# Chinese Open Relation Extraction for Knowledge Acquisition

**Yuen-Hsien Tseng[1], Lung-Hao Lee[1,2], Shu-Yen Lin[1], Bo-Shun Liao[1], Mei-Jun Liu[1], Hsin-Hsi Chen[2], Oren Etzioni[3], Anthony Fader[4]**

[1]Information Technology Center, National Taiwan Normal University
[2]Dept. of Computer Science and Information Engineering, National Taiwan University
[3]Allen Institute for Artificial Intelligence, Seattle, WA
[4]Dept. of Computer Science and Engineering, University of Washington
{samtseng, lhlee, sylin, skylock, meijun}@ntnu.edu.tw,
hhchen@ntu.edu.tw, OrenE@allenai.org, afader@cs.washington.edu

## Abstract

This study presents the Chinese Open Relation Extraction (CORE) system that is able to extract entity-relation triples from Chinese free texts based on a series of NLP techniques, *i.e.,* word segmentation, POS tagging, syntactic parsing, and extraction rules. We employ the proposed CORE techniques to extract more than 13 million entity-relations for an open domain question answering application. To our best knowledge, CORE is the first Chinese Open IE system for knowledge acquisition.

## 1 Introduction

Traditional Information Extraction (IE) involves human intervention of handcrafted rules or tagged examples as the input for machine learning to recognize the assertion of a particular relationship between two entities in texts (Riloff, 1996; Soderland, 1999). Although machine learning helps enumerate potential relation patterns for extraction, this approach is often limited to extracting the relation sets that are predefined. In addition, traditional IE has focused on satisfying pre-specified requests from small homogeneous corpora, leaving the question open whether it can scale up to massive and heterogeneous corpora such as the Web (Banko and Etzioni, 2008; Etzioni et al., 2008, 2011).

Open IE, a new domain-independent knowledge discovery paradigm that extracts a diverse set of relations without requiring any relation-specific human inputs and a pre-specified vocabulary, is especially suited to massive text corpora, where target relations are unknown in advance. Several Open IE systems, such as TextRunner (Banko et al., 2007), WOE (Wu and Weld, 2010), ReVerb (Fader et al., 2011), and OLLIE (Mausam et al., 2012) achieve promising performance in open relation extraction on English sentences. However, application of these systems poses challenges to those languages that are very different from English, such as Chinese, as grammatical functions in English and Chinese are realized in markedly different ways. It is not sure whether those techniques for English still work for Chinese. This issue motivates us to extend the state-of-the-art Open IE systems to extract relations from Chinese texts.

The relatively rich morpho-syntactic marking system of English (e.g., verbal inflection, nominal case, clausal markers) makes the syntactic roles of many words detectable from their surface forms. A tensed verb in English, for example, generally indicates its main verb status of a clause. The pinning down of the main verb in a Chinese clause, on the other hand, must rely on other linguistic cues such as word context due to the lack of tense markers. In contrast to the syntax-oriented English language, Chinese is discourse-oriented and rich in ellipsis – meaning is often construable in the absence of explicit linguistic devices such that many obligatory grammatical categories (e.g., pronouns and BE verbs) can be elided in Chinese. For example, the three Chinese sentences “蘋果營養豐富” (‘Apples nutritious’), “蘋果是營養豐富的” (‘Apples are nutritious’), and “蘋果富含營養”



*Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 12–16,
Gothenburg, Sweden, April 26-30 2014. 

('Apples are rich in nutrition') are semantically synonymous sentences, but the first one, which lacks an overt verb, is used far more often than the other two. Presumably, an adequate multilingual IE system must take into account those intrinsic differences between languages.

This paper introduces the Chinese Open Relation Extraction (CORE) system, which utilizes a series of NLP techniques to extract relations embedded in Chinese sentences. Given a Chinese text as the input, CORE employs word segmentation, part-of-speech (POS) tagging, and syntactic parsing, to automatically annotate the Chinese sentences. Based on this rich information, the input sentences are chunked and the entity-relation triples are extracted. Our evaluation shows the effectiveness of CORE, and its deficiency as well.

## 2 Related Work

TextRunner (Banko et al., 2007) was the first Open IE system, which trains a Naïve Bayes classifier with POS and NP-chunk features to extract relationships between entities. The subsequent work showed that employing the classifiers capable of modeling the sequential information inherited in the texts, like linear-chain CRF (Banko and Etzioni, 2008) and Markov Logic Network (Zhu et al., 2009), can result in better extraction performance. The WOE system (Wu and Weld, 2010) adopted Wikipedia as the training source for their extractor. Experimental results indicated that parsed dependency features lead to further improvements over TextRunner.

ReVerb (Fader et al., 2011) introduced another approach by identifying first a verb-centered relational phrase that satisfies their pre-defined syntactic and lexical constraints, and then split the input sentence into an Argument-Verb-Argument triple. This approach involves only POS tagging for English and "regular expression"-like matching. As such, it is suitable for large corpora, and likely to be applicable to Chinese.

For multilingual open IE, Gamallo et al. (2012) adopts a rule-based dependency parser to extract relations represented in English, Spanish, Portuguese, and Galician. For each parsed sentence, they separate each verbal clause and then identify each one's verb participants, including their functions: subject, direct object, attribute, and prepositional complements. A set of rules is then applied on the clause constituents to extract the target triples. For Chinese open IE, we adopt a similar general approach. The main differences are the processing steps specific to Chinese language.

## 3 Chinese Open Relation Extraction

This section describes the components of CORE. Not requiring any predefined vocabulary, CORE's sole input is a Chinese corpus and its output is an extracted set of relational tuples. The system consists of three key modules, i.e., word segmentation and POS tagging, syntactic parsing, and entity-relation triple extraction, which are introduced as follows:

Chinese is generally written without word boundaries. As a result, prior to the implementation of most NLP tasks, texts must undergo automatic word segmentation. Automatic Chinese word segmenters are generally trained by an input lexicon and probability models. However, it usually suffers from the unknown word (i.e., the out-of-vocabulary, or OOV) problem. In CORE, a corpus-based learning method to merge the unknown words is adopted to tackle the OOV problem (Chen and Ma, 2002). This is followed by a reliable and cost-effective POS-tagging method to label the segmented words with part-of-speeches (Tsai and Chen, 2004). Take the Chinese sentence "愛迪生發明了燈泡" ('Edison invented the light bulb') for instance. It was segmented and tagged as follows: 愛迪生/Nb 發明/VC 了/Di 燈泡/Na. Among these words, the translation of a foreign proper name "愛迪生" ('Edison') is not likely to be included in a lexicon and therefore is extracted by the unknown word detection method. In this case,

the special POS tag 'Di' is a tag to represent a verb's tense when its character "了" follows immediately after its precedent verb. The complete set of part-of-speech tags is defined in the technical report (CKIP, 1993). In the above sentence, "了" could represent a complete different meaning if it is associate with other character, such as "了解" meaning "understand". Therefore, "愛迪生發明了解藥" ('Edison invented a cure') would be segmented incorrectly once "了" is associated with its following character, instead of its precedent word.

We adopt CKIP, the best-performing parser in the bakeoff of SIGHAN 2012 (Tseng et al., 2012), to do syntactic structure analysis. The CKIP solution re-estimates the context-dependent probability for Chinese parsing and improves the performance of probabilistic context-free grammar (Hsieh et al., 2012). For the example sentence above, '愛迪生/Nb' and '燈泡/Na' were annotated as two nominal phrases (i.e., 'NP'), and '發明/VC 了/Di' was annotated as a verbal phrase (i.e., 'VP').

CKIP parser also adopts dependency decision-making and example-based approaches to label the semantic role "Head", showing the status of a word or a phrase as the pivotal constituent of a sentence (You and Chen, 2004). CORE adopts the *head-driven principle* to identify the main relation in a given sentence (Huang et al., 2000). Firstly, a relation is defined by both the "Head"-labeled verb and the other words in the syntactic chunk headed by the verb. Secondly, the noun phrases preceding/preceded by the relational chunk are regarded as the candidates of the head's arguments. Finally, the entity-relation triple is identified in the form of (entity1, relation, entity2). Regarding the example sentence described above, the triple (愛迪生/Edison, 發明了/invented, 燈泡/light bulb) is extracted by this approach.

Figure 1 shows the parsed tree of a Chinese sentence for the relation extraction by CORE. The Chinese sentence "白宮預算委員會的民主黨星期一發佈報告" ('Democrats on the House Budget Committee released a report on Monday') is the manual translation of one of the English sentences evaluated by ReVerb (Fader et al., 2011). The first step of CORE involves word-segmentation and POS-tagging, thus returning eight word/POS pairs: 白宮/Nc, 預算/Na, 委員會/Nc, 的/DE, 民主黨/Nb, 星期一/Nd, 發佈/VE, 報告/Na. Next, "星期一/Nd 發佈/VE" is identified as the verbal phrase that heads the sentence. This verbal phrase is regarded as the center of a potential relation. The two noun phrases before and after the verbal phrase, *i.e.*, the NP "白宮 預算 委員會 的 民主黨" and NP "報告" are regarded as the entities that complete the relation. A potential entity-relation-entity triple (*i.e.*, 白宮預算委員會的民主黨 / 星期一發佈 / 報告, 'Democrats on the House Budget Committee / on Monday released / a report') is extracted accordingly. This triple is chunked from its original sentence fully automatically. Finally, a filtering process, which retains "Head"-labeled words only, can be applied to strain out from each component of this triple the most prominent word: "民主黨 / 發佈 / 報告" ('Democrats / released / report').

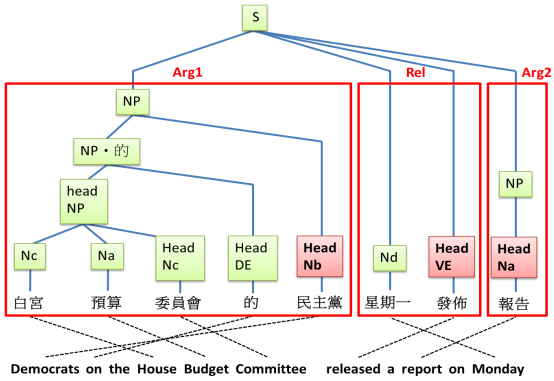


Figure 1: The parsed tree of a Chinese sentence.

## 4 Experiments and Evaluation

We adopted the same test set released by ReVerb for performance evaluation. The test set consists of 500 English sentences randomly sampled from the Web and were annotated using a pooling method. To obtain "gold standard" relation triples in Chinese, the 500 test sentences were manually translated from English to Chinese by a

trained native Chinese speaker and verified by another. Additionally, two other native Chinese speakers annotated the relation triples for each Chinese sentence. In total, 716 Chinese entity-relation triples with an agreement score of 0.79 between the two annotators were obtained and regarded as gold standard.

Performance evaluation of CORE was conducted based on: 1) exact match; and 2) relation-only match. For exact match, each component of the extracted triple must be identical with the gold standard. For relation-only match, the extracted triple is regarded as a correct case if an extracted relation agreed with the relation of the gold standard.

Without another Chinese Open IE system for performance comparison, we compared CORE with a modification of ReVerb system capable of handling Chinese sentences. The modification of ReVerb's verb-driven regular expression matching was kept to a minimum to deal with language-specific processing. As such, ReVerb remains mostly the same as its English counterpart so that a bilingual (Chinese/English) Open IE system can be easily implemented.

Table 1 shows the experimental results. Our CORE system obviously performs better than ReVerb when recall is considered for both exact and relation-only match. The results suggest that utilizing more sophisticated NLP techniques is effective to extract relations without any specific human intervention. In addition, there is a slight decrease in the precision of exact match for CORE. This reveals that ReVerb's original syntactic and lexical constraints are also useful to identify the arguments and their relationship precisely. In summary, CORE achieved relatively promising F1 scores. These results imply that CORE method is more suitable for Chinese open relation extraction.

| Chinese Open IE | | **Precision** | **Recall** | **F1** |
|---|---|---|---|---|
| Exact Match | ReVerb | **0.5820** | 0.0987 | 0.1688 |
| | CORE | 0.5579 | **0.3291** | **0.4140** |
| Relation Only | ReVerb | 0.8361 | 0.1425 | 0.2435 |
| | CORE | **0.8463** | **0.5000** | **0.6286** |

Table 1: Performance evaluation on Chinese Open IE.

We also analyzed the errors made by the CORE model. Almost all the errors resulted from incorrect parsing. Enhancing the parsing effectiveness is most likely to improve the performance of CORE. The relatively low recall rate also indicates that CORE misses many types of relation expression. Ellipsis and flexibility in Chinese syntax are so difficult not only to fail the parser, but also the extraction attempts to bypass the parsing errors.

To demonstrate the applicability of CORE, we implement a Chinese Question-Answering (QA) system based on two million news articles from 2002 to 2009 published by the United Daily News Group (udn.com/NEWS). CORE extracted more than 13 million unique entity-relation triples from this corpus. These extracted relations are useful for knowledge acquisition. Take the question "什麼源自於中國？" ('What is originated from China?') as an example, the relation is automatically identified as "源" ('originate') that heads the following entity "中國" ('China'). Our open QA system then searched the triples and returned the first entity as the answers. In addition to the obvious answer "中醫" ('Chinese medicine'), which is usually considered as common-sense knowledge, we also obtained those that are less known, such as the traditional Japanese food "納豆" ('natto') and the musical instrument "手風琴" ('accordion').

## 5 Conclusions

This work demonstrates the feasibility of extracting relations from Chinese corpus without the input of any predefined vocabulary to IE systems. This work is the first to explore Chinese open relation extraction to our best knowledge.

## Acknowledgments

This research was partially supported by National Science Council, Taiwan under grant NSC102-2221-E-002-103-MY3, and the "Aim for the Top University Project" of National Taiwan Normal University, sponsored by the Ministry of Education, Taiwan.

## References


Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. *Proceedings of EMNLP'11*, pages 1535-1545.

Chu-Ren Huang, Feng-Yi Chen, Keh-Jiann Chen, Zhao-Ming Gao, and Kuang-Yu Chen. 2000. Sinina Treebank: design criteria, annotation guidelines, and on-line interface. *Proceedings of SIGHAN'00*, pages 29-37.

Chinese Knowledge Information Processing (CKIP) Group. 1993. Categorical analysis of Chinese. *ACLCLP Technical Report* # 93-05, Academia Sinica.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. *Proceedings of ACL'10*, pages 118-127.

Jia-Ming You, and Keh-Jiann Chen. 2004. Automatic semantic role assignment for a tree structure. In *Proceedings of SIGHAN'04*, pages 1-8.

Jun Zhu, Zaiqing Nie, Xiaojiang Lium Bo Zhang, and Ji-Rong Wen. 2009. StatSnowball: a statistical approach to extracting entity relationships. In *Proceedings of WWW'09*, pages 101-110.

Keh-Jiann Chen and Wei-Yun Ma. 2002. Unknown word extraction for Chinese documents. In *Proceedings of COLING'02*, pages 169-175.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. *Proceedings of IJCAI'07*, pages 2670-2676.

Michele Banko, and Oren Etzioni. 2008. The tradeoffs between open and traditional relation extraction. *Proceedings of ACL'08*, pages 28-26.

Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam. 2011. Open information extraction: the second generation. In *Proceedings of IJCAI'11*, pages 3-10.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68-74.

Pablo Gamallo, Marcos Garcia, and Santiago Fernández-Lanza. 2012. Dependency-based open information extraction. In *Proceedings of ROBUS-UNSUP'12*, pages 10-18.

Elleen Riloff. 1996. Automatically constructing extraction patterns from untagged text. In *Proceedings of AAAI'96*, pages 1044-1049.

Stephen Soderland. 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34(1-3):233-272.

Yu-Ming Hsieh, Ming-Hong Bai, Jason S. Chang, and Keh-Jiann Chen. 2012. Improving PCFG Chinese Parsing with Context-Dependent Probability Re-estimation. *Proceedings of CLP'12*, pages 216-221.

Yu-Fang Tsai, and Keh-Jiann Chen. 2004. Reliable and cost-effective pos-tagging. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(1):83-96.

Yuen-Hsien Tseng, Lung-Hao Lee, and Liang-Chih Yu 2012. Traditional Chinese parsing evaluation at SIGHAN Bake-offs 2012. *Proceedings of* CLP'12, pages 199-205.