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.

На **актуальность** настоящей работы указывают следующие факторы. Во-первых, активное изучение темы автоматического определения семантических изменений. В последние годы в работах использовались различные методы, включая статические эмбеддинги, контекстуальные эмбеддинги и заканчивая генерацией определений с помощью языковых моделей в новейших исследованиях [1—3] . При этом, абсолютное большинство исследований, посвященных моделированию определений, проводятся с использованием материала английского языка [4]. Для русского языка вопрос анализа семантических изменений на основе автоматически сгенерированных определений недостаточно изучен. Во-вторых, неудовлетворительное качество традиционных методов для основных потенциальных пользователей таких технологий, таких как лексикографы, историки языка и социологов. Например, лексикографам недостаточно данных только о факте сдвига значения, им хотелось бы получать

описания старых и новых значений слов в пригодной для чтения форме, возможно, даже с дополнительными пояснениями. Данная проблема может решаться моделированием определений с использованием языковых моделей, при использовании которых исследователи смогут получить более наглядные результаты [3].

Новизна настоящей работы состоит в том, что для детектирования семантических изменений значений слов применяется на материале русского языка и с использованием SOTA-моделей.

Практическая значимость данной работы заключается в том, что результаты настоящей работы 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 для построения новых словарей [3]. Кроме того, модель, позволяющая автоматически генерировать качественные словарные определения, может быть полезна в таких задачах обработки естественного языка, как анализ тональности, машинный перевод и разграничение семантической неоднозначности [4].

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 датасет слов для валидации с изменившимся и неизменившимся значением, использующийся для оценки алгоритма, основан на данном корпусе [2]. Корпус был получен по запросу к авторам НКРЯ.

Глава 1. Теоретические аспекты автоматического выявления семантических изменений

1.1. Понятия и классификации

TBD

1.2. Обзор существующих методов

1.2.1. Статические эмбеддинги

TBD

1.2.2. Контекстуализированные эмбеддинги

TBD

1.2.3. Трансформеры

Одни из последних работ по теме автоматического выявления семантических сдвигов для русского языка были написаны в рамках соревнования RuShiftEval, прошедшего в 2021 году. В ходе него участники должны были рассмотреть три исторических периода русского языка и общества: предсоветский (1700-1916), советский (1918-1990) и постсоветский (1992-2016). Исследование базировалось на наборе данных RuShiftEval, который состоит из 111 русских существительных (99 в тестовом наборе и 12 в наборе для разработки), вручную аннотированных по степени изменения их значения в трех парах временных периодов.

Аннотаторам предлагалось оценить семантическую связь значений целевого слова в двух предложениях из разных временных периодов. Оценки (от 1 до 4) отражали степень семантического родства между значениями слова, где 1 обозначало отсутствие связи между значениями, а 4 – их совпадение. Затем индивидуальные оценки усреднялись, формируя общую меру семантической

родственности между употреблениями слова в разные временные периоды. Такая задача как правило называется Word-in-Context или WiC.

Для каждого из 99 целевых русских слов участники должны были представить три значения, соответствующих семантическому изменению в упомянутых парах временных периодов. Эти значения использовались для построения трех ранжирований: RuSemShift1, RuSemShift2 и RuSemShift3. В качестве метрики оценки использовалась ранговая корреляция Спирмена между ранжированием слов, сгенерированным системой, и золотым ранжированием, полученным в ручной аннотации.

Победители вышеупомянутого соревнования (команда GlossReader) указывают, что проблемой в существующих решениях являлось то, что эмбеддинги несут в основном информацию о форме слова, а не значении [5]. Чтобы решить это, они дообучали модель XLM-R на задаче генерации эмбеддингов, максимально близким к таким, какие получены на соответстующим использованиям слов словарным определениям [6].

При дообучении их система включает в себя два отдельных энкодера на основе XLM-R: Энкодер контектов для кодирования предложения с целевым словом и энкодер глоссов для кодирования определения слова. Система оценивает возможные значения смысла слова путём сравнения векторных представлений слова и его определений. При этом для обучения использовались данные только по английскому языку, но модель также показала хорошие результаты для русского языка.

Далее, исследователи получали эмбеддинги контекстов слов с помощью дообученного энкодера контекстов, высчитывали расстояние с помощью различных метрик расстояния, самым эффективным из которых были евклидово расстояние с нормализацией, после чего логистическая регрессия приводила значения к формату в датасете.

Авторы статьи предоставляют доступ к части исходного кода их исследования [7].

Так, были опубликованы следующие компоненты:

1. Код, предназначенный для генерации прогнозов на основе заранее вычисленных эмбеддингов, полученных с использованием модели.
2. Код для оценки результатов.

В то же время, авторы исследования не представили в открытый доступ следующие части:

1. Код для предварительного обучения модели.
2. Код, позволяющий осуществлять генерацию контекстуализированных эмбеддингов.

В соответствии с инструкциями, данными авторами, мы запустили доступный код, в следствие чего были полуены высокие результаты, совпадающие с тем, что сообщают авторы в своей работе:

Таблица 1.1.. Коэффициэнты корреляции

Пары периодов	Коэффициент корреляции
Среднее	0.8021
pre-Soviet:Soviet	0.7808
Soviet:post-Soviet	0.8032
pre-Soviet:post-Soviet	0.8223

Среди недостатков работы можно отметить неспособность модели корректно выявлять значения тех слов, которые отличаются от ближайших аналогов в английском, например, ”пионер”, связанный с коммунистической идеологией и не соответствующий в полной мере словам ”scout” или ”pioneer”.

Кроме того, команда DeepMistake представила решение, занявшее в соревнование второе место [8]. Однако, они смогли доработать его и повысить результаты до первого уже после окончания соревнования.

Исследователи обучали модель XLM-R на обширном многоязычном датасете Word-in-Context, а затем дооубчали ее на наборе данных RuSemShift для настоящей задачи, приводит к наилучшим результатам. В отношении архитектуры авторы утверждают, что применение линейного слоя на верхнем уровне, основанного на объединении L1-метрики и скалярного произведения между контекстуализированными эмбеддингами XLM-R, показывает лучшую производительность по сравнению с более традиционными подходами, такими как конкатенация эмбеддингов и использование нелинейных классификаторов.

Исследователи выложили исходный код полностью и предлагают возможность воспроизвести их результат [9]. Значени метрик, сообщенные исследователями, воспроизводятся.

Таблица 1.2.. Коэффициэнты корреляции с использованием IsoReg

Пары периодов	Коэффициент корреляции
Среднее	0.8494
pre-Soviet:Soviet	0.8563
Soviet:post-Soviet	0.841
pre-Soviet:post-Soviet	0.8511

Среди недостатков статьи можно выделить то, что авторы не предоставляют возможность визуализации или интерпретации результатов, кроме непосредственно получившегося значения метрики.

TBD

Самой актуальной работой по теме автоматического выявления семантических изменений является статья Giulianelli et al. [3], в которой

исследователи предложили использовать автоматически сгенерированные определения для задачи анализа семантических изменений.

Авторы определяют задачу генерации определений следующим образом: для заданного слова w и примера использования s (предложения, содержащего w) необходимо сгенерировать определение d на естественном языке, которое будет грамматически корректным и точно передавать значение слова w в контексте его использования. Для генерации определений они используют модель Flan-T5, версию трансформера T5, дополнительно обученную на 1,8 тысячах задач по обработке естественного языка.

Первым шагом исследователи выбирают, используя метрики BLEU, NIST, BERTScore, наиболее подходящий под задачу промпт из нескольких вариантов, например ”what is the definition of <trg>?” или ”define the word <trg>”.

Для дообучения модели авторы используют три датасета, каждый из которых содержит определения слов, сопровождаемые примерами употребления: WordNet, данные Оксфордского словаря и CoDWoE, основанный на определениях и примерах, извлеченных из Викисловаря.

Для оценки качества модели исследователи используют метрики SacreBLEU, ROUGE-L и BERT-F1.

Для демонстрации работы со сгенерированными определениями авторы работы используют датасет, в котором слова представлены в графах диахронного использования слов (Diachronic Word Usage Graphs, DWUG), взвешенных, ненаправленных графах, созданных в результате ручной аннотации, узлами которых служат примеры использования слов, а веса рёбер отражают семантическую близость пар употреблений.

Прежде всего, авторы исследования проводят анализ корреляции между близостью пар слов в DWUG и контекстуальными эмбеддингами токенов, эмбеддингами предложений примеров использования, а также сгенерированными определениями. Результаты показали, что сгенерированные определения обладают

более высокой степенью корреляции с данными из DWUG, чем традиционно полученные эмбеддинги.

Далее исследователи анализируют пространство эмбеддингов определений слов, чтобы выяснить, как они могут помочь в различении разных значений слов. Они обнаружили, что эмбеддинги определений образуют более плотные и четко определенные кластеры по сравнению с эмбеддингами токенов и примеров предложений, что делает их подходящими для представления значений слов.

Позже авторы присваивали кластерам, полученным на основе данных из DWUG, соответствующие им определения. Для обобщения определений в одном кластере авторы использовали самое прототипическое из них. Они представляли все определения с помощью их эмбеддингов предложений и выбирали в качестве прототипичного определение, эмбеддинг которого наиболее похож на среднее значение всех эмбеддингов в кластере.

Авторы приходят к выводу, что сгенерированные определения слов могут играть роль семантического представления слов, аналогичному традиционным эмбеддингам. Они утвердлают, что большие языковые модели достаточно развитые для генерации определений через промпты с контекстом. При этом полученные таким образом определения превосходят по качеству традиционные эмбеддинги и являются более наглядными.

Глава 2. Имплементация автоматического выявления семантических изменений

2.1. 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 [10]. Для выбора модели мы использовали бенчмарк для оценки продвинутого понимания русского языка ”RussianSuperGLUE” [11].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 [12], а ROUGE-L измеряет схожесть между сгенерированным текстом и эталонным текстом на основе наибольшей общей последовательности слов [13]. Также использовалась метрика BERT-F1, учитывающая семантику сравниваемых текстов благодаря использованию контекстуальных эмбеддингов при подсчете значения метрики [14]. 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 [15]. 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 [4]. 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 [3].

Список литературы

1. *Kutuzov A.*, *Øvrelid L.*, *Szymanski T.*, *Velldal E.* Diachronic word embeddings and semantic shifts: a survey // Proceedings of the 27th International Conference on Computational Linguistics. — Santa Fe, New Mexico, USA : Association for Computational Linguistics, 08.2018. — С. 1384—1397.
2. *Rodina J.*, *Trofimova Y.*, *Kutuzov A.*, *Artemova E.* ELMo and BERT in semantic change detection for Russian. — 2020.
3. *Giulianelli M.*, *Luden I.*, *Fernández R.*, *Kutuzov A.* Interpretable Word Sense Representations via Definition Generation: The Case of Semantic Change Analysis. — 2023.
4. *Gardner N.*, *Khan H.*, *Hung C.-C.* Definition modeling: literature review and dataset analysis // Applied Computing and Intelligence. — 2022. — Т. 2. — С. 83—98.
5. *Rachinskiy M.*, *Arefyev N.* Zero-shot Cross-lingual Transfer of a Gloss Language Model for Semantic Change Detection //. — 06.2021. — С. 578—586.
6. *Conneau A.*, *Khandelwal K.*, *Goyal N.*, *Chaudhary V.*, *Wenzek G.*, *Guzmán F.*, *Grave E.*, *Ott M.*, *Zettlemoyer L.*, *Stoyanov V.* Unsupervised Cross-lingual Representation Learning at Scale // CoRR. — 2019. — Т. abs/1911.02116.
7. GlossReader. — URL: https://github.com/myrachins/RuShiftEval (дата обр. 18.01.2024).
8. *Arefyev N.*, *Fedoseev M.*, *Protasov V.*, *Panchenko A.*, *Homskiy D.*, *Davletov A.* DeepMistake: Which Senses are Hard to Distinguish for a Word-in-Context Model //. — 06.2021. — С. 16—30.
9. DeepMistake. — URL: https://github.com/Daniil153/DeepMistake (дата обр. 18.01.2024).

10. 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).

11. *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.

12. *Papineni K.*, *Roukos S.*, *Ward T.*, *Zhu W. J.* BLEU: a Method for Automatic Evaluation of Machine Translation. — 2002. — Окт.

13. *Lin C.-Y.* ROUGE: A Package for Automatic Evaluation of summaries //. — 01.2004. — С. 10.

14. *Zhang T.*, *Kishore V.*, *Wu F.*, *Weinberger K. Q.*, *Artzi Y.* BERTScore: Evaluating Text Generation with BERT. — 2020.

15. Evaluate. — URL: https://github.com/huggingface/evaluate (дата обр. 15.11.2023).