УДК 004.89

РАЗРАБОТКА СЕРВИСА ДЛЯ ЛЕММАТИЗАЦИИ РУССКОЯЗЫЧНЫХ ТЕКСТОВ С ПОМОЩЬЮ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Копырин Е.А. (ТюмГУ), **Сагадеев А.Р.** (ТюмГУ)
Научный руководитель – к.т.н., доцент Глазкова А.В. (ТюмГУ)

Введение. Лемматизация играет важную роль в обработке естественного языка, так как позволяет учитывать различные грамматические формы одного и того же слова как эквивалентные. Лемматизация применяется в различных областях, например: в поисковых системах для улучшения релевантности и полноты поиска, в чат-ботах и голосовых помощниках для понимания запроса пользователя без учёта грамматических форм слова, при анализе тональности текста для правильного определения эмоциональной окраски независимо от формы слов. Модели, основанные на правилах и эвристиках, часто демонстрируют низкую точность лемматизации для текстов социальных сетей [1], поскольку в таких текстах постоянно появляются новые слова, сокращения, аббревиатуры. В связи с этим, в качестве решения предлагается рассмотреть большие языковые модели, основанные на управлении с помощью запросов (промптов), оценить их перспективность в данной ситуации, провести сравнение с текущими решениями, и создать сервис для лемматизации русскоязычных текстов социальных сетей с помощью больших языковых моделей.

Основная часть. Перед исследованием больших языковых моделей в рамках задачи лемматизации русскоязычных текстов социальных сетей были взяты замеры точности (Accuracy) с современных инструментов для выполнения задачи лемматизации, таких как: PyMorphy2 [2], PyMystem3 [3], Stanza [4]. В качестве набора данных для замера точности был выбран корпус с русскоязычными текстами социальных сетей GramEval-2020 [1], имеющий формат CoNLL-U: корпус состоит из предложений, каждое из которых разбито на токены, при этом для каждого токена прописаны морфологические характеристики, например: число, лицо, род, часть речи, начальная форма. Всего токенов в корпусе – 1122, текстов – 114. При лемматизации без учёта контекста и части речи были получены следующие результаты: PyMorphy2 – 89,39%, PyMystem3 – 75,58%, Stanza – 89,13%. Были выбраны большие языковые модели для дальнейших исследований: T-Lite-Instruct-0.1 и Saiga-LLaMa-3 (8b). Данные модели были предложены для сравнения поскольку являются одними из самых актуальных среди больших языковых моделей, работающих с русскоязычными текстами. Сравнение моделей по точности проводилось по нескольким подходам к лемматизации: без учета контекста и части речи, с учетом контекста, с учетом части речи. Для каждого подхода к лемматизации были использованы следующие техники подбора шаблона запроса:

1. Zero-shot промптинг - техника создания шаблона запроса без указания примеров.
   Пример созданного шаблона:
   Приведи слово '{word}' к начальной форме (лемме) с учетом части речи ({pos}). В ответе напиши только одно слово - лемму, без комментариев и предложений. Все символы по типу ')', '-', '/' а также знаки препинания сохраняй в изначальном виде.
2. Few-shot промптинг - техника создания шаблона запроса с примерами, т.е. с использованием нескольких примеров входных и выходных данных.
   Пример созданного шаблона:
   "user": "Лемматизируй вводимые токены с учетом контекста. Знаки пунктуации при лемматизации не изменяются. В ответ выводи только одно слово или токен. Точку в конце токена не ставь. Букву 'ё' во всех словах заменяй на 'е'"
   "user": "Приведи существительное 'собаку' к именительному падежу, единственному числу.”
   "system": "собака"

…
"user": “{базовый запрос, определенный выше, с подставленным словом из корпуса}”

3. Role-based промптинг - техника создания шаблона запроса, в основе которого лежит задание роли / точки зрения. Пример такого подхода с двумя ролями: system и user приведен выше, но в конечном варианте был реализован подход с тремя ролями. Пример созданного шаблона:
"system": "Лемматизируй вводимые токены с учетом контекста. Знаки пунктуации при лемматизации не изменяются. В ответ выводи только одно слово или токен. Точку в конце токена не ставь. Букву 'ё' во всех словах заменяй на 'е'"
"user": "Приведи глагол 'играли' к инфинитиву, сохранив его вид (совершенный или несовершенный), так чтобы он отвечал на вопрос 'что делать?' или 'что сделать?’
"assistant": "играть"
…
"user": “{базовый запрос, определенный выше, с подставленным словом из корпуса}”

Наилучшие результаты по всем подходам к лемматизации были получены от модели Saiga-LLaMa-3: 85,88% при лемматизации без учета части речи и контекста, 80,39% при лемматизации без учета контекста, 91,1% при лемматизации с учетом части речи.

По результатам исследований больших языковых моделей в контексте задачи лемматизации, был разработан прототип сервиса по лемматизации русскоязычных текстов социальных сетей с помощью большой языковой модели Saiga-LLaMa-3 (8b), состоящий из API, WEB-приложения и Telegram Mini App. Помимо трех вышеописанных подходов к лемматизации, для API и WEB-приложения был реализован подход к лемматизации датасетов формата CoNLL-U, основанный на подходе к лемматизации с учетом части речи.

Выводы. Исследован потенциал больших языковых моделей в рамках задачи лемматизации русскоязычных текстов социальных сетей. Реализован прототип сервиса для лемматизации русскоязычных текстов социальных сетей с использованием большой языковой модели, позволяющий использовать несколько подходов к лемматизации: без учёта контекста и части речи, с учетом контекста, с учетом части речи. Для таких компонентов сервиса как API и WEB-приложение реализован четвертый подход: лемматизация токенов в датасетах формата CoNLL-U. Реализованный сервис позволит ускорить работу разметчиков, работающих с корпусами, содержащими русскоязычные тексты социальных сетей.

Список использованных источников:

1. Lyashevskaya O. N. et al. GramEval 2020: Russian full morphology and universal dependencies parsing // Proc. Dialogue. – 2020. – С. 553-569.
2. Segalovich I. A fast morphological algorithm with unknown word guessing for web search // Proc. ICML MTA. – 2003. – С. 273-280.
3. Qi P. et al. Stanza: A Python NLP toolkit for many languages // Proc. ACL System Demonstrations. – 2020. – С. 101-108..
4. Korobov M. Morphological analyzer for Russian and Ukrainian // Proc. AIST. – Springer, 2015. – С. 320-332.