Научная статья
Original article
УДК 004.032.2

АНАЛИЗ МОДЕЛИ BERT КАК ИНСТРУМЕНТА ОПРЕДЕЛЕНИЯ МЕРЫ СМЫСЛОВОЙ БЛИЗОСТИ ПРЕДЛОЖЕНИЙ ЕСТЕСТВЕННОГО ЯЗЫКА

ANALYSIS OF THE BERT MODEL AS A TOOL FOR DETERMINING THE MEASURE OF SEMINAL PROXIMITY OF NATURAL LANGUAGE SENTENCES

Салып Богдан Юрьевич, студент 2 курса магистратуры, факультет "Информатика и системы управления", МГТУ им. Н.Э. Баумана, Россия, г. Москва

Смирнов Александр Андреевич, студент 2 курса магистратуры, факультет "Информатика и системы управления", МГТУ им. Н.Э. Баумана, Россия, г. Москва

Научный руководитель: Ничушкина Татьяна Николаевна

Salyp Bogdan Yurievich, 2nd year master's student, Faculty of Informatics and Control Systems, Moscow State Technical University. N.E. Bauman, Russia, Moscow

Smirnov Alexander Andreevich, 2nd year master's student, Faculty of Informatics and Control Systems, Moscow State Technical University. N.E. Bauman, Russia, Moscow

Scientific adviser: Nichushkina Tatyana Nikolaevna

Аннотация: В статье проводится анализ модели BERT как инструмента определения меры смысловой близости предложений естественного языка.

Рассматриваются специфика обработки естественного языка, строение, особенности и факторы успешности нейросетевой архитектуры трансформеров. Приводятся описание, структура и характерные черты модели BERT, отмечаются её преимущества, недостатки и возможности для определения меры смысловой близости предложений естественного языка.

Annotation: The article analyzes the BERT model as a tool for determining the measure of semantic similarity of natural language sentences. The specifics of natural language processing, the structure, features and success factors of the neural network architecture of transformers are considered. The description, structure and characteristic features of the BERT model are given, its advantages, disadvantages and possibilities for determining the measure of semantic similarity of natural language sentences are noted.

Ключевые слова: естественный язык, смысловая близость предложений, языковая модель, архитектура трансформеров, модель BERT.

Keywords: natural language, semantic similarity of sentences, language model, transformer architecture, BERT model.

Одной из наиболее заметных современных технологических тенденций является развитие автоматического анализа информации на естественном языке (далее – ЕЯ) [1]. Данная область представляет огромный интерес для маркетинга, социологии, психологии, лингвистики и прочих сфер человеческой деятельности, в особенности в условиях стремительного расширения сети Интернет. В число наиболее актуальных задач обработки ЕЯ входит разработка методов и алгоритмов установления и формального описания смысла сущностей ЕЯ и последующей численной оценки их попарной семантической близости [2]. Под мерой семантической близости пары текстов или предложений понимается числовая оценка семантической близости этой пары, от точности которой зависит полнота использования потенциальных возможностей семантического ресурса. Это повышает значимость исследования отдельных моделей определения меры смысловой

близости предложений ЕЯ, в частности – нейросетевой архитектуры трансформеров и её разновидности BERT, показавшей наилучшие результаты работы и скорости обучения [3].

Целью работы является анализ модели BERT как инструмента определения меры смысловой близости предложений ЕЯ. Для её достижения были использованы методы анализа и синтеза научных публикаций и литературных источников по рассматриваемой теме.

Центральным элементом в обработке ЕЯ является представление текста в виде отражающих его семантику набора численных характеристик [4]. Поскольку любой текст имеет чётко выделенную секвенциальную структуру, ему в соответствие обычно ставится последовательность числовых векторов фиксированной размерности, характеризующих отдельные элементы текста. Для языков индоевропейской группы наиболее популярны модели векторных представлений слов, в которых каждому слову соответствует вектор (эмбеддинг), а каждому тексту – последовательность векторов слов, образующих его. Одной их таких моделей является архитектура трансформеров.

Трансформер представляет собой модель глубокого обучения, состоящую из стэков равного количества декодеров и энкодеров [5]. Энкодеры обрабатывают входные последовательности данных и кодируют их для отражения сведений о них и их признаках, а декодеры обрабатывают полученные от энкодеров данные и генерируют выходные последовательности. Все энкодеры построены по единой структуре, включающей два слоя: внутреннее внимание и нейронная сеть с прямой связью. Входная текстовая последовательность сперва проходит через слой внутреннего внимания, что позволяет энкодеру проанализировать другие слова во входном предложении в случае кодирования конкретного слова. Выходные данные данного слоя поступают в нейронную сеть с прямой связью.

Декодер содержит такие же слои, между которыми включён дополнительный слой внимания, позволяющей определять релевантные части

входного предложения. Внутреннее внимание даёт модели возможность обнаруживать зависимости между обрабатываемым словом и остальными словами входной последовательности, помогающими лучше закодировать слово. После всех декодеров применяется полносвязный слой Softmax, преобразующий полученные значения в вероятности, из которых осуществляется выборка наибольшего значения, которому в соответствие ставится слово, становящееся выходом для данного временного шага.

Трансформеры обучаются на данных большого объёма, что позволяет им выдавать векторные представления слов, богатые семантическими и синтаксическими данными [6]. Благодаря этому возможно избежать процесса подбора низкоуровневых признаков и сразу переходить к обнаружению и классификации анализируемых семантических отношений.

Успешность архитектуры трансформеров и её вариаций обусловлена следующими разработками [7]:

- улучшенные алгоритмы токенизации, использующиеся для обработки редких или маскированных слов;
- контекстно-векторное представление слов;
- механизмы внимания, устраняющие необходимость кодировать все данные предложения в единый эмбеддинг контекста;
- использование в предложениях дополнительных токенов (слов) запоминания;
- позиционное кодирование порядка слов;
- механизм многопоточного самовнимания;
- сжатие моделей;
- параллелизируемые архитектуры, которые ускоряют точную настройку и обучение;
- перенос обучения (многоязычность и многозадачность).

На архитектуре трансформеров основана модель BERT, обучающаяся методом стохастического градиентного спуска на базе входных данных [8]. Данная модель не требует обработки текста в строгом порядке следования в

документе, благодаря чему она легко распараллеливается и может быть быстро обучена. Во время обучения для каждого представленного во входном наборе данных документа BERT учится обнаруживать векторное представление, в последующем отправляемое на вход уровня классификации. Уровнем классификации в этом методе выступает нейронная сеть прямого распространения.

В отличие от прочих подобных моделей, BERT предназначена для предобучения двунаправленных представлений на неразмеченном текстовом корпусе посредством обучения на контексте и справа, и слева [9]. Под двунаправленностью понимается обучение предсказанию токенов в зависимости как от префикса, так и от суффикса, которые окружают маскированное слово. В качестве последнего случайным образом выбирается любой токен входной последовательности, который маскируется, после чего модели необходимо предсказать его исходное значение, исходя из контекста. По сравнению с однонаправленными, двунаправленные модели предлагают существенно более широкие возможности предобучения, что приводит к увеличению эффективности языковых моделей в решении задач обработки ЕЯ.

Схема преобразования данных в модели BERT при решении задач обработки ЕЯ имеет следующий вид [10]:

$S = (w_1, \dots, w_n)$ – токены входного предложения

↓ обучение эмбеддингов вместе с трансформером

$X = (x_1, \dots, x_n)$ – эмбеддинги токенов входного предложения

↓ трансформер кодировщика

$Z = (z_1, \dots, z_n)$ – трансформированные эмбеддинги

↓ дообучение на конкретную задачу

Y – выходные данные

Ключевой особенностью BERT является возможность рассмотрения в ходе обучения отдельного слова в контексте слов, окружающих его [11]. Предобучение BERT на большом объёме текстов даёт возможность

сформировать контекстуализированные векторные представления слов для некоторого ЕЯ, которые в дальнейшем могут быть использованы для определения меры смысловой близости предложений.

BERT является автоэнкодером, что обуславливает недостатки модели, в том числе индивидуальное предсказание каждого скрытого слова, приводящее к потере информации о возможных связях между замаскированными словами [12]. При попытке самостоятельного предсказания таких слов в контексте не будут учитываться взаимосвязи между ними. Другим недостатком является несоответствие между фазами обучения модели BERT и использованием предварительно обученной системы. Так, при обучении модели используются скрытые токены, а при применении предобученной модели такие токены уже не подаются на вход.

Также, несмотря на очень высокий процент верных решений серии различных задач, продемонстрированный моделью BERT, исследования показали, что после обучения данная система ориентируется в том числе на ложные эвристики, которые обусловлены неудачным подбором данных в используемых датасетах [13]. Это свидетельствует о необходимости дальнейшего совершенствования рассматриваемой языковой модели.

Таким образом, модель BERT отличается простотой и высокой скоростью обучения и позволяет рассматривать отдельные слова в контексте окружающих их слов. Она опирается на применение механизма внимания и принцип переноса знаний, совместное функционирование которых дают возможность эффективно обнаруживать закономерности между использованием слов в значительной обучающей выборке неразмеченных текстов. Обнаруживаемые закономерности позволяет результативно решать разнообразные задачи, сопряжённые с автоматической обработкой ЕЯ. Модель BERT обладает способностью к вычислению контекстуализированных векторных представлений слов входного предложения, а выходы её слоёв кодирования являются готовыми эмбеддингами, что делает модель перспективным инструментом для решения

задачи определения семантической схожести предложений ЕЯ. При этом качество работы модели можно увеличивать за счёт усложнения архитектуры и расширения обучающей выборки.

Список литературы

1. Лыченко Н.М., Сороковая А.В. Сравнение эффективности методов векторного представления слов для определения тональности текстов // Математические структуры и моделирование. – 2019. – № 4 (52). – С. 97-110.
2. Городецкий В.И., Тушканова О.Н. Семантические технологии для семантических приложений. Часть 2. Модели сравнительной семантики текстов // Искусственный интеллект и принятие решений. – 2019. – № 1. – С. 49-61.
3. Николенко С. Transformer: внимание на себя // Лаборатория математической логики. – СПб.: НИУ ВШЭ, 2019. – 77 с.
4. Гринчук А.В. Матричные и тензорные разложения в задачах обработки естественного языка: дис. … канд. физ.-мат. наук: 05.13.18 / Алексей Валерьевич Гринчук; Москов. физ.-тех. инст. – М., 2021. – 98 с.
5. Гусенков А.М., Ситтикова А.Р. Применение машинного обучения к задаче генерации поисковых запросов // Электронные библиотеки. – 2021. – Т. 24, № 2. – С. 272-293.
6. Батура Т.В. Семантический анализ научных текстов: опыт создания корпуса и построения языковых моделей / Т.В. Батура, Е.П. Бручес, А.Е. Паульс, В.В. Исаченко, Д.Р. Щербатов // Программные продукты и системы. – 2021. – № 1. – С. 132-144.
7. Йылдырым С., Асгари-Ченаглу М. Осваиваем архитектуру Transformer. Разработка современных моделей с помощью передовых методов обработки естественного языка / пер. с анг. В.С. Яценкова. – М.: ДМК Пресс, 2022. – 318 с.
8. Саврасов А.Е. Построение модели классификации обращений в службу поддержки //Математика, информационные технологии, приложения:

межвузовская научная конференция молодых учёных и студентов. – Воронеж, 2021. – С. 211-216.

9. Понкин Д.И. Концепт предобученных языковых моделей в контексте инженерии знаний // International Journal of Open Information Technologies. – 2020. – Т. 8, № 9. – С. 18-29.
10. Devlin J., Chang M.-W., Lee K., Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*. Available at: https://arxiv.org/abs/1810.04805 (accessed 20.04.2022).
11. Ломов П.А., Малоземова М.Л. Аугментация обучающего набора при обучении нейросетевой языковой модели для наполнения онтологии // Труды Кольского научного центра РАН. Информационные технологии. – 2021. – Т. 12, № 5. – С. 22-34.
12. Solomin A.A., Ivanova Yu.A. Modern approaches to multiclass intent classification based on pre-trained transformers. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2020, vol. 20, no. 4, pp. 532-538.
13. Качков Д.И. Моделирование языка и двунаправленные представления кодировщиков: обзор ключевых технологий // Информатика. – 2020. – Т. 17, № 4. – С. 61-72.

Bibliography

1. Lychenko N.M., Sorokovaya A.V. Comparison of the effectiveness of methods of vector representation of words for determining the tonality of texts // Mathematical structures and modeling. - 2019. - No. 4 (52). - S. 97-110.
2. Gorodetsky V.I., Tushkanova O.N. Semantic technologies for semantic applications. Part 2. Models of comparative semantics of texts // Artificial intelligence and decision making. - 2019. - No. 1. - P. 49-61.
3. Nikolenko S. Transformer: attention to yourself // Mathematical Logic Laboratory. - St. Petersburg: NRU HSE, 2019. - 77 p.

4. Grinchuk A.V. Matrix and tensor expansions in natural language processing problems: dis. … cand. Phys.-Math. Sciences: 05.13.18 / Aleksey Valerievich Grinchuk; Moscow. fiz.-tech. inst. - M., 2021. - 98 p.
5. Gusenkov A.M., Sittikova A.R. Application of machine learning to the problem of generating search queries // Digital Libraries. - 2021. - T. 24, No. 2. - S. 272-293.
6. Batura T.V. Semantic analysis of scientific texts: the experience of creating a corpus and constructing language models / T.V. Batura, E.P. Bruches, A.E. Pauls, V.V. Isachenko, D.R. Shcherbatov // Software products and systems. - 2021. - No. 1. - P. 132-144.
7. Yildirim S., Asgari-Cenaglu M. Mastering the architecture of Transformer. Development of modern models using advanced methods of natural language processing / transl. from eng. V.S. Yatsenkov. – M.: DMK Press, 2022. – 318 p.
8. Savrasov A.E. Building a classification model for calls to the support service //Mathematics, Information Technology, Applications: Interuniversity Scientific Conference of Young Scientists and Students. - Voronezh, 2021. - S. 211-216.
9. Ponkin D.I. The concept of pre-trained language models in the context of knowledge engineering // International Journal of Open Information Technologies. - 2020. - T. 8, No. 9. - S. 18-29.
10. Devlin J., Chang M.-W., Lee K., Toutanova K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. Available at: https://arxiv.org/abs/1810.04805 (accessed 04/20/2022).
11. Lomov P.A., Malozemova M.L. Augmentation of the training set when training a neural network language model for ontology filling // Proceedings of the Kola Scientific Center of the Russian Academy of Sciences. Information Technology. - 2021. - T. 12, No. 5. - S. 22-34.
12. Solomin A.A., Ivanova Yu.A. Modern approaches to multiclass intent classification based on pre-trained transformers. Scientific and Technical

Journal of Information Technologies, Mechanics and Optics, 2020, vol. 20, no. 4, pp. 532-538.

13. Kachkov D.I. Language Modeling and Encoder Bidirectional Representations: A Review of Key Technologies // Informatics. - 2020. - V. 17, No. 4. - S. 61-72.

© Салып Б.Ю., Смирнов А.А. 2022 Научно-образовательный журнал для студентов и преподавателей «StudNet» №5/2022.

Для цитирования: Салып Б.Ю., Смирнов А.А. АНАЛИЗ МОДЕЛИ BERT КАК ИНСТРУМЕНТА ОПРЕДЕЛЕНИЯ МЕРЫ СМЫСЛОВОЙ БЛИЗОСТИ ПРЕДЛОЖЕНИЙ ЕСТЕСТВЕННОГО ЯЗЫКА // Научно-образовательный журнал для студентов и преподавателей «StudNet» №5/2022.