DOI 10.37882/2223–2966.2021.07.09

ПРОБЛЕМА АВТОМАТИЧЕСКОЙ ОБРАБОТКИ НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ СООБЩЕНИЙ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ[1]

Гагарина Лариса Геннадьевна
Д.т.н., профессор, Национальный исследовательский университет «МИЭТ»; Директор, Институт системной и программной инженерии и информационных технологий
gagar@bk.ru
Капитанов Андрей Иванович
Ассистент, Национальный исследовательский университет «МИЭТ»
andrey@kapdx.ru

THE PROBLEM OF AUTOMATIC PROCESSING OF UNSTRUCTURED TEXT MESSAGES IN REAL TIME[2]

L. Gagarina
A. Kapitanov

Summary. Regardless of the task of automatically processing unstructured text messages, one of the first steps is to present the data in a form suitable for machine processing. This paper analyzes the ways of presenting text messages and pays special attention to the possibility of processing messages in the context of time constraints. An algorithm is described that allows real-time extraction of facts from unstructured text.

Keywords: GLR parser, NLP, real-time, text analysis, text normalization.

Аннотация. Независимо от задачи автоматической обработки неструктурированных текстовых сообщений, одним из первых этапов является представление данных в виде, пригодном для машинной обработки. В данной работе приводится анализ способов представления текстовых сообщений, а также уделяется особое внимание возможности обработки сообщений в разрезе временных ограничений. Приводится описание алгоритма, позволяющего в режиме реального времени извлекать факты из неструктурированного текста.

Ключевые слова: анализ текста, нормализация текста, обработка естественного языка, режим реального времени, GLR-парсер.

Введение

С не прекращающимся ростом количества информации становится всё более актуальной задача машинной обработки неструктурированных текстовых сообщений на естественном языке. В зависимости от цели обработки документов (определение тональности отзывов, фильтрация спама, выявление дубликатов, машинный перевод, и т.д.) применяют различные методы компьютерной лингвистики. Однако независимо от задачи, одним из первых этапов является представление данных в виде, пригодном для машинной обработки.

В данной работе уделяется особое внимание возможности обработки сообщений в режиме реального времени — это позволяет оперативно решать такие задачи как: реализация голосового помощника, обнаружение радикального или экстремистского контента и другие. Целью работы являются аналитический обзор основных этапов обработки неструктурированной текстовой информации и анализ существующих проблемных ситуаций в разрезе временных ограничений.

Способы представления текстовых сообщений

В зависимости от необходимой глубины понимания смысла и сложности анализа выделяют следующие способы представления текстовых сообщений:

- лексико-морфологический анализ;
- синтаксический анализ;
- семантический анализ;
- прагматический уровень анализа (онтологии).

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19–37–90144.
[2] The reported study was funded by RFBR, project number 19–37–90144.

Лексико-морфологический анализ. Сообщение разбивается на абзацы, предложения и слова (токенизация). Особое внимание необходимо уделить обработке сообщений для флективных языков. В случае обработки текстов на флективном языке (например, русском) необходимо привести все слова в начальную форму. Для определения нормальной формы зачастую используются готовые словари, например модификацию грамматического словаря русского языка А.А. Зализняка, содержащую приблизительно 100 тысяч начальных форм и около 2,3 миллиона словоформ русского языка [2]. Для слов, не содержащихся в словаре, используется алгоритм стемминга — нахождения основы слова. Различные алгоритмы стемминга различаются производительностью и точностью. Один из примитивных стеммеров ищет флективную форму по таблице поиска. Одним из главных преимуществ данного подхода заключается в его скорости и простоте.

Алгоритмы, основанные на усечении окончаний, как правило, используют небольшой список правил, по которым алгоритм учитывает форму слова для нахождения его основы [3, с. 138]. Использование справочной таблицы, состоящей из флективных форм и отношений корня и формы, не применяется.

Данные алгоритмы на порядок эффективнее, чем алгоритмы полного перебора. Однако существенной проблемой является сложность разработки такого алгоритма: разработчик должен быть компетентен в области лингвистики, в частности морфологии. Так как не все части речи имеют формализованные «правила усечения», решения, полученные данными алгоритмами, ограничиваются только частями речи, имеющими определенные суффиксы и окончания.

Синтаксический анализ. Происходит выделение предложений в тексте, а также определение структуры и связи между словами внутри предложения. Для реализации синтаксического этапа в рамках компьютерной лингвистики предложено большое число разных идей и методов, отличающихся способом описания синтаксиса языка, способом использования этой информации при анализе или синтезе предложений, а также способом представления синтаксической структуры предложения [4, с. 500]. Можно выделить три основных подхода: генеративный подход, восходящий к идеям порождающих грамматик Н. Хомского [5, с. 88]; подход, восходящий к идеям И. Мельчука и представленный в лингвистической модели «Смысл ⇔ Текст», а также подход, в рамках которого делаются те или иные попытки преодолеть ограничения первых двух подходов, в частности, теория синтаксических групп [1, с. 5].

Семантический анализ. Определяются значения всех слов, строится семантическая структура предложения на основе связей, которые были выделены на предыдущих этапах. В течение последних 30 лет были предложены и протестированы различные методы вычисления семантической близости слов, начиная от методов на основе лексических источников, заканчивая подходами на основе векторного представления. Векторное представление, в свою очередь, эволюционировало от гиперпространственного аналога языка к латентно-семантическому анализу, тематическому моделированию, дистрибутивной семантике, и, в итоге, к нейронным языковым моделям. Одним из важнейших факторов, влияющих на качество обработки текстовых сообщений на естественном языке, является устранение семантической омонимии. Наиболее популярным методом определения многозначных слов является поиск по заранее определенным значениям: например, спискам слов, найденные в словарях. Данный метод позволяет достаточно точно определить, является ли слово многозначным, однако не может сказать, к какой семантической группе принадлежит слово. Также данный метод имеет свойство со временем устаревать, т.к. по мере развития языка данный словарь необходимо пополнять неологизмами. Тезаурусы, семантические сети и другие специализированные структуры позволяют установить связи между значениями, однако их создание и поддержание в актуальном состоянии требует больших трудозатрат. К наиболее простым методам разрешения полисемии можно отнести выбор значения, наиболее приближенного по смыслу к контексту встречаемого многозначного термина.

Прагматический уровень анализа. Применение терминологий для определенной предметной области анализа (онтологий) и правил извлечения нужных объектов, использование прагматического слоя анализа текста. Итогом анализа является универсальное представление информации, что позволяет структурировать данные в нужном виде. Таким образом, данная технология позволяет эффективно решать задачи, связанные с интеллектуальным поиском и классификацией текстовой информации на естественном языке [7].

Особенности обработки текстовых сообщений в режиме реального времени

Для систем, работающих в режиме реального времени, важной характеристикой является скорость ответа. При обработке текстовых сообщений алгоритм должен выдавать результаты своей работы по мере продвижения вглубь текста. Такой особенностью обладает GLR-парсер (*Generalized Left-to-right Rightmost derivation parser*) — алгоритм, предназначенный для разбора по недетерминированным и неоднозначным грамматикам. В отличие от LR-анализатора, GLR-парсер

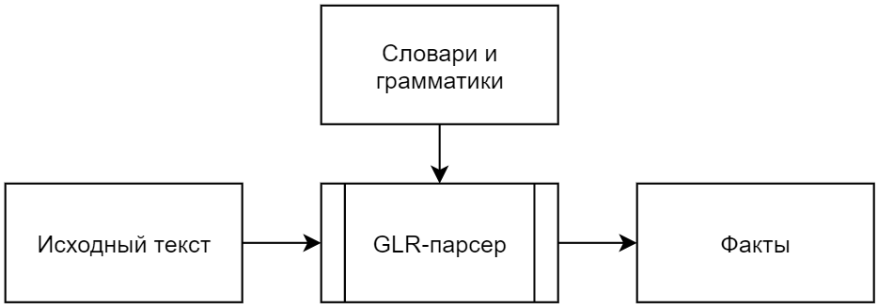

Рис. 1. Упрощенная схема работы сервиса обработки сообщений на основе GLR-парсера

позволяет работать с естественным языком. Несмотря на то, что в худшем случае алгоритм имеет сложность $O(n^3)$, у GLR-парсера есть следующие преимущества.

1. GLR-парсер работает за $O(n)$ при условии полной детерминированной грамматики.
2. GLR-парсер позволяет работать в «режиме реального времени» — то есть он выполняет как можно больше анализа в процессе считывания последовательности токенов.

GLR-парсер лег в основу Томита-парсера, используемого в сервисах компании Яндекс [6]. Его упрощенная схема представлена на рис. 1.

Парсер получает исходный текст, а также словарь и грамматику, по которым производится извлечение фактов из текста. Файл грамматики содержит шаблоны на формальном языке, описывающие возможные цепочки слов в тексте и определяющие формат вывода извлекаемых фактов. Словарь содержит понятия, ключевые слова, показывающие, как понятие может быть отражено в тексте, и лемму, к которой эти ключевые слова должны будут приводиться при отображении результатов.

Заключение

Важным этапом в автоматической обработке неструктурированных текстовых сообщений является представление данных в виде, пригодном для машинной обработки. Для этого в зависимости от необходимой глубины понимания текста применяются такие способы представления, как лексико-морфологический анализ, синтаксический анализ, семантический анализ, прагматический уровень анализа. При обработке текстовых сообщений в режиме реального времени алгоритм должен выдавать результаты своей работы по мере продвижения вглубь текста. GLR-парсер, обладающий данной особенностью, позволяет в режиме реального времени, используя словари и грамматики, извлекать факты из неструктурированного текста.

ЛИТЕРАТУРА

1. Большакова Е.И. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / Е.И. Большакова, К.В. Воронцов, Н.Э. Ефремова, Э.С. Клышинский, Н.В. Лукашевич, А.С. Сапин — М.: Изд-во НИУ ВШЭ, 2017. — 269 с. ISBN978–5–9909752–1–7.
2. Зализняк А.А. Грамматический словарь русского языка. Словоизменение. Изд. 5-е, испр. — М.: Аст-пресс, 2008.
3. Кузнецов П.С. Морфологическая классификация языков // Материалы к курсам языкознания. — М.: 2016. — 138 с. ISBN978–5–397–05350–1.
4. Лукашевич Н.В., Рубцова Ю.В. Объектно-ориентированный анализ твитов по тональности: результаты и проблемы //Аналитика и управление данными в областях с интенсивным использованием данных. 2015. С. 499–507.
5. Марчук А.А., Уланов А.В., Макеев И.В., Чугреев, А.А. Автоматическое извлечение параметров продуктов из текстов отзывов при помощи интернет-статистик // Труды Международной конференции Компьютерная лингвистика и информационные технологии Диалог-2013. 2013. т. 2. С. 81–91.
6. Томита-парсер [Электронный ресурс]: Томита-парсер — Технологии Яндекса — Режим доступа: https://yandex.ru/dev/tomita/
7. ABBYY FlexiCapture [Электронный ресурс]: ABBYY FlexiCapture. Решение для потокового ввода данных — Режим доступа: https://www.abbyy.com/ru/flexicapture/

Журнал «Современная наука: актуальные проблемы теории и практики»