ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет гуманитарных наук

Максим Дмитриевич Татаринов

ОЦЕНКА ПРИМЕНИМОСТИ МЕТОДА ДЕТЕКТИРОВАНИЯ СЕМАНТИЧЕСКИХ ИЗМЕНЕНИЙ СЛОВ НЕЙРОСЕТЕВОЙ ЯЗЫКОВОЙ МОДЕЛЬЮ НА ОСНОВЕ ГЕНЕРИРУЕМЫХ ОПРЕДЕЛЕНИЙ

Выпускная квалификационная работа

по направлению подготовки

45.03.03. Фундаментальная и прикладная лингвистика

Рецензент
TBD
П.П. Петров

Руководитель
доцент факультета информатики, математики и компьютерных наук ВШЭ
А. В. Демидовский

Нижний Новгород 2023

Аннотация

TODO

Оглавление

Введение

Целью настоящей работы является оценка применимости метода детектирования семантических изменений слов нейросетевой языковой моделью на основе генерируемых определений.

Из поставленной цели были сформулированы следующие **задачи**:

1. Провести анализ существующей литературы и решений по задаче детектирования семантических изменений на основе генерируемых определений.
2. Собрать тезаурус русского языка в качестве материала для обучения модели, а также диахронический корпус текстов на основе НКРЯ.
3. Обучить языковую модель на данных тезауруса для того, чтобы генерировать определения.
4. Провести анализ метрик и качества обученной языковой модели и сравнить их с существующими решениями.
5. Создать алгоритм автоматического определения семантических сдвигов на основе их векторного представления.
6. Провести комплексный лингвистический анализ результатов работы компьютерной программы.
7. Разработать прототип системы, позволяющей проводить анализ семантических изменений сторонним исследователям, используя разработанный в настоящей работе алгоритм.

Объектом исследования является метод детектирования семантических изменений слов.

Предметом исследования является применимость метода детектирования семантических изменений слов с использованием нейросетевой языковой модели на основе генерируемых определений.

Для решения поставленных задач были использованы следующие **методы**:

1. Метод анализа и синтеза для создания теоретической базы для данного исследования на основе литературы.
2. Компьютерный метод для написания алгоритмов программы и обучения модели.
3. Методы обработки естественного языка для предобработки текстов.
4. Методы машинного обучения для алгоритма автоматического определения семантических сдвигов на основе их векторного представления.
5. Метод комплексного лингвистического анализа результатов работы алгоритма.

Актуальность настоящей работы состоит в том, что, во-первых, вопрос анализа семантических изменений в русском языке на основе автоматически сгенерированных определений недостаточно изучен. Так, в настоящее время представление слов с помощью сгенерированных определений является перспективной темой для поиска семантических изменений, с еще небольшим количеством статей на данную тему на английском языке и отсутствием таких для русского. Во-вторых, традиционные методы поиска семантических изменений недостаточно информативны для основных потенциальных пользователей, таких как лексикографы или историки языка [1]. Им хотелось бы получать описания старых и новых значений слов в пригодной для чтения форме, возможно, даже с дополнительными пояснениями.

Новизна настоящей работы состоит в создании компьютерной программы, позволяющей автоматически определять семантические изменения,

с использованием автоматически сгенерированных определений, а также применением этого метода на русском языке.

Практическая значимость данной работы заключается в том, что результаты работы программы можно применять для определения степени семантического сдвига лексем, с наличием визуализаций и определений для каждого выявленного значения, что может быть использовано в лексикологии, где необходимы актуальные данные построения новых словарей.

В качестве **материала исследования** используется диахронический корпус НКРЯ, охватывающий три периода (1700—1916, 1918—1991 и 1992—2016 годы) и имеющий в совокупности 250 миллионов словоупотреблений. Данный корпус выбран, поскольку золотой датасет слов с изменившимся и неизменившимся значением, использующийся для оценки модели, основан на данном корпусе. Корпус был получен по запросу к авторам НКРЯ.

Глава 1. Имплементация автоматического выявления семантических изменений

1.1. Обучение языковой модели на данных тезауруса

В качестве модели была выбрана FRED-T5-1.7B, являющаяся одной из новейших языковых моделей, выпущенных SberDevices и обученных с нуля на материале русского языка [2]. Для выбора модели мы использовали бенчмарк для оценки продвинутого понимания русского языка ”RussianSuperGLUE” [3]. В бенчмарке присутствуют шесть групп задач, охватывая общую диагностику языковых моделей и различные лингвистические задачи: понимание здравого смысла, логическое следование в естественном языке, рассуждения, машинное чтение и знания о мире. FRED-T5-1.7B занимает самое высокое место в лидерборде данного бенчмарка, со значением 0.762, уступая лишь результатам выполнения данных заданий людьми со значением 0.811, что свидетельствует о ее способности к выдающемуся языковому пониманию и анализу. Таким образом, FRED-T5-1.7B представляется нам наиболее подходящей языковой моделью для задачи генерации определений.

В качестве материала, используемого для обучения модели, выступила русская версия Викисловаря. Материал получен с помощью самостоятельно написанного скрипта на языке Python, позволяющего извлечь данные из выгрузки Викисловаря в формат JSONL, где в каждом вхождении присутствовали идентификатор статьи, лексема, про которую написана данная статья, а также определения с примерами использования.

FRED-T5-1.7B была дообучена на полученном из Викисловаря материале, где на вход модель принимает лексему и контекст, в которой она употреблялась, а на выход ожидается сгенерированное определение.

Для оценки качества обучения модели используются метрики BLEU и ROUGE-L, которые оценивают формальную схожесть текста: BLEU оценивает точность совпадений n-грамм в сгенерированном тексте по сравнению с эталонным текстом [4], а ROUGE-L измеряет схожесть между сгенерированным текстом и эталонным текстом на основе наибольшей общей последовательности слов [5]. Также использовалась метрика BERT-F1, которая учитывает семантику сравниваемых текстов, так как использует модель BERT, представитель семейства моделей Transformer, обученные на больших объемах текста и имеющие глубокое понимание семантики [6]. Использование нескольких метрик позволяет получить более полную картину качества модели, поскольку каждая из них оценивает разные аспекты сгенерированного текста. Как традиционные BLEU и ROUGE-L, так и более современный BERT-F1 активно используются в задачах обработки естественного языка, в том числе в задачах генерации текста. В данной работе использовались версии этих инструментов, взятые из библиотеки evaluate [7]. Так, в обзорной статье по моделированию определений утверждается, что на момент выпуска статьи BLUE использовался в 9 научных публикациях, ROUGE-L и BERTScore – в 3 [8]. Кроме того, данные метрики используются и в более новых работах. Так, в настоящей статье результаты данных метрик будут сравниваться с таковыми из статьи Giulianelli M. et al., где сообщаются результаты трёх вышеперечисленных метрик при обучении модели T5 для задаче генерации определений на английском языке [1].

Список литературы

1. *Giulianelli M.*, *Luden I.*, *Fernández R.*, *Kutuzov A.* Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. — 2023.
2. FRED-T5. Новая SOTA модель для русского языка от SberDevices. — 2023. — URL: https://habr.com/ru/companies/sberdevices/articles/730088/ (дата обр. 15.11.2023).
3. *Shavrina T.*, *Fenogenova A.*, *Anton E.*, *Shevelev D.*, *Artemova E.*, *Malykh V.*, *Mikhailov V.*, *Tikhonova M.*, *Chertok A.*, *Evlampiev A.* RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Association for Computational Linguistics, 2020.
4. *Papineni K.*, *Roukos S.*, *Ward T.*, *Zhu W. J.* BLEU: a Method for Automatic Evaluation of Machine Translation. — 2002. — Окт.
5. *Lin C.-Y.* ROUGE: A Package for Automatic Evaluation of summaries //. — 01.2004. — С. 10.
6. *Zhang T.*, *Kishore V.*, *Wu F.*, *Weinberger K. Q.*, *Artzi Y.* BERTScore: Evaluating Text Generation with BERT. — 2020.
7. Evaluate. — URL: https://github.com/huggingface/evaluate (дата обр. 15.11.2023).
8. *Gardner N.*, *Khan H.*, *Hung C.-C.* Definition modeling: literature review and dataset analysis // Applied Computing and Intelligence. — 2022. — Т. 2. — С. 83—98.