НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
январь–февраль 2018 Том 18 №1 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
January–February 2018 Vol. 18 No 1 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 519.68

МЕТОД АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ОТКРЫТЫХ ОТНОШЕНИЙ ИЗ КИТАЙСКИХ ТЕКСТОВ

Юй Чуцяо[a]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
Адрес для переписки: yuchuqiao123@gmail.com

Информация о статье
Поступила в редакцию 12.12.17, принята к печати 30.12.17
doi: 10.17586/2226-1494-2018-18-1-163-165
Язык статьи – русский

Ссылка для цитирования: Юй Чуцяо. Метод автоматического извлечения открытых отношений из китайских текстов // Научно-технический вестник информационных технологий, механики и оптики. 2018. Т. 18. № 1. С. 163–165. doi: 10.17586/2226-1494-2018-18-1-163-165

В работе исследована проблема извлечения открытых отношений в форме субъект-предикат-объект из китайских текстов. В отличие от общепринятых многофазных методов, включающих сегментацию слов, частеречный и синтаксический анализ, предлагается ролевой подход к выявлению членов предложения без предварительного разбиения последовательности иероглифов на отдельные слова. В основе подхода лежит использование служебных слов, предлогов и послелогов в качестве признаков частей речи и членов предложения. В сочетании со словарем небольшого размера этого достаточно для извлечения фактов по запросу. Проведенные эксперименты на реальном техническом тексте показывают удовлетворительные результаты, сопоставимые с традиционным подходом.

Ключевые слова
извлечение фактов, китайский язык, ролевой подход, анализ текстов, словарь, сегментация предложений, частеречный анализ

A METHOD OF AUTOMATIC OPEN RELATION EXTRACTION FROM CHINESE TEXTS

Yu Chuqiao[a]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
Corresponding author: yuchuqiao123@gmail.com

Article info
Received 12.12.17, accepted 30.12.17
doi: 10.17586/2226-1494-2018-18-1-163-165
Article in Russian
For citation: Yu Chuqiao. A method of automatic open relation extraction from Chinese texts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 1, pp. 163–165 (in Russian). doi: 10.17586/2226-1494-2018-18-1-163-165

Abstract
The paper considers the problem of Chinese Open Relation Extraction represented in a form of subject-predicate-object. In contrary to well-known multi-phase methods including word segmentation, part-of-speech tagging, and syntactic analysis, we propose a role approach to detection of parts of sentences without preliminary word segmentation. The key idea is to use syntactic words, prepositions and postpositions as part of speech and member of sentence attributes. Coupled with a small dictionary, it is enough for facts extraction by a query. The experiments conducted on a real technical text show satisfactory results comparable to a traditional approach.

Keywords
facts extraction, Chinese language, role approach, texts analysis, dictionary, phrase segmentation, part-of-speech tagging

Извлечение фактов из китайских текстов (Chinese Open Relation Extraction) в последние годы является предметом исследования многих авторов. Разработки, предназначенные для алфавитных языков, такие как TextRunner [1], не подходят для китайского языка в силу его особенностей. В китайском языке отсутствуют пробелы между словами, и почти любое сочетание иероглифов может быть интерпретировано тем или иным способом, а выбор варианта сегментации обычно делается на основе контекста. Вторая проблема обусловлена полисемией иероглифов, каждый из которых может иметь десятки смыслов и быть разными членами предложения, в результате чего многозначность устраняется лишь после анализа всего текста. Третья проблема вызвана тем, что, несмотря на простую грамматику, в китайском языке сущест-

вует тенденция к максимальному упрощению речи, в результате чего могут опускаться отдельные части речи. Наконец, еще одна проблема, не свойственная другим языкам, – это отсутствие заимствованных слов, включая имена собственные.

Процесс извлечения информации из китайских текстов обычно включает в себя следующие фазы: сегментацию слов (word segmentation) [2, 3], выявление частей речи (lexical processing) [4], извлечение терминов [5, 6], поверхностный синтаксический анализ (shallow parsing) [7], семантический анализ (domain knowledge analysis). Особенности китайского языка диктуют широкое вовлечение контекста во все перечисленные фазы, вследствие чего даже сегментация слов должна делаться с учетом семантики.

В настоящей работе предлагается ролевой подход к анализу текста, традиционно используемый при извлечении именованных сущностей [8]. Несмотря на отсутствие падежных окончаний, спряжения глаголов и других признаков, являющихся маркерами для синтаксического анализа, в китайском языке есть служебные иероглифы, позволяющие извлекать полезную информацию о частях речи. Помимо служебных иероглифов, в китайском языке есть ограниченный набор слов, сопутствующих именам людей (названия должностей, воинских званий, ученых степеней, профессий и др.), географическим названиям (провинция, область, район, море, река), названиям организаций (завод, университет, совет, музей), которые могут служить выявлению имен собственных. Кроме того, есть ограниченный набор модальных глаголов, которые позволяют выявить сказуемые в предложении. В китайском языке, как и в любом другом, имеется достаточно узкий набор очень часто используемых слов, которые можно включить в состав универсального словаря, использование которого позволит существенно повысить полноту распознавания предложений. Также для извлечения фактов из текста необходимо располагать словарем терминов, для формирования которого в рамках настоящего исследования используется подход, описанный в работах [9, 10].

В данной работе предлагается многофазный процесс анализа китайского текста без предварительной сегментации слов с постепенным устранением неоднозначностей. Процесс анализа текста с целью извлечения фактов выглядит следующим образом.

1. Разбиение текста на отдельные предложения по терминальным символам.
2. Первичная сегментация предложений по символам, отличным от иероглифов.
3. Выделение в тексте предлогов, послелогов, частиц, модальных глаголов.
4. Сегментация оставшихся в тексте цепочек иероглифов с помощью словаря.
5. Выявление в тексте имен собственных с помощью служебных слов.
6. Выявление в тексте числительных с помощью служебных слов.
7. Назначение словам, соседствующим с выявленными предлогами, послелогами, частицами и модальными глаголами, атрибутов в соответствии с их ролями.
8. Выбор моделей предложений, не противоречащих выявленным словам, и назначение им частей речи.
9. Извлечение фактов, релевантных запросу.

Субъект	Предикат	Объект	Перевод
岩石 горная порода	组成 включает	硅	кремний
		同矿物特别	различные минералы
		使地表	чтобы поверхность Земли *
		单位	компонент
		陨石	метеорит
		火成岩	вулканические породы
		酸	кислота
		一定类型	определенный тип*
		矿物全部结晶	все кристаллические минералы
		矿物部	набор минералов
		云母	слюда
		枚岩片岩	филлит-сланцы
		粒	фракция
		又硬又脆	и жесткий и ломкий *
		结构构造变形	тектонические деформации структуры

* отмечены результаты, заведомо не релевантные запросу. Им соответствуют не полностью идентифицированные члены предложения.

Таблица. Результаты извлечения открытых отношений из книги «Основы геологии»

Результатом обработки текста будет представление каждого предложения в виде цепочек иероглифов, часть из которых снабжена атрибутами (член предложения, произношение, перевод, признак имени собственного, признак притяжательного и др.). Полученные частично разобранные предложения сопоставляются с запросом пользователя, в результате чего формируются факты вида субъект-предикат-объект.

Исследовательский прототип программы, реализующей предложенный алгоритм, разработан на языке SWI-Prolog[1] и занимает около 900 строк вместе с внутренним словарем. В качестве целевого корпуса документов использовался учебник по основам геологии[2]. В таблице приведены результаты выполнения поискового запроса «Что включают в себя горные породы», где субъект – 岩石 (горная порода), предикат – 组成 (включать в себя, состоять из), а объекты должны быть извлечены из текста.

Таким образом, предлагаемый ролевой подход к анализу китайских текстов с целью извлечения сущностей продемонстрировал свою работоспособность. Точность на множестве запросов составила 75–85%, что сопоставимо с результатами, показанными в работах [1] (82%) и [2] (0,55–0,84). Дальнейшие исследования должны быть нацелены на улучшение качества идентификации моделей предложений на основе расширения грамматики и разумного увеличения состава словаря общеупотребительных слов.

Литература

1. Banko M., Cafarella M.J., Soderland S., Broadhead M., Etzioni O. Open information extraction from the Web // Proc. 20th Int. Joint Conf. on Artificial Intelligence (IJCAI'07). Hyderabad, India, 2007. P. 2670–2676.
2. Tseng Y.H., Lee L.H., Lin S.Y, Liao B.S., Liu M.J., Chen H.H., Etzioni O., Fader A. Chinese open relation extraction for knowledge acquisition // Proc. 14th Conf. of the European Chapter of the Association for Computational Linguistics (EACL). Gothenburg, Sweden, 2014. V. 2. P. 12–16. doi: 10.3115/v1/e14-4003
3. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information // Information Systems Frontiers. 2011. V. 13. N 1. P. 115–125. doi: 0.1007/s10796-010-9278-5
4. Zhao J., Qiu X., Zhang S., Ji F., Huang X. Part-of-speech tagging for Chinese-English mixed texts with dynamic features // Proc. 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island, Korea, 2012. P. 1379–1388.
5. Basili R. A contrastive approach to term extraction // Proc. 4th Terminological and Artificial Intelligence Conference (TIA2001). Nancy, France, 2001.
6. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency – TF-DCF // Knowledge-Based Systems. 2016. V. 97. P. 237–249. doi: 10.1016/j.knosys.2015.12.015
7. Zhu Q., Cheng X.Y. The overview of Chinese information extraction // IJCSNS International Journal of Computer Science and Network Security. 2010. V. 10. N 9. P. 171–174.
8. Wong W. Determination of unithood and termhood for term recognition / In: Text and Web Mining Technologies. IGI Global, 2008. P. 500–529. doi: 10.4018/978-1-59904-990-8.ch030
9. Nugumanova A., Bessmertny I.A., Baiburin Y., Mansurova M. A new operationalization of contrastive term extraction approach based on recognition of both representative and specific terms // Communications in Computer and Information Science. 2016. V. 649. P. 103–118. doi: 10.1007/978-3-319-45880-9_9
10. Бессмертный И.А., Юй Чуцяо, Ма Пенюй. Статистический метод извлечения терминов из китайских текстов без словаря // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 6. С. 1096–1102. doi: 10.17586/2226-1494-2016-16-6-1096-1102

References

1. Banko M., Cafarella M.J., Soderland S., Broadhead M., Etzioni O. Open information extraction from the Web. *Proc. 20th Int. Joint Conf. on Artificial Intelligence, IJCAI'07*. Hyderabad, India, 2007, pp. 2670–2676.
2. Tseng Y.H., Lee L.H., Lin S.Y, Liao B.S., Liu M.J., Chen H.H., Etzioni O., Fader A. Chinese open relation extraction for knowledge acquisition. *Proc. 14th Conf. of the European Chapter of the Association for Computational Linguistics, EACL*. Gothenburg, Sweden, 2014, vol. 2, pp. 12–16. doi: 10.3115/v1/e14-4003
3. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information. *Information Systems Frontiers*, 2011, vol. 13, no. 1, pp. 115–125. doi: 0.1007/s10796-010-9278-5
4. Zhao J., Qiu X., Zhang S., Ji F., Huang X. Part-of-speech tagging for Chinese-English mixed texts with dynamic features. *Proc. 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Jeju Island, Korea, 2012, pp. 1379–1388.
5. Basili R. A contrastive approach to term extraction. *Proc. 4th Terminological and Artificial Intelligence Conference, TIA2001*. Nancy, France, 2001.
6. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency – TF-DCF. *Knowledge-Based Systems*, 2016, vol. 97, pp. 237–249. doi: 10.1016/j.knosys.2015.12.015
7. Zhu Q., Cheng X.Y. The overview of Chinese information extraction. *IJCSNS International Journal of Computer Science and Network Security*, 2010, vol. 10, no. 9, pp. 171–174.
8. Wong W. Determination of unithood and termhood for term recognition. In *Text and Web Mining Technologies*. IGI Global, 2008, pp. 500–529. doi: 10.4018/978-1-59904-990-8.ch030
9. Nugumanova A., Bessmertny I.A., Baiburin Y., Mansurova M. A new operationalization of contrastive term extraction approach based on recognition of both representative and specific terms. *Communications in Computer and Information Science*, 2016, vol. 649, pp. 103–118. doi: 10.1007/978-3-319-45880-9_9
10. Bessmertny I.A., Yu Chuqiao, Ma Pengyu. Statistical method of term extraction from Chinese texts without preliminary segmentation of phrases. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 6, pp. 1096–1102. doi: 10.17586/2226-1494-2016-16-6-1096-1102

Авторы

Юй Чуцяо – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, Scopus ID: 57195374429, ORCID ID: 0000-0003-4592-7611, yuchuqiao123@gmail.com

Authors

Yu Chuqiao – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, Scopus ID: 57195374429, ORCID ID: 0000-0003-4592-7611, yuchuqiao123@gmail.com

[1] www.swi-prolog.org

[2] www.baike.com/wiki/地质学基础