Л. В. Шимчук, А. К. Дисько

СОВРЕМЕННЫЕ ТЕНДЕНЦИИ В АНАЛИЗЕ МЕТАТЕКСТОВОЙ И МОРФОЛОГИЧЕСКОЙ РАЗМЕТКИ В ПАРАЛЛЕЛЬНОМ КОРПУСЕ ТЕКСТОВ

Анализ метатекстовой и морфологической разметки в параллельном корпусе текстов является актуальной темой в контексте современного развития технологий обработки языка. В настоящее время наблюдаются определенные изменения в подходах к рассмотрению данного вопроса. В первую очередь речь идет о все более активном использовании интенсивно развивающихся методов глубокого обучения нейронных сетей. Исследователи все чаще начинают применять рекуррентные нейронные сети (RNN), свёрточные нейронные сети (CNN), и особенно часто трансформеры, для улучшения методов морфологического анализа и метатекстовой разметки. Это объясняется тем, что глубокие нейронные сети отлично подходят для обучения на больших объемах текстовых данных, что позволяет им улавливать сложные зависимости и структуры в языке, включая морфологические особенности и метатекстовые параметры. С их помощью возможно извлекать высокоуровневые скрытые признаки из текста, такие как словоформы, окончания слов, синтаксическая информация и другие морфологические особенности, которые могут быть полезны для метатекстовой разметки и анализа. Современные архитектуры нейронных сетей имеют возможность учитывать контекстуальные зависимости в тексте, что позволяет более глубоко анализировать смысловые аспекты текста. Их способность быть обученными на данных нескольких языков делает их эффективными для работы с мультиязычными параллельными корпусами, что чрезвычайно важно для переводов и исследований в области межъязыковых взаимодействий.

Корпусный анализ метатекстовой и морфологической разметки с использованием глубоких нейронных сетей может быть успешно использован для разработки и оценки методов машинного перевода, а также для исследований в области сопоставительного анализа между языками, экстракции параллельных выражений и других лингвистических задач.

Так, например, использование методов глубокого обучения нейронных сетей помогает эффективно решить проблему многозначности слов в корпусах текстов, поскольку они позволяют моделировать контекст и учиться на основных статистических закономерностях языка, могут создавать контекстуальные эмбеддинги слов, которые отражают их значения в данном контексте. Это позволяет более точно определить семантическое значение многозначных слов. Нейронные сети способны учитывать сложные зависимости между словами в предложении. Это позволяет им выявлять семантические связи и использовать их для разрешения многозначности. Методы глубокого обучения могут моделировать вероятностные распределения для различных значений слова в конкретных контекстах. Это позволяет определять вероятности различных значений слов и выбирать наиболее вероятное значение. Глубокое обучение позволяет использовать большие корпуса текстов для обучения моделей, что позволяет изучать широкий спектр контекстов, в которых встречаются многозначные слова.

Современные модели глубокого обучения, такие как BERT, GPT-3, трансформеры и другие, демонстрируют высокую эффективность в разрешении проблемы многозначности слов, обеспечивая более точное понимание и использование слов в соответствии с их значениями в конкретных контекстах.

Примером использования техник машинного обучения и контекстуальных эмбеддингов является попытка определить смысл многозначного глагола "turn" в контексте его окружения. При использовании контекстуальных эмбеддингов, модель учитывает окружение слова "turn" и создает векторное представление, учитывающее семантику слова в данном конкретном контексте. Пример использования многозначного глагола "turn" в различных контекстах:

1. "She decided to turn left at the intersection." (Она решила повернуть налево на перекрестке.)
2. "After the first chapter, the plot of the story began to take an unexpected turn." (После первой главы сюжет истории начал принимать неожиданный поворот.)
3. "He always turns to his friend for advice." (Он всегда обращается к своему другу за советом.)
4. "The economy is expected to turn around next year." (Ожидается, что экономика пойдет вверх в следующем году.)

В каждом из контекстов слово "turn" приобретает свое уникальное значение, и контекстуальные эмбеддинги могут учесть эту семантическую вариативность при представлении слова "turn" в контексте. Векторное представление контекстуальных эмбеддингов многозначного глагола "turn" в контексте его окружения будет зависеть от используемой модели. Для демонстрации примера рассмотрим, как могут выглядеть векторные представления для различных значений глагола "turn" с использованием модели контекстуального эмбеддинга, такой как BERT. Допустим, мы используем контекст "She decided to turn left at the intersection". Для каждого значения "turn" в данном контексте модель BERT создаст свой уникальный вектор. Предположим, что "turn" в этом предложении относится к действию поворота на перекрестке. Модель BERT, из-за контеста, закодирует это значение

"turn" в свой уникальный векторный представитель. Аналогично, когда мы используем другой контекст для "turn", например, "After the first chapter, the plot of the story began to take an unexpected turn", модель BERT создаст векторное представление, учитывающее значение "turn" как неожиданное сюжетное развитие. Векторное представление контекстуальных эмбеддингов многозначного глагола "turn" будет отличаться в зависимости от контекста, в котором он используется. Каждое значение "turn" удетб представлено собственным уникальным вектором, отражающим его семантическое значение в данном контексте. Этот вектор является результатом обучения модели на большом корпусе текстов и извлечении семантических связей между словами на основе их контекста. В результате обучения модель выявляет семантические ассоциации между словами и формирует числовые векторы для каждого слова, которые отражают его семантику на основе контекста, в котором оно встречается.

Таким образом, векторное представление контекстуальных эмбеддингов многозначных глаголов в контексте его окружения может быть использовано для анализа метатекстовой и морфологической разметки в параллельном корпусе текстов для разрешения многозначности, так как позволяет определить эквивалентные значения многозначных слов в различных контекстах, проводить семантический анализ слов в контексте их окружения и может помочь в выявлении семантических отношений между словами, такими как синонимы, антонимы, гиперонимы и гипонимы, и тем самым улучшить качество систем машинного перевода, поскольку этот прием способствует более точному сопоставлению слов и фраз в различных языках на основе их семантики в контексте.

Очевидно, что современные техники глубокого обучения нейронных сетей существенно улучшили процесс морфологического анализа и разметки метатекстов, способствуя развитию современных методов обработки текста и машинного перевода. Рекуррентные и сверточные нейронные сети, трансформеры способны обучаться на больших объемах данных и обобщать сложные зависимости в тексте, что позволяет им эффективно моделировать морфологические и синтаксические особенности языка. С развитием глубокого обучения достигнуты значительные успехи в задачах классификации, извлечения информации, определения частей речи, а также в машинном переводе. Трансформеры и другие архитектуры нейронных сетей значительно улучшили качество автоматического перевода за счет эффективного учета контекста и глубокого понимания семантики текста. Модели глубокого обучения значительно сократили потребность в ручной разметке текста и семантических аннотациях, позволяя автоматизировать и ускорить обработку больших объемов текста. Это стимулирует развитие методов обработки текста и перевода на различных языках. В целом, методы глубокого обучения нейронных сетей существенно улучшили качество морфологического анализа и метатекстовой разметки, способствуя развитию современных методов обработки текста и перевода за счет их способности учиться сложным зависимостям в тексте, учитывать контекст и автоматизировать процессы обработки текста.