Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ_____

КАФЕДРА _____СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ_____

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:
ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ТЕХНОЛОГИИ ЯЗЫКОВОГО АНАЛИЗА

Студент __ИУ5И-32М_____ _____ _____**Лу Жуньда**_____
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель курсовой работы _____ ___**Ю.Е. Гапанюк**____
(Подпись, дата) (И.О.Фамилия)

2024 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой ___ИУ5____
(Индекс)
_____ ___В.И. Терехов__
(И.О.Фамилия)
«_25_» ____декабря_____ 2024 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме ___ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ТЕХНОЛОГИИ ЯЗЫКОВОГО АНАЛИЗА_____
_____

Студент группы _ИУ5И-32М__

_____**Лу Жуньда**_____
(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)
_____ИССЛЕДОВАТЕЛЬСКАЯ_____
Источник тематики (кафедра, предприятие, НИР) _____КАФЕДРА_____

График выполнения НИР: 25% к __._ нед., 50% к _._ нед., 75% к _.__ нед., 100% к __._ нед.

Техническое задание _рассматривает применение, методы и вызовы искусственного интеллекта в_ обработке естественного языка (NLP), исследуя, как технологии глубокого обучения и другие подходы помогают компьютерам понимать и генерировать человеческий язык, а также решать проблемы разреженности данных и семантического понимания для продвижения общественного прогресса._____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на _22_ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
_____
_____
_____
_____

Дата выдачи задания «_25_» ____декабря_____ 2024 г.

Руководитель НИР _____ _____Ю.Е. Гапанюк_____
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____Лу Жуньда_____
(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

Содержание

Введение

Обработка естественного языка (NLP) — это технология, позволяющая компьютерам понимать и генерировать человеческий язык. Она всегда была одной из ключевых тем исследований в области искусственного интеллекта. В настоящее время технологии искусственного интеллекта широко применяются в сферах, таких как интеллектуальная служба поддержки, интеллектуальные вопросы и ответы, автоматическое резюмирование, интеллектуальные рекомендации и т.д. В этих приложениях искусственный интеллект играет все более важную роль. Например, в интеллектуальных службах поддержки искусственный интеллект может автоматически генерировать текст ответов на основе понимания и анализа вопросов пользователей, помогая им решать проблемы. В интеллектуальных рекомендациях искусственный интеллект может предлагать пользователям соответствующие продукты или услуги на основе их интересов и поведенческих данных. Эти приложения не только улучшают пользовательский опыт, но и способствуют развитию технологий искусственного интеллекта.

1. Определение обработки естественного языка

Обработка естественного языка (NLP) — это обработка естественного языка человеком компьютерами. Она включает в себя анализ, понимание и генерацию текстов, чтобы обеспечить эффективное взаимодействие между людьми и машинами. Исследования в области NLP охватывают множество дисциплин, включая лингвистику, информатику и искусственный интеллект.

В исследованиях NLP процесс обработки языка обычно делится на три уровня: синтаксическая обработка, семантическая обработка и прагматическая обработка. Синтаксическая обработка фокусируется на формальной структуре языка, например, на грамматическом составе предложений и порядке слов; семантическая обработка включает в себя понимание смысла языка, такого как значения слов и фраз; прагматическая обработка касается контекста и среды использования языка, например, контекста диалога и вежливости языка.

Применение NLP очень обширно и включает, но не ограничивается, следующими аспектами: машинный перевод, автоматическое резюмирование, классификация текстов, системы вопросов и ответов, чат-боты, анализ настроений и анализ общественного мнения. Эти приложения в той или иной степени

решают проблемы коммуникации между людьми и машинами, улучшая понимание и выражение естественного языка.

2. Связь между искусственным интеллектом и обработкой естественного языка

Основная цель искусственного интеллекта — это создание машин, способных мыслить и действовать как люди, реализуя некоторые функции человеческого интеллекта. В области обработки естественного языка искусственный интеллект играет решающую роль. С одной стороны, искусственный интеллект предоставляет мощную техническую поддержку для NLP. Например, алгоритмы машинного обучения могут быть использованы для классификации текстов и анализа настроений, алгоритмы глубокого обучения — для вычисления векторных представлений слов и семантического понимания, а алгоритмы обучения с подкреплением — для обучения диалоговых систем. С другой стороны, NLP также предоставляет широкие перспективы для развития искусственного интеллекта. Например, NLP может применяться в интеллектуальных службах поддержки, интеллектуальных рекомендациях, умных домах и других областях, повышая уровень интеллекта и качество обслуживания систем искусственного интеллекта.

Конкретные применения искусственного интеллекта в NLP можно разделить на следующие аспекты:

(1) Вычисление векторных представлений слов: векторные представления слов — это способ представления слов в виде

вещественных векторов, который может использоваться для вычисления схожести текстов и семантического понимания. Алгоритмы глубокого обучения в искусственном интеллекте могут автоматически обучаться методам вычисления векторных представлений слов, повышая точность и эффективность обработки текстов.

(2) Языковые модели: языковые модели — это вероятностные модели, используемые для предсказания последовательности текстов, которые могут применяться в машинном переводе, генерации текстов и диалоговых системах. Модели, такие как рекуррентные нейронные сети (RNN) и трансформеры, могут использоваться для построения языковых моделей, улучшая их производительность и точность.

(3) Семантическое понимание: семантическое понимание относится к анализу и пониманию значений текстов. Технологии обработки естественного языка могут применяться в классификации текстов, анализе настроений, системах вопросов и ответов и т.д., повышая уровень интеллектуального взаимодействия между людьми и машинами.

(4) Диалоговые системы: диалоговые системы — это технологии для реализации взаимодействия между людьми и машинами, которые требуют знания в нескольких областях, включая обработку естественного языка, распознавание речи и компьютерное зрение.

Алгоритмы глубокого обучения могут использоваться для построения диалоговых систем, повышая плавность и интеллект общения.

Искусственный интеллект и обработка естественного языка находятся в состоянии взаимного стимулирования и совместного развития. Искусственный интеллект предоставляет мощную техническую поддержку и перспективы для обработки естественного языка, тогда как обработка естественного языка открывает новые пространства для развития искусственного интеллекта. С постоянным прогрессом технологий и ростом потребностей в приложениях взаимодействие между искусственным интеллектом и обработкой естественного языка будет становиться все более тесным, принося людям больше удобства и ценности в их жизни и работе.

3. Применение технологий обработки естественного языка на основе искусственного интеллекта

3.1 Применение моделей глубокого обучения

Глубокое обучение является важной ветвью искусственного интеллекта, и его применение в области обработки естественного языка также принесло значительные результаты. Модели глубокого обучения, такие как рекуррентные нейронные сети (RNN), сети длительной краткосрочной памяти (LSTM) и сверточные нейронные сети (CNN), широко применяются.

(1) Классификация текстов

Классификация текстов направлена на распределение текстов по различным категориям. Модели глубокого обучения могут автоматически извлекать ключевую информацию из текста, обучаясь его семантическим характеристикам, что позволяет осуществлять классификацию текстов. Например, в классификации новостей модели глубокого обучения могут по семантическим признакам заголовков и содержания новостей распределять их по категориям: политика, экономика, культура и т.д.

(2) Анализ настроений

Анализ настроений направлен на определение эмоциональной окраски текста. Модели глубокого обучения могут анализировать семантические характеристики текста и его эмоциональную

направленность для выполнения анализа настроений. Например, в анализе настроений в социальных медиа модели глубокого обучения могут определять эмоциональное отношение пользователей к продукту, анализируя семантику пользовательских комментариев.

(3) Машинный перевод

Машинный перевод направлен на преобразование одного языка в другой. Модели глубокого обучения могут автоматически генерировать переводы на целевой язык, обучаясь семантическим характеристикам исходного и целевого языков.

3.2 Применение генерации естественного языка

Генерация естественного языка является важным приложением искусственного интеллекта, целью которого является создание текстов на естественном языке на основе заданной семантической информации [4]. Генерация естественного языка находит широкое применение в интеллектуальных службах поддержки, интеллектуальных вопросах и ответах, автоматических аннотациях и т.д.

(1) Интеллектуальная служба поддержки

Интеллектуальная служба поддержки нацелена на автоматическую генерацию ответов на вопросы пользователей. Технология генерации естественного языка может автоматически создавать текст ответов на основе вопросов пользователей и

контекстной информации, обеспечивая быструю реакцию на запросы пользователей. Например, в интеллектуальной службе поддержки на торговой платформе пользователи могут задавать вопросы, и служба может автоматически генерировать ответы на основе контекста и информации о товарах, помогая пользователям решать их проблемы.

(2) Интеллектуальные вопросы и ответы

Интеллектуальные вопросы и ответы направлены на автоматическую генерацию ответов на вопросы пользователей. Технология генерации естественного языка может создавать текст ответов на основе вопросов пользователей и информации из базы знаний, обеспечивая быстрые ответы на запросы. Например, в интеллектуальных вопросах и ответах поисковых систем пользователи могут задавать вопросы, и система может автоматически генерировать ответы на основе информации из базы знаний и контекста.

(3) Автоматическая аннотация

Автоматическая аннотация подразумевает извлечение ключевой информации из длинных текстов и создание кратких аннотаций. Технология генерации естественного языка может автоматически создавать аннотации, основываясь на семантической и структурной информации длинного текста, помогая пользователям быстро понять основное содержание. Например, в приложениях для аннотации

новостей технология может автоматически создавать краткие аннотации на основе ключевых данных заголовков и содержания новостей, что помогает пользователям быстро ознакомиться с основными моментами.

3.3 Применение понимания естественного языка в интеллектуальных рекомендациях, анализе настроений, системах вопросов и ответов

Технологии понимания естественного языка широко применяются в интеллектуальных рекомендациях, анализе настроений, системах вопросов и ответов и т.д.

(1) Интеллектуальные рекомендации

Интеллектуальные рекомендации направлены на предложение пользователям релевантных продуктов или услуг на основе их интересов и поведенческих данных. Технологии понимания естественного языка могут анализировать историю поиска, просмотра и покупок пользователей, чтобы понять их интересы и потребности, и соответственно предлагать им соответствующие продукты или услуги. Например, на торговых платформах технологии понимания естественного языка могут использовать информацию о истории просмотров и покупок для создания персонализированных рекомендаций.

(2) Анализ настроений

Анализ настроений подразумевает анализ и оценку эмоциональной направленности текста. Технологии понимания естественного языка могут осуществлять такую оценку с помощью семантического и эмоционального анализа текста. Например, в анализе настроений в социальных медиа такие технологии могут определять эмоциональное отношение пользователей к продукту, анализируя семантику комментариев.

(3) Системы вопросов и ответов

Системы вопросов и ответов предназначены для автоматического предоставления ответов на вопросы пользователей. Технологии понимания естественного языка могут анализировать и интерпретировать вопросы пользователей, чтобы генерировать релевантные ответы. Например, в системах вопросов и ответов поисковых систем такие технологии могут извлекать информацию из базы знаний и формировать ответы на основе анализируемых вопросов.

3.4 Вычисление семантической схожести слов

3.4.1 Метод вычисления семантической схожести слов на основе гипотезы распределения естественного языка

Метод вычисления семантической схожести слов на основе гипотезы распределения естественного языка — это статистический

подход для оценки семантической схожести между двумя словами. Основная идея заключается в том, что если два слова часто встречаются вместе в тексте, то они могут быть семантически связаны. Этот метод обладает простотой, эффективностью и возможностью масштабирования; для его применения не требуется большого количества размеченных данных — достаточно использовать информацию о совместном вхождении слов. Поэтому он нашел широкое применение в области обработки естественного языка, включая классификацию текстов, кластеризацию, вопросы и ответы, рекомендации и другие задачи.

Однако метод также имеет свои ограничения, например, он может игнорировать семантические связи и контекстуальную информацию между словами, что может привести к неточным результатам. Кроме того, для редких или неучтенных слов могут не быть получены удовлетворительные результаты. Поэтому на практике необходимо сочетать этот метод с другими подходами для его оптимизации и улучшения.

3.4.2 Применение семантической схожести в извлечении имен собственных, переформулировании запросов и других областях

В извлечении имен собственных семантическая схожесть может помочь в идентификации имен собственных в тексте и определении их семантических категорий. Например, в новостных отчетах может

встречаться множество имен компаний, людей, географических названий и т.д. Вычисляя семантическую схожесть этих имен собственных с другими словами, можно определить, к каким типам сущностей они относятся, что облегчит последующий анализ данных и извлечение знаний.

В переформулировании запросов семантическая схожесть может преобразовывать естественно-языковые запросы пользователей в другую форму для повышения эффективности поиска. Например, когда пользователь задает вопрос "Что такое искусственный интеллект?", используя семантическую схожесть, можно найти ключевые слова, связанные с "искусственным интеллектом", такие как "машинное обучение", "глубокое обучение", и переформулировать запрос в "Связь между машинным обучением и искусственным интеллектом" или "Применение глубокого обучения в искусственном интеллекте", что позволяет более точно находить связанные документы и повышать эффективность поиска.

Важно отметить, что вычисление семантической схожести на практике часто требует сочетания с другими технологиями, такими как методы на основе правил и статистические методы. Также, учитывая сложность естественного языка, вычисление семантической схожести представляет собой определенные трудности и вызовы. Поэтому в реальных приложениях необходимо постоянно

оптимизировать и улучшать алгоритмы для повышения точности и надежности вычислительных результатов.

3.5 Применение других технологий искусственного интеллекта в обработке естественного языка

3.5.1 Применение обучения с подкреплением в обработке естественного языка

Обучение с подкреплением — это метод обучения оптимальной стратегии через пробу и ошибку. В области обработки естественного языка обучение с подкреплением может быть использовано для обучения систем диалогов, систем машинного перевода и других задач. Устанавливая функцию вознаграждения, можно направлять процесс обучения моделей, достигая более интеллектуальных задач обработки естественного языка.

3.5.2 Применение генеративных состязательных сетей в обработке естественного языка

Генеративные состязательные сети (GAN) все более активно применяются в области обработки естественного языка. GAN, создавая противоречивые отношения между генератором и дискриминатором, могут генерировать высококачественные тексты на естественном языке или описания изображений.

В обработке естественного языка GAN могут быть использованы для генерации текста, аннотаций, машинного перевода и других задач.

Например, с помощью GAN можно генерировать тексты с определенной темой или стилем, а также переводить текст с одного языка на другой. Обучая генератор на создании текста и дискриминатор на оценке качества сгенерированного текста, можно постепенно повысить способности генератора, что приведет к более интеллектуальным задачам обработки естественного языка.

Кроме того, GAN также могут быть применены в задачах описания изображений. Обучая генератор на создании описаний изображений и используя дискриминатор для оценки качества сгенерированных описаний, можно добиться более точных и живых текстов, связанных с изображениями.

4. Проблемы применения искусственного интеллекта в обработке естественного языка

4.1 Проблемы разреженности и несбалансированности данных

В обработке естественного языка разреженность и несбалансированность данных представляют собой две важные проблемы. Разреженность данных означает, что из большого объема текстовых данных только небольшая часть действительно полезной информации затрудняет эффективное обучение модели важным признакам. Для решения проблемы разреженности данных можно использовать различные технологии, такие как отбор признаков и увеличение данных.

Несбалансированность данных относится к ситуации, когда в обучающем наборе данных количество образцов некоторых классов значительно превышает количество образцов других классов, что приводит к смещению модели при обработке этих классов. Для решения проблемы несбалансированности данных можно применять техники, такие как увеличенное или уменьшенное выборки. Кроме того, можно использовать методы ансамблевого обучения, такие как Bagging и Boosting, для повышения устойчивости модели.

4.2 Сложность семантического понимания и определения намерений

Семантическое понимание и определение намерений — это две важные задачи в обработке естественного языка. Однако из-за разнообразия и сложности естественного языка обе эти задачи сталкиваются с большими трудностями.

Семантическое понимание означает необходимость для компьютера понять истинный смысл текста, что требует от модели глубокого понимания языка и обширных знаний о контексте. В настоящее время модели на основе глубокого обучения добились значительного прогресса в области семантического понимания, но все еще существуют проблемы, такие как недостаточное понимание скрытых значений в тексте и зависимость от контекста.

Определение намерений означает, что компьютеру необходимо определить намерение или цель текста. Это требует от модели способности понимать и рассуждать о контекстной информации текста. Хотя модели на основе глубокого обучения достигли определенных успехов в определении намерений, все еще существуют проблемы, такие как недостаточное понимание скрытых намерений в тексте и неадекватная обработка двусмысленностей.

Для решения сложности семантического понимания и определения намерений можно использовать различные методы,

такие как правила, статистические методы и методы глубокого обучения. Кроме того, можно сочетать многомодальные данные, многозадачное обучение и другие методы для улучшения производительности модели.

4.3 Проблемы межъязыковой и междоменной обработки

Проблемы межъязыковой и междоменной обработки в области обработки естественного языка также представляют собой два важных вызова. Межъязыковая проблема возникает при обработке естественного языка на различных языках. Различия в грамматике и лексике разных языков делают межъязыковую обработку более сложной. Для решения межъязыковой проблемы можно использовать технологии, такие как машинный перевод и межъязыковой информационный поиск. Также можно использовать многоязычные корпуса и модели для повышения производительности модели. Проблема домена возникает при обработке естественного языка в различных областях. Из-за того, что текстовые данные в разных областях имеют различные характеристики и контексты, становится важной адаптация к домену. Для решения этой проблемы можно применять технологии, такие как адаптивное обучение к домену и трансферное обучение. Также можно использовать доменно-специфические корпуса и модели для повышения производительности модели.

Заключение

Таким образом, в области обработки естественного языка применение искусственного интеллекта постепенно углубляется, предоставляя человечеству беспрецедентные удобства. От распознавания речи до машинного перевода, от анализа настроений до интеллектуальных вопросов и ответов — технологии искусственного интеллекта постоянно расширяют свои сферы применения, внося новую жизнь в развитие человеческого общества. В будущем, с постоянным развитием технологий искусственного интеллекта, обработка естественного языка станет более интеллектуальной, персонализированной и гуманной. Мы ожидаем, что искусственный интеллект сможет лучше понимать человеческий язык, более эффективно взаимодействовать с людьми и лучше служить обществу.