МЕТОДЫ РАСПОЗНАВАНИЯ СМЫСЛОВОЙ НАГРУЗКИ ТЕКСТОВ С ЭЛЕМЕНТАМИ НЕОДНОЗНАЧНОСТИ

Голубев Р.О., Мамонова Т.Е.
ТПУ, группа: А4-36, 1 курс, rog3@tpu.ru
Доцент ОИС ИШИнЭС ТПУ, stepte@tpu.ru

Введение

Современные технологии обработки естественного языка (Natural Language Processing, NLP) активно развиваются и находят применение в самых различных областях – от анализа социальных медиа до автоматизированных систем поддержки клиентов. Одной из ключевых задач, стоящих перед NLP, является распознавание смысловой нагрузки текстов, особенно когда они содержат элементы неоднозначности. В этой статье будут рассмотрены основные методы и подходы, позволяющие справляться с подобными задачами.

Понятие смысловой нагрузки и неоднозначности

Одним из ключевых свойств человеческого языка является произвольность, то есть возможность разной трактовки сказанного в зависимости от контекста – узкого или широкого [1].

Эта возможность различных трактовок, то есть неоднозначность, присутствует на всех языковых уровнях, и именно в ее широком распространении и состоит одно из важных отличий естественного языка от искусственных семантических систем [2].

Факт существования у некоторой единицы более одного значения обозначается наиболее широким термином – «многозначность» (которая противопоставляется омонимии, предполагающей звуковое и графическое совпадение различных языковых единиц, значения которых не связаны друг с другом). Наличие же у языкового выражения нескольких различных смыслов одновременно предлагается обозначать термином «неоднозначность». [3].

Методы распознавания смысловой нагрузки

1. Нейронные сети и глубокое обучение.

Рекуррентная нейронная сеть (Recurrent Neural Network, RNN) – популярный вид нейронных сетей, используемых в обработке естественного языка (NLP). Рекуррентная нейросеть оценивает произвольные предложения на основе того, насколько часто они встречались в текстах. Это даёт меру грамматической и семантической корректности, что позволяет использовать такие модели для перевода текстов [4]. Они могут учитывать контекст слов в тексте, что особенно важно для обработки неоднозначностей.

2. Структура нейросети.

Сети с многослойными перцептронами (MLP) или радиальными базовыми функциями (RBF) зависят от предикторов (называемых также входными или независимыми переменными), которые минимизируют погрешность переменных назначения (их называют также выходными переменными) [5].

На рис.1 представлена архитектура прямой связи, соединения в сети направлены прямо от входного слоя на выходной слой без контуров обратной связи. На данном рисунке [5]:

1. Входной слой содержит предикторы.
2. Скрытый слой содержит ненаблюдаемые узлы, или нейроны. Значение каждого скрытого нейрона – это некоторая функция предикторов; точная форма этой функции частично зависит от типа сети, а частично - от управляемых пользователем спецификаций.
3. Выходной слой содержит отклики. Каждый выходной нейрон - это некоторая функция скрытых нейронов. Точная форма этой функции также частично зависит от типа сети, а частично – от управляемых пользователем спецификаций [5].

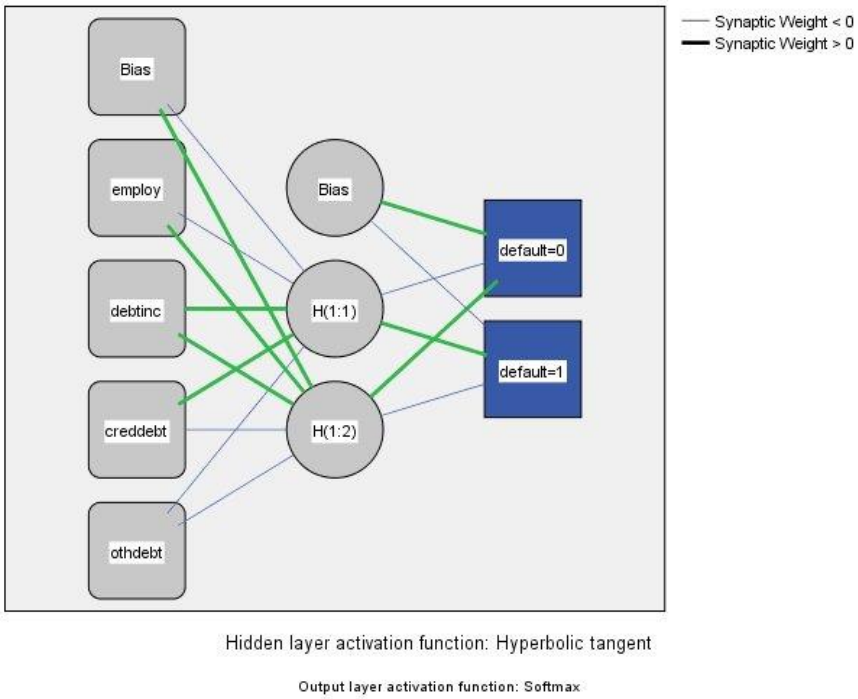

Рис. 1. Структура прямой связи с одним скрытым слоем

BERT (англ. Bidirectional Encoder Representations from Transformers) – языковая модель, предназначенная для предобучения языковых представлений с целью их последующего применения в широком спектре задач обработки естественного языка. BERT представляет собой нейронную сеть, основу которой составляет композиция кодировщиков трансформера. BERT является автокодировщиком. В каждом слое кодировщика применяется двустороннее внимание, что позволяет модели учитывать контекст с обеих сторон от рассматриваемого токена, а значит, точнее определять значения токенов [6].

Трансформер BERT обладает способностью к контекстуализации слов, позволяя лучше справляться с многозначностью. Он позволяет учитывать окружение слова, что позволяет различать его значения в зависимости от контекста.

Структура BERT

При подаче текста на вход сети сначала выполняется его токенизация. Токенами служат слова, доступные в словаре, или их составные части — если слово отсутствует в словаре, оно разбивается на части, которые в словаре присутствуют (рис. 2)

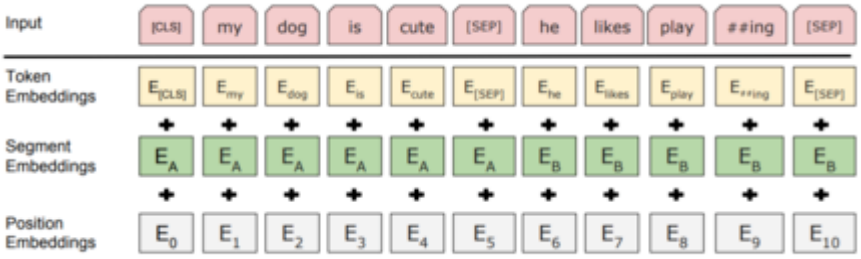

Рис. 2. Представление входных данных модели

В самой нейронной сети токены кодируются своими векторными представлениями (англ. embeddings), а именно, соединяются представления самого токена (предобученные), номера его предложения, а также позиции токена внутри своего предложения. Входные данные поступают на вход и обрабатываются сетью параллельно, а не последовательно, но информация о взаимном расположении слов в исходном предложении сохраняется, будучи включённой в позиционную часть эмбеддинга соответствующего токена [6].

Выходной слой основной сети имеет следующий вид: поле, отвечающее за ответ в задаче предсказания следующего предложения, а также токены в количестве, равном входному. Обратное преобразование токенов в вероятностное распределение слов осуществляется полносвязным слоем с количеством нейронов, равным числу токенов в исходном словаре [6].

Контекстуальное векторное представление

Векторное представление слов (англ. word embedding) – общее название для различных подходов к моделированию языка и обучению представлений в обработке естественного языка, направленных на сопоставление словам из некоторого словаря векторов небольшой размерности [7].

word2vec – способ построения сжатого пространства векторов слов, использующий нейронные сети. Принимает на вход большой текстовый корпус и сопоставляет каждому слову вектор. Сначала он создаёт словарь, а затем вычисляет векторное представление слов. Недостатком word2vec является то, что с его помощью не могут быть представлены слова, не встречающиеся в обучающей выборке [7].

fastText решает эту проблему с помощью NN-грамм символов. Например, 33-граммами для слова «яблоко» являются «ябл», «бло», «лок», «око». Модель fastText строит векторные представления NN-грамм, а векторным представлением слова является сумма векторных представлений всех его NN-грамм. Части слов с большой вероятностью встречаются и в других словах, что позволяет выдавать векторные представления и для редких слов [7].

Использование данных методов, таких как word embeddings позволяет создать векторные представления слов, учитывающие их семантическую схожесть. Это помогает в различии значений многозначных слов в зависимости от контекста.

Семантический анализ и расширенные алгоритмы

Семантический анализ – важная подзадача обработки естественного языка (Natural language processing, NLP), этап в последовательности действий алгоритма автоматического понимания текстов, заключающийся в выделении семантических отношений, формировании семантического представления текстов.

В общем случае семантическое представление является графом, семантической сетью, отражающей бинарные отношения между двумя узлами – смысловыми единицами текста. В ходе анализа текст проходит через несколько этапов обработки: токенизация для идентификации словоформ, морфологический, синтаксический анализ. Последним этапом идёт вторичный семантический анализ (первичный анализ в основном происходит параллельно морфологическому), в ходе которого устанавливаются взаимосвязи между сущностями, происходит извлечение мнений и анализ тональности текста. Основной целью анализа тональности является не только определение настроений, но также уровень объективности высказывания [8]. Методы семантического анализа, могут помочь в распознании значений слов и фраз. Эти подходы фокусируются на понимании отношений между различными элементами языка.

Анализ качества работы интеллектуальных систем (ИНС)

Анализ качества работы интеллектуальных систем (ИНС) в области обработки текстов с неоднозначностью требует детального подхода, учитывающего различные настройки и методы. Неоднозначность может происходить из-за контекста, многозначности слов, грамматических структур и других факторов.

Основные направления и предложения для улучшения работы:

1. Понимание контекста.

Контекстуальные модели: использование современных языковых моделей, таких как BERT, GPT и других, которые способны учитывать контекст, в котором используется слово или фраза.

Очерёдность обработки: для многозначных слов рассмотрение предыдущих и последующих предложений или фраз для определения правильного значения.

2. Разметка и аннотирование данных. Обогащение данных: создание дополнительных примеров, в которых неоднозначные слова используются в различных контекстах.

3. Адаптация и дообучение. Машинное обучение с подкреплением: внедрение подходов машинного обучения, где система обучается на основе обратной связи пользователей, улучшая свои решения в контексте неоднозначности.

4. Непрерывная обратная связь. Оценка качества: внедрение метрик и тестов для оценки работы системы в условиях неоднозначности, таких как точность, полнота и F1-мера.

Проблемы и вызовы

Несмотря на достижения в области обработки естественного языка, задачи распознавания смысловой нагрузки текстов с элементами неоднозначности остаются сложными. Основные проблемы включают:

1. Высокий уровень контекстуальной зависимости, особенно в разговорной речи. Значение слов и фраз может существенно варьироваться в зависимости от сценария, эмоциональной окраски и культурного контекста. Например, слово «блестящий» может означать «ярко сверкающий», но в другом контексте оно может означать «умный» или «выдающийся». Машине часто не хватает интуиции для распознавания таких нюансов, как наличие контекста, порой требующего глубоких культурных или социальных знаний.

2. Неполнота обучающих данных и ограниченная лексическая база, что может привести к недостаточной точности. Для обучения моделей машинного обучения необходимы качественные и разнообразные обучающие данные. Однако в реальности многие корпуса текстов имеют свои ограничения: они могут быть неполными, не учитывать редкие слова и фразы или не отражать все возможные контексты использования. Это создаёт пробелы в знаний моделей, что приводит к ошибкам и снижению точности в распознавании смысловой нагрузки.

3. Культурные и языковые различия, влияющие на интерпретацию текста. Язык – это нечто большее, чем просто набор правил и слов; он является отражением культуры и социальных норм. Модели, обученные на данных, относящихся к одной культуре, могут сталкиваться с трудностями при работе с текстами, генерируемыми в другой культурной среде. Например, идиоматические выражения, сленг и фразеологии могут не быть понятны без понимания культурного контекста, что представляет значительную сложность для моделей.

Заключение

Задачи распознавания смысловой нагрузки текстов с элементами неоднозначности многообразны и требуют комплексного подхода для их решения.Методы распознавания смысловой нагрузки текстов с элементами неоднозначности развиваются вместе с технологией обработки естественного языка. Нейронные сети, контекстуальные представления слов и правила семантического анализа становятся важными инструментами для решения этой задачи. Однако перед исследователями и разработчиками всё ещё стоят значительные вызовы, требующие дальнейших исследований и экспериментов. Интеграция различных подходов и методов, поможет достичь более высоких результатов в распознавании и интерпретации текстов в будущем.

Список использованных источников

1. Черниговская Т.В. Чеширская улыбка кота Шредингера: язык и сознание. / Т.В. Черниговская. – М. : Языки славянской культуры, 2013. – 448 с.

2. Маслов Ю.С. Омонимы в словарях и омонимия в языке (к постановке вопроса) / Ю.С. Маслов // Вопросы теории и истории языка: сб. в честь проф. Б.А. Ларина. – Л. : Изд–во ЛГУ, 1963. – С. 198–202.

3. Зализняк А.А. Феномен многозначности и способы его описания. / А. А. Зализняк // Вопросы языкознания. – 2004. – № 2. – С. 20–45.

4. Рекуррентная нейронная сеть (RNN): виды, обучение, примеры: : сайт. – 2024. – URL: https://neurohive.io/ru/osnovy-data-science/rekurrentnye-nejronnye-seti/ (дата обращения 22.10.2024).

5. Структура нейросети - IBM Documentation: сайт. – 2024. – URL: https://www.ibm.com/docs/ru/spss-statistics/saas?topic=networks-neural-network-structure (дата обращения 22.10.2024).

6. BERT (языковая модель) – Викиконспекты: сайт. – 2024. – URL: https://neerc.ifmo.ru/wiki/index.php?title=BERT (дата обращения 22.10.2024).

7. Векторное представление слов – Викиконспекты: сайт. – 2024. – URL: https://neerc.ifmo.ru/wiki/index.php?title= (дата обращения 22.10.2024)

8. Семантический анализ для автоматической обработки естественного языка — Научно-технический центр ФГУП "ГРЧЦ" (НТЦ): сайт. – 2024. – URL: https://rdc.grfc.ru/2021/09/semantic_analysis/ (дата обращения 22.10.2024).