УДК 81'33

EDN HWTSRQ

Компьютерная лингвистика

М.И. Кондратьева[1*], В.В. Бронская[1], К.С. Бронская[2]

[1]Казанский национальный исследовательский технологический университет, ул. Карла Маркса, 68, Казань, 420015, Россия
[2]Казанский (Приволжский) федеральный университет, ул. Кремлевская, 19, Казань, 420008, Россия

*E-mail: kondrattteva@yandex.ru

Аннотация. Область компьютерной лингвистики и обработки естественного языка за последние десятилетия претерпела значительные изменения, превратившись в важную научную дисциплину и область технологического развития. Увеличение объема текстовой информации в Интернете и растущая потребность в её автоматизированной обработке стимулировали создание инновационных решений в лингвистике и IT. В статье рассматриваются ключевые задачи и приложения компьютерной лингвистики, такие как машинный перевод, автоматическая редактура текста, извлечение информации и разработка чат-ботов. Также обсуждаются актуальные проблемы, включая сложности распознавания речи, вызванные различиями в акцентах, и проблемы синтаксической и морфологической неоднозначности. Влияние полярных и неполярных растворителей на результаты анионной полимеризации служит дополнительным фактором, влияющим на точность автоматических систем обработки. Примечательно, что такие сложности, как определение точного значения неоднозначных слов, остаются открытыми вопросами, влияющими на работу поисковых систем и другие приложения. Хотя достижения в области машинного перевода и автоматизации текстов значительны, они остаются несовершенными и требуют дальнейшего развития. Таким образом, сфера компьютерной лингвистики продолжает сталкиваться с комплексными вызовами, решение которых важно для улучшения эффективности взаимодействия с текстовой информацией и её анализа в различных приложениях.

Ключевые слова: компьютерная лингвистика, обработка естественного языка, машинный перевод, распознавание речи.

Computer linguistics

M.I. Kondrateva[1*], V.V. Bronskaya[1], K.S. Bronskaya[2]

[1] Kazan National Research Technological University, 68 Karl Marx str., Kazan, 420015, Russia
[2] Kazan Federal University, 18 Kremlyovskaya str., Kazan, 420008, Russia

* E-mail: kondrattteva@yandex.ru

Abstract. The field of computational linguistics and natural language processing has undergone significant changes over the past decades, transforming into an important scientific discipline and a key area of technological development. The increasing volume of textual information on the Internet and the growing need for its automated processing have driven the creation of innovative solutions in linguistics and IT. This article addresses the key tasks and applications of computational linguistics, such as machine translation, automatic text editing, information extraction, and chatbot development. Current challenges, including the difficulties of speech recognition due to accent variations, as well as syntactic and morphological ambiguity, are also discussed. The impact of polar and non-polar solvents on the outcomes of anionic polymerisation serves as an additional factor influencing the accuracy of automated processing systems. Notably, issues such as the precise identification of ambiguous word meanings remain unresolved and affect the performance of search engines and other applications. Although advancements in machine translation and text automation are significant, they remain imperfect and require further development. Thus, the field of computational linguistics continues to face complex challenges, the resolution of which is crucial for enhancing the efficiency of textual information processing and analysis across various applications.

Keywords: computational linguistics, natural language processing, machine translation, speech recognition.

1. Введение

За последние три десятилетия область компьютерной лингвистики (КЛ) вместе с ее инженерной областью обработки естественного языка (ЕЯ) превратилась из относительно малоизвестного дополнения к искусственному интеллекту (ИИ) [1-4] в процветающую научную дисциплину и стала важной областью промышленного развития. Это связано с большим потоком текстовой информации в Интернете, а также необходимостью её обработки, что и подтолкнуло ученых-разработчиков на создание новейших технологий в сфере лингвистики и IT [5-7]. Целью нашего исследования является изучение области компьютерной лингвистики, описание её специфики и изложение её основных задач, а также составление краткой характеристики существующих приложений КЛ.

Однако существует множество актуальных проблем, которые еще предстоит решить компьютерным лингвистам, среди них:

- Трудности в системах распознавания речи в основном вызваны различиями в акцентах; использование спонтанной речи; различия в артикуляции, громкости, скорости и т.д.; акустические условия и многие другие.
- Уровень понимания; Морфологические и синтаксические явления, такие как многоточие и разрешение анафоры, представляют собой проблемы на уровне понимания ЕЯ и являются активными областями исследований.
- Выбор одного единственного значения для неоднозначного слова, является одним из самых популярных открытых вопросов в лингвистике, поскольку он оказывает огромное влияние на точность поисковых систем.
- Устранение неоднозначности контекста, социальный интеллект, интерпретация спонтанных жестов и т. д. – вот некоторые из текущих проблем (в разной степени) в области КЛ.

2. Инструменты компьютерной лингвистики

Каждое разработанное приложение для КЛ должно включать в себя все необходимые знания(лингвистические, синтаксические, морфемные и т.д.) об исследуемом языке. Всю эту информацию можно найти в интернет(компьютерных) – словарях, за основу которых берутся текстовые словари разных языков, исходным

материалом которых является текстовый корпус – это большой и структурированный набор текста, собранный систематически.

Текстовые корпуса используются в КЛ для статистического анализа, проверки гипотез, поиска моделей использования языка, исследования языковых изменений и вариаций, а также обучения владению языком. На сегодняшний день чаще всего исходным материалом компьютерной лингвистики становятся тексты из Интернета.

Для решения практических задач КЛ используются такие инструменты как:

Машинный перевод – это раздел компьютерной лингвистики, который исследует использование программного обеспечения для перевода текста или речи с одного языка на другой. Некоторые системы машинного перевода могут переводить тексты широких категорий, такие как новости и технические документы, и производить переводы приемлемого качества. Однако в технологии МП все еще есть пробелы – например, она часто не распознает идиомы и каламбуры, а также может быть достаточно не точной, когда речь идет о переводе узкоспециализированных текстов, таких как юридические или технические документы, или сложный лингвистический текст. Наконец, многие пользователи не уверены в том, что окончательный машинный перевод является правильным или даже имеет смысл.

Поиск и кластеризация документов часто служат предварительными шагами в извлечении информации или анализе текста, двух пересекающихся областях, связанных с извлечением полезных знаний из документов, таких как основные характеристики именованных объектов (категория, роли по отношению к другим объектам, местоположение, даты и т. д.) или отдельных типов событий, или вывод правильных корреляций между относительными терминами (например, что покупка одного типа продукта коррелирует с покупкой другого).

Ответы на вопросы (Question answering) – это задача в области поиска информации и обработки естественного языка, которая связана с созданием систем, которые автоматически отвечают на вопросы, заданные людьми на ЕЯ. Эта задача решается путем определения типа вопроса, поиском текстов, потенциально содержащих ответ на этот вопрос, и извлечением ответа из этих текстов.

Редактирование текста и исправление предложений являются важными понятиями лингвистической структуры и анализа. Программами, по решению данной задачи, являются автокорректоры и программы по автоматическому переносу слов, что

вызывает свою сложность в том плане, что для правильного решения приложению необходимо знать морфемную структуру слов ЕЯ, то есть знание целого словаря.

Извлечение знаний или составление краткого изложения текста(конспекта) из неструктурированного текста становятся все более важными приложениями, учитывая поток документов, исходящих от средств массовой информации, организаций любого типа и отдельных лиц. Этот непрекращающийся поток информации затрудняет обзор элементов, имеющих отношение к какой-либо конкретной цели, таких как основные данные о лицах, организациях и потребительских товарах, или сведения о несчастных случаях, землетрясениях, преступлениях, обслуживании продуктов, результаты медицинских исследований и так далее.

Извлечение информации создает структурированное представление знаний из неструктурированного текста, чтобы полученные сведения можно было в дальнейшем использовать для поиска, вывода и анализа. Учитывая спецификацию выбранных типов сущностей, семантических отношений и событий, приложение строит базу данных из экземпляров этой информации в тексте.

Еще одно приложение, которое стоит упомянуть – **чат-бот** – компьютерная программа или виртуальный помощник, используемый для ведения онлайн-чата с помощью текста или преобразования текста в речь вместо обеспечения прямого контакта с живым человеком. Разработанные для убедительной имитации того, как человек ведет себя в качестве собеседника, системы чат-ботов обычно требуют постоянной настройки и тестирования, и многие из них по-прежнему не могут адекватно вести диалог с человеком.

3. Выводы

В то время как другие отрасли работают на основе очень точных принципов, КЛ стремится воссоздать или смоделировать естественные языки общения. Направления, в которых движется компьютерная лингвистика, довольно ошеломительны, поскольку эти решения связаны с инструментами распознавания голоса и дальнейшего перевода, и редактирования грамматики и правописания. Несмотря на то, что каждая задача представляет собой настоящую проблему, некоторые из недавних инноваций в этой области, начинают комбинировать варианты их решений, чтобы разнообразить рентабельность приложения и лучше исполнять запросы потребителей.

Список литературы

1. Ковалев И.В. Системные аспекты организации и применения мультилингвистической адаптивно-обучающей технологии / И.В. Ковалев, М.В. Карасева, Е.А. Суздалева // Образовательные технологии и общество. – 2002. – Т. 5, № 2. – С. 198-212. – EDN GJLTGN.
2. Бронская В.В. Применение искусственной нейронной сети для создания картины из фотографии / В.В. Бронская, В.В. Плющев, Т.В. Игнашина, К.С. Бронская, М.И. Кондратьева, А.В. Шипин // Научно-технический вестник Поволжья. – 2024. – № 7. – С. 123-126.
3. Кондратьева М.И. Расчет сопротивления насадочной колонны с помощью искусственной нейронной сети / М.И. Кондратьева, В.В. Бронская // В сборнике: Развитие современной науки и технологий в условиях трансформационных процессов. Сборник материалов VIII Международной научно-практической конференции. – Санкт-Петербург, 2023. – С. 268-271.
4. Brester C. Evolutionary feature selection for emotion recognition in multilingual speech analysis / C. Brester, E. Semenkin, I. Kovalev et al. // IEEE Congress on Evolutionary Computation (CEC 2015), Sendai, Japan, 25–28 мая 2015 года. – Sendai, Japan: Institute of Electrical and Electronics Engineers Curran Associates, Inc. (Nov 2015). – 2015. – P. 2406-2411. – DOI 10.1109/CEC.2015.7257183. – EDN WSYCXP.
5. Ковалев И.В. Внутриязыковые ассоциативные поля в мультилингвистической адаптивно-обучающей технологии / И.В. Ковалев, О.В. Лесков, М.В. Карасева // Системы управления и информационные технологии. – 2008. – № 3-1(33). – С. 157-160. – EDN JURFJT.
6. Зеленков П.В. Мультилингвистическая модель распределенной системы на основе тезауруса / П.В. Зеленков, И.В. Ковалев, М.В. Карасева, С.В. Рогов // Вестник Сибирского государственного аэрокосмического университета им. академика М.Ф. Решетнева. – 2008. – № 1(18). – С. 26-28. – EDN IPVDMZ.
7. Польщикова О.Н. Становление и формирование терминологии компьютерной лингвистики / О.Н. Польщикова // Вопросы журналистики, педагогики, языкознания. – 2022. – Т. 41. – № 3. – С. 590-607.