probability statistic model as well as CLAWS, VOLSUNGA and the corresponding transmutation algorithms.

Unfortunately, we found there still exist numerous word segmentation and POS tagging errors when we make use of this system to process our

texts on the sports domain. Obviously, these errors will have an effect on the consequent recognition for the NEs.

Some word segmentation and POS tagging errors related with six NE types are shown as follows:

– PN is segmented into PN and general noun or PN is tagged as a LN;
– TW and conjunction are segmented together or TW is tagged as a general noun;
– LN is segmented into general noun and verb or LN is tagged as a PN;
– TN is segmented into some parts with verb and general noun tag or TN is tagged as a PN;
– CT is segmented into some parts with general noun tag or CT keyword is tagged as a verb;
– PI is segmented into PN and general noun or PI is tagged as a verb;

Against these errors, we define some features for machine learning, such as error and correct word segmentation position, error and correct POS tag as well as context including word and POS tag for the rules. On the basis of that, we further define the error correction rule:

rectify_segmentation_error (**concat**, old_word1 | old_tag1 | old_word2 | old_tag2 | …, concat_number, new_tag, preceding_word | preceding_tag, following_word | following_tag)
rectify_segmentation_error (**split**, old_word | old_tag, split_position1 | split_position2 | …, new_tag1 | new_tag2 | …, preceding_word | preceding_tag, following_word | following_tag)
rectify_segmentation_error (**slide**, old_word1 | old_tag1 | old_word2 | old_tag2 | …, slide_direction_length1 | slide_direction_length2 | …, new_tag1 | new_tag2 | …, preceding_word | preceding_tag, following_word | following_tag)
rectify_tag_error (old_word | old_tag, new_tag, preceding_word | preceding_tag, following_word | following_tag)

In the error correction rule of word segmentation, **concat** means some words (or characters) that have been separated will be put together; **split** represents some words (or characters) that have been put together will be separated and **slide** denotes some words (or characters) whose word segmentation positions are not correct will be segmented newly. That is, the new position will be moved to the left or right side of the original position.

The machine learning's procedure includes detecting error positions, producing error correction rules, selecting higher-score rules, ordering rules etc. The concrete algorithm is explained as follows:

a) Compare automatic word segmentation and POS tagging with manual word segmentation and POS tagging in a sentence.

b) If they are different, record word segmentation and POS tagging environments. Otherwise transfer to f).
c) Build a new transformation rule that consists of transformation condition and action. The condition presents all triggering environment including error word segmentation position, error POS and context. The action executes correcting action that transforms word segmentation positions and POS tags.
d) Examine whether this new transformation rule is at odds with the transformation rules in the candidate rule library. If it is true, either merge rules or delete this new rule depending on both conditions. Otherwise add the new rule into the library.
e) Test the rules in the candidate rule library and choose some higher-score transformation rules that can reduce more errors. Then determine whether they are added into the final rule library.
f) If there is still sentence to be processed in a text, transfer to a).
g) Order the rules depending on their score.

The following are some examples from error correction rules:

Ex1. rectify_segmentation_error (**concat**, 莫晨|N4|月|N, 1, N4, 前锋|N, 在|P)
Ex2. rectify_segmentation_error (**split**, 本周日和|N5, 1|3, R|T|C, 参加|V, 阿曼|N7)
Ex3. rectify_segmentation_error (**slide**, 宏|G|远门|N|将|D, right1|right1, N|N, 使|P, 猝不及防|I)
Ex4. rectify_tag_error (赛|V, N, 小组|N, 时|N)

Here note that D, G, I, N, N4, N5, N7 and P represent an adverb, a morpheme, an idiom, a general noun, a PN, a LN, a transliterated PN or LN and a proposition respectively.

Considering the requirements of context constraints for different rules, we divide the rules into three rule types, that is, whole context sensitive, preceding context sensitive and whole context free, manually. Hence, this prevents new errors caused by using error correction rules. The algorithm applied to correct errors is given as follows:
a) Input a sentence by automatic word segmentation and POS tagging.
b) Retrieve the transformation rule library in such sequence: whole POS context constraints, preceding POS context constraints and without context constraints. If one of rules in rule library is matched, execute the corresponding transformation action.
c) Correction of word segmentation errors precedes correction of POS tagging errors.
d) If there is still sentence to be processed, transfer to a).

For example, the above rule for correcting segmentation error (Ex.3) is applied to the sentence with the corresponding errors, the rectified sentence is shown as follows:

使|P|宏远|N|门将|N|猝不及防|I

3. NAMED ENTITY RECOGNITION

We make use of Finite-State Cascades (FSC) [1] as analysis mechanism for NER in our system. FSC is automatically constructed by the POS rule set with the semantic constraints. It consists of three levels. Each level has a NE recognizer, that is TN, CT and PI recognizer. In recognition, if two rules all are matched, we select maximum length match as final match.

The basic recognition procedure is described in a following example:

上海|N5|申花|N|队|N|在|P|百事可乐|N|甲|N|A|QT|联赛|N|中|F|击败|V|对手|N|吉林|N5|敖东|N|队|N|。|W|

Shanghai Shenhua Team defeated the opponent – Jilin Aodong Team in the Pepsi First A League Matches.

```
L3   ------------TN  P   -----------------CT  N   V     PI     -----------TN  W
L2   ------------TN  P   -----------------CT  N   V     N      -----------TN  W
L1   ------------TN  P    N     N  QT  N   N   V     N      -----------TN  W
L0    N5   N   N    P    N     N  QT  N   N   V     N     N5    N    N  W
     上海申花 队 在 百事可乐甲 A联赛中 击败 对手 吉林 敖东 队 。
```

L_i is a level of FSC, which corresponds to a NE recognizer. That is, TN, CT and PI recognizer are located on L_1, L_2 and L_3 respectively. The shadow under the word or phrase means that it is a NE.

Sometimes there is no keywords combined with TN or CT. For such situation, domain verbs are collected and verb valency [7] is applied to analyze the constituents in sentences. Additionally, we use TN and CT context clues to determine whether the current entity is one of them.

4. EXPERIMENTAL RESULT

The Chinese Named Entity Recognition System (CNERS) has been implemented with Java 2 (ver.1.4.0) under Windows 2000. The recognized text can be entered from disk or directly downloaded from WWW. HowNet Knowledge Dictionary [4] is used to provide English and concept explanation of Chinese words in the recognized results. The system, which has been tested on a big corpus in the sports domain, is fast and reliable.

Moreover, 20 Web news about football sports from Jie Fang Daily (http://www.jfdaily.com/) in May 2002 have randomly been chosen and tested. Recognition results are compared with and without error correction. Average recall and precision are shown in Figure 2 and 3.

The experimental result has indicated that the performance for most of NEs in our system has been improved, the average recall and precision of six NEs are increased by 14%.

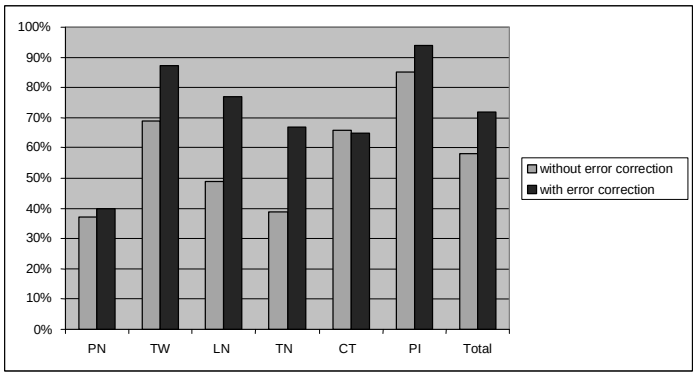

Figure -2. Recall Comparison

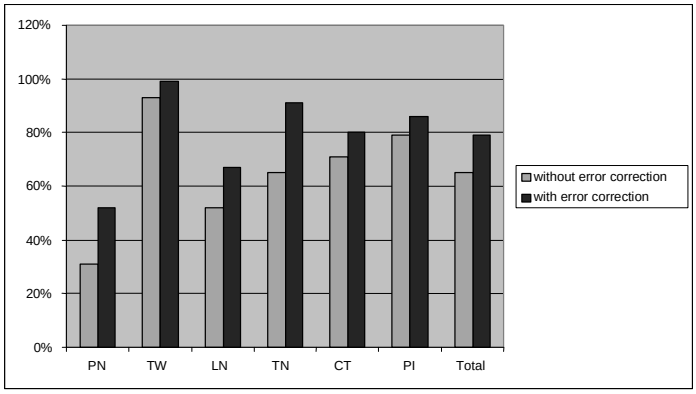

Figure -3. Precision Comparison

But the result has also revealed the recall and precision of PN are still lower than other NEs. The reason is that a Chinese name can be combined with nearly every Chinese character, so that the right boundary of a name is difficult to determine. Therefore, the rules from machine learning can not cover most of name's errors.

5. CONCLUSIONS

In Chinese IE investigation, we note that the errors from word segmentation and POS tagging have adversely affected the performance of NER to a certain extent. We utilize a machine learning technique to perform error correction for word segmentation and POS tagging in Chinese texts before NER is done and improve the recognition performance for most NEs. In addition, FSC is used as an analysis mechanism for Chinese NER, it is suitable and reliable. Such a hybrid approach used in our system synthesizes the advantages of knowledge engineering and machine learning.

ACKNOWLEDGEMENTS

This work is a part of the COLLATE (Computational Linguistics and Language Technology for Real World Applications) project under contract no. 01INA01B, which is being supported by the German Ministry for Education and Research.

REFERENCES

1. Abney S. Partial Parsing via Finite-State Cascades. In Proceedings of the ESSLLI '96 Robust Parsing Workshop. Prague, Czech Republic, 1996.
2. Brill E. Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics. Vol. 21, No. 4, 1995.
3. Chen H.H. et al. Description of the NTU System Used for MET2. Proceedings of 7th Message Understanding Conference, Fairfax, VA, U.S.A., 1998.
4. Dong Z.D. and Dong Q. HowNet. http://www.keenage.com/zhiwang/e_zhiwang.html, 2000.
5. Hockenmaier J. and Brew C. Error-Driven Learning of Chinese Word Segmentation. Communications of COLIPS 8 (1), 1998.
6. Kameyama M. Information Extraction across Linguistic Barriers. In AAAI Spring Symposium on Cross-Language Text and Speech Processing, 1997.
7. Lin X.G. et al. Dictionary of Verbs in Contemporary Chinese. Beijing Language and Culture University Press. Beijing China, (In Chinese), 1994.
8. Liu K.Y. Automatic Segmentation and Tagging for Chinese Text. The Commercial Press. Beijing, China. (In Chinese), 2000.
9. Palmer D. and Burger J. Chinese Word Segmentation and Information Retrieval. In AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Electronic Working Notes, 1997.
10. Wilks Y. Information Extraction as a Core Language Technology. In Maria Teresa Pazienza editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, LNAI 1299, pages 1-9. Springer, 1997.