Многовариантность в синтаксических анализаторах русского языка

Шамаева Елена Денисовна

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: derinhelm@yandex.ru

Научный руководитель *— Волкова Ирина Анатольевна*

Синтаксический анализ русского языка принципиально неоднозначен. С помощью многовариантного синтаксического анализатора можно выявить предложения с несколькими вариантами разбора. Он может быть использован и как вспомогательный инструмент для систем машинного перевода, извлечения сущностей и других.

Несколько вариантов разбора предложения могут быть вызваны омонимией на морфологическом или синтаксическом уровнях. Синтаксические анализаторы на вход получают результат работы морфологического анализатора. Однако каждая словоформа может иметь несколько вариантов морфологического разбора. В корпусе СинТагРус 51% словоформ имеет 1 вариант разбора, 26% — 2 варианта разбора, 13% — 3 варианта разбора [1].

Существующие синтаксические анализаторы либо основаны на правилах и грамматиках, либо используют машинное обучение и статистику. Некоторые синтаксические анализаторы, основанные на правилах, снимают морфологическую омонимию в процессе синтаксического разбора. Многие статистические синтаксические анализаторы (например, MaltParser [1]) используют для каждой словоформы один вариант морфологического разбора, то есть снимают морфологическую омонимию до начала синтаксического разбора. В данной работе рассматривается, каким образом статистический синтаксический анализатор снимает морфологическую омонимию.

В синтаксическом анализаторе по моделям управления [2] на каждом этапе синтаксического разбора связываются два слова. Для выбора пары слов используется база данных обобщенных моделей управления. Обобщенная модель управления показывает, насколько часто данная комбинация словоформ или морфологических слов (то есть морфологических характеристик) встречалась в корпусе. На

[1] Использовался морфологический анализатор pymorphy2 [3], были отброшены маловероятные (по мнению pymorphy2) результаты и результаты, полученные с помощью эвристических правил pymorphy2.

каждом этапе разбора связываются слова с «наилучшей» моделью управления.

В начале синтаксического анализа все словоформы считаются «неразобранными»: каждой словоформе соответствует несколько вариантов морфологического разбора. В процессе синтаксического разбора словоформа становится «разобранной». Она добавляется в дерево зависимостей (которое строится при разборе), фиксируется один вариант морфологического разбора этой словоформы. На данном этапе исследования применяется следующая эвристика: на каждом шаге моделью управления связываются разобранное и неразобранное слово (неразобранное слово становится разобранным).

На каждом этапе синтаксического разбора можно применить несколько моделей управления (могут быть связаны разные пары слов). В результате может быть получено несколько вариантов разбора предложения. Причем в каждом варианте разбора каждой словоформе будет соответствовать ровно один морфологический разбор, то есть будет снята морфологическая омонимия.

В процессе синтаксического разбора не будут потеряны варианты разбора, как это может случиться при снятии морфологической омонимии до синтаксического анализа. Для оптимального получения нескольких вариантов разбора планируется исследовать применимость алгоритмов поиска в графе. Эти варианты разбора могут использоваться автоматическими системами обработки текста, такими как машинный перевод, выделение терминов и другими.

Литература

1. Nivre J., Hall J., Nilsson J., Chanev A., Eryigit G., Kubler S., et al. MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering, P. 95–135.
2. Шамаева Е. Д. Синтаксический анализатор предложений русского языка на основе моделей управления // XXVII МЕЖДУНАРОДНАЯ НАУЧНАЯ КОНФЕРЕНЦИЯ СТУДЕНТОВ, АСПИРАНТОВ И МОЛОДЫХ УЧЕНЫХ «ЛОМОНОСОВ»: `https://olymp.msu.ru/archive/Lomonosov_2020_2/data/19259/uid432441_report.pdf`
3. Морфологический анализатор pymorphy2: `https://pymorphy2.readthedocs.io/en/stable/`