УДК 519.767.6

Обучение в контексте для разметки семантических ролей на русском языке

Казачёнок Г.А.[1], Ларионов Д.С.[2], Смирнов И.В.[2]
[1]ФПМИ МФТИ
[2]ФИЦ ИУ РАН

Аннотация

В работе исследуется возможность использования языковых моделей (LLM) для автоматической разметки семантических ролей (Semantic Role Labeling, SRL) в русскоязычных текстах. Предложен подход на основе few-shot обучения с использованием корпуса FrameBank. Проведено сравнение эффективности модели Gemini 2.5 Flash с традиционными методами SRL.

Введение

Семантические роли (агенс, пациенс, инструментатив и др.) являются важным аспектом лингвистического анализа. Большинство современных подходов к SRL основаны на статистических методах и специализированных нейросетевых архитектурах. В данной работе исследуется потенциал современных LLM для этой задачи, что особенно актуально для русского языка с его богатой морфологией.

В этой работе используется few-shot подход: модель получает инструкцию, примеры разметки и целевое предложение, и на основе их предсказывает семантические роли. Примеры и тестовые данные берутся из размеченного русскоязычного корпуса FrameBank. Моделью решаются сразу две задачи: выделение аргументов предиката и разметка их семантических ролей. Результаты для этих задач оцениваются раздельно. Для оценки качества разметки используются классические метрики машинного обучения: precision, recall, F1. Предложенный подход сравнивается с результатами из [4].

Результаты

Таблица 1: Сравнение эффективности моделей

Модель	Precision (арг.)	Recall (арг.)	F1 (арг.)	F1 (роли)
Базовая [4]	74%	85%	79%	83%
Gemini 2.5 Flash	88%	85%	86%	83%

Как итог, 62% предложений получили разметку, полностью совпадающую с корпусной. В задаче нахождения аргументов LLM справляется значительно лучше, чем базовая модель, основанная на синтаксических признаках.

Основные проблемы:

- Трудности с извлечением аргументов из сложных синтаксических конструкций
- Путаница между семантически близкими ролями (напр. адресат/контрагент)

Заключение

LLM демонстрируют сопоставимую со специализированными моделями эффективность в задаче SRL. Большинство ошибок связаны со сложностью системы ролей, а не с пониманием языка. Предложенный подход может быть использован для автоматической и полуавтоматической разметки корпусов. Перспективным направлением является разработка более простых систем ролей, которые могут быть размечены даже неподготовленным человеком, и применение этого метода к ним.

Финансирование Исследование выполнено при финансовой поддержке Кирилла Пупкова.

Список литературы

[1] Gildea D., Jurafsky D. Automatic labeling of semantic roles // Computational linguistics. – 2002. – Т. 28, № 3. – С. 245-288.

[2] Strubell E. et al. Linguistically-informed self-attention for semantic role labeling // Proceedings of EMNLP. – 2018. – С. 5027-5038.

[3] Kuznetsov I. Semantic role labeling for Russian language based on Russian framebank // International Conference on Analysis of Images, Social Networks and Texts. – 2015. – С. 333-338.

[4] Lartonov D. et al. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates // Proceedings of RANLP. – 2019. – С. 619-628.