АНАЛИЗ СЕГМЕНТАЦИИ КИТАЙСКОГО ЯЗЫКА

Цянь Лунвэй

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Ломако С. В. – ст. преподаватель

В статье рассмотрено понятие языковой сегментации применительно к проблеме компьютерной обработки китайского языка.

Основными языками в мире являются английский, китайский, русский, французский и арабский. Каждый язык имеет свои особенности в лексическом и синтаксическом аспектах, в наличии морфологических категорий, а также грамматических форм. При использовании существующих методов обработки естественного языка необходимо учитывать характеристики каждого языка [1]. В китайском языке нет естественных границ для разделения каждого слова. У китайского слова есть только одна форма, и нет никаких изменяемых форм, таких как единственное и множественное число, временная форма, вид, залог, а также падежи. Эти характеристики китайского языка создают трудности в технологии машинной обработки китайского языка в сравнении с другими языками. Для обработки текста естественного языка надо проанализировать каждое слово в тексте. Из-за особенностей китайского языка, первым шагом в обработке текста является сегментация. Сегментация – это разделение последовательности китайских иероглифов в отдельное слово.

В обработке текста китайского языка сегментация стала самой основной проблемой, а также основой последующего лексического и синтаксического анализа. Ранее для обработки текстов китайского языка использовали методы на основе правил, которые были написаны вручную. Из-за субъективности написанных вручную правил и несовместимости правил в разных системах наблюдалось огромное количество ошибок при переводе. В настоящее время большинство методов обработки китайского языка сочетают метод на основе статистики и правил. Институт компьютерных исследований Национальной академии наук Китая разработал систему сегментации текста китайского языка – NLPIR-ICTCLAS. Однако до сих пор эта проблема полностью не решена. Проблема сегментации текста китайского языка не полностью решена в основном из-за следующих двух факторов: идентификация неоднозначностей, идентификация новых слов.

Сегментация китайского текста существует в нескольких формах в разных контекстах. Рассмотрим два примера китайского предложения: (1) 面的/价格 (Цена вида транспорта); (2) 面/的/价格 (Цена муки). В данной ситуации иероглифы 面 и 的 не являются отдельно взятыми словами и не несут законченной смысловой (понятийной) нагрузки. Только в сочетании с конкретными иероглифами они

будут переводиться как «Цена вида транспорта». В ситуации (2) эти же иероглифы (面 и 的), но в сочетании с другими иероглифами уже будут обозначать «Цена муки».

В процессе сегментации китайского текста, помимо идентификации неоднозначностей, существует еще одна проблема – идентификация новых слов. Новые слова – это слова, которых нет в словаре. Новые слова в целом подразделяются на две категории: 1) общеупотребительные слова и технические термины; 2) собственные существительные, такие как китайские имена, иностранные переведенные названия. 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 возникла проблема идентификации новых слов. С точки зрения 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 частота может отражать то, что смежные китайские символы могут стать одним словом.

Эти методы реализуют сегментацию китайских текстов и повышают точность сегментации слов в разных аспектах, но не идеальны для идентификации неоднозначности.

Мы предлагаем использовать комбинацию методов на основе словаря и на основе онтологии [2]. Онтологию можно понимать как концептуальную модель, описывающую понятия и отношения между понятиями. Эта модель сочетает в себе преимущества метода на основе словаря (высокая скорость сегментации) и метода на основе онтологии (повышение точности). Это способствует улучшению идентификации слов и устранению неоднозначности.

Китайский язык является языком, в котором порядок слов и функциональные слова используются для выражения грамматического значения. В отличие от акцента на формальности в английском и русском языках, китайский язык уделяет больше внимания согласованности значения. Различные компоненты предложения больше полагаются на значение.

Технологии обработки китайского языка получили большое развитие в последние годы. Большинство современных моделей обработки естественного языка основаны на исследованиях английского языка. В настоящее время из-за разнообразия языков исследование специальной модели обработки китайского языка особенно важно.

Список использованных источников:

1. Chengqing Zong. Chinese Language Processing: Achievements and Problems / Zong Chengqing // Chinese Journal of Language Policy and Planning. – 2016. – № 1(06). – P. 19-26.

2. Голенков, В. В. Проект открытой семантической технологии компонентного проектирования интеллектуальных систем. Часть 1: принципы создания. / В. В. Голенков, Н. А. Гулякина // Онтология проектирования. – 2014. – № 01. – С. 42-64.