Д. Смольская

АКТУАЛЬНОСТЬ АВТОМАТИЧЕСКОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА ТЕКСТОВ КОРЕЙСКОГО ЯЗЫКА

За последнее десятилетие 한류 «корейская волна», куда входят фильмы, сериалы, музыка, кухня и т.д., приобрела популярность во всем мире. Вместе с тем начал набирать популярность и корейский язык. Поэтому исследований на тему корейской лингвистики должно быть больше. В данной работе затронута морфология корейского языка, а именно существующие методы и инструменты автоматического морфологического анализа текстов корейского языка.

Такой анализ, выполняемый модулем автоматической морфологической обработки текста естественного языка, может быть следующим: нормализация словоформ (лемматизация, т.е. приведение различных словоформ к какому-то единому представлению – к исходной форме, или лемме); стемминг – еще один вид нормализации, когда разные словоформы приводятся к одной основе, а точнее «псевдооснове»; частеречевое тегирование (pos-тегирование), т.е. указание части речи для каждой словоформы в тексте; полный морфологический анализ – приписывание грамматических характеристик словоформе.

Основными методами, применяемыми для проведения автоматического анализа текста, являются лингвистический метод, статистический анализ, нейросетевой метод. К основным инструментам автоматического анализа относятся Tree Tagger, модуль Tied Sequence-to-Sequence Multi-Task, Two-level representation of Korean words и Hannanum.

Для примера практического применения модулей автоматического морфологического анализа текстов корейского языка был выбран модуль Tree Tagger.

В ходе проводимых экспериментов анализатор каждому слову предложения обрабатываемого текста присвоил POS-tag. Помимо того, что tree tagger определил часть речи каждого слова, он также определил грамматические показатели всех окончаний, добавленных к леммам. Однако анализатором были допущены морфологические, грамматические и стилистические ошибки, проанализировав которые, можно сделать вывод, что Tree Tagger может быть использован для морфологического анализа отдельных слов, но не для анализа целых текстов, так как он анализирует все слова отдельно от контекста и использует только первые их словарные значения.