## **Extracting Chinese Entity Names and Their Relations**

Yimin ZHANG

NLP/IR, Intel China Research Center 6F1, North Kerry Center 100020 Beijing, China, yimin.zhang@intel.com

## Abstract

We built an entity extraction tool that uses machine-learned rules to locate and determine entity names in natural running text, such as people, companies, products, and places. It also uses machine-learned rules to identify relations between those Entities, e.g., the relation between a product and a company, a person and a company, etc. The key techniques used are the identification of Chinese NP chunks and machine learning methods adapted from dictionary induction systems for information extraction, such as Crystal and Rapier.

The Chinese NP chunking is based on cascaded finite state automata combined with dependency relations extracted from corpus. In many cases, structural ambiguities inherent in the Chinese language are correctly handled using dependency relations between words. Joe F ZHOU

NLP/IR, Intel China Research Center 6F1, North Kerry Center 100020 Beijing, China, joe.f.zhou@intel.com

The extraction of entity names and relations is rule-based. The rules are learned using machine learning techniques. Coreference resolution is introduced to identify more implicit relations. For example, from "该公司的总裁 (CEO of this company)", we first resolve what "该公司 (this company)" is referred to, then use the rules to extract the relation between the people and the company.

## References

- Califf M. and Mooney R. (1997) Relational learning of pattern-match rules for information extraction. Working Papers of the ACL-97 Workshop in Natural Language Learning, pp9-15.
- Huffman S. (1995) *Learning information extraction patterns from examples*. IJCAI-95 Workshop on new approaches to learning for natural language processing, pp127-142.
- Soderland S., Fisher D., Aseltine, J. and Lehnert W. (1995) *Crystal: Inducing a conceptual dictionary.* Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95), pp1314-1319.
- Dekang Lin. (1997) Using syntactic dependency as local context to resolve word sense ambiguity. In Proceedings of ACL/EACL-97, Madrid, Spain, pp64-71.

A sample screen shot:

LEGEND NP COMPANY PEOPLE TIME LOCATION MONEY	•
微軟 进军 B2B 电子窗务领域	
4月1日	
《华尔街日报》: 拥挤的 企业对企业 互联网市场有了新的参与者: 微软公司(Microsoft Corp.).	
正当投資者和信葉服許公司沉醉于通过互联同的 企业对企业 (B2B) 商务而产生的潜在利益时, 微軟 公司发示说它不想更身性外。 总裁 史谱夫 已含恶(Stave Ballaer)复带设在主要能的Redaund公司将参 和党工业 方面的基于互联问的80次总, 其中设有公司(Bonna Cop.)考定出来和武仪公司期关考部杯4.	
到目前为止,在这个舞台上还很少见到教教的身影。但鲍尔默表示该公司将发表更多的类似声明。教 教已经开始将火力集中在Oracle公司及其长期的对手Redwood Shores公司,该公司为B2B 方面的领先者。	
B28是科技领域中最热门的部分,微软在该领域的参与表明,该公司在努力了结政府关于教软在互联网 浏览器和J ava编程语言 方面的反托拉斯条的同时,已经完全投入了今天的这场窗业大战中。	
The extracted relations:	
微軟公司 总裁 史蒂夫. 巴尔默	
Oracle公司 对手 Redwood Shores公司	-
微软公司 别称 微软	