ОНТОЛОГИЧЕСКИЙ ПОДХОД К ОБРАБОТКЕ ТЕКСТОВ КИТАЙСКОГО ЯЗЫКА

Цянь Лунвэй, Ли Вэньцзу
Кафедра интеллектуальных информационных технологий,
Белорусский государственный университет информатики и радиоэлектороники
Минск, Республика Беларусь
E-mail: qianlw1226@gmail.com, wzzggml@gmail.com

Как типичный представитель аналитического языка, китайский язык сильно отличается от других языков. Не существует границ между словами в китайских языках, и нет морфологических изменений про слова. Этими характеристиками китайского языка обусловлены существенные различия в технологиях обработки китайского языка по сравнению с другими языками. В данной работе кратко рассмотрены основные трудности обработки китайского языка и проанализированы существующие методы обработки.

Введение

Разнообразие естественного языка и его неоднозначность создает большие трудности в обработке естественного языка. Основными языками в мире являются английский, китайский, русский, французский и арабский. Каждый язык имеет свои особенности в лексическом и синтаксическом аспектах, такие как, например, морфологические изменения в английском и изменение по падежами в русском языке, что приводит к большим различиям в методах обработки языка. В китайском языке нет естественных границ для разделения каждого слова [1]. У каждого китайского слова есть только одна форма, и нет изменяемых форм, таких как множественное число, часть речи, форма времени и т. д. Далее кратко рассмотрим эти трудности и методы их преодоления в технологии обработки китайского языка.

I. Сегментация текстов китайского языка

Сегментация текста – процесс объединения последовательностей иероглифов в последовательности слов в соответствии с определенными спецификациями. В китайском предложении не каждый иероглиф имеет значение, но слово выражает определенное значение. В английском и русском языке между словами пробелы используются в качестве естественных разделителей слов. В китайском языке можно легко разделить текст на отдельные иероглифы, предложения и абзацы, однако на уровне слов в китайском языке нет формального разделителя. Таким образом, в сегментации китайских текстов существует проблема идентификации неоднозначности.

Например, если есть фраза вида ABC, то можно выделить слово AB, но также можно выделить слово BC:

- 老板/有意/见他 (босс встречает его с некоторой целью);
- 老板/对/他/有/意见 (у босса есть мнение о нём).

В свою очередь, слово вида AB может рассматриваться как одно слово, но может быть разделено на два слова:

- 语 言 学/是/以/人 类/语 言/为/研 究/对 象/的/学科 (Лингвистика является предметом, который изучает человеческий язык);
- 其它/语言/学/起来/很/难 (Другие языки трудно выучить).

В процессе сегментации китайского текста, помимо идентификации неоднозначности, существует еще одна проблема – идентификация новых слов. Новые слова – это слова, которые на данный момент нет в словаре. Новые слова в целом подразделяются на две категории: 1. Появляющиеся общие слова или технические термины и т. д. 2. Собственные существительные, такие как китайские имена, иностранные переведенные названия, топонимы. В английском и русском языках эти слова могут быть легко оценены по разнице между заглавными и строчными буквами. Из-за отсутствия морфологических изменений в китайском языке возникает проблема идентификации новых слов.

В настоящее время основные методы сегментации на слова делятся на три категории [2].

Сегментация слов на основе словаря. Для набора китайских символов происходит сопоставление его фрагментов с терминами в «достаточно большом» машинном словаре в соответствии с определенной стратегией.

Сегментация слов на основе статистики. Слова состоят из китайских иероглифов, то есть слова представляют собой комбинацию устойчивых китайских иероглифов. Частота комбинаций смежных китайских символов в фразах может быть подсчитана и может отражать вероятность того, что смежные китайские символы могут быть одним словом.

Сегментация слов на основе понимания является одном из методов на основе статистики. С появлением моделей глубокого обучения исследователи применяют векторные мо-

дели слов, которые содержат синтаксическую и семантическую информацию.

Рассмотренные методы реализуют сегментацию китайских текстов и повышают точность сегментации слов в разных аспектах, но не являются идеальными для идентификации неоднозначности. При улучшении точности эффективность сегментации китайских текстов невелика.

II. Определение частей речи в китайском языке

Одна и та же часть речи может выполнять множество синтаксических ролей без морфологических изменений в китайском языке. В английском и русском словаре, кроме толкования слова и примеров использования каждого слова, приводится также часть речи для каждого слова. Даже без просмотра словаря, часть речи определенных слов можно различить по морфологическим характеристикам слова. Например, в русском языке есть суффикс, который точно определяет часть речи. В китайском языке при определении части речи могут возникать трудности:

这篇文章的重点是第三段 (Внимание в этой статье сфокусировано на третий абзац).

Слово «重点» является существительным.

这篇文章重点讲述机器翻译问题 (В этой статье внимание уделено вопросам машинного перевода).

Слово «重点» является наречием.

В настоящее время методы определения частей речи (частеречевого тегирования) делятся на две категории [3]. **Метод на основе правил** был самым ранним методом, использованным для маркировки части речи. По контексту и отношению в словосочетании между многофункциональном словом и другом словом строятся правила устранения неоднозначности при определении части речи. Ранее правила маркировки части речи обычно создавались вручную, однако по мере увеличения размера аннотированного корпуса текстов извлечение правил вручную становится непрактичным. Для решения данной задачи предложен **метод на основе статистики**. На основе заданной последовательности слов с соответствующей им маркировкой частей речи может быть определена наиболее вероятная часть речи для следующего слова.

Метод на основе статистики позволяет получить более высокую точность маркировки части речи, но требует высокого качества корпуса.

III. Онтологический подход

Классические методы имеют высокую скорость в обработке китайского текста, однако точность обработки не может быть сильно улучшена без использования новых методов. Мы предлагаем использовать комбинацию классических методов сегментации и частеречевого тегирования и методов на основе онтологий. В данной работе предлагается онтологический подход, разработанных на основе технологии OSTIS [4]. Технология OSTIS ориентирована на разработку совместимых компьютерных систем, управляемых знаниями. Одним из ключевых принципов OSTIS технология является использование онтологического подхода.

При сегментации китайского текста комбинация методов на основе словаря и на основе онтологий может сочетать в себе преимущества метода на основе словаря – высокую скорость сегментации, и использовать метод на основе онтологии для повышения точности. При использовании такого комбинированного подхода результаты сегментации сопоставляются с онтологией семантики китайского языка. Это способствует повышению качества идентификации неоднозначности и ее устранения.

При частеречевом тегировании в соответствии с характеристиками китайской грамматики создается онтология синтаксиса китайского языка. Онтология синтаксиса китайского языка используется для проверки результатов выделения частей речи в китайских предложениях. Это позволяет уменьшить зависимость статистических методов от корпуса.

Заключение

Китайский язык является языком, в котором порядок слов и функциональные слова используются для выражения грамматического значения. В настоящее время большинство современных моделей обработки естественного языка основаны на исследованиях английского языка. Из-за разнообразия языков необходимо учитывать особенности китайского синтаксиса и семантики, что делает актуальным построение онтологии синтаксиса и семантики китайского языка, которые могут в дальнейшем использоваться для устранения неоднозначности при обработке текстов. Особенно важной задачей становится разработка специальной модели обработки китайского языка.

Список литературы

1. Huang, C. N. Chinese Word Segmentation: A Decade Review /C. N. Huang. // Journal of Chinese Information Processing. – 2007. – vol.21, № 03. – p. 8–19.
2. Long, S. Q. Zhao, Z. G. Overview on Chinese Segmentation Algorithm /S. Q. Long, Z. W. Zhao. // Computer Knowledge and Technology. –2009. – vol.5, № 10. –p. 2605–2607.
3. Liu, H. M. Research on Chinese parts of speech tagging and POS guessing over unknown word / H. M. Liu. – Nanjing Normal University. 2015.
4. Голенков, В. В. Гулякина, Н. А. Проект открытой семантической технологии компонентного проектирования интеллектуальных систем. Часть 1: Принципы создания. /В. В. Голенков, Н. А. Гулякина // Онтология проектирования. — 2014. № 1. – с.42–64.