

Системы автоматической обработки текстов



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

Ключевые слова: автоматический перевод, неструктурированные данные, компьютерная лингвистика, извлечение смысла
Keywords: machine translation, unstructured data, computational linguistics, semantic analysis

Дмитрий Ильвовский,
Екатерина Черняк

Задачи обработки текстов — неструктурированной документации, историй болезни, патентов и диссертаций и т. п. — можно разбить на две условные категории. К первой относятся задачи, с которыми ежедневно сталкивается любой пользователь: проверка орфографии, фильтрация спама, автоматический перевод небольших фрагментов текста (несколько предложений) и др. С точки зрения исследователей в области **автоматической обработки текстов (АОТ)**, все эти задачи почти решены, и сегодня более актуальны задачи из второй категории, требующие обработки больших текстовых массивов: нахождение релевантных ответов на вопросы (задачи «вопрос-ответ»), полноценный машинный перевод целостных текстов, анализ мнений и отзывов, конструирование рекомендательных систем, работающих с большими массивами неструктурированных данных. Отличительная особенность таких задач — их сложность и отсутствие формализации, приводящие к тому, что для них пока еще нет полноценного набора решений, а применяются вспомогательные методы выделения ключевых слов и словосочетаний, суммаризации (автоматического реферирования) текстов и классификации текстов.

Теоретическую основу автоматической обработки текстов составляет компьютерная лингвистика, наиболее востребованы в которой методы машинного обучения, статистического анализа, модели Маркова, логические модели и модификации этих методов с учетом специфики Больших Данных [1]. Существует несколько подходов к такой модификации: распараллеливание алгоритмов, применение методов снижения раз-

мерности, предобработка данных, в ходе которой целостные тексты заменяются их отдельными элементами. Несмотря на различие между национальными языками, лингвистические методы могут быть универсальными — некоторые морфологические и синтаксические модели удастся использовать для анализа текстов как на английском, так и на русском языке.

Появление компьютеров, на которых можно было целиком хранить, обрабатывать корпуса, содержащие наборы эталонных текстов, и проводить сложные вычисления, позволило активно использовать статистические методы и методы машинного обучения для работы с текстами [2]. В целом в начале 90-х годов в области компьютерной лингвистики произошел переход к статистическим методам и, затем, методам машинного обучения и анализа данных [3], которые применяют к уже написанным и существующим текстам.

В настоящее время в области автоматической обработки текстов в России значительная часть работ посвящена переносу методов, разработанных для английского языка, на русский, и, к сожалению, оригинальных разработок очень мало.

Системы АОТ можно классифицировать по виду лицензирования (проприетарные — как правило, принадлежащие известным производителям, и академические разработки — распространяемые бесплатно); открытости (системы могут быть либо доступны только узкому кругу людей, либо находиться в открытом доступе); целевой аудитории (исследователи в области компьютерной лингвистики, разработчики, рядовые пользователи и т. п., что определяет интерфейс системы); мультиязычности (различие по числу поддерживаемых языков); характеру (готовые системы или библи-

отеки инструментов обработки текстов); универсализму (решение конкретных задач, обработка текстов в целом); используемым данным (тип и объемы обрабатываемых данных); применяемым экспертным правилам и математическим моделям; ориентации на конкретную прикладную область.

Мультиязычные системы часто более коммерчески привлекательны и просты в использовании. В свою очередь, системы, ориентированные на конкретный язык или подмножество языков, обеспечивают пусть и небольшой, но очень важный во многих задачах прирост качества за счет учета специфики языка. Классический пример мультиязычной системы — переводчик Google.

Системы, рассчитанные на достаточно широкое (и, как правило, коммерческое) использование, обладают хорошо развитым интерфейсом для конечного пользователя (например, Microsoft Bing Translator и Google translator, «ОРФО», программа для автоматического переключения между различными раскладками клавиатуры Punto Switcher, различные утилиты «Яндекса» и т. д.). Ряд этих систем обладает также своим собственным программным интерфейсом («ОРФО», Microsoft Bing Translator). Но в данном случае он является скорее приятным дополнением, чем основным способом использования систем. Напротив, для систем, рассчитанных только на исследователей или являющихся составной частью более объемных проектов, программный интерфейс становится главным (а часто и единственным) способом взаимодействия. Интерфейсы для конечных пользователей в этих системах рассчитаны скорее на работу в тестовом режиме и часто являются консольными. В качестве примеров такого рода систем можно назвать *mystem*,

Задачи компьютерной лингвистики

В сфере обработки текстов на сегодняшний день сформировалось два подхода: на основе моделей языка и правил, составленных экспертами; на базе машинного обучения. Первый позволяет достичь лучших результатов, однако составление моделей и правил настолько трудоемкий процесс, что уступающие по качеству методы машинного обучения практически его вытеснили. Повышение качества достигается не за счет совершенствования математических методов, а за счет увеличения и улучшения обучающей выборки. Оба подхода направлены сегодня на решение следующих задач.

- **Анализ и грация мнений.** Соотнесение текста, написанного от первого лица, с дискретной шкалой оценок: плохо, хорошо, очень хорошо и т. д. Используется для анализа отзывов в интернет-магазинах и высказываний в социальных сетях.
- **Анализ тональности высказываний.** Выявление позитивного или негативного отношения к обсуждаемому предмету. Используется для анализа отзывов, генерации диалога и т. д.
- **Классификация текстов по темам.** Отнесение текста к той или иной тематике. Используется во многих приложениях — в частности, в рекомендательных системах, для рубрикации текстов в онлайн-библиотеках и для организации новостных потоков.
- **Генерация речи.** Используется в робототехнике, смартфонах, навигаторах.
- **Ведение диалога.** Анализ реплик собеседника и формирование на их основе ответов. Используется в робототехнике, экспертных системах — например, Королевский банк Шотландии частично заменил контакт-центры роботами, поддерживающими диалог с пользователем.
- **Проверка правописания.** Используется в текстовых редакторах, поисковых системах.
- **Извлечение смысла из текста.** Выделение ключевых слов и словосочетаний, трендов, суммаризация. Применяется в новостных системах для агрегирования серии новостных сообщений, базах знаний для организации хранения знаний и вывода новых фактов.
- **Поиск ответов на вопросы.** Подборка по вопросу и, возможно, контексту наиболее релевантного ответа. Применяется в поисковых и экспертных системах.
- **Машинный перевод.**

АОТ, ruMorphy 1 и 2, «Томита парсер» [4], OpenXerox, Snowball. Почти все они предназначены для решения конкретных задач, возникающих на различных этапах анализа текстов: выделения слов из текста (токенизация), морфологического анализа (определения частей речи и других грамматических характеристик), построения синтаксической структуры предложений и т. д.

Корпусы — неотъемлемая часть многих систем обработки текстов. Каждое слово в корпусах снабжено исчерпывающими грамматическими характеристиками: к какой части речи оно принадлежит, в какой форме оно находится, какова его синтаксическая роль. Корпусы служат входными данными для обучения в задачах классификации текстов по темам и жанрам, для обучения синтаксических парсеров и программ, используемых для снятия омонимии и разрешения анафоры. Параллельные корпуса, состоящие из одинаковых текстов на разных языках, используют для обучения машинных переводчиков. Как правило, корпуса собираются десятилетиями, и в их создании участвуют большие исследовательские группы — например, проект «Национальный корпус русского языка» существует уже 13 лет и поддерживается компанией «Яндекс».

Важный тип входных данных любой системы АОТ — *морфологические словари*. Например, библиотека «АОТ», используемая во многих исследовательских и коммерчес-

ких проектах, представляет собой словарь Зализняка в цифровой форме. *Тезаурусы* (или семантические сети) — другой тип широко востребованных входных данных. Пожалуй, самый известный тезаурус — это WordNet, представляющий собой ресурс, в котором слова связаны с помощью так называемых семантических отношений: синонимии, гиперонимии (частное — обобщение), гипонимии (обобщение — частное), меронимии (часть — целое) и др. WordNet полезен в задачах машинного перевода, генерации текстов, классификации текстов. К сожалению, русского аналога WordNet пока нет.

Решение практически любой задачи АОТ так или иначе включает в себя проведение анализа текста на нескольких уровнях представления.

1. **Графематический анализ.** Выделение из массива данных предложений и слов (токенов).
2. **Морфологический анализ.** Выделение грамматической основы слова, определение частей речи, приведение слова к словарной форме.
3. **Синтаксический анализ.** Выявление синтаксических связей между словами в предложении, построение синтаксической структуры предложения.
4. **Семантический анализ.** Выявление семантических связей между словами и синтаксическими группами, извлечение семантических отношений.

Каждый такой анализ — самостоятельная задача, не имеющая собственного практического применения, но активно используемая для решения более общих задач. Многие исследовательские системы предназначены для решения именно вспомогательных задач. Такие системы применяются либо для апробации методов и проведения числительных экспериментов, либо в качестве составных частей (или библиотек) для систем, решающих ту или иную прикладную задачу. Примером таких систем могут служить средство NLTK для графематического анализа и токенизации, морфологический анализатор mystem и синтаксический парсер «ЭТАПЗ».

Универсализм в АОТ подразумевает наличие в системе набора взаимосвязанных методов и подходов. Существует два класса таких систем. К первому относятся системы, разрабатываемые исследовательскими департаментами крупных компаний: IBM, Intel, SAS, ABBYY, Microsoft, Xerox и т. д. В качестве примеров систем, предназначенных для обработки текстов на английском языке, можно назвать IBM Content Analytics, SAS Text Miner и IBM Watson. Ко второму классу относятся открытые интегрированные программные пакеты, созданные в университетах и представляющие собой множество методов и моделей, построенных на единой программной и математической платформе. Для английского языка можно назвать системы Apache OpenNLP, StanfordNLP, NLTK, GATE. Систем для работы с русским языком, претендующих на универсализм, пока нет, более того, в случае русского языка отсутствуют даже доступные для конечного пользователя системы, решающие основные лингвистические задачи: выделение ключевых слов, классификация текстов по темам, определение тональности текстов. В таблице перечислены программные системы, работающие с русским языком.

Некоторые системы АОТ направлены на анализ текстов определенных жанров или тематики. Например, система Watson применяется в медицине для диагностирования и облегчения процедуры принятия врачами решений. Рекомендательная система новостных сообщений News360 представляет собой приложение для мобильных устройств, с помощью которого пользователь может читать и выбирать наиболее интересную для него информацию. На основе предпочтений пользователя система предлагает новые статьи, собранные с разных новостных порталов и отвечающие конкретной тематике. В некоторых случа-

Системы АОТ для русского языка

Название	Применение	Языки	Интерфейсы	Целевая аудитория	Доступность
NLTK	Разработка систем анализа текстов	Английский + поддержка обучения классификаторов для остальных языков	Командная строка Python	Студенты, исследователи и разработчики в области NLP	Бесплатно
PyMorphy2	Морфологический словарь для исследовательских и коммерческих проектов	Русский	Командная строка Python	Студенты, исследователи и разработчики в области NLP	Бесплатно
«Томита Парсер» («Яндекс»)	Выделение именованных сущностей	Русский	API	Студенты, исследователи и разработчики в области NLP	Бесплатно
«Яндекс.Спеллер»	Проверка орфографии	Русский	Онлайн-версия и API	Широкий круг онлайн-пользователей и разработчиков сервисов и мобильных приложений	Бесплатная онлайн-версия и платный доступ к API
OntosMiner	Извлечение знаний из текстовых коллекций и их структурирование в виде онтологии	Английский, русский, немецкий, французский	Пользовательский	Студенты, исследователи и разработчики в области NLP	Бесплатная демо-версия
«ЭТАПЗ»	Синтаксический анализ и визуализация деревьев разбора	Русский	Пользовательский, API	Студенты, исследователи и разработчики в области NLP	Бесплатно
«Антиплагиат»	Проверка текстов на наличие заимствований	Русский	Специальный ресурс в Интернете	Рядовые пользователи и пользователи специальной версии для университетов	Бесплатно
«ОРФО»	Проверка орфографии	Русский	Пользовательский и API	Пользователи персональных компьютеров и разработчики онлайн-сервисов и мобильных приложений	Платно
Microsoft Word	Текстовый процессор с проверкой орфографии и синтаксиса	Почти все языки	Пользовательский и API	Пользователи персональных компьютеров и мобильных устройств, разработчики	Платно
«Наносемантика»	Интернет-роботы, помогающие при регистрации на ресурсах (ведение диалога)	Русский	Пользовательский	Посетители соответствующих ресурсов	Бесплатно
«Генон»	Поиск ответа на вопрос на популярных вики- и интернет-ресурсах	Русский	Пользовательский	Пользователи Интернета	Бесплатно
2long2read	Выделение ключевых предложений в связном тексте	Русский	Пользовательский и API	Пользователи Интернета	Бесплатно

ях эти системы умеют определять тональность новостного сообщения — например, пользователь может просматривать только хорошие новости и исключить из своей ленты все плохие. Рекомендательные системы, работающие с текстовыми данными, особенно востребованы в интернет-магазинах. С точки зрения АОТ отзыв пользователя интернет-магазина — это текст, имеющий явную тональную окраску и посвященный конкретному предмету. По отзыву пользователя необходимо определить, остался ли он доволен купленным товаром или нет, а если ему что-то не понравилось, то понять, что именно. Кроме того, перед интернет-магазинами встает задача выявления поддельных отзывов, написанных производителем товара.

Сегодня многие модели, разработанные в недрах научных сообществ, взяты на вооружение крупными игроками рынка ИТ

(Google, IBM, Microsoft), однако в секторе, ориентированном на работу с русским языком, наблюдается ошутимое отставание от английского, китайского, арабского и от европейских языков. Существующие системы решают либо совсем простые (проверка орфографии, базовая корректировка поискового запроса), либо вспомогательные (выделение основы слов, приведение слова к начальной форме), либо специальные задачи (автоматическое составление резюме, анализ компетенций, анализ профиля среднестатистического пользователя социальной сети). Сравнение с рядом славянских и восточно-европейских языков также оказывается не в пользу русского. ■

ЛИТЕРАТУРА

1. Sergei O. Kuznetsov/ Fitting Pattern Structures to Knowledge Discovery in Big Data. ICFA 2013. P. 254–266.

2. Christopher Manning, Hinrich Schuetze. Foundations of Statistical Natural Processing. MIT Press, 1999.

3. Boris Mirkin/ Core Concepts in Data Analysis: Summarization, Correlation and Visualisation, DOI 10.1007/978-0-85729-287-2. Springer, 2011.

4. Константин Селезнев, Александр Владимиров. Лингвистика и обработка текстов // Открытые системы. — 2013. — № 04. — С. 46–49. URL: <http://www.osp.ru/os/2013/04/13035562> (дата обращения: 05.02.2014).

Дмитрий Ильвовский (dilvovsky@hse.ru) — сотрудник лаборатории интеллектуальных систем и структурного анализа, Екатерина Черняк (echernyak@hse.ru) — сотрудник международной лаборатории анализа и выбора решений, НИУ-ВШЭ (Москва). Работа проведена в рамках Программы фундаментальных исследований НИУ ВШЭ.