УДК 519.68
DOI: 10.15827/2311-6749.17.1.4

РОЛЕВОЙ ПОДХОД К АВТОМАТИЧЕСКОМУ ИЗВЛЕЧЕНИЮ ФАКТОВ ИЗ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ КИТАЙСКИХ ТЕКСТОВ

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

В работе обсуждается проблема извлечения фактов из китайских текстов. Китайский язык достаточно сложен для машинной обработки, что обусловлено отсутствием пробелов между словами и многозначностью иероглифов, поэтому синтаксический анализ текстов невозможен без семантического анализа, поскольку любое сочетание иероглифов может быть интерпретировано неоднозначно. Существующие статистические методы сегментации предложений на слова и синтаксического анализа не обладают достаточной полнотой и точностью, вследствие чего многофазный процесс (сегментация фраз, синтаксический анализ, извлечение фактов) приводит к накоплению ошибок.

В статье предлагается ролевой подход к выявлению членов предложения на основе служебных слов, предлогов и послелогов, а также достаточно ограниченного словаря. Эти служебные слова и символы позволяют не только сегментировать последовательности символов, но и выявлять роли слов, а значит, и части речи в предложении. Даже такой небольшой набор слов позволяет в большинстве случаев успешно выявлять роль слов в предложении, в частности, имена собственные, существительные, глаголы, что делает возможным извлечение из текстов сущностей, а также фактов в виде субъект–предикат–объект. Проведенные на реальных текстах эксперименты показывают удовлетворительные результаты даже при ограниченном словаре. Предложенный подход демонстрирует высокую скорость, поскольку отсутствуют синтаксический разбор и сегментация фраз, использующие переборные методы.

Ключевые слова: *извлечение фактов, китайский язык, ролевой подход, анализ текстов, словарь, сегментация предложений, поверхностный синтаксический анализ, частеречный анализ.*

Извлечение фактов из китайских текстов (Chinese Open Relation Extraction, CORE) 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, почти любое сочетание иероглифов может быть интерпретировано тем или иным способом, а выбор варианта сегментации обычно делается на основе контекста. Из этого следует, что сегментация предложений в китайском языке неотделима от семантического анализа, что существенно усложняет задачу автоматической сегментации фраз. Схожая ситуация наблюдается в языках, широко использующих сложные слова. Например, немецкое слово *Süsswasserkrokodil* может быть интерпретировано как *Süsswasser-krokodil* (пресноводный крокодил) или *Süss-wasserkrokodil* (сладкий водяной крокодил).

Вторая проблема обусловлена полисемией иероглифов, каждый из которых может иметь десятки смыслов и быть разным членом предложения, в результате чего многозначность устраняется лишь после анализа всего текста. Аналогичная проблема существует и в других языках, но в значительно меньшем масштабе.

Третья проблема вызвана тем, что, несмотря на простую грамматику, в китайском языке существует тенденция к максимальному упрощению речи, в результате чего могут опускаться отдельные части речи. В других языках это явление также наблюдается. Например, в испанском языке нормой считается опускание местоимений (*Voy de compras* – иду за покупками), но это компенсируется спряжением глаголов.

Наконец, еще одна проблема, не свойственная другим языкам, – это отсутствие заимствованных слов, включая имена собственные, при том, что состав иероглифов зафиксирован около 400 лет назад. Термины в китайском языке обозначаются через сочетания понятий. Например, термин 电子 (электрон) состоит из двух иероглифов: 电 (электричество, включая атмосферное) и 子 (ребенок). Для имен собственных подбираются наборы иероглифов, похожие по произношению и, возможно, осмысленные. Так, фамилия Трамп пишется 特朗普 (tè lǎng pǔ), а город Санкт-Петербург как 圣彼得堡 (shèng bǐdébǎo), где 圣 – совершенный, чудодейственный, святой, монарший, 彼得 – Пётр, 堡 – крепость, форт, поселение. При этом, в отличие от алфавитных языков, имена собственные никак не выделяются в иероглифическом тексте.

Перечисленные факторы делают проблематичными даже несложные задачи обработки китайских текстов, такие, как извлечение сущностей или несложных фактов в виде субъект–предикат–объект, что позволяет считать актуальными все частные задачи: сегментацию фраз, извлечение терминов и именованных сущностей, синтаксический и семантический анализ.

Состояние проблемы и текущие исследования

Процесс извлечения информации из китайских текстов обычно включает в себя следующие фазы: сегментация (segmentation), выявление частей речи (lexical processing), синтаксический анализ (shallow parsing), семантический анализ (domain knowledge analysis) [2, 3].

Сегментация текстов. Несмотря на наличие правил сегментации [4], исследователи в области обработки естественно-языковых текстов предпочитают статистические методы. Это, видимо, обусловлено тем, что упомянутые правила в большей степени ориентированы на носителей языка. Автоматическая сегментация предложений чаще всего базируется на методе взаимной информации [5], где анализируются частоты совместной и раздельной встречаемости пар иероглифов. Статистические методы, естественно, не гарантируют 100 %-ной полноты и точности. В работе [5] величины точности и полноты сегментации не превышают 90 %.

Извлечение терминов. Для формирования словарей предметной области используются контрастные методы с использованием двух коллекций документов – целевого и контрастного корпусов. Метод TF-IDF (Term Frequency – Inverse Document Frequency) [6] ориентирован на извлечение часто используемых слов, к которым относятся ключевые слова, и плохо извлекает редкие термины. Вообще говоря, контрастный подход очень популярен среди исследователей, и на его базе разработано множество разнообразных техник извлечения терминов [7, 8].

В работе [9] для автоматического формирования тезауруса предлагается гибридный подход, в котором для извлечения часто используемых слов используется мера средней взаимной информации MI (Mutual Information), а для редких терминов – PMI (Pointwise Mutual Information в большей степени – точечная взаимная информация). Применительно к китайскому языку авторами предложен эвристический подход [10], основанный на выявлении в составе кандидатов в термины последовательностей иероглифов, составляющих общеупотребительные слова, что позволяет повысить точность извлечения редких терминов. Применительно к китайскому языку точечная взаимная информация используется в работе [11], авторы которой утверждают, что их метод позволяет достичь точности не менее 81 % без понижения полноты отбора терминов.

Синтаксический анализ. В китайском языке отсутствуют падежи, спряжение глаголов, время и род, что делает невозможным использование данных морфности слов для синтаксического анализа, как это предложено в работе [12]. Несмотря на строгий порядок слов в предложении, распознать конкретную грамматическую структуру фразы часто невозможно в связи с тем, что одно и то же слово может выступать в качестве существительного, глагола или прилагательного, а отдельные части речи могут опускаться. Все это порождает множество вариантов интерпретации предложений. Вместо полноценного синтаксического анализа чаще используется частеречный анализ (part of speech tagging), как это делается в работе [13] для смешанных англо-китайских текстов.

Извлечение имен собственных. Как уже было сказано, в китайском языке имена собственные, включая иностранные, представляются иероглифами и ничем не выделяются в основном тексте. Например, США записывается в виде 美国(měi guó), что может быть переведено как *красивое государство*. Для географических названий можно воспользоваться словарями, но для имен людей или названий организаций это невозможно. В работе [14] для выявления имен собственных предлагается ролевой подход, основанный на том, что в текстах имена собственные обычно хотя бы один раз соседствуют с такими словами, как *председатель, студент, генерал, сказал, увидел* (для имен людей), *компания, университет, завод, банк* (для названий организаций), *река, город, область, гора* (для географических названий). Данный подход представляется перспективным не только для выявления имен собственных, но и для всех стадий анализа текстов.

Ролевой подход к анализу китайских текстов

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. К таким служебным иероглифам относятся следующие:

– все предлоги, определяющие, как и в других языках, положение в пространстве (например, 在, zài – в), во времени (之后, zhīhòu – после), направление движения (过来, guòlai – приближение к чему-либо) и др.;

– послелоги:
吗, ma – признак вопросительного предложения;
了, le – признак однократного прошедшего времени глагола;
过, guo – признак действия в прошлом;
的, de – признак притяжательного слова, принадлежность к кому- или чему-нибудь;
们, men – признак множественного числа;
– признаки числительных:
个, ge – универсальное счетное слово;
支, zhī – счетное слово для длинных предметов;
把', bǎ – счетное слово для предметов с ручкой;
辆', liàng – счетное слово для машин;
– частицы и союзы:
和, hē – и;
或者, huòzhě – или;
不, bu – не;
…

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 для выявления 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.

Алгоритм ролевого анализа китайского текста

В данной работе предлагается многофазный процесс анализа китайского текста с постепенным устранением неоднозначностей. Для анализа используются следующие признаки в порядке убывания приоритета:

- терминальные символы (точка, восклицательный, вопросительный знак, конец абзаца);
- символы, отличные от иероглифов (цифры, кавычки, спецсимволы);
- слова из внутреннего словаря;
- предлоги, послелоги, частицы, модальные глаголы;
- служебные слова, сопутствующие именам собственным;
- служебные слова, сопутствующие числительным.

Указанная приоритетность, в частности, в отношении служебных слов, обусловлена тем, что иероглифы, обозначающие служебные слова, могут входить в общеупотребительные слова и имена собственные, что может привести к слишком мелкой сегментации таких слов с полной потерей их смысла.

С учетом выбранных приоритетов процесс анализ текста с целью извлечения фактов выглядит следующим образом:

– разбиение текста на отдельные предложения по терминальным символам;
– первичная сегментация предложений по символам, отличным от иероглифов;
– выделение в тексте предлогов, послелогов, частиц, модальных глаголов;
– сегментация оставшихся в тексте цепочек иероглифов с помощью словаря;
– выявление в тексте имен собственных с помощью служебных слов;
– выявление в тексте числительных с помощью служебных слов;
– назначение словам, соседствующим с выявленными предлогами, послелогами, частицами и модальными глаголами, атрибутов в соответствии с их ролями;
– выбор моделей предложений, не противоречащих выявленным словам, и назначение им частей речи;
– извлечение фактов, релевантных запросу.

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.). Кроме того, для каждого предложения возможно множество его интерпретаций. Все это делает проблематичным дальнейшее использование результатов анализа текста в практических целях.

Поскольку целью анализа текста является извлечение из него сущностей или фактов в виде субъект–предикат–объект, это делает возможным существенное сокращение сложности анализа. Во-первых, текст можно отфильтровать, отбросив предложения, не содержащие слова из запроса. Во-вторых, можно извлечь семантику из запроса, в котором в явном виде устанавливается принадлежность слов к подлежащему (субъект), сказуемому (предикат) или дополнению (объект). В таком случае размерность задачи может быть существенно снижена.

Результаты экспериментального исследования разработанного алгоритма

Исследовательский прототип программы, реализующей предложенный алгоритм, разработан на языке SWI-Prolog (www.swi-prolog.org) и занимает около 900 строк вместе с внутренним словарем. Фрагмент словаря приведен ниже (первый аргумент содержит иероглифическое написание, второй – произношение (пиниинь), третий – смысл, четвертый – роль):

```
preposition('从' , 'cóng',   'из',            'направление движения')
preposition('在' , 'zài',    'в',          'местоположение')
preposition('过来', 'guòlai','приближение к,  'направление движения ')
afterlog('之后',    'zhīhòu','после',         'время')
afterlog('们',      'men',     'мн.ч.',       'число существительного')
afterlog('州' ,    'zhōu',     'провинция',     'местоположение')
name('主席',        'zhǔxi', 'председатель',       'атрибут имени')
name('上校',        'shàngxiào',  'полковник',  'атрибут имени')
modal_verb('要' ,   'yào',   'намереваться')
modal_verb('作' ,   'zuò',   'делать')
```

Внутренний словарь также содержит набор самых часто используемых глаголов, в числе которых следующие:

```
verb('说' ,  'shuō',    'говорить');
verb('看' ,  'kàn',     'смотреть');
verb('去' ,  'qù ',     'идти');
verb('工作' , 'gōngzuò', 'работать').
```

Попытка включить в словарь такие очень популярные слова, как 人,rén – человек, 大, dà – большой, 子, zi – ребенок, только ухудшила качество анализа текста, поскольку эти иероглифы часто входят в состав более сложных понятий (大学 – университет, 王子 – принц, 人物 – характер).

Предположим, что запрос на извлечение факта выглядит следующим образом.

Где работает Wang Shu? В виде триплета это может быть сформулировано так:

subject: *Ван Шу* (王书), predicate: *работать* (工作), object: *Где* (?x).

Из текста отобрано следующее предложение, содержащее паттерны *Ван Шу* и *работать*.

王书10岁在中国饭店工作。 (Ван Шу работает в китайском ресторане 10 лет).

После выявления предлогов и послелогов, подстановки слов из словаря, а также разметки слов-кандидатов на субъект и предикат получим промежуточный результат:

```
[(_,王书,_,Wang Shu),
   (number,10/岁,sui,лет),
   (местоположение,在/中国,zài/zhōngguó,в/Китай),
   (_,饭店,_,_),
 (predicate,工作,_,работать)].
```

Здесь знаки подчеркивания замещают отсутствующие значения. К полученной промежуточной структуре предложения может быть применена следующая модель предложения:

```
sentence,[subject,object,predicate].
object,[attribute,object].
```

Результат применения данной модели предложения:

```
[(subject, 王书,_,Ван Шу),
 (number, 10/岁,sui,лет),
 (location*,在/中国,zài/zhōngguó,в/Китай),
 (object,饭店,fàndiàn,ресторан),
 (predicate,工作,gōngzuò,работать)].
```

Таким образом, получаем, что искомый объект в запросе получает значения, содержащие название объекта (китайский ресторан) и даже срок (10 лет).

Разумеется, такой искусственный пример дает оптимистическую оценку качества извлечения фактов из текстов. Ниже приведена фраза из газеты China Daily, извлеченная по запросу, в котором нас интересовала любая информация по персоне с именем *Бонго* (邦戈):

全国人大常委会委员长张德江8日在北京人民大会堂会见了加蓬总统邦戈。

В результате анализа получаем следующую структуру фразы:

```
[(subject,全国人大常委会/委员长,/wěiyuánzhǎng,/председатель комитета),
 (_,张德江,_,_),
 (number,8日,rì,день),
 (location,在/北京,zài/běijīng,в/Пекин),
 (dict,人民,_,_),
 (dict,大会,_,_),
 (object,堂会,_,_),
 (predicate,见/了,jiàn/le,see/однокр.прошедшее время),
 (object, 加蓬,_,_),
 (object,总统/邦戈,zǒngtǒng/,президент/)].
```

Здесь видно, что *Бонго* является президентом объекта 加蓬 (*Габон*), восьмого числа выступал в качестве объекта встречи в Пекине с председателем комитета 全国人大常委会 (*Всекитайского собрания народных представителей*). Имя председателя указанного комитета 张德江 (Джан Децзян) оказалось нераспознанным в силу слишком сложной конструкции данного речевого оборота, в котором служебное слово *председатель* находится между наименованием объекта председательства и субъектом председательства.

Заключение

Таким образом, предлагаемый ролевой подход к анализу китайских текстов с целью извлечения сущностей продемонстрировал свою работоспособность. 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., and Etzioni O. Open information extraction from the web. Proc. IJCAI'07, 2007, pp. 2670–2676.

2. Yuen-Hsien Tseng, Lung-Hao Lee, Shu-Yen Lin, BoShun Liao, Mei-Jun Liu, Hsin-Hsi Chen, Etzioni O., and Fader A. Chinese open relation extraction for knowledge acquisition. Proc. EACL, 2014, pp. 12–16.

3. Zhu Qian and Cheng Xian Yi. The Overview of Chinese Information Extraction. IJCSNS, 2010, vol. 10, no. 9, pp. 171–174.

4. Taiwanese Principles of Text Segmentation. URL: http://ip194097.ntcu.edu.tw/TG/CompLing/hunsu/hunsu.htm (дата обращения: 09.12.2016).

5. Zeng D., Wei D., Chau M., Wang F. Domain-specific Chinese word segmentation using suffix tree and mutual information. Inform. Systems Frontiers, 2011, vol. 13, no. 1, pp. 115–125; DOI: 0.1007/s10796-010-9278-5.

6. Huang Lei, Wu Yan-Peng, Zhu Qun-Feng. Research and improvement of TFIDF feature weighting method. Comp. Sc., 2014, vol. 41, no. 6, pp. 204–208.

7. Basili R. A contrastive approach to term extraction. Proc. 4th Terminological and Artificial Intelligence Conf. 2001, pp. 119–128.

8. Lopes L., Fernandes P., Vieira R. Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. Knowledge-Based Systems, 2016, pp. 237–249.

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 Comp. and Inform. Sc., 2016, vol. 649, pp. 103–118.

10. Bessmertny I.A., Platonov A.V., Poleschuk E.A., Pengyu Ma. Syntactic text analysis without a dictionary. Application of Information and Communication Technology, 2016, pp. 100–105.

11. Hua-Ping Zhang, Qun Liu, Hong-Kui Yu. Chinese Named Entity Recognition Using Role Model. Comp. Linguistics and Chinese Language Processing. 2003, vol. 8, no. 2, pp. 29–60.

12. Бессмертный И.А., Юй Чуцяо, Ма Пенюй. Статистический метод извлечения терминов из китайских текстов без словаря // Науч.-технич. вестн. информ. технологий, механики и оптики. 2016. Т. 16. № 6. С. 1096–1102; DOI: 10.17586/2226-1494-2016-16-6-1096-1102.

13. Jiayi Zhao, Xipeng Qiu, Shu Zhang, Feng Ji, Xuanjing Huang. Part-of-speech tagging for Chinese-English mixed texts with dynamic features. Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1379–1388.

14. Wong W., Liu W., Bennamoun M. Determination of unithood and termhood for term recognition. Handbook of Research on Text and Web Mining Technologies, IGI Global, 2008.