М. И. Святощик (*Минск, МГЛУ*)

СИСТЕМЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ И ИХ УРОВНЕВОЕ ПОСТРОЕНИЕ

Данная статья посвящена анализу систем автоматической обработки текстов и их уровневому построению. Большое внимание уделяется технологиям визуализации больших объемов текстовой информации. Рассматривается многообразие систем автоматической обработки неструктурированных текстов. Выдвигается предложение об оперативном создании приложений (прикладных программных сетей) для автоматизированной

или автоматической обработки текстов на естественном языке. Предлагаются решения задач автоматической обработки текстов (АОТ), возникающих при создании высокотехнологичных интеллектуальных систем, обеспечивающих замену человеческого труда в интеллектуальной сфере, опирающейся на использование естественного языка.
Ключевые слова: естественный язык; автоматическая обработка текстов; компьютерная лингвистика; корпус текстов; естественный язык; текст.

This article is devoted to the analysis of automatic text processing systems and their level construction. Much attention is paid to technologies for visualizing large amounts of text information. The variety of systems for automatic processing of unstructured texts is considered. A proposal is put forward for the rapid creation of applications (application software networks) for automated or automatic text processing in natural language. Solutions to the problems of automatic text processing (ATP) that arise when creating high-tech intelligent systems that provide a replacement for human labor in the intellectual sphere based on the use of natural language are proposed.
Key words: natural language; automatic text processing; computational linguistics; text corpus; natural language; text.

Когда речь идет о создании перспективных информационных технологий, то проблемы автоматической обработки текстовой информации выступают на передний план. Это определяется тем, что естественный язык (ЕЯ) является не только инструментом мышления, но и универсальным средством общения – средством восприятия, накопления, хранения, обработки и передачи информации [1, с. 43]. Более того, ЕЯ становится также универсальным средством описания действительности и коммуникации с вычислительной системой. В наше время, когда пользователем может оказаться практически каждый, проблема взаимодействия человека с ЭВМ на естественном языке стала важной практической задачей.

Сегодня автоматическая обработка ЕЯ (в том числе и текста ЕЯ) – это бурно развивающаяся область научных исследований и коммерческих разработок, ставящих целью создание промышленных систем обработки ЕЯ, которые должны быстро и эффективно обрабатывать в режиме реаль-ного времени огромные потоки информации, циркулирующие в информа-ционных сетях [2, с. 17].

Автоматическая обработка текстов (АОТ) предполагает решение многих задач, которые условно можно разбить на два уровня. Задачи высокого уровня представлены задачами распознавания речи, реферирования текстов, генерации документов, машинного перевода, извлечения информации, обучения языку, т.е. приложениями. К задачам низшего уровня относят грамматический разбор, снятие смысловой многозначности, корректировку орфографии и синтаксический разбор, т.е. задачи собственно лингвистической обработки ЕЯ [1, с. 43]. К настоящему времени этот круг задач значительно расширился и в целом охватывает всю индустрию развития и поддержки компьютерной формы существования ЕЯ.

Главная проблема при решении указанных задач состоит в необходимости обрабатывать неструктурированные тексты. Единый типовой алгоритм их автоматической обработки создать не удается, поскольку конкретный вид алгоритма, в первую очередь, определяется строем языка [3, P. 24].

Многообразие систем автоматической обработки неструктурированных текстов сегодня вызывает необходимость их систематизации и классификации с целью упрощения выбора решения, наиболее адекватного для конкретной задачи [4, с. 5].

Обработка текстов, которые представляют собой неструктурированную информацию, например, патенты, истории болезней, преследует задачи, которые условно делятся на категории:

1. Распространенные пользовательские задачи, с которыми потребители сталкиваются постоянно. Здесь можно отметить фильтрацию спама, проверку орфографии, автоперевод небольших текстовых фрагментов.

2. Обработка внушительных массивов текста, например, полноценный автоперевод целостных текстов, поиск релевантных ответов на вопросы, построение рекомендательных систем, которые работают с большим объемом неструктурированной информации, аналитика отзывов и мнений [5, с. 21].

Отличительной особенностью данных задач является отсутствие формализации и сложность. В реально работающих современных системах данные проблемы не решены. Вместо полноценного набора решений используются вспомогательные методы, например, такие как:

- выделение ключевых словосочетаний и слов;
- классификация текстов;
- суммаризация (автоматическое реферирование).

Здесь большое внимание уделяется технологиям визуализации больших объемов текстовой информации.

Неотъемлемой частью многих систем обработки текстов являются корпусы. Слова в корпусах наделены полными грамматическими характеристиками, например, часть речи, форма, синтаксическая роль.

Корпусы – это входные данные для обучения в задачах классификации текстов по жанрам и темам, синтаксических программ и парсеров, которые применяются для снятия омонимии и допуска анафоры. При обучении машинных переводчиков применяются параллельные корпусы, которые состоят из одинаковых текстов на разных языках. Сбор корпусов осуществляется десятилетиями. Это очень трудоемкое исследование с участием больших групп научных специалистов. В качестве примера можно привести проект под названием «Национальный корпус русского языка». Данный проект реализуется уже тринадцать лет при поддержке компании «Яндекс» [6, с. 14].

Морфологические словари являются важным типом входных данных любой АОТ. Здесь можно упомянуть библиотеку «АОТ», которая применяется во многих коммерческих и исследовательских проектах. Библиотека представляет собой словарь Зализняка в цифровом варианте.

Еще одним распространенным типом входных данных являются семантические сети (тезаурусы). WordNet – самый известный тезаурус. WordNet – ресурс связанных между собой слов. Связь между словами осуществляется по типу семантических отношений. Например, гипонимия (обобщение – частное), синонимия, гиперонимия (частное – обобщение), меронимия (часть – целое). WordNet эффективен при решении задач классификации текста, машинного перевода, генерации текстов. Стоит отметить, что пока, к сожалению, русский аналог WordNet не разработан.

Развитие АОТ-систем уже в наши дни представляет коммерческий интерес и используется при решении следующих прикладных задач:

1. Machine Translation and Translation Aids – машинный перевод;
2. Text Generation – генерация текста;
3. Localization and Internationalization – локализация и интернационализация;
4. Contolled language – работа на ограниченном языке;
5. Word Processing and Spelling Correction – создание текстовых документов (ввод, редактирование, исправление ошибок);
6. Information Retrieval – информационный поиск и связанные с ним задачи.

Нужно отметить, что это деление достаточно условное, и в реальных системах часто встречается объединение функций. Так, для машинного перевода требуется генерация текста, а при исправлении ошибок приходится заниматься поиском вариантов словоформы и т.д.

На современном этапе существует острая необходимость в оперативном создании приложений (прикладных программных сетей) для автоматизированной или автоматической обработки текстов на ЕЯ. Это обусловлено активным ростом объемов текстовой информации. В качестве примеров подобной обработки можно привести фильтрацию и сбор данных, которые находятся в разных источниках, реферирование, извлечение знаний, аннотирование. При разработке приложений часто возникают такие сложности, как интеграция огромного числа программных компонентов, которые выполняют алгоритмы текстов на естественном языке, работают на разных уровнях текста, например, обработка абзацев, слов, предложений.

Для решения задач АОТ необходимо выполнить аналитику текста на различных уровнях представления. Виды анализа:

1. Графематический, в ходе которого из массива данных выделяются предложения и слова.

2. Морфологический, в ходе которого выделяется грамматическая основа, определяется часть речи, слова приводятся к словарной форме.

3. Синтаксический, в ходе которого выявляются синтаксические связи между словами в предложениях, строится синтаксическая структура предложений.

4. Семантический, в ходе которого выявляются семантические связи между синтаксическими группами и словами, извлекаются семантические отношения.

Каждый из описанных выше анализов представляет собой самостоятельную задачу. Она не имеет своего практического применения, но активно используется в качестве составной части более глобальных задач [7, с. 62].

Компьютерная лингвистика (КЛ) существует уже более полувека и известна также под названиями «машинная лингвистика», «автоматическая обработка текстов на естественном языке». В рамках КЛ исследователями и разработчиками предложено множество решений, которые являются достаточно перспективными. Стоит также отметить, что далеко не все эти решения были воплощены в жизнь в виде программных продуктов. Несмотря на этот недостаток, компьютерная лингвистика показывает вполне реальные результаты. Это видно по различным приложениям по автоматической обработке текстов на естественном языке. Дальнейшее развитие КЛ зависит от разработки новых приложений, различных языковых моделей, в которых пока не решены многие задачи.

Неотъемлемой частью многих систем обработки текстов являются корпусы. Все слова в корпусах имеют исчерпывающие грамматические характеристики, в перечень которых входят часть речи, форма слова, синтаксическая роль. Корпусы представляют собой входные данные, которые служат для обучения в задачах классификации текстов по жанрам и темам. Кроме того, они применяются для обучения синтаксических программ, применяемых в процессе снятия омонимии и разрешения анафоры.

Учитывая все вышесказанное, можно сделать вывод о том, что КЛ (компьютерная лингвистика) появилась на стыке математики, лингвистики, информатики и искусственного интеллекта.

На современном этапе наиболее разработаны модели морфологического синтеза и анализа. Необходимо уточнить, что модели синтаксиса пока не являются устойчиво и эффективно функционирующими моделями. Они не смогли достичь этого уровня, несмотря на присутствие большого количества предложенных методов и формализмов. Модели семантики и прагматики исследованы и формализованы еще меньше, но нужно учитывать, что автоматической обработки дискурса уже требует ряд приложений. Решение существующих проблем может активизировать используемые инструменты КЛ, а также применение корпусов текстов и машинного обучения.

ЛИТЕРАТУРА

1. Компьютерная лингвистика и перспективные информационные технологии / Г. Г. Белоногов [и др.]. // Научно-техническая информация. – 2004. – № 8. – С. 30–43.
2. *Забежайло, М. И.* Интеллектуальный анализ данных – новое направление развития информационных технологий / М. И. Забежайло // Науч.-технич. информация. Сер. 2. – 1998. – № 8. – С. 6–17.
3. *Brill, E.* An Overview of Empirical Natural Language Processing / E. Brill, R. J. Mooney // AI magazine, – 1997. – Vol. 18. – № 4. – P. 13–24.
4. *Луканин, А. В.* Автоматическая обработка естественного языка : учеб. пособие / А. В. Луканин. – Челябинск : Издат. центр ЮУрГУ, 2011. – 70 с.
5. *Луканин, А. В.* Инструментарий прикладного лингвиста / А. В. Луканин // Современные направления прикладной лингвистики : материалы I Студенч. науч.-практ. конф., Челябинск, 7–9 сент. 2008 г. / Челябин. гос. ун-т ; редкол. : Ф. Г. Самсонов [и др.]. – Челябинск, 2008. – С. 34.
6. *Козлова, Н. В.* Лингвистические корпуса : определение основных понятий и типология / Н. В. Козлова // Вестн. НГУ. Сер. Лингвистика. – Новосибирск, 2013. – С. 95.
7. *Апресян, Ю. Д.* Идеи и методы современной структурной лингвистики / Ю. Д. Апресян. – М. : Просвещение, 1966. – 301 с.