УДК 20.19.27
DOI: 10.15827/0236-235X.117.138-142
Дата подачи статьи: 23.11.16
2017. Т. 30. № 1. С. 138–142

АВТОМАТИЧЕСКИЙ СИНТАКСИЧЕСКИЙ АНАЛИЗ КИТАЙСКИХ ПРЕДЛОЖЕНИЙ ПРИ ОГРАНИЧЕННОМ СЛОВАРЕ

Юй Чуцяо, аспирант, yuchuqiao123@gmail.com;
И.А. Бессмертный, д.т.н., профессор, bia@cs.ifmo.ru
(Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), Кронверкский просп., 49, г. Санкт-Петербург, 197101, Россия)

В работе обсуждается проблема автоматического анализа естественно-языковых текстов на китайском языке. Одной из актуальных задач в этой области является автоматическое извлечение из текстовых документов фактов по запросу, поскольку автоматические переводчики здесь практически бесполезны.

Целью работы является прямое извлечение фактов из текстов на языке оригинала без его перевода. Для этого предлагается подход на основе синтаксического анализа предложений анализируемого текста с последующим сопоставлением найденных частей речи с формализованным запросом в форме субъект–предикат–объект.

Отличительная особенность предложенного алгоритма синтаксического анализа – отсутствие фазы сегментации последовательности иероглифов, составляющих предложения, на слова. Узким местом при решении данной задачи является словарь, поскольку при отсутствии слова в словаре правильная интерпретация фразы может быть невозможна. Для преодоления этой проблемы в работе предлагается идентификация модели предложения по служебным словам, а ограниченность словаря устраняется предварительным автоматическим построением тезауруса предметной области и словаря общеупотребительных слов на основе статистической обработки корпуса документов.

Апробация предложенного подхода выполнена на небольшой предметной области и с ограниченным словарем, где данный метод показал свою работоспособность. Проведен также анализ временных характеристик разработанного алгоритма.

Поскольку для синтаксического анализа используется метод простого перебора, скорость работы парсера на реальных задачах может оказаться неприемлемо низкой, что должно стать темой дальнейших исследований.

Ключевые слова: *естественные языки, синтаксический анализ, извлечение фактов, тезаурус, дерево поиска.*

В последнее время наблюдается рост потребности в поиске информации в китайских текстах. В частности, в силу расширения кооперационных связей между нашими странами возникает необходимость мониторинга сайтов госзакупок, научных фондов, производителей товаров и услуг. Существующие программы автоматического перевода, например Google Translate, здесь не очень полезны, поскольку дают только перевод, который еще нужно интерпретировать, чтобы извлечь нужную информацию. В связи с этим целесообразно прямое извлечение фактов из исходного текста на языке оригинала. Поставленная цель достигается решением следующих основных задач. Во-первых, необходимо выполнить синтаксический анализ естественно-языкового текста. Во-вторых, найти предложение в тексте, содержащее интересующую нас информацию. В-третьих, провести унификацию – сопоставить утверждения запроса с частями речи предложения. Результатом унификации как раз и будет подстановка в переменные запроса искомых значений.

Китайский язык, несмотря на кажущуюся сложность, имеет чрезвычайно простую грамматику. Главная проблема при анализе китайских текстов обусловлена двумя особенностями. В китайском языке отсутствуют пробелы между словами, и практически любое сочетание иероглифов может быть интерпретировано тем или иным образом. Это порождает проблему сегментации предложений на слова. Даже наличие полного словаря не гарантирует правильную интерпретацию последовательностей символов [1]. Статистические методы, в частности метод взаимной информации [2], позволяют обойтись без словарей, но также не дают 100 %-ную полноту и точность, поскольку в подавляющем большинстве случаев сегментация определяется контекстом, то есть в процессе семантического анализа.

Другой проблемой интерпретации китайских текстов является словарь. В отличие от алфавитных языков, где одна буква ничего не означает, в иероглифических языках каждому иероглифу соответствует определенное, достаточно широкое понятие. Конкретный смысл иероглиф приобретает только в контексте, из чего следует, что синтаксический анализ в китайском языке неотделим от семантического.

Постановка задачи

В целях извлечения фактов задача синтаксического анализа китайских текстов может быть конкретизирована следующим образом. Поисковый запрос должен быть формализован в виде *субъект–предикат–объект*, где *субъект, предикат* и *объект* могут быть либо словами, либо переменными. Для исполнения запроса необходимо найти в тексте фразу с заданными словами в качестве указанных членов предложений. Структура предложения

описывается его грамматической моделью. Следовательно, для каждого предложения необходимо найти его модель и подставить в него слова из поискового запроса. Полный синтаксический анализ в таком случае делать не требуется.

Состояние проблемы

Существует большое количество инструментальных средств синтаксического анализа, включая инструменты с открытыми кодами, например Томита-парсер [3], но все они ориентированы на алфавитные языки. Среди синтаксических анализаторов (парсеров), поддерживающих китайский язык, следует выделить Stanford CoreNLP [4] и CKIP Chinese Parser [5]. Основной проблемой всех парсеров, в том числе перечисленных, является структурная неоднозначность синтаксического разбора, обусловленная возможным наличием многих ролей слов в предложении. Для сокращения числа вариантов интерпретации предложений применяются методы машинного обучения, в частности [6], где учитываются частоты совместной встречаемости пар слов, а также принцип «разделяй и властвуй» (divide and conquer), при котором фраза разбивается на последовательности, анализируемые раздельно [7]. При этом качество синтаксического разбора существенно зависит от словаря. Таким образом, синтаксический анализ китайских текстов является сложной задачей, приемлемого решения которой на данный момент не существует.

Предлагаемый подход к синтаксическому анализу китайских текстов

Синтаксический анализ опирается на грамматику и словарь, однако их недостатки ухудшают качество парсинга. Если отсутствие адекватной структурной модели приводит к искаженной интерпретации фразы, то отсутствие хотя бы одного слова в словаре может сделать синтаксический анализ невозможным. Во флективных языках, в частности русском, имеется большая избыточность за счет окончаний, суффиксов, предлогов, приставок и др., что позволяет в принципе вообще обходиться без словаря [8]. К сожалению, китайский язык не обладает свойством флективности, более того, в нем отсутствуют даже времена глаголов, а одно и то же слово может использоваться в качестве глагола, существительного, наречия или прилагательного.

В работе предлагается не подключать к парсеру мощные словари, а наоборот, ограничить словарь небольшим количеством слов, в наибольшей степени определяющих структуру фразы. К таким словам относятся модальные глаголы, предлоги, послелоги, такие как 吗 (*ma* – аналог частицы *ли* в русском языке – признак вопроса), 的 (*de* – признак притяжательного прилагательного или аналога родительного падежа), 了 и 过 (*le* и *guo* – две разновидности прошедшего времени), счетные слова и некоторые другие.

С учетом жесткого порядка слов в китайском языке выделение в предложении служебных слов с определенной достоверностью может позволить выявить его структуру. Однако этого недостаточно без сегментации последовательностей иероглифов на слова. Здесь можно использовать вышеупомянутые статистические методы [1, 2], позволяющие фиксировать границы слов на редко встречающихся парах иероглифов. В работе [9] авторами предложена модификация данного подхода, заключающаяся не в выявлении границ слов в конкретной фразе, а в составлении списка слов, используемых в тексте, на основе статистической обработки корпуса документов. Такой обезличенный словарь (без перевода и свойств каждого слова) может использоваться для сегментации фраз. Отличие данного подхода состоит в том, что он позволяет в первую очередь подбирать наиболее длинные слова, что позволяет сокращать неоднозначности. Например, последовательность 交换式局 может рассматриваться как два термина: 交换式 (*переключение*) и 局 (*офис, помещение, служба,…*). В свою очередь, 交换式 распадается на слова 交换 (*обмен*) и 式 (*образец, правило*, *стандарт*, *формула?...*). Все вместе это означает *коммутационный центр*. Если при сегментации фразы в приоритетном порядке выделять наиболее длинные последовательности, в том числе из сгенерированного обезличенного словаря, то можно существенно сократить число вариантов сегментации фразы.

Реализация предлагаемого подхода

В данной работе не преследовалась цель полного синтаксического анализа, а только апробировалась возможность извлечения фактов, поэтому был создан несложный парсер на языке SWI-Prolog объемом 200 строк, опирающийся на грамматику и небольшой словарь. Грамматика в формате Бэкуса–Наура на языке Prolog имеет вид, представленный следующим фрагментом:

```
group(sentence,[subject,predicate,
object]).
group(sentence,[subject,object,
predicate]).
group(sentence,[subject,predicate,
object,afterlog]).
group(subject,[nouns]).
group(subject,[pronouns]).
group(subject,[attribute,nouns]).
group(subject,[nouns,attribute]).
group(subject,[subject,link,subject]).
group(nouns,[noun]).
group(nouns,[noun,noun]).
…
```

Словарь содержит минимум атрибутов слов, необходимых для синтаксического разбора. В частности, для предлогов указывается, с какими частями речи они сочетаются. Приведем примеры частей речи (рис. 1). Первый аргумент предикатов частей речи содержит иероглиф(ы), второй – произношение (*пиньинь*), затем перевод на русский язык и атрибуты (число, лицо, …).

Алгоритм синтаксического разбора предложения является рекурсивным и выглядит следующим образом.

1. Выбрать модель предложения (предикат group(sentence,[List]) из грамматики).
2. Выбрать первый элемент из списка членов предложения List.
3. Отделить N (максимально возможное число) иероглифов от предложения.
4. Провести синтаксический разбор члена предложения.
5. Если разбор удачный, перейти к п. 7, иначе $N := N - 1$.
6. Если N=0, выбрать следующую модель предложения и перейти к п. 2.
7. Провести синтаксический анализ оставшейся части предложения.

Синтаксический разбор члена предложения выполняется рекурсивно с помощью того же самого алгоритма, но с использованием модели не предложения (`sentence`), а более мелкой структурной единицы (`subject, object, predicate, attribute, noun, verb, preposition,…`). Если на самом нижнем уровне грамматики парсер не находит словарного слова, он пытается подставить слово из обезличенного словаря, полученного путем статистической обработки корпуса документов. Естественно, слова из запроса также используются для синтаксического анализа в качестве обезличенных словарных слов. В предложенном алгоритме перебор вариантов сегментации фразы выполняется начиная с наиболее длинных последовательностей символов (п. 3 алгоритма) с целью предпочтительного выбора наиболее длинных слов.

Оценка сложности алгоритма

Пусть N – число иероглифов в предложении, b – коэффициент ветвления дерева грамматики, d – средняя глубина дерева синтаксического разбора, s – среднее число иероглифов в слове. Тогда число шагов спуска по дереву решений для парсинга первого слова составит $M_1 = \frac{(N-s)b^d}{2}$, а для каждого последующего i-го слова $M_i = \frac{(N-is)b^d}{2}$.

Таким образом, общее число шагов алгоритма M составит $M = \frac{1}{2}\sum_{i=1}^{N/s}(N-is)b^d$.

Для фразы длиной 12 иероглифов потребуется 1,44 млн шагов алгоритма. Для более длинных фраз и с более сложной грамматикой время поиска решений может стать ощутимым и, возможно, потребует применения методов редуцирования дерева решений.

Результаты экспериментов

Исследование работоспособности предлагаемого подхода проводилось на простой грамматике и небольшом словаре, сформированных на основе телевизионного курса китайского языка Д. Петрова «Китайский язык за 16 часов» (http://16polyglot.ru/chinese/). На рисунке 2 приведен пример синтаксического разбора фразы 我在中国饭店工作 (*Я работаю в китайском ресторане*).

Здесь каждое слово описано следующими атрибутами: иероглиф, произношение, перевод на русский язык, число, падеж (для существительных). Все слова из данного предложения присутствовали в словаре, поэтому синтаксический разбор выполнен на 100 %. Заменим теперь в исходной фразе слово 中国 (*Китай*) на 意大利 (*Италия*) и получим следующую фразу: 我在意大利饭店工作. Слово *Италия* отсутствует в словаре, поэтому оно было извлечено из списка обезличенных слов и не снабжено переводом, но атрибуты (число и падеж) взяты из свойств предлога 在 (*в*) и приписаны этому обезличенному слову (см. рис. 3).

Таким образом, продемонстрирована работоспособность алгоритма синтаксического анализа китайских текстов в условиях ограниченного словаря.

Заключение

В результате проведенного исследования апробирован алгоритм синтаксического анализа китайских предложений, опирающийся на ограниченный словарь. Результаты синтаксического анализа могут использоваться для последующего извлечения фактов из текстовых документов. Поскольку предложенный алгоритм основан на переборе всех элементов грамматики и словаря, сложность алгоритма на реальных задачах может оказаться неприемлемой. Если речь идет о синтаксическом анализе для поиска и извлечения фактов из текстов, то каждую фразу в первую очередь следует проверять на наличие искомых паттернов в составе запроса; если хотя бы один из них отсутствует, фразу можно сразу пропустить. Для редуцирования дерева поиска также можно использовать подход, предложенный в работе [10] и основанный на том, что дерево решений часто образовано повторяющимися фрагментами. Применительно к рассматриваемой задаче это может означать следующее: если разные модели предложений содержат одну и ту же часть,

```
pronoun(     '我',   'wǒ',      'я', singular, '1st').
modal_verb(  '要',   'yào',     'намереваться').
verb(        '工作', 'gōngzuò', 'работать').
adjective(   '大',   'dà',      'большой, noun').
preposition( '在',   'zài',     'в', noun).
afterlog(    '了',   'le',      '*однократное прошедшее время', verb).
link(        '和',   'hē',      'и').
```

Рис. 1. Примеры частей речи

Fig. 1. The examples of parts of speech

```
Model: [subject,object,predicate]
     subject
        nouns
           pronoun
              我,wǒ,я,singular,1st
     object
        preposition
           在,zài,в,singular,prepositional
        noun
           中国,zhōngguó,Китай,singular,prepositional
           饭店,fàndiàn,ресторан,singular,prepositional
     predicate
        verb
           工作,gōngzuò,работать
```

Рис. 2. Пример синтаксического разбора

Fig. 2.The syntactic analysis example

```
Model: [subject,object,predicate]
     subject
        nouns
           pronoun
              我,wǒ,я,singular,1st
     object
        preposition
           在,zài,в,singular,prepositional
        noun
           意大利,singular,prepositional
           饭店,fàndiàn,ресторан,singular,prepositional
     predicate
        verb
           工作,gōngzuò,работать
```

Рис. 3. Пример работы алгоритма синтаксического анализа китайских текстов в условиях ограниченного словаря

Fig. 3. The example of the syntactic analysis algorithm for Chinese texts in the context of restricted dictionary

например группу подлежащего, то успешный синтаксический разбор этой группы может быть подставлен в другие модели предложений, в которых группа подлежащего также присутствует.

Литература

1. Xue N. Chinese word segmentation as character tagging. Intern. Jour. Computational Linguistics and Chinese Language Processing, February 2003, vol. 8, no. 1, pp. 29–48.
2. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information. Intern. Jour. Information Systems Frontiers. March 2011, vol. 13, iss. 1, pp. 115–125.
3. Томита-парсер. URL: http://tech.yandex.ru/tomita/ (дата обращения: 18.11.2016).
4. Christopher M.D., Surdeanu M., Bauer J., Finkel J., Bethard S.J., and McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014, pp. 55–60.
5. CKIP Chinese Parser. URL: http://140.109.19.112/ (дата обращения: 21.11.2016).
6. Hsieh Yu-M., Yang D.-Ch., Chen K.-J. Improve parsing performance by self-learning. Intern. Jour. Computational Linguistics and Chinese Language Processing, June 2007, vol. 12, no. 2, pp. 195–216.
7. Yang D.Ch., Hsieh Yu-M., Chen K.-J. Resolving ambiguities of chinese conjunctive structures by divide-and-conquer approaches. URL: http://godel.iis.sinica.edu.tw/CKIP/paper/Resolving_Ambiguities_of_Chinese_Conjunctive_Structures_by_Divide_and_conquer_Approaches.pdf (дата обращения: 21.11.2016).
8. Bessmertny I.A., Platonov A.V., Poleschuk E.A., Pen-

gyu Ma. Syntactic text analysis without a dictionary. Proc Conf. Application of Information and Communication Technology (AICT-2016). 2016, pp. 100–105.

9. Бессмертный И.А., Юй Чуцяо, Ма Пенюй. Статистический метод извлечения терминов из китайских текстов без сегментации фраз // Науч.-технич. вестн. информ. технологий, механики и оптики. 2016. Т. 16. № 6. С. 1096–1102.

10. Бессмертный И.А. Методы поиска информации с использованием интеллектуального агента // Изв. вузов. Приборостроение. 2009. Т. 52. № 12. С. 26–31.

Software & Systems
DOI: 10.15827/0236-235X.117.138-142

Received 23.11.16
2017, vol. 30, no. 1, pp. 138–142

AUTOMATIC SYNTACTIC ANALYSIS OF CHINESE SENTENCES BY A RESTRICTED DICTIONARY

Yu Chuqiao[1], *Postgraduate Student, yuchuqiao123@gmail.com*
I.A. Bessmertny, *Dr.Sc. (Engineering), Professor, bia@cs.ifmo.ru*

[1] *The National Research University of Information Technologies, Mechanics and Optics, Kronverksky Ave. 49, St. Petersburg, 197101, Russian Federation*

Abstract. The paper considers a problem of natural language processing of Chinese texts. One of the relevant tasks in this area is automatic fact acquisition by a query since existing automatic translators are useless for this task. The suggested approach includes a syntactic analysis of phrases and matching parts of speech founded with a formalized query.

The purpose of the study is direct fact extracting from original texts without translation. For this purpose the paper suggests to use an approach based on syntactic analysis of sentences from a text with further comparison of the found parts of speech with a formalized subject–object–predicate query. A key feature of the proposed approach is a lack of a segmentation phase of a hieroglyph sequence in a sentence by words. The bottleneck at this task is a dictionary because interpretation of a sentence is impossible without even a single word in the dictionary. To eliminate this problem the authors propose to identify a sentence model by function words while restraint of the dictionary could be compensated by automatic building of a thesaurus using statistical processing of a document corpus. The suggested approach is tested on a small topic where it demonstrates its robustness. There is also an analysis of temporal properties of the developed algorithm.

As the proposed algorithm uses a direct-search method, the parsing speed for real tasks could be unacceptably low and this is a subject for further research.

Keywords: natural language, syntactic analysis, fact extraction, thesaurus, search tree.

References

1. Xue N. Chinese Word Segmentation as Character Tagging. *Computational Linguistics and Chinese Language Processing*. 2003, vol. 8, no. 1, pp. 29–48.

2. 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, iss. 1, pp. 115–125.

3. *Tomita-parser*. 2015. Available at: https://tech.yandex.ru/tomita/ (accessed November 18, 2016).

4. Christopher M.D., Surdeanu M., Bauer J., Finkel J., Bethard S.J., McClosky D. The Stanford CoreNLP Natural Language Processing Toolkit. *Proc. 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2014, pp. 55–60.

5. *CKIP Chinese Parser*. Available at: URL:http://140.109.19.112/ (accessed November 21, 2016).

6. Hsieh Yu-M., Yang D.-Ch., Chen K.-J. Improve Parsing Performance by Self-Learning. *Computational Linguistics and Chinese Language Processing*. 2007, vol. 12, no. 2, pp.195–216.

7. Yang D.Ch., Hsieh Yu-M., Chen K.-J. Resolving Ambiguities of Chinese Conjunctive Structures by Divide-and-conquer Approaches. *IJCNLP200*. Available at: http://godel.iis.sinica.edu.tw/CKIP/paper/Resolving_Ambiguities_of_Chinese_Conjunctive_Structures_by_Divide_and_conquer_Approaches.pdf (accessed November 21, 2016).

8. Bessmertny I.A., Platonov A.V., Poleschuk E.A., Ma P. Syntactic Text Analysis Without a Dictionary. *Application of Information and Communication Technology (AICT-2016)*. 2016, pp. 100–105.

9. Bessmertny I.A., Yu Ch., Ma P. Statistical method of term extraction from chinese texts without preliminary segmentation of phrases. *Nauch.-tehnich. vestn. inform. tekhnology, mekhaniki i optiki* [Scientific and Technical Jour. of Information Technologies, Mechanics and Optics]. 2016, vol. 16, no. 6, pp. 1096–1102 (in Russ.).

10. Bessmertny I.A. Methods of information retrieval by intelligent agent. *Izv. vuzov. Priborostroenie* [Jour. of Instrument Engineering]. 2009, vol. 52, no. 12, pp. 26–31 (in Russ.).