УДК 004.031.42

М.Ю. ПОЛЯКОВА, студентка НТУ "ХПИ", Харьков,
Б.Н. СУДАКОВ, к.т.н., проф. НТУ "ХПИ", Харьков

РАЗРАБОТКА ПОДХОДА К СОЗДАНИЮ АЛГОРИТМА СИНТАКСИЧЕСКОГО АНАЛИЗА ЕСТЕСТВЕННО-ЯЗЫКОВОГО ТЕКСТА ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

Рассмотрены существующие методы синтаксического анализа естественно-языкового текста и выделены основные преимущества и недостатки. Разработан усовершенствованный алгоритм синтаксического анализа. Показано, что параллельное использование синтаксического и семантического анализа позволяет сократить временные затраты на обработку естественно-языкового текста. Ил.: 1. Библиогр.: 10 назв.

Ключевые слова: синтаксический анализ, семантический анализ, естественно-языковый текст.

Постановка проблемы. В настоящее время существует множество методик информационного поиска, условно разделенных на три группы. К первой группе относятся статистические методы, которые являются наиболее распространенными методами информационного поиска. Основной их особенностью является качественная математическая модель, позволяющая получать хорошие оценки релевантности файлов поиска. Поисковые машины, основанные на данных методах, отличаются простотой интерфейса. Основным минусом данного метода является тот факт, что не учитывается смысловая нагрузка текста документов коллекции и текста запроса. Отсутствие учета смысловой нагрузки текстов зачастую приводит к нерелевантным результатам. Примерами поисковых машин такого типа являются поисковые машины Google, Yandex, Rambler, Yahoo и т.д.

Основная идея второй группы методов информационного поиска заключается в том, что все исходные данные представлены в виде онтологий, а поиск ведется путем указания свойств искомого объекта. Данные методы, в отличие от статистических, учитывают смысловую нагрузку информации, поскольку информация представлена в виде онтологий. Однако данные методы имеют ряд недостатков:

– сложность пользовательского интерфейса, требующая от пользователя дополнительных затрат на конкретизацию объектов и свойств;

– большинство информации в Интернет представлено в виде простых HTML-страниц и не содержат семантического описания

контента. В качестве примера подобной системы можно рассматривать систему АСНИ (Автоматизированная система научных исследований).

К третьей группе методов информационного поиска относятся методы, которые помимо статистических методов поиска используют методы семантического анализа текстов. Данная группа методов развивается в настоящее время наиболее интенсивно. Основным плюсом систем комбинированного типа является комбинация качественной статистической модели поиска и учета семантических конструкций. Основные минусы подобных систем, существующих в настоящее время:

– большое время отклика;

– мало где используются механизмы логического вывода;

– ограничения на структуру запроса (при использовании простого пользовательского интерфейса);

– необходимость установки дополнительных параметров поиска (при использовании сложных пользовательских интерфейсов);

– большинство систем подобного типа используют в качестве исходной информации стандартные тексты, проводя семантический анализ на конечном этапе задачи поиска, что приводит к медлительности данных систем.

Несмотря на то, что третья группа методов наиболее полно отвечает требованиям, предъявляемым к системам информационного поиска на основе семантики, все системы данного типа имеют недостатки.

Анализ литературы. В [1 – 3] изложены основные принципы проектирования экспертных систем. Рассматриваются базовые концепции технологии экспертных систем. Освещаются основные схемы представления проблемно-ориентированных знаний в программах и методы применения этих знаний к построению искусственного интеллекта. В [4, 5] освещены теоретические концепции и практические методы автоматической обработки естественно-языкового текста на всех уровнях лингвистического анализа. В [6, 7] представлены методы синтаксического анализа. В [8] рассматривается решение по синтаксическому разбору структуры вводимого текста, применяемых в средах разработки для повышения эффективности работы программиста. В электронных источниках приведены базовые понятия синтаксического компьютерного анализа.

Цель статьи. Разработать подход к построению алгоритма синтаксического анализа естественно-языкового текста.

Актуальность вопроса. В последние годы большое распространение получили различного рода интеллектуальные системы, выполняющие обработку текстов на естественном языке (ЕЯ). В

простейшем случае они используются в информационно-поисковых системах (ИПС), ориентированных на естественно-языковое общение с пользователем.

Одним из основных элементов ИПС является лингвистический процессор (ЛП), выполняющий роль посредника между пользователем и базой данных, в которой хранится интересующая его информация. Классическая структура лингвистического процессора содержит три последовательных блока – для морфологического, синтаксического и семантического анализа текста. Синтаксический анализ, является главным блоком, определяющим качество работы ЛП в целом.

Задача синтаксического анализа. Используя морфологическую информацию о словоформах, необходимо построить синтаксическую структуру входного предложения (осуществить разбор предложения). Морфологический анализ – обработка отдельных словоформ, в результате которой каждой словоформе относится в соответствие ее информация – характеристика, которая отображает те свойства словоформы, которые необходимы для следующего синтаксического анализа.

К началу синтаксического анализа весь текст представляется в виде последовательности характеристик к словоформам, так что алгоритм синтаксического анализа имеет дело не со словоформами, а лишь с соответствующими характеристиками.

Программа синтаксического анализа, как правило, состоит из двух компонентов: сегментации предложения и установления связей между словами. Компоненты работают параллельно или последовательно, в зависимости от архитектуры синтаксического модуля.

Сегментация соединяет простые предложения в составе сложного. В любое простое предложение могут быть вставлены причастные или деепричастные обороты, придаточные предложения, которые, в свою очередь, тоже могут быть "разбиты" другими оборотами. Существуют примеры, когда части цельного высказывания находятся на значительном расстоянии друг от друга, а глубина вложения небольших предложений теоретически не ограничена.

На следующем этапе происходит установление связей между словами в построенных сегментах. На этом этапе появляется проблема морфологической омонимии, то есть неоднозначности. Морфологическая омонимия возникает, когда одна и та же форма может выражать разные морфологические значения. Пример: форма "весла" может быть родительным падежом единственного числа, именительным падежом множественного числа.

Явление морфологической омонимии весьма негативно отражается на скорости работы программы синтаксического анализа. На "длинных" предложениях количество комбинаторных вариантов иногда достигает нескольких сотен, поэтому используются разного рода математические и лингвистические ухищрения, позволяющие избежать анализа всех комбинаторно возможных вариантов.

Разработка основных принципов построения алгоритма синтаксического анализа. Синтаксический анализ может рассматриваться как процесс поиска дерева синтаксического анализа. Существуют два противоположных способа задания соответствующего пространства поиска. Во-первых, можно начать с вершины и искать дерево, листьями которого являются соответствующие слова. Такой способ называется нисходящим синтаксическим анализом (поскольку символ вершины изображается в верхней части рисунка, на котором показано дерево, повернутое корнем вверх). Во-вторых, можно начать со слов и выполнять поиск дерева начиная с корня. Такой метод называется восходящим синтаксическим анализом.

Традиционным методом построения синтаксической структуры является метод фильтров. В данном методе построение дерева зависимостей начинается с построения наборов всевозможных связей (синтаксических отношений) между словами. При этом для большинства слов устанавливается несколько потенциально возможных связей с различными управляющими словами. Применение фильтров позволяет отбросить многие из первоначально установленных связей и, в идеале, получить количество связей, равное числу слов (при условии, что между всеми словами установлены синтаксические отношения). В чистом виде метод фильтров для практической реализации неприменим, так как число всевозможных связен между словами весьма велико, а число всевозможных способов выбора из них конкретного дерева зависимостей огромно. На практике для получения эффективных алгоритмов необходимо применять методы, направляющие и ускоряющие выбор правильных вариантов анализа. Одним из таких методов является метод параллельного синтаксического и семантического анализов, в котором последний выступает в качестве фильтра тупиковых вариантов.

Создание алгоритма синтаксического анализа естественно-языкового текста. Алгоритм представляет собой структуру построения синтаксических групп (СГ), которая состоит из существительного с зависимыми от него словами. Присоединяя к существительному словоформы, находящиеся в предложении слева от него, отыскиваются все СГ (В1 – В2) (см. рис.).

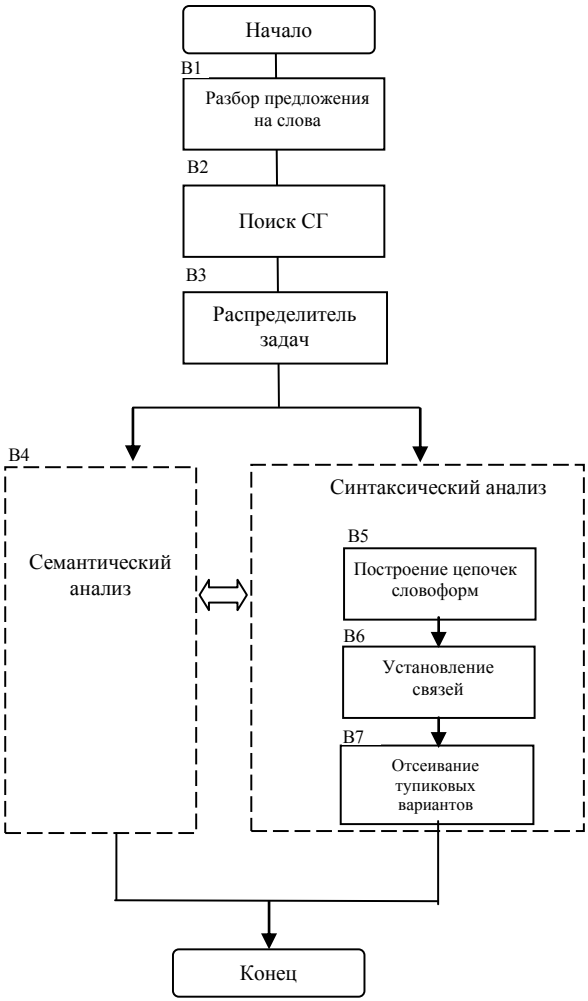

Рис. Алгоритм синтаксического анализа

Для установления поверхностно-синтаксических связей в группах анализ начитается с конца СГ. При этом анализируются пары слов с целью установления связей между ними. Если связь установлена, то переходят к следующей паре. Если для соотношения слов найдено

главное слово, то записывается его номер. Анализ продолжается до тех пор, пока в анализируемой группе все соотношения слов не будут иметь главные слова, кроме одной словоформы. Если этого сделать не удается, то уточняется вариант разбиения на группы. Строятся цепочки словоформ и устанавливается связь между ними в предложении (В5 – В6).

Для отсеивания тупиковых вариантов параллельно используется семантический анализ (В4), направленный на решение задач, связанных с возможностью понимания смысла фразы и выдачи запроса поисковой системе в необходимой форме. Используя модуль семантического анализа текстов, повышают эффективность лингвистических систем (программ автоматического перевода, информационно-поисковых систем, рубрикаторов и рефераторов текстов) на основе реализации извлечения и обработки смысловой информации. Таким образом, время, затрачиваемое на синтаксический анализ, сокращается за счет отброса заранее не существующих соотношений между словоформами (В7). Описанная схема алгоритма представлена на рис.

Выводы. Предложен новый метод синтаксического анализа для информационно-поисковой системы, суть которого заключается в более быстром нахождении связей между словоформами и подготовке качественного материала для семантического уровня. Это позволяет в целом улучшить качество работы лингвистического процессора поисковой системы, поскольку семантический и синтаксический анализы проходят параллельно. За счет этого заранее отбрасываются тупиковые варианты.

Список литературы: **1**. *Джексон П.* Введение в экспертные системы / *П. Джексон*. – М.: Основа, 2001. – 624 с. **2**. *Карпова Г.Д.* Компьютерный синтаксический анализ: описание моделей и направлений разработок / *Г.Д. Карпова, Ю.К. Пирогова, Т.Ю. Кобзарева, Е.В. Микаэлян*. – М.: ВИНИТИ, 1991. – 304 с. **3**. *Поспелова Д.А.* Искусственный интеллект / *Д.А. Поспелова*. – М.: Наука, 1990. – 246 с. **4**. *Романенко В.Н.* Сетевой информационный поиск / *В.Н. Романенко*. – М.: Профессия, 2005. – 290 с. **5**. *Фостер Дж.* Автоматический синтаксический анализ / *Дж. Фостер*. – М.: Мир, 1975. – 71 с. **6**. *Попов Э.В.* Общение с ЭВМ на естественном языке / *Э.В. Попов*. – К.: Наука, 1992. – 254 с. **7**. *Анно Е.И.* Типологии алгоритмов синтаксического анализа (для формальных моделей естественного языка) / *Е.И. Анно*. – СПб.: Питер, 1989. – 152 с. **8**. *Омар А.Х. Авадала* Подход к синтезу естественно-языковых сообщений по формализованному представлению базы знаний: автореф. дис. на соискание ученой степени канд. техн. наук / *Омар А.Х. Авадала*. – Х., 2001. – 20 с. **9**. *Руских И.В.* Инкрементальный синтаксический анализ в средах разработки и текстовых редакторах // Нижегородский университет. – 2007. – С. 277 **10**. Автоматическая обработка текста [Электронний ресурс] / *Леонтьев Н.Н.* – 2003. – С. 5. – Режим доступа к статье: http:\\www.aot.ru/technology/html.

Статья представлена д.т.н., проф. НТУ "ХПИ" Серковым А.А.

УДК 004.031.42

Розробка підходу до створення алгоритму синтаксичного аналізу природно-мовного тексту інформаційно-пошукових систем/ Судаков Б.М., Полякова М.Ю. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2011. – № 17. – С. 128 – 134.

Розглянуто існуючі методи синтаксичного аналізу природно-мовного тексту та виділено основні переваги та недоліки. Розроблено удосконалений алгоритм синтаксичного аналізу. Показано, що паралельне використання синтаксичного та семантичного аналізу дозволяє скоротити часові витрати на обробку природно-мовного тексту. Іл.: 1. Бібліогр.: 10 назв.

Ключові слова: синтаксичний аналіз, семантичний аналіз, природно-мовний текст.

UDC 004.031.42

Developing an approach to creating an algorithm parsing natural language text of information retrieval systems / Sudakov B.N., Poliakova M.U. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2011. – №. 17. – P. 128 – 134.

The article deals with the existing methods of parsing natural language text. The main advantages and disadvantages were identified. Developed an improved algorithm for parsing. It is shown that the parallel use of syntactic and semantic analysis can reduce the time required for processing natural language text. Figs.: 1. Refs.: 10 titles.

Key words: syntactic analysis, semantic analysis, natural language text.

Поступила в редакцию 26.01.11