УДК 81.32

В. В. Сушко

Методы обработки текстов в современной лингвистике

Аннотация:

При работе с текстом современная лингвистика все больше склоняется к использованию искусственного интеллекта. Возникают новые задачи, которые становится возможно решить лишь с помощью информационных технологий, благодаря чему появляется возможность работать с большим объемом текстов. Автор рассматривает различные инновационные методы обработки текстов, описывает их разнообразие и вариативность.

Ключевые слова: компьютерная лингвистика, искусственный интеллект, обработка информации, автоматическая обработка текста, семантический анализ.

Об авторе: Сушко Виктория Вадимовна, Государственный университет «Дубна», аспирант кафедры лингвистики, эл. почта: viktoriawadimovna@yandex.ru

Научный руководитель: Шимон Наталья Владимировна, Государственный университет «Дубна», кандидат филологических наук, доцент кафедры иностранных языков и русского как иностранного, эл. почта: kutepovan@mail.ru

С каждым днем наука все сильнее увеличивает темпы своего развития. Столкнувшись с появлением вычислительной техники, она поставила вопрос и об электронной обработке текстов. Преимущества такой обработки очевидны: компьютер сможет распознавать и расшифровывать тексты гораздо быстрее и в большем объеме, чем человек. Постепенно эта идея переросла в целый раздел науки – компьютерную лингвистику, изучающую применение математических моделей для выявления и описания лингвистических закономерностей текста.

Во всем мире, и в России в том числе, существует направление в науке, занимающееся обработкой текстов на естественном языке (Natural Language Processing). Автоматической обработке подвергаются уже не отдельные тексты, а целые коллекции

документов на разных языках [9, с. 16]. Существует несколько систем обработки текстов на естественном языке. Общий для всех языков способ проходит три этапа: анализ отдельных слов, анализ отдельных предложений и семантический анализ [2, с. 14-17]. Рассмотрим каждый этап более подробно.

На первом этапе проводится морфологический и морфемный анализ слова, определяются морфологические характеристики слова и его основная словоформа. Этот этап во многом идентичен для большинства естественных языков. После того, как проведен анализ каждого слова, начинается анализ предложений, а именно – определение взаимосвязи между словами. Результаты обычно записываются схематично: как рисунок, напоминающий цепь. Позднее по этим рисункам сверяют стиль написания разных авторов. Система запоминает свой, особенный рисунок, чаще всего повторяющийся у конкретного автора.

Впоследствии такая база данных помогает установить, как минимум, предположительное авторство неизвестного текста. Например, для текстов А. С. Пушкина, как для стихотворных, так и для прозаических, характерна линейная, развернутая схема, обладающая длительной протяженностью. Особую сложность представляют предложения, обладающие двойным смыслом: с точки зрения синтаксиса, слова в них могут быть связаны друг с другом двумя и более способами. Например: «*Мне необходимо забрать старые книги и билеты*». Задача ученого – научить компьютер распознавать контекст и правильно понимать такие предложения.

Семантический анализ – финальная стадия обработки текста. Его основные задачи – это поиск текстовой информации, ее извлечение из текста и представление в формате, удобном для хранения и дальнейшего использования, а также построение нового текста. Семантический анализ используется для автоматического реферирования и перевода текстов [3]. Т. В. Батура и М. В. Чаринцева выделяют графематический анализ (на уровне символов) и фрагментационный (на уровне фраз, частей текста). Графематический анализ отвечает за сегментацию текста в языках, подобных китайскому и японскому, где текст не имеет пробелов. Фрагментационный анализ отвечает за деление предложения на неразрывные фрагменты, которые называются синтаксическими единствами, и установление иерархии на множестве этих единств. Синтаксические единства обычно равны или превосходят по объему словосочетания, являющиеся синтаксическими группами [2, с. 4-5].

Существует проект «Автоматической обработки текста». Исследователи выделяют в нем три этапа: определение типа фрагмента текста, снятие омонимии и применение устанавливающих иерархию правил. Типы фрагмента существуют следующие: глагол в личной форме, краткое причастие, краткое прилагательное, предикативное слово, причастие, деепричастие, инфинитив, вводное слово или пустое значение (если ни один из типов не определяется).

Если нельзя однозначно определить тип фрагмента, существуют два правила для прояснения ситуации. Первое правило гласит, что, если омоним слова – краткое причастие или прилагательное, и в тексте нет ни одного слова, которое согласовалось бы с ним грамматически, омоним уничтожается. Второе правило заключается в том, что, если во фрагменте есть неомонимичный предикат (глагол в личной форме, краткое прилагательное, краткое причастие, предикативное слово, причастие или деепричастие), омонимы этих частей речи уничтожаются во всех остальных словах данного фрагмента. Третий этап – правила, устанавливающие иерархию – отвечает за объединение фрагментов [2, с. 14-17].

Существуют и другие способы обработки текстов. Во многом способ зависит от поставленной задачи. Для разработки классификации текстов и отнесения текста в конкретный класс используется так называемый метод регулярных выражений. Компьютер ищет в тексте клише, устойчивые слова и обороты, позволяющие определить, что за текст перед нами. С этой задачей лучше всего справляются нейросети, поскольку они охватывают наибольшее количество классов [4].

Отдельный метод анализа текстов направлен на выявление именованных сущностей (Named Entity Recognition). К ним относятся различные имена собственные, денежные суммы, бренды и т.д. Таким методом чаще всего пользуются компании, специализирующиеся на массовых услугах. Создаются базы именованных сущностей, отдельные списки синонимичных слов, включая все парадигматические изменения той или иной лексической единицы [7].

Важной процедурой выступает распознавание текста, необходимое для его перевода с бумажного носителя в электронный вид без набора. Задача искусственного интеллекта – расшифровать текст на бумаге и в точности запомнить его для дальнейшего воспроизведения. Сегодня эта задача полностью решена, однако в дальнейшем планируется научить искусственный интеллект редактировать текст: исправлять ошибки,

выявлять и запоминать нужные данные, передавать их в другие системы – иначе говоря, переходить от классификации текстов к классификации информации.

Решение всех перечисленных выше задач невозможно без главной – научить компьютер понимать текст. Для большинства естественных языков это остается наиболее сложным процессом ввиду их особенностей. Например, во многих языках существуют такие лингвистические явления, как эллипсис, омонимия и т.д.

Эллипсис – это пропуск в устной или письменной речи элемента предложения, который подразумевается в контексте. Например: «*Он не умеет плавать, а я умею».* В этом предложении слово «плавать» второй раз опускается во избежание повторения, однако слушателю или читателю понятно, о чем идет речь, из первой части предложения. Омонимия – это совпадение двух слов в написании и звучании при различии их значений. Например: «*ключ (родник) – ключ (инструмент, открывающий замок)»*. Человеческий мозг быстро улавливает эти особенности и в контексте распознает их значение. Чтобы научить этому машину, необходима разработка новых моделей анализа текста. Такие модели подразумевают переработку данных в знания.

Следующий уровень понимания – анализ тональности текста. Эта область компьютерной лингвистики занимается изучением эмоций и мнений, определением эмоциональной окраски текстов, а ее результаты активно применяются в социологии, медицине, маркетинге, политологии. Здесь определяются три признака: автор, тема и тональная оценка. К определению тональности существует три подхода: основанные на правилах, словарях, использующие машинное обучение с учителем и без учителя.

Наиболее точным является подход, основанный на определенном наборе правил, однако создание необходимой базы данных требует существенных временных затрат. В тональных словарях представлены списки слов с обозначением в цифрах тональности для каждой лексемы. В подходах, основанных на таких словарях, тональность вычисляется математически как среднее арифметическое всех значений – это наиболее простой в применении подход. Вместе с тем создание базы данных остается сложной процедурой, а ее терминология сильно зависит от контекста [2, с. 27-35].

Для обработки текстов создаются специальные анализаторы. Один из таких – Link Grammar Parser, разработанный еще в 1990-х гг. Его словари насчитывают около 60 000 форм. Этот анализатор способен пропускать недоступные ему фрагменты текста и анализировать оставшиеся. Анализ проходит в два этапа: построение множества

синтаксических представлений одного предложения и постобработка получившихся структур [10].

Однако исследователи сталкиваются с рядом проблем, которые пока не могут разрешить. Одна из них – морфологическая омонимия. Это совпадение одной или нескольких грамматических форм слов разных частей речи при различии их значений. К примеру, трудно распознать смысл следующего предложения: *«За песчаной косой лопоухий косой пал под острой косой косой девки с косой».* Даже носителю языка понадобится время, чтобы вспомнить и распределить по тексту пять значений одной словоформы.

Синтаксическая омонимия, о которой мы упоминали, также остается проблемой для искусственного интеллекта. Пример: «*мать любит дочь*». Трудность представляет и лексическая омонимия и полисемия. Лексическая омонимия – это совпадение по звучанию и написанию слов во всех их грамматических формах при различии их значений. Например: *лук (овощ) – лук (оружие)*. Полисемия – это многозначность одного слова. Отличие ее от омонимии заключается в том, что значения многозначного слова близки по своей семантике. Например: *тихий голос – тихий нрав.*

Таким образом, развитие компьютерной лингвистики не останавливается на достигнутых результатах. Однако обработка, реферирование и перевод текстов вручную продолжают существовать, поскольку искусственный интеллект все еще находится на этапе обучения.

Библиографический список:

1. Автоматическая обработка текста [Электронный ресурс]. Режим доступа: http://www.aot.ru (дата обращения: 19.01.2023).

2. Батура Т. В. Основы обработки текстовой информации. Учебное пособие / Т. В. Батура, М. В. Чаринцева. Новосибирск: Институт систем информатики им. А. П. Ершова СО РАН, 2016. 45 с.

3. Методы и системы семантического анализа текстов [Электронный ресурс] // Software Journal: Theory and Applications. Режим доступа: http://swsys-web.ru/methods-and-systems-of-semantic-text-analysis.html (дата обращения: 19.01.2023).

4. Обработка документов и текстов на естественном языке [Электронный ресурс] // Tadviser. Режим доступа: https://www.tadviser.ru/index.php (дата обращения: 19.01.2023).

5. Симанков В. С. Подходы к автоматизации процедур получения и обработки экспертных знаний на основе моделей интеллектуального анализа данных / В. С. Симанков, Е. С. Тарасов // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2012. № 84. С. 383-394.

6. Чебанов А. С. Разработка подходов к оптимизации сложных организационно-технических систем на основе адаптивных моделей / А. С. Чебанов, А. В. Власенко, Е. С. Тарасов // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2015. № 112. С. 850-861.

7. Bing L. Sentiment Analysis and Opinion Mining. California: Morgan and Claypool Publishers, 2012. 168 p.

8. Pang B. Opinion Mining and Sentiment Analysis / B. Pang, L. Lee // Foundations and Trends in Information Retrieval. 2008. Vol. 2. Pp. 1–135.

9. Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition / D. Jurafsky, J. Martin. New Jersey: Prentice-Hall, 2008. 1024 p.

10. Link Grammar Parser [Electronic resource] // AbiWord Available at: http://abisource.com/projects/link-grammar/ (access date: 19.01.2023).

Sushko V.V. **Text processing methods in modern linguistics**

There are various text processing methods in modern linguistics. Artificial intelligence is getting more and more popular in science. There are new challenges, which only PC is able to decide, so we can work with higher amount of texts. This article deals with various innovative methods of text processing. The author was to dedicate the reader to the variety and variability of methods. The article may be useful for beginners in the field of computational linguistics.

Keywords: computer linguistics, artificial intelligence, Natural Language Processing, automatic text processing, semantic analysis.