Статистические методы снятия морфологической омонимии

Смирнова Александра Сергеевна

Студент

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

E-mail: a.smirnova94@gmail.com

Одной из основных сложностей автоматического морфологического анализа текстов является омонимия - языковое явление, при котором у различающихся по смыслу слов совпадают некоторые словоформы. Например, «мыла» может быть словоформой глагола «мыть» или существительного «мыло». Такая омонимия называется частеречной, т.к. одной словоформе соответствуют начальные формы слов разных частей речи. Если же неоднозначность проявляется в определении падежа, числа, рода и других морфологических признаков, то омонимия называется грамматической. Например, словоформа «машины» имеет либо единственное число и родительный падеж, либо множественное число и именительный падеж.

Выделяют два основных подхода к снятию морфологической омонимии: основанный на правилах и статистический. Суть первого подхода заключается в применении набора правил, на основе которых удаляются неподходящие или выбираются правильные варианты морфологического разбора слов. Статистический подход опирается на статистическую информацию, собранную по размеченному корпусу текстов со снятой омонимией.

Если сравнивать эти подходы, то статистический обладает двумя важными достоинствами: он не требует экспертных знаний, необходимых для составления правил, и практически не зависит от предметной области анализируемых текстов.

Существует множество работ, посвященных использованию статистических методов при решении задачи снятия морфологической омонимии, например [2], и для достижения высоких результатов все они учитывают контекст вокруг омонимичного слова. Например, морфологические признаки для словоформы «машины» контекст определяет следующим образом: «на парковке стояли машины» («стояли» указывает на множественное число) или «у него не было машины» («было» указывает на родительный падеж). Однако, как правило, очень высокие результаты достигаются только при снятии частеречной омонимии, а задача снятия грамматической омонимии представляется более сложной, а иногда просто не решается.

Данная работа посвящена исследованию различных статистических методов снятия как частеречной, так и грамматической омонимии в текстах на русском языке. Выбраны методы, учитывающие контекст вокруг омонима: условные случайные поля [1], технология word2vec [3] и скрытые модели Маркова [2]. В работе планируется программно реализовать соответствующие методы и сравнить результаты их работы.

Литература

1. Антонова А. Ю., Соловьев А. Н. Использование метода условных случайных полей для обработки текстов на русском языке // Информационные технологии и системы, Калининград, 2013, С. 321–325.
2. Порохнин А. А. Анализ статистических методов снятия омонимии в текстах на русском языке // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. 2013. № 2. С. 168–174.
3. Технология word2vec: `https://code.google.com/archive/p/word2vec`