УДК 004.91

Разработка методов уменьшения семантического пространства для обработки специализированных текстов

Першинов А.В. (ИТМО), **Сапельникова К.С.** (ИТМО), **Басилаев Д.И.** (ИТМО), **Воскресенский А.С.** (ИТМО)
Научный руководитель – инженер, аспирант Ковальчук М.А.
(ИТМО)

Введение. При работе с промышленными данными часто возникает проблема обработки различных видов документов. Эта проблема связана с тем, что за время работы предприятия накапливается массив разнородных текстовых данных, таких как документы (doc, pdf), таблицы (excel) и логи (txt). Извлечение данных из этих файлов – основной этап при реализации технологий индустрии 4.0. Помимо сложности и разнородности форматов данных, перед разработчиками систем обработки данных встаёт проблема неструктурированности их содержимого. Решением для обработки неструктурированных текстовых данных могут быть языковые модели. Применение больших языковых моделей в обработке документов включает в себя широкий спектр задач. Например, они могут автоматически генерировать резюме текстов, что значительно экономит время при работе с большими объемами информации. Также модели способны классифицировать документы по темам, упрощая их сортировку и поиск необходимых материалов. Кроме того, они могут выявлять ключевые моменты и сущности в текстах, такие как имена, даты и места, что полезно для аналитической работы и исследований.

В силу того, что предприятия накапливает информацию в течение нескольких лет – характер и структура этих данных может изменяться. Множество таких документов заполняются специалистами из отрасли, что порождает достаточно узкоспециализированный вокабуляр, из-за чего языковые не могут корректно обработать информацию.

В качестве решения этой проблемы предлагается методика уменьшения семантического пространства за счет замещения неизвестных терминов на известные синонимы или определения.

Основная часть. В настоящее время существует множество решений для работы с доменными текстами. Большие языковые модели используются в медицине (1), туризме , юриспруденции [3] и в социальных медиа [4]. В приведенных выше примерах используются предобученные на доменных текстах модели. Целью этой разработки является создание единого пайплайна для работы с различными доменами без необходимости дообучать модель. Такая задача обусловлена тем, что дообучение модели достаточно затратный процесс, который к тому же требует большой объем данных из доменной области.

Перед обработкой языковыми моделями содержимое документов нуждается в нескольких этапах предобработки.

1. Преобразование документов к единому формату
2. Поиск специализированных терминов
3. Приведение терминов к пространству знаний модели

Для того, чтобы выявить специализированные термины необходимо провести лемматизацию, стемминг и просканировать текст на наличие слов, которых нет в базе знаний. Найденные термины удаляются из текста и заменяются на подходящие по контексту слова из базы знаний модели.

Такой подход обеспечивает объединение доменных текстов в единую базу знаний модели.

Выводы. Проведен анализ методов решения проблемы Out-Of-Vocabulary слов [5]. Сформулирована задача и построен пайплайн обработки текстов.

Список использованных источников:

1. Yalunin A, Nesterov A, Umerenkov D. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. 2022 Apr 8; Available from: http://arxiv.org/abs/2204.03951
2. Arefieva V, Egger R. TourBERT: A pretrained language model for the tourism industry [Internet]. Available from: https://www.kaggle.com/jiashenliu/515k-
3. Huang Q, Tao M, Zhang C, An Z, Jiang C, Chen Z, et al. Lawyer LLaMA Technical Report. 2023 May 24; Available from: http://arxiv.org/abs/2305.15062
4. Sun Z, Zemel R, Xu Y. Semantically Informed Slang Interpretation [Internet]. Available from: https://github.com/
5. Andrusenko AY, Romanenko AN. Improving out of vocabulary words recognition accuracy for an end-to-end Russian speech recognition system. Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2022 Nov 1;22(6):1143–9.
6. Haque S, Eberhart Z, Bansal A, McMillan C. Semantic Similarity Metrics for Evaluating Source Code Summarization. In: IEEE International Conference on Program Comprehension. IEEE Computer Society; 2022. p. 36–47.
7. Haque S, Eberhart Z, Bansal A, McMillan C. Semantic Similarity Metrics for Evaluating Source Code Summarization. In: IEEE International Conference on Program Comprehension. IEEE Computer Society; 2022. p. 36–47.
8. Ke Z, Kachuee M, Lee S. Domain-Aware Contrastive Knowledge Transfer for Multi-domain Imbalanced Data. 2022 Apr 4; Available from: http://arxiv.org/abs/2204.01916
9. Färber M, Popovic N. Vocab-Expander: A System for Creating Domain-Specific Vocabularies Based on Word Embeddings. 2023 Aug 7; Available from: http://arxiv.org/abs/2308.03519

Першинов А.В. (автор) Подпись _____
Сапельникова К.С. (автор) Подпись _____
Басилаев Д.И. (автор) Подпись _____
Воскресенский А.С. (автор) Подпись _____
Ковальчук М.А. (научный руководитель) Подпись _____