Chinese Named Entity and Relation Identification System

Tianfang Yao
Department of Computer Science and Engineering
Shanghai Jiao Tong University
Shanghai, 200030, China
yao-tf@cs.sjtu.edu.cn

Hans Uszkoreit
Department of Computational Linguistics and Phonetics
Saarland University
Saarbrücken, 66041, Germany
uszkoreit@coli.uni-sb.de

Abstract

In this interactive presentation, a Chinese named entity and relation identification system is demonstrated. The domain-specific system has a three-stage pipeline architecture which includes word segmentation and part-of-speech (POS) tagging, named entity recognition, and named entity relation identitfication. The experimental results have shown that the average F-measure for word segmentation and POS tagging after correcting errors achieves 92.86 and 90.01 separately. Moreover, the overall average F-measure for 6 kinds of name entities and 14 kinds of named entity relations is 83.08% and 70.46% respectively.

1 Introduction

The investigation for Chinese information extraction is one of the topics of the project COLLATE (DFKI, 2002) dedicated to building up the German Competence Center for Language Technology. The presented work aims at investigating automatic identification of Chinese named entities (NEs) and their relations in a specific domain.

Information Extraction (IE) is an innovative language technology for accurately acquiring crucial information from documents. NE recognition is a fundamental IE task, that detects some named constituents in sentences, for instance names of persons, places, organizations, dates, times, and so on. Based on NE recognition, the identification of Named Entity Relation (NER) can indicate the types of semantic relationships between identified NEs. e.g., relationships between person and employed organization; person and residing place; person and birthday; organization and seat, etc. The identified results for NEs and NERs can be provided as a resource for other application systems such as question-answering system. Therefore, these two IE tasks are selected as our investigation emphases.

Chinese has a very different structure from western languages. For example, it has a large character set involving more than 48,000 characters; there is no space between words in written texts; and Chinese words have fewer inflections, etc. In the past twenty years there have been significant achievements in IE concerning western languages such as English. Comparing with that, the research on the relevant properties of Chinese for IE, especially for NER, is still insufficient.

Our research focuses on domain-specific IE. We picked the sports domain, particularly, texts on soccer matches because the number and types of entities, relations and linguistic structures are representative for many applications.

Based on the motivations above mentioned, our goals for the design and implementation of the prototype system called CHINERIS (Chinese Named Entity and Relation Identification System) are:

- Establishing an IE computational model for Chinese web texts using hybrid technologies, which should to a great extent meet the requirements of IE for Chinese web texts;
- Implementing a prototype system based on this IE computational model, which extracts information from Chinese web texts as accurately and quickly as possible;
- Evaluating the performance of this sys tem in a specific domain.

Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 37–40,
Sydney, July 2006.

2 System Design

In the model, the IE processing is divided into three stages: (i) word segmentation and part-of-speech (POS) tagging; (ii) NE recognition; (iii) NER identification. Figure 1 demonstrates a Chinese IE computational model comprised of these three stages. Each component in the system corresponds to a stage.

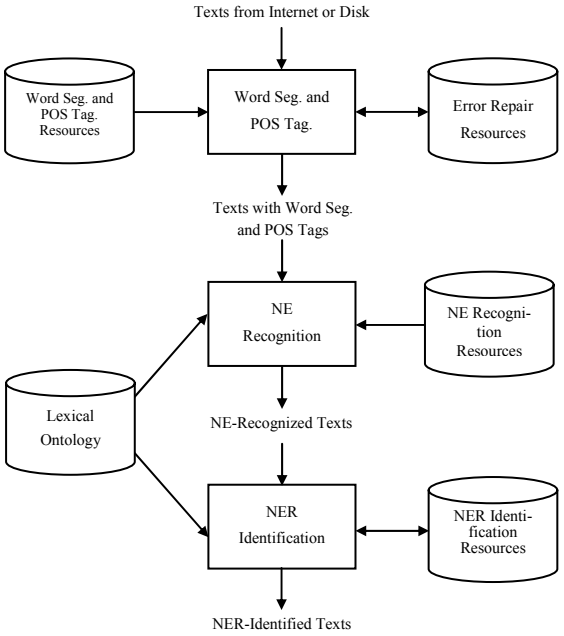

Figure 1. A three-stage Chinese IE computational model.

In general, the accuracy of the first stage has considerable influence on the performance of the consequent two stages. It has been demonstrated by our experiments (Yao et al., 2002). In order to reduce unfavorable influence, we utilize a trainable approach (Brill, 1995) to automatically generate effective rules, by which the first component can repair different errors caused by word segmentation and POS tagging.

At the second stage, there are two kinds of NE constructions to be processed (Yao et al., 2003). One is the NEs which involve trigger words; the other those without trigger words. For the former NEs, a shallow parsing mechanism, i.e., finite-state cascades (FSC) (Abney, 1996) which are automatically constructed by sets of NE recognition rules, is adopted for reliably identifying different categories of NEs. For the latter NEs, however, some special strategies, such as the valence constraints of domain verbs, the constituent analysis of NE candidates, the global context clues and the analysis for preposition objects etc., are designed for identifying them.

After the recognition for NEs, NER identification is performed in the last stage. Because of the diversity and complexity of NERs, at the same time, considering portability requirement in the identification, we suggest a novel supervised machine learning approach called positive and negative case-based learning (PNCBL) used in this stage (Yao and Uszkoreit, 2005).

The learning in this approach is a variant of memory-based learning (Daelemans et al., 2000). The goal of that is to capture valuable information from NER and non-NER patterns, which is implicated in different features. Because not all features we predefine are necessary for each NER or non-NER, we should select them by a reasonable measure mode. According to the selection criterion we propose - self-similarity, which is a quantitative measure for the concentrative degree of the same kind of NERs or non-NERs in the corresponding pattern library, the effective feature sets - General-Character Feature (GCF) sets for NERs and Individual-Character Feature (ICF) sets for non-NERs are built. Moreover, the GCF and ICF feature weighting serve as a proportion determination of feature's degree of importance for identifying NERs against non-NERs. Subsequently, identification thresholds can also be determined.

Therefore, this approach pursues the improvement of the identification performance for NERs by simultaneously learning two opposite cases, automatically selecting effective multi-level linguistic features from a predefined feature set for each NER and non-NER, and optimally making an identification tradeoff. Further, two other strategies, resolving relationship conflicts and inferring missing relationships, are also integrated in this stage.

Considering the actual requirements for domain knowledge, we defined a hierarchical taxonomy and constructed conceptual relationships among Object, Movement and Property concept categories under the taxonomy in a lexical sports ontology (Yao, 2005). Thus, this ontology can be used for the recognition of NEs with special constructions - without trigger words, the determination of NE boundaries, and the provision of feature values as well as the computation of the semantic distance for two concepts during the identification of NERs.

3 System Implementation

During the implementation, object-oriented design and programming methods are thoroughly

used in the system development. In order to avoid repeated development, we integrate other application system and resource, e.g., Modern Chinese Word Segmentation and POS Tagging System (Liu, 2000) and HowNet (Dong and Dong, 2000) into the system. Additionally, we utilize Protégé-2000 (version 1.9) (Stanford Medical Informatics, 2003) as a development environment for the implementation of lexical sports ontology.

The prototype system CHINERIS has been implemented in Java. The system can automatically identify 6 types of NEs[1] and 14 types of NERs[2] in the sports domain. Furthermore, its run-time efficiency is acceptable and the system user interfaces are friendly.

4 Testing and Evaluation

We have finished three experiments for testing three components. Table 1 shows the experimental results for the performance of these components.

Stage	Task	(Total) Ave. Rec.	(Total) Ave. Pre.	(Total) Ave. F-M
1st	Word Seg.	95.08	90.74	92.86
	POS Tag.	92.39	87.75	90.01
2nd	NE Ident.	83.38	82.79	83.08
3rd	NER Ident.	78.50	63.92	70.46

Table 1. Performance for the System CHINERIS.

In the first experiment, the training set consists of 94 texts including 3473 sentences collected from the soccer matches of the Jie Fang Daily (http://www.jfdaily.com/) in 2001. During manual error-correction, we adopted a double-person annotation method. After training, we obtain error repair rules. They can repair at least one error in the training corpus. The rules in the rule library are ranked according to the errors they correct. The testing set is a separate set that contains 20 texts including 658 sentences. The texts in the testing set have been randomly chosen from the Jie Fang Daily from May 2002. In the testing, the usage of error repair rules with context constraints has priority over those without context constraints, and the usage of error repair rules for word segmentation has priority over those for POS tagging. Through experimental observation, this processing sequence can ensure that the rules repair many more errors. On the other hand, it can prevent new errors occurring during the repair of existing errors. The results indicate that after the correction, the average F-measure of word segmentation has increased from 87.75 % to 92.86%; while that of POS tagging has even increased from 77.47% to 90.01%. That is to say, the performance of both processes has been distinctly enhanced.

In the second experiment, we utilize the same testing set for the error repair component to check the named entity identification which includes regular and special entity constructions. The rule sets provided for TN, CT, and PI recognition have 35, 50, and 20 rules respectively. In lexical sports ontology, there are more than 350 domain verbs used for the identification of TN with special constructions. Among six NEs, the average F-measure of DT, PI, and CT exceeds 85%. Therefore, it specifies that the identification performance of named entities after adding the special recognition strategies in this component has reached a good level.

In the third experiment, both pattern libraries are established in terms of the annotated texts and lexical sports ontology during learning. They have 142 (534 NERs) and 98 (572 non-NERs) sentence groups respectively. To test the performance of our approach, we randomly choose 32 sentence groups from the Jie Fang Daily in 2002 (these sentence groups are out of either NER or non-NER pattern library), which embody 117 different NER candidates. Table 1 shows the total average recall, precision, and F-measure for 14 different NERs by positive and negative case-based learning and identification. Among 14 types of NERs, the highest total average F-measure is 95.65 from the relation LOC_CPC and the lowest total average F-measure is 34.09 from TM_CPC. The total average F-measure is 70.46. In addition, we also compared the performance between the total average recall, precision, and F-measure for all NERs only by positive and by positive and negative case-based learning and identification separately. It shows the total average F-measure is enhanced from 63.61% to 70.46% as a whole,

[1] Personal Name (PN); Date or Time (DT); Location Name (LN); Team Name (TN); Competition Title (CT); Personal Identity (PI).

[2] Person ↔ Team (PS_TM); Person ↔ Competition (PS_CP); Person ↔ City / Province / Country (PS_CPC); Person ↔ Identification (PS_ID); Home Team ↔ Visiting Team (HT_VT); Winning Team ↔ Losing Team (WT_LT); Draw Team ↔ Draw Team (DT_DT); Team ↔ Competition (TM_CP); Team ↔ City / Province / Country (TM_CPC); Identification ↔ Team (ID_TM); Competition ↔ Date (CP_DA); Competition ↔ Time (CP_TI); Competition ↔ Location (CP_LOC); Location ↔ City / Province / Country (LOC_ CPC).

due to the adoption of both positive and negative cases.

From the result, we also realize that the selection of relation features is critical. First, they should be selected from multiple linguistic levels, e.g., morphology, syntax and semantics. Second, they should also embody the crucial information of Chinese language processing, such as word order, the context of words, and particles etc. Moreover, the proposed self-similarity is a reasonable measure for selecting GCF and ICF for NERs and non-NERs identification respectively.

5 Conclusion

This three-stage IE prototype system CHINERIS is appropriate and effective for Chinese named entity and relation identification in sports domain.

In the first component, it is a beneficial exploration to develop an error repairer which simultaneously enhances the performance of Chinese word segmentation and POS tagging.

In the second component, we theoretically extend the original definition of Finite State Automata (FSA), that is, we use complex constraint symbols rather than atomic constraint symbols. With this extension, we improve the practicability for the FSC mechanism. At the same time, the new issue for automatically constructing FSC also increases the flexibility of its maintenance. In order to improve the NE identification performance, some special strategies for the identification of NEs without trigger words are added in this stage, which cannot be recognized by FSC.

In the third component, automatically selecting effectual multi-level linguistic features for each NER and non-NER and learning two opposite types of cases simultaneously are two innovative points in the PNCBL approach.

The lexical sports ontology plays an important role in the identification of NEs and NERs, such as determination of the boundary of NEs, identification for NE with special constructons and calculation of similarity for the features (e.g. semantic distance).

The experimental results for the three components in the prototype system show that the system CHINERIS is successful for the sample application.

Acknowledgement

This work is a part of the COLLATE project under contract no. 01INA01B, which is supported by the German Ministry for Education and Research.

References

S. Abney. 1996. *Partial Parsing via Finite-State Cascades*. In Proceedings of the ESSLLI '96 Robust Parsing Workshop, pages 8-15. Prague, Czech Republic.

E. Brill. 1995. *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging*. Computational Linguistics, 21(4): 543-565.

W. Daelemans, A. Bosch, J. Zavrel, K. Van der Sloot, and A. Vanden Bosch. 2000. *TiMBL: Tilburg Memory Based Learner*, Version 3.0, Reference Guide. Technical Report ILK-00-01, ILK, Tilburg University. Tilburg, The Netherlands. http://ilk.kub.nl/~ilk/papers/ilk0001.ps.gz.

DFKI. 2002. *COLLATE: Computational Linguistics and Language Technology for Real Life Applications*. DFKI, Saarbrücken, Germany. http://collate.dfki. de/.

Z. Dong and Q. Dong. 2000. *HowNet*. http://www.keenage.com/zhiwang/e_zhiwang.html.

K. Liu. 2000. *Automatic Segmentation and Tagging for Chinese Text*. The Commercial Press. Beijing, China.

Stanford Medical Informatics. 2003. *The Protégé Ontology Editor and Knowledge Acquisition System*. The School of Medicine, Stanford University. Stanford, USA. http://protege.stanford.edu/.

T. Yao, W. Ding, and G. Erbach. 2002. *Correcting Word Segmentation and Part-of-Speech Tagging Errors for Chinese Named Entity Recognition*. In G. Hommel and H. Sheng, editors, The Internet Challenge: Technology and Applications, pages 29-36. Kluwer Academic Publishers. The Netherlands.

T. Yao, W. Ding and G. Erbach. 2003. *CHINERS: A Chinese Named Entity Recognition System for the Sports Domain*. In: Proc. of the Second SIGHAN Workshop on Chinese Language Processing (ACL 2003 Workshop), pages 55-62. Sapporo, Japan.

T. Yao and H. Uszkoreit. 2005. *A Novel Machine Learning Approach for the Identification of Named Entity Relations*. In: Proc. of the Workshop on Feature Engineering for Machine Learning in Natural Language Processing (ACL 2005 Workshop), pages 1-8. Michigan, USA.

T. Yao. 2005. *A Lexical Ontology for Chinese Information Extraction*. In M. Sun and Q. Chen, editors, Proc. of the 8th National Joint Symposium on Computational Linguistics (JSCL-2005), pages 241-246. Nanjing, China.