Задачи построения программных морфологических моделей для русского языка

Большакова Елена Игоревна[1], **Сапин Александр Сергеевич**[2]

[1] Кафедра алгоритмических языков, e-mail: eibolshakova@gmail.com

[2] Кафедра алгоритмических языков, e-mail: alesapin@gmail.com

Методы и средства автоматической обработки текстов на естественном языке (АОТ) относятся к числу актуальных научных направлений в области искусственного интеллекта. В большинстве систем АОТ применяется морфологический анализ текста, результаты которого необходимы для последующих этапов его обработки. Также морфологический анализ используется при построении распределенного векторного представления слов высоко флективного языка, каковым является русский. Известные открытые морфологические процессоры для русского языка реализуют в первую очередь такие виды анализа (разбора) словоформы, как определение ее части речи, нормальной формы (леммы) и морфологических тегов (характеристик: падеж, род и т.д.). Однако не менее важным является решение сопутствующих задач: выбор нужного варианта анализа в случае неоднозначности (разрешение морфологической омонимии), разбор новых слов, отсутствующих в словаре процессора, а также определение морфемного состава слова (разбиения его на составляющие морфы). Последнее не реализовано ни в одном морфопроцессоре, однако необходимо для предсказания смысла новых слов языка.

В докладе рассматриваются программные модели, разработанные при создании многофункционального морфологического процессора XMorphy [1]. Процессор базируется на словарной модели морфологии русского языка, его словарь словоформ русского языка основан на размеченных данных проекта Opencorpora (opencorpora.ru) и представлен компактной структурой данных (DAFSA) с эффективным доступом к хранимым элементам. Для морфологического разбора несловарных словоформ применяется ряд эвристик, включая отсечение известного префикса и аналогию по финальным буквам словоформы. Дополнительно реализован модуль автоматической конвертации исходных морфологических тегов в универсальную систему разметки UD (universaldependencies.org), все более востребованную при разработке современных систем АОТ.

Для разрешения морфологической омонимии были рассмотрены и реализованы два подхода. В программной модели бесконтекстного разрешения осуществляется выбор наиболее вероятного варианта разбора для заданной словоформы, который определяется на базе статистики, собранной по нескольким размеченным корпусам русского языка (НКРЯ, ГИКРЯ, SynTagRus).

Для создания модели контекстного разрешения рассматривались методы машинного обучения с учителем по размеченным данным: метод условных случайных полей (CRF) и обучение свёрточной нейронный сети (CNN).

Уникальной особенностью процессора XMorphy является встроенный модуль, выполняющий сегментацию заданного слова на морфы (минимальные значимые единицы языка) с определением их типа. Данная задача решается как классификация букв слова по соответствующим типам морфем (приставка, корень, суффикс, окончание, постфикс). При помощи машинного обучения были построены несколько программных моделей сегментации: на основе метода CRF, на базе деревьев решений с градиентным бустингом (GBDT) и на базе рекуррентных нейронных сетей (LSTM). В качестве обучающих выборок применялись словари морфемного разбора русского языка.

Список литературы

[1] Bolshakova E. I., Sapin A. S. A morphological processor for Russian with extended functionality. // Analysis of Images, Social Networks and Texts. Lecture Notes in Computer Science, 10716. Springer, 2018, pp. 22-33.