Автоматический анализ текстов. Синтаксический и семантический анализ

Аношин Павел Игоревич
Магистрант ИКБСП,
Россия, г. Москва
E-mail: pasha.a.505@gmail.com

Научный руководитель: **Капалин Владимир Иванович**
д.т.н. профессор. Кафедра автоматизирова

Автоматический анализ текста представляет собой операцию, которая из заданного текста на естественном языке извлекает грамматическую и семантическую информацию, содержащуюся в тексте. Автоматический анализ выполняется по некоторому алгоритму в соответствии с заранее разработанным описанием данного языка. Обратная операция называется автоматическим синтезом текста.

Автоматический анализ является одним из важнейших этапов в различных видах автоматической обработки текстов:

· автоматического реферирования;

· автоматического перевода;

· информационного поиска и т.п. [2].

Автоматический анализ не стоит путать с автоматическим исследованием текстов, в котором практически полностью отсутствуют данные о языке обрабатываемого текста, и обработка текста осуществляется алгоритмом с целью создания описания языка. В алгоритмах автоматического анализа, как правило, имеются сведения о языке (его «грамматика») и сведения о самом процессе анализа («механизм», т.е. алгоритм автоматического анализа).

Любая современная система анализа текста, в том числе поисковые машины, осуществляющие поиск документов в сети Интернет, содержит те или иные модули автоматического лингвистического анализа. Необходимыми этапами лингвистического анализа практически в любой современной системе являются:

· токенизация (разбиение на орфографические слова и выделение границ предложений);

· морфологический анализ (разбор слова как части речи).

Некоторые системы могут включать и иные модули:

· модуль синтаксического анализа (синтаксический парсер), главной задачей которого является представление предложения в качестве синтаксической структуры, такой как дерево зависимостей или дерево непосредственных составляющих или частичного синтаксического

анализа, или модуль выделения отдельных словосочетаний внутри текста;

· модуль семантического анализа, устанавливающий семантические отношения между словами текста и объединяющий языковые выражения, которые относятся к одному и тому же понятию. Семантический модуль не может работать без различного рода лексикографических ресурсов, таких как информационно-поисковые тезаурусы или лингвистические онтологии;

· модуль разрешения анафоры и т.д.

Как уже говорилось, целью синтаксического анализа является автоматическое построение дерева фразы, нахождение взаимозависимостей между разными элементами предложения. Если функциональное дерево фразы успешно построено, то из предложения можно выделить смысловые элементы, такие как: логический субъект, логический предикат, прямые и косвенные дополнения, а также различные виды обстоятельств [5].

Пример синтаксического дерево предложения «Мама мыла раму» в упрощенном графическом виде, изображен на рисунке 1:

Рисунок 1. Синтаксическое дерево предложения «Мама мыла раму»

Зная структуру предложения, можно сделать достаточно глубокий анализ и в дальнейшем использовать это на практике, например, создать систему автоматического перевода. В упрощенном виде это будет выглядеть так: выполнить каждого слова по словарю, а после сгенерировать новое предложение из синтаксического дерева.

Основной проблемой синтаксического анализа текста является разрешение неоднозначностей синтаксиса. Эта проблема решается двумя подходами: формально-графическим или вероятностно-статистическим. С помощью первого подхода создаются сложные системы правил, с помощью которых в каждом конкретном случае можно принимать решение в пользу какой-либо синтаксической структуры. Второй подход основан на сборе статистики встречаемости различных структур в похожих текстах, на основе которой затем происходит выбор варианта структуры [3].

Современные разработки в области синтаксического анализа имеют тенденцию к тому, что формально-грамматические методы анализа планомерно вытесняются методами, ориентирующимися на вероятностные оценки. Методы вероятностного характера однозначно не способны обеспечить полную точность анализа, но их результаты работы с реальными текстами показывают весьма удовлетворительные результаты для многих применений. Что касается затрат на разработку, то здесь однозначно выигрывают вероятностные анализаторы: стоимость разработки из значительно ниже, чем стоимость разработки структурных моделей естественного языка.

Семантический (смысловой) анализ необходим для оценивания смысла передаваемой информации, соотношения ее с информацией, которая хранилась до появления обрабатываемой информации. Семантические связи между словами или другими единицами языка отражаются в семантических словарях.

Задачами семантического анализа являются:

· построение семантической интерпретации слов и конструкций;

· установление семантических отношений между различными элементами текста.

При семантическом анализе предложений используют падежные грамматики и семантические валентности, а семантика предложения задается через связи главного слова (глагола) с его семантическими актантами [1].

Основой семантического анализа является утверждение, что конкретное значение слова не является элементарной семантической единицей. Оно, в свою очередь, делится на более мелкие единицы — единицы словаря семантического языка, являющиеся своеобразными атомами, комбинации которых складываются в «молекулы» — значения слов естественного языка. Именно семантический анализ дает возможность решить проблемы многозначности (омонимии), которая часто возникает при автоматическом анализе на разных языковых уровнях.

Семантический анализ текста является одной из наиболее сложных проблем таких областей как искусственный интеллект и компьютерная лингвистика. Результаты семантического анализа текстов могут быть применены для решения задач диагностирования больных в психиатрии, предсказания результатов выборов в политологии. Однако, несмотря на свою востребованность, семантический анализ остается одной из сложнейших математических задач. Главная проблема заключается в том, как «научить» компьютер однозначно верно трактовать образы, которые пытался передать автор текста [4].

В заключении стоит отметить, что ценность автоматического анализа текста на данный момент особенно высока, поскольку человек уже не в состоянии самостоятельно обработать современные объемы информации. Автоматический анализ текста находит применение в самых различных сферах, таких как бизнес (автоматическая обработка и классификация документов), политология и социология (предсказание результатов выборов или будущих общественных волнений на основе записей пользователей в социальных сетях), филология (определение авторства произведений, авторского стиля), в экспертных системах, системах машинного перевода, поисковых системах, а также во многих других.

1. Барышникова Надежда Юрьевна Обработка запросов на естественном языке на основе семантических сетей и шаблонов // Вестник АГТУ. Серия: Управление, вычислительная техника и информатика. 2016. № 4. URL: http://cyberleninka.ru/article/n/obrabotka-zaprosov-na-estestvennom-yazyke-na-osnove-semanticheskih-... (дата обращения: 11.06.2017).
2. Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. — СПб: НИУ ИТМО, 2013. — 72 с.
3. Кагиров Ильдар Амирович, Леонтьева Анастасия Борисовна Автоматический синтаксический анализ русских текстов на основе грамматики составляющих // Приборостроение. 2008. № 11. URL: http://cyberleninka.ru/article/n/avtomaticheskiy-sintaksicheskiy-analiz-russkih-tekstov-na-osnove-gr... (дата обращения: 11.06.2017).
4. Мочалова Анастасия Викторовна Алгоритм семантического анализа текста, основанный на базовых семантических шаблонах с удалением // Научно-технический вестник информационных технологий, механики и оптики. 2014. № 5 (93). URL: http://cyberleninka.ru/article/n/algoritm-semanticheskogo-analiza-teksta-osnovannyy-na-bazovyh-seman... (дата обращения: 11.06.2017).
5. Чапайкина Н. Е. Семантический анализ текстов. Основные положения // Молодой ученый. — 2012. — № 5. — С. 112-115.