О влиянии синтаксической структуры предложения на его векторизацию с использованием модели Bert

Э.С. Клышинский [1], В.В. Васильева [2], О.В. Карпик [1], Ю.А. Белобокова [3]

[1] *Институт прикладной математики им. М.В. Келдыша, Миусская пл., д.4, Москва, 125047, Россия*
[2] *Национальный исследовательский университет «Высшая школа экономики», ул. С. Басманная, 21/4с1, Москва, 105066, Россия*
[3] *МРСЭИ, ул. Школьная д. 55а к. 1, Видное, 142700, Россия*

Аннотация
Эксперименты показывают, что векторизация Bert отражает синтаксическую структуру предложения. В данной работы мы провели эксперименты по определению косинусной меры сходства между векторами Bert для слов, занимающих в предложении близкие позиции. Эксперименты показали, что векторизация Bert зависит от количества синтаксических составляющих, которые завершаются или начинаются между этими словами — с увеличением числа составляющих косинусное сходство падает. Более значительный эффект достигается для синтаксических составляющих, которые завершились между этими словами. При увеличении расстояния между словами до промежутка в три слова косинусное сходство также падает. Вообще, Bert присваивает словам в тексте близкие векторы, косинусное сходство которых выше 0,6. Для Word2Vec, которая формирует векторы без учета контекста, подобное поведение не характерно.

Ключевые слова
Модель Bert, векторное представление слов, синтаксический анализ, составляющие.

On Dependence of Bert Embeddings on the Syntactic Structure of a Sentence

E.S. Klyshinsky [1] V.V. Vasilyeva [2], O.V. Karpik [1], Yu.A. Belobokova [3]

[1] *Keldysh Institute of Applied Mathematics, Miusskaya sq. 4, Moscow, 125047, Russia*
[2] *National Research University «Higher School of Economics», S. Basmannaya str., 21/4 bld.1, Moscow, 105066, Russia*
[3] *MRSEI, Shkolnaya str. 55a bld. 1, Vidnoe, 142700, Russia*

Abstract
Experiments demonstrate that Bert embeddings are reflecting syntactic structure of a sentence. In this paper we conduct some experiments demonstrating dependency between cosine similarity of Bert embeddings and position of neighbouring words with consideration of a consistency structure of a sentence. Our experiments demonstrated that Bert embeddings are dependent on the number of started and finished consistencies between those words. The more the number of borders of consistencies, the less is the cosine similarity; finished borders take more influence than started ones. Increasing the distance between words up to 3 cosine similarity decreases. Moreover, we have found that Bert model assigns very close vectors which cosine similarity does not fall less than 0.6. The Word2Vec model, which takes not a context of a word into consideration, does not demonstrate such a behaviour.

Keywords
Bert model, word embeddings, natural text parsing, consistency.

ГрафиКон 2023: 33-я Международная конференция по компьютерной графике и машинному зрению, 19-21 сентября 2023 г., Институт проблем управления им. В.А. Трапезникова Российской академии наук, г. Москва, Россия
EMAIL: eklyshinsky@hse.ru (Э.С. Клышинский); varvaravasilyeva22@gmail.com (В.В. Васильева); parlak@mail.ru (О.В. Карпик); yulya.belobokova@mail.ru (Ю.А. Белобокова)
ORCID: 0000-0002-4020-488X (Э.С. Клышинский); 0009-0007-0814-7874 (В.В. Васильева); 0000-0002-0477-1502 (О.В. Карпик); 0009-0002-4169-7847 (Ю.А. Белобокова)

1. Введение

Одной из наиболее известных моделей, используемых для обработки естественного языка, является BERT (Bidirectional Encoder Representations from Transformers) [1]. BERT является предобученной нейронной сетью, которая может использоваться для широкого спектра задач, таких как вопросно-ответные системы, классификация текстов и многие другие. Преимущества Bert над другими векторными моделями состоит в том, что он не только разбивает слова на части и приписывает вектора именно им, но и приписывает слову контекстный вектор. Если слово обладает несколькими значениями, контекст может помочь определить конкретное значение употребления этого слова.

Многие работы по исследованию возможностей Bert посвящены семантике, в отличие от синтаксиса. В [2] исследовалась возможность использовать Bert для выделения параграфов и глав в тексте. Для этого авторы приводят сравнение векторных представлений различных предложений, путем расчета косинусных близостей между ними, на предмет поиска, так называемых, splitting points или переломных моментов в текстах. По результатам исследования модель неплохо справилась с задачей разбиения на абзацы, и векторные репрезентации переломных моментов достаточно четко отражались по метрике косинусной близости. В рамках другого исследования [3] проводился большой эксперимент, в рамках которого было выявлено, что Bert часто кодирует важную информацию в определенных токенах, что позволяет модели лучше различать синтаксические и семантические аномалии. Также рассматривалось, в векторных представлениях каких токенов предложения модель кодирует информацию о множественном числе объекта и времени смыслового глагола. Примером синтаксической аномалии выступали случаи инверсии. Тестирование показало, что коэффициент корреляция Спирмена вырос в первых нескольких слоях, что свидетельствует о повышенной чувствительности модели к измененному порядку слов.

Наши наблюдения показали, что косинусная мера сходства для векторов слов, находящихся рядом, является экстремально большой. При этом по мере удаления от слова косинусная мера несколько уменьшается. В связи с тем, что Bert учитывает информацию о позиции слова в предложении и, определяя взаимодействия между словами в предложении, создает вектора контекста для каждого слова, можно предположить, что Bert может в своих векторных представлениях отражать и синтаксическую структуру предложения, а именно принадлежность слова к синтаксической группе. Если это так, то можно будет рассматривать Bert как инструмент для выделения составляющих. [4]

2. Метод исследования результатов, возвращаемых моделью Bert

Прежде, чем изложить собственно метод оценки близости векторного представления слов, введем несколько определений. Под линейным расстоянием между словами будем понимать разницу между их позициями в предложении. Заметим, что при нашем анализе мы убирали все знаки препинания после синтаксического анализа, поэтому линейное расстояние между ними считалось без них.

Составляющие определялись из результата разбора с использованием деревьев зависимости. Расстоянием между словами в синтаксическом дереве мы будем называть количество границ составляющих, которые надо пересечь пройдя по пути, который соединяет два слова в дереве зависимостей. При этом мы не считаем терминальные вершины, полученные после удаления знаков препинания, за самостоятельные составляющие. Начало составляющей и ее завершение являются самостоятельными границами. Если между двумя словами находится полная вложенная составляющая, то она не засчитывается.

Разберем пример определения расстояния между словами в дереве. Пусть дано предложение «*Это связано с тем, что работа каких-то инструкций алгоритма может быть зависима от других инструкций или результатов их работы.*» После разделения на составляющие получим для него следующую скобочную запись, отражающую границы составляющих: «[*Это связано* [[*с тем*], [*что* [*работа* [*каких-то инструкций алгоритма*]] *может* [*быть зависима* [*от других*

инструкций или [*результатов* [*их работы*]]]]]]].» Дерево зависимостей для предложения приведено на рисунке 1. Линейное расстояние между словами «*алгоритма*» и «*от*» равно 4. Расстояние по дереву также равно 4: завершаются две составляющие (переходы вверх по дереву) и начинаются еще две составляющие (переходы вниз по дереву. При этом переход от листовой вершины или к ней не учитывается.

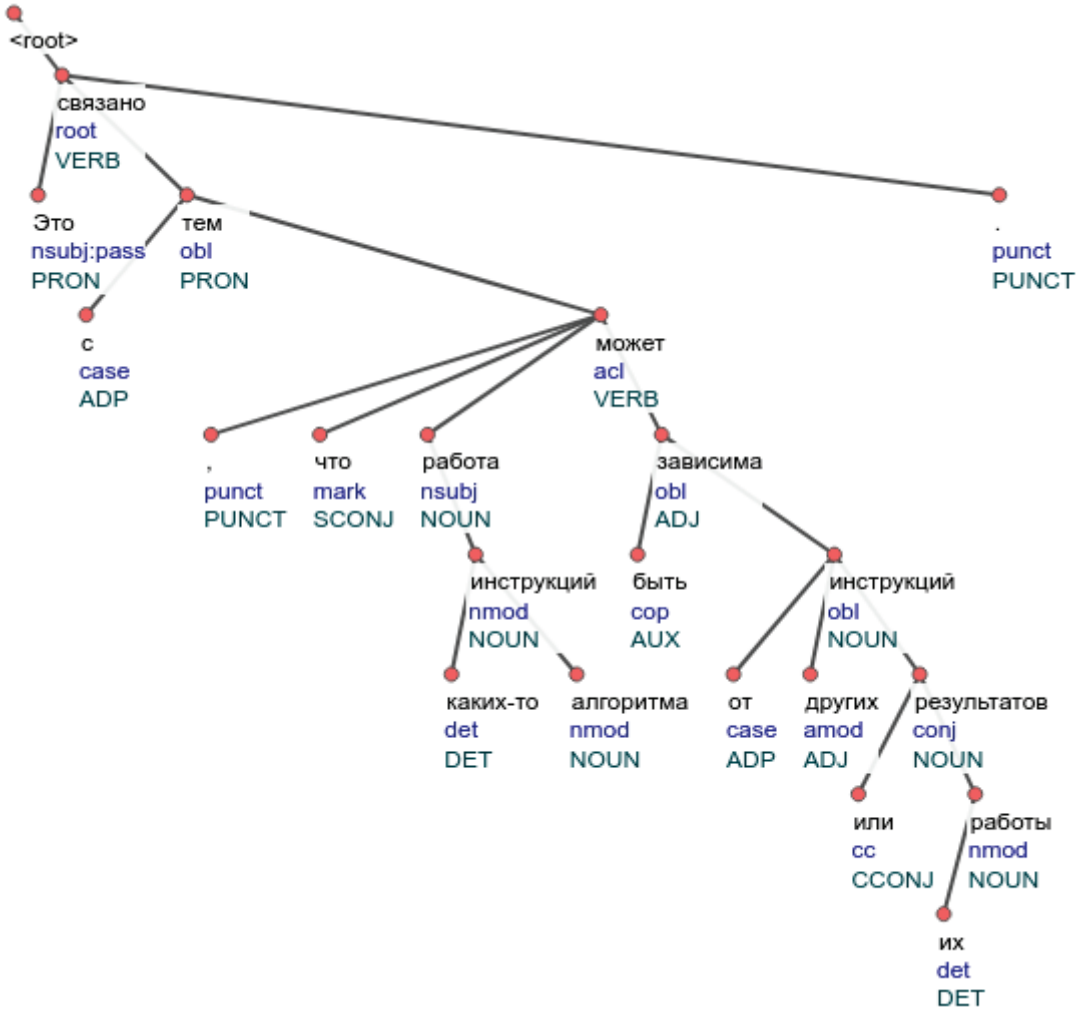

Рисунок 1 – Дерево зависимостей для предложения «[*Это связано* [[*с тем*], [*что* [*работа* [*каких-то инструкций алгоритма*]] *может* [*быть зависима* [*от других инструкций или* [*результатов* [*их работы*]]]]]]]»

Теперь опишем метод проверки гипотезы о том, что косинусное расстояние между векторными представлениями слов Bert зависит от линейного расстояния и расстояния по дереву между словами.

Входом является синтаксическая разметка предложения, например, в формате CONLLU. Предложение в таком формате преобразуется в скобочную запись. Слова в скобочной записи могут быть заменены на их позиции в предложении. Знаки препинания при этом удаляются, позиции считаются после удаления знаков препинания. Полученная запись предложения позволяет проверить, является ли оно проективным; непроективные предложения пропускались полностью.

Мы пользовались определением проективности, сформулированным А. В. Гладким, смысл которого заключается в том, что «дерево подчинения называется проективным, если для любых трех узлов α, β, γ из того, что β зависит от α и γ лежит между α и β следует, что γ зависит от α» [5: 19]. Именно проективные предложения, в отличие от непроективных и слабо-проективных, можно представить в линейной скобочной записи, при этом не нарушив порядок слов в

предложении. Непроективность предложения можно показать на примере дерева зависимостей, если строить его над анализируемым предложением, располагая точки на горизонтальной оси в линейной последовательности узлов. В нашем случае непроективность дерева будет означать, что в скобочной записи предложения будет наблюдаться разрыв в нумерации слов.

Далее слова векторизовались. Известно, что токенизация Bert разделяет длинные слова на части - сабтокены. Сабтокен маркируется определенным префиксом — ##. Например, слово *моет* токенизатор разобьет следующим образом: *мое* и *##т*. Деление на сабтокены не регулярное, и, как уже было видно, не обязательно соответствует выделению морфем, так, для слова *пьет* разбиение будут таким: *пь* и *##ет*, а для слова *готовит* токенизатор не выделит сабтокены. Такая особенность токенизатора осложняла задачу по обращению к определенному слову по индексу, поскольку одному слову мог соответствовать больше, чем один вектор. Чтобы учитывать этот факт, мы рассчитывали среднее значение векторов, полученных для сабтокенов одного слова.

Следующий этап работы был связан с представлением распределения косинусных близостей. Для ее визуализации использовался прямой обход дерева зависимостей, при прохождении дерева от первого до последнего слова в интервале расставлялись открывающие и закрывающие скобки, встретившиеся на пути. Подобная строка использовалась как графический ключ для обозначения разницы глубин слов в дереве. Так, в примере на Рис. 1 паре *алгоритм* и *от* приписывалась метка]][[. Для слов, находящихся внутри одной составляющей была введена граница типа 0, символизирующая отсутствие границы между словами. Для слов на стыке составляющих собрались различные виды границ.

В итоге была собрана статистика данных о значениях метрики косинусной метрики сходства в зависимости от типа границы. Визуализация и анализ полученных данных показаны в следующем разделе.

3. Результаты экспериментов

Эксперименты проводились на корпусе SynTagRus в разметке Universal Dependencies (https://universaldependencies.org/). Данный корпус размечен вручную и содержит почти 86 000 предложений (около 1,5 млн слов). На первом этапе по собранному материалу были построены диаграммы размаха для линейного расстояния между словами от 1 до 4 (см. рисунки 2-5).

Сортировка отображения границ при их визуализации показала, что косинусное расстояние между векторами соседних слов резко уменьшается при наличии границ составляющих. Причем значительным является не столько изменение среднего значения, сколько нижняя граница — от 0,9 до 0,75. Для слов на расстоянии от 2 этот эффект заметен уже не так сильно. Также видна корреляция между расстоянием по дереву и средним значением косинусной меры. Однако, расчеты показали, что зависимость наблюдается слабо.

После этого было принято решение проанализировать количество закончившихся и начавшихся составляющих по отдельности. Для этого был использован другой способ визуализации результатов: по оси абсцисс откладывалось количество начавшихся составляющих, по оси ординат — завершившихся. Результаты показаны на рисунках 6-9.

Из рисунков становится очевидно, что косинусное расстояние падает с ростом количества начавшихся или закончившихся составляющих, то есть расстояния, пройденного по дереву вверх или вниз. При этом, количество завершившихся составляющих влияет несколько сильнее, чем количество открывшихся. Аналогичная ситуация наблюдается и для минимального и максимального значений косинуса: с удалением от нулевого расстояния по дереву и увеличением линейного расстояния минимальный косинус с некоторыми колебаниями растет. Максимальное значение практически не изменяется с изменением линейного расстояния, но немного падает с ростом расстояния по дереву (колебания в районе 0,01). Заметим, что среднее значение косинуса для трех произвольных пар слов в каждом предложении примерно равно 0,9586.

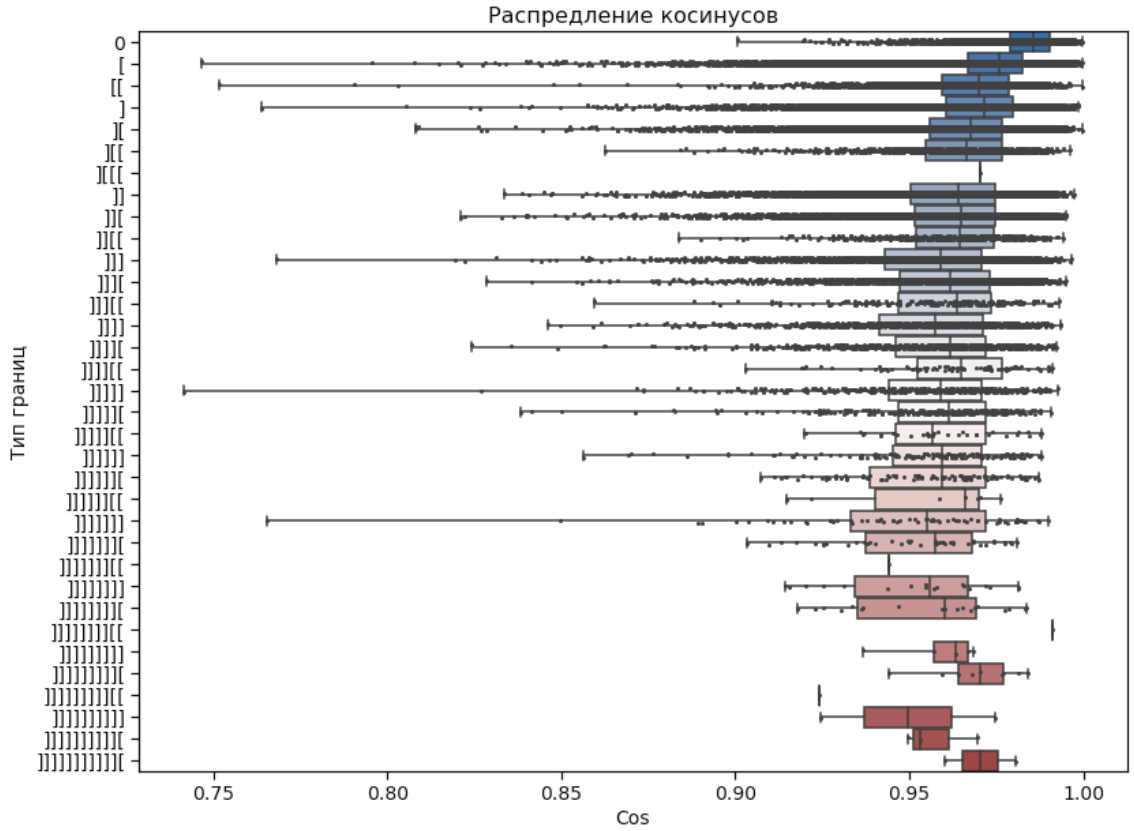

Рисунок 2 – Диаграмма размаха для значений косинусной меры сходства для слов на линейном расстоянии 1 в зависимости от пересечения границ клауз

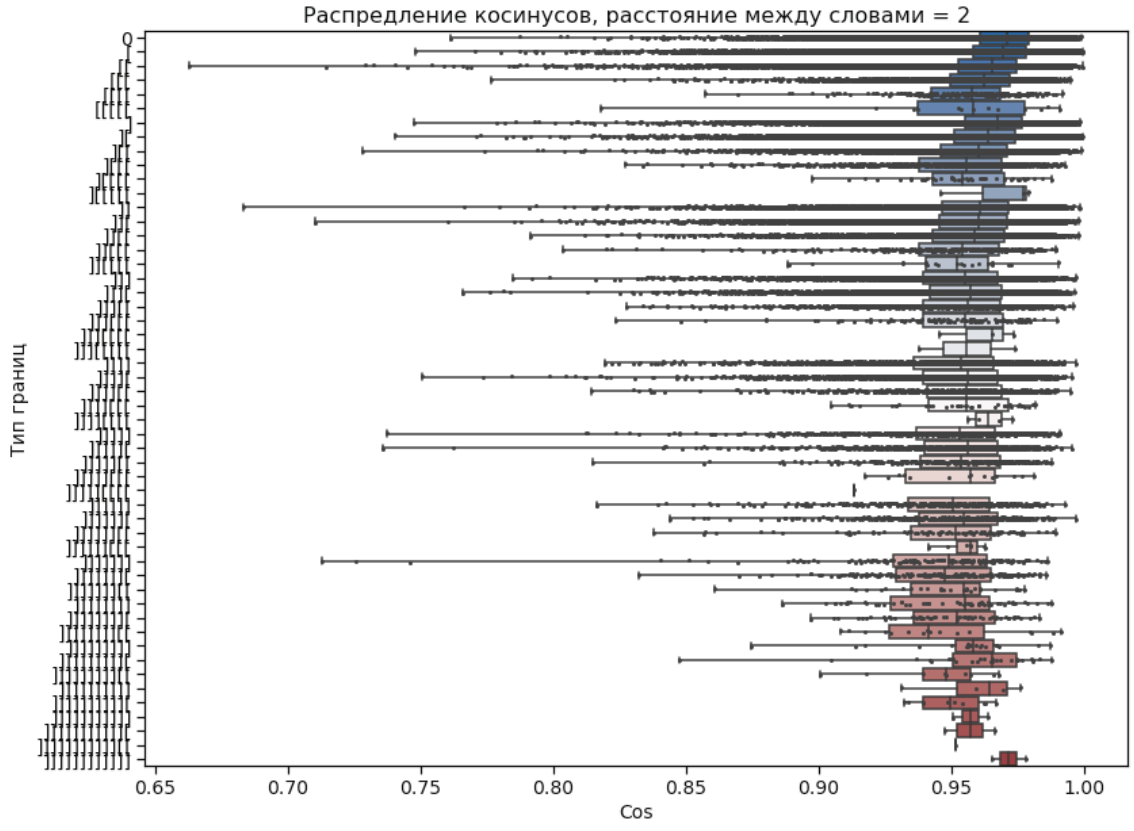

Рисунок 3 – Диаграмма размаха для значений косинусной меры сходства для слов на линейном расстоянии 2 в зависимости от пересечения границ клауз

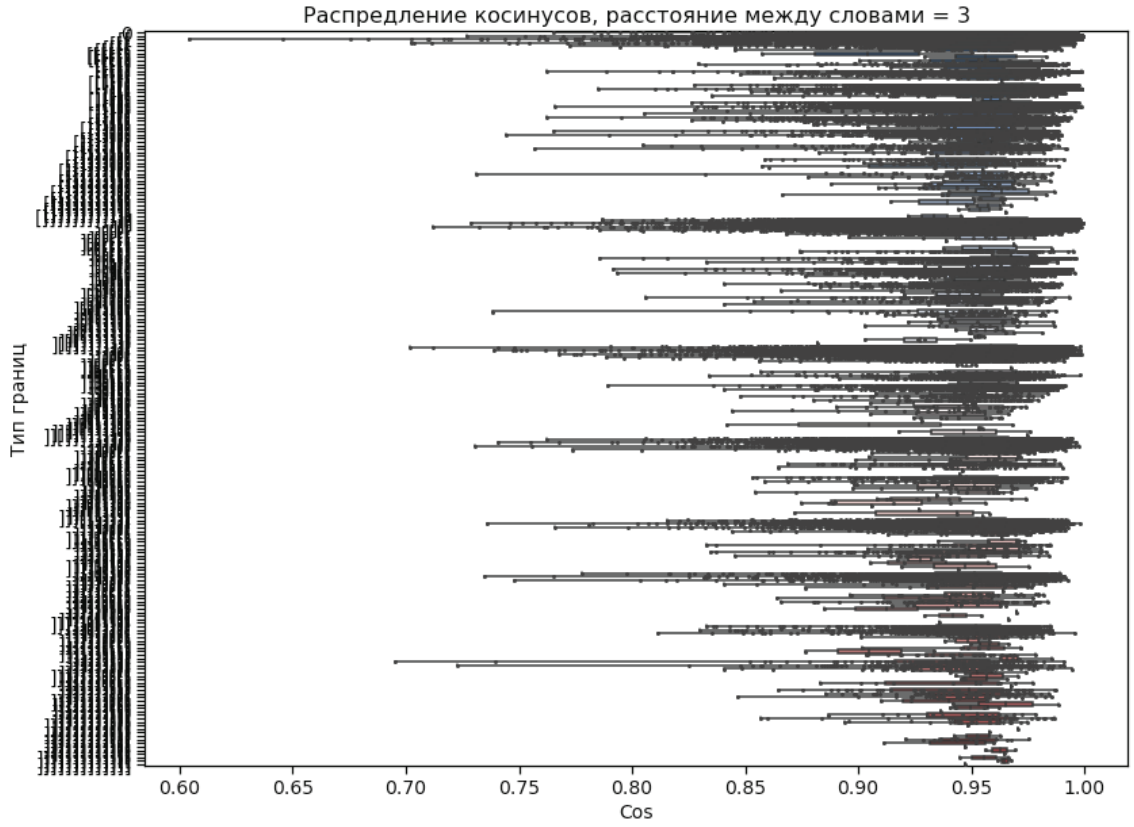

Рисунок 4 – Диаграмма размаха для значений косинусной меры сходства для слов на линейном расстоянии 3 в зависимости от пересечения границ клауз

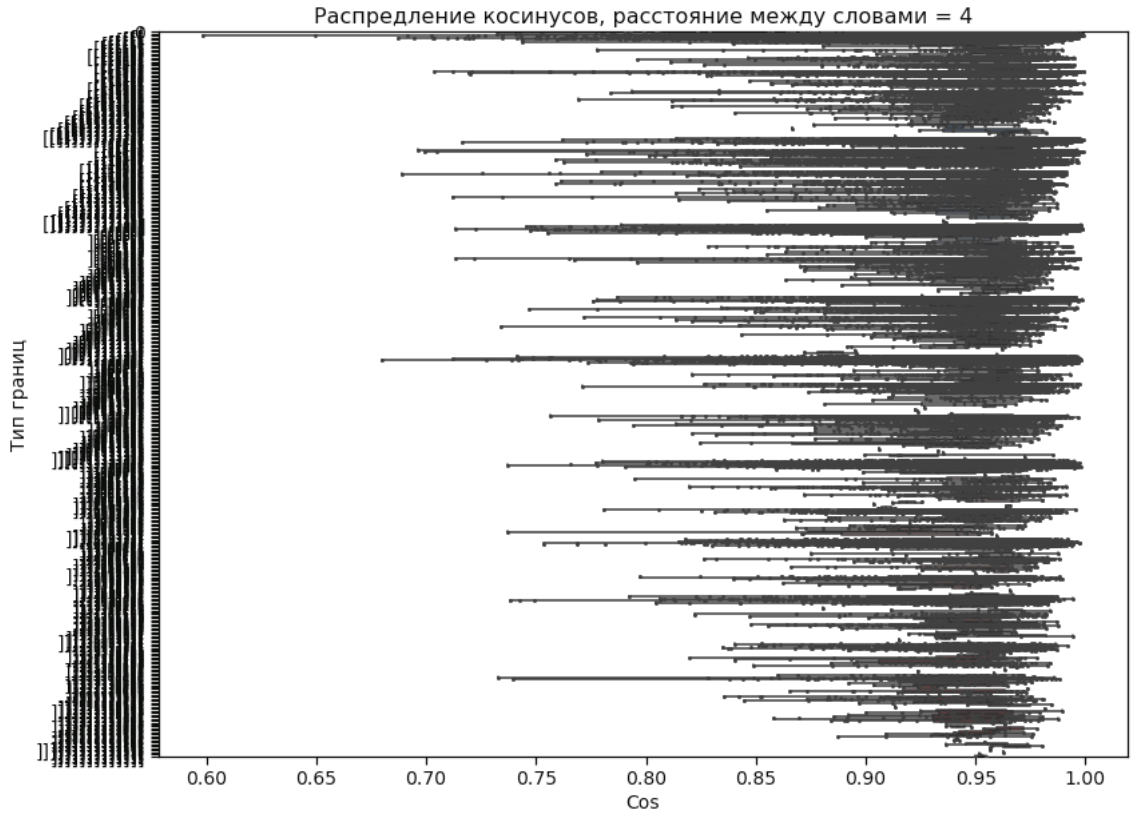

Рисунок 5 – Диаграмма размаха для значений косинусной меры сходства для слов на линейном расстоянии 4 в зависимости от пересечения границ клауз

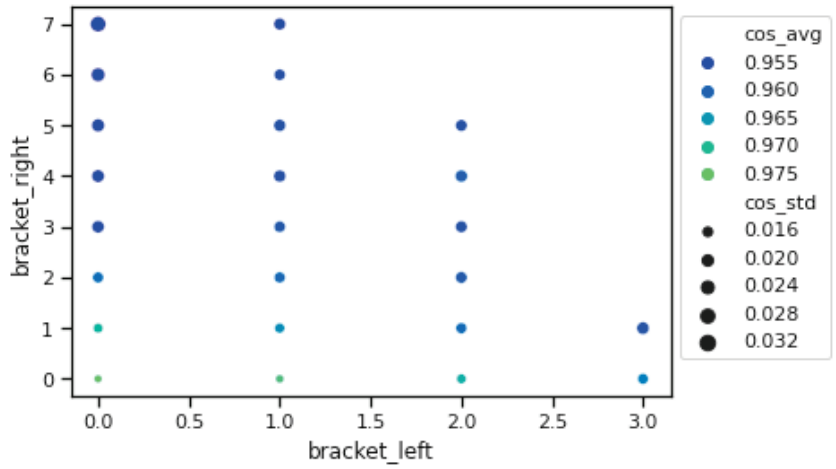

Рисунок 6 – Зависимость среднего косинусного расстояния и его дисперсии в зависимости от количества переходов по дереву зависимостей, линейное расстояние между словами = 1

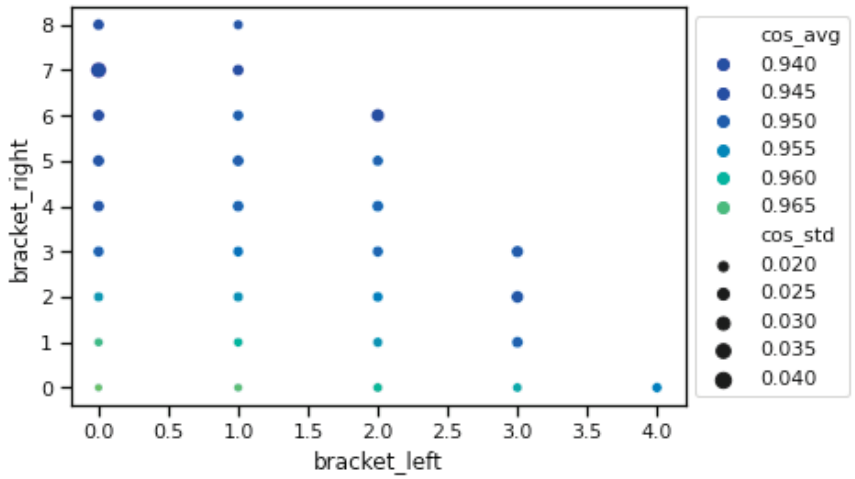

Рисунок 7 – Зависимость среднего косинусного расстояния и его дисперсии в зависимости от количества переходов по дереву зависимостей, линейное расстояние между словами = 2

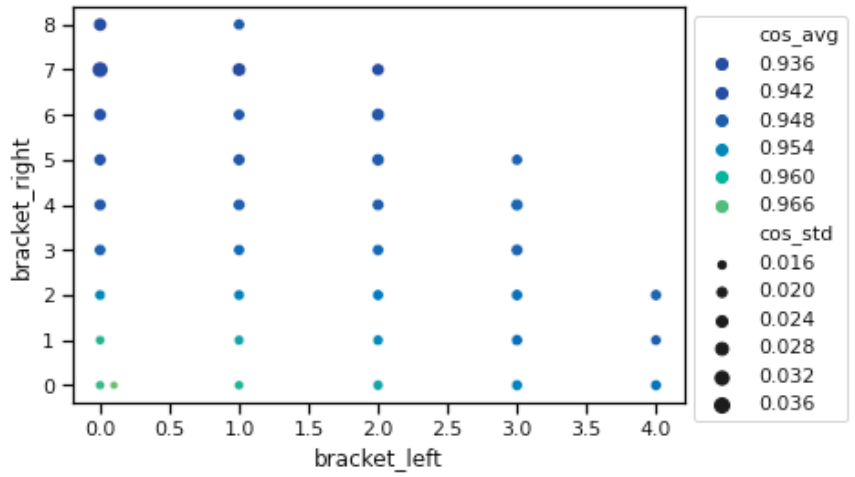

Рисунок 8 – Зависимость среднего косинусного расстояния и его дисперсии в зависимости от количества переходов по дереву зависимостей, линейное расстояние между словами = 3

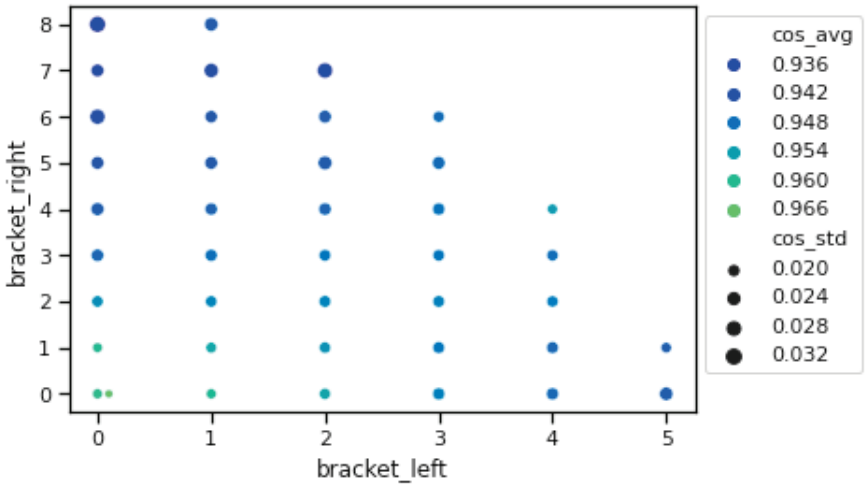

Рисунок 9 – Зависимость среднего косинусного расстояния и его дисперсии в зависимости от количества переходов по дереву зависимостей, линейное расстояние между словами = 4

Поведение косинусного сходства для Bert кардинально отличается от аналогичного для Word2Vec [6]. Так, среднее значение косинусной меры колеблется между 0,15 и 0,25, причем оно растет по мере увеличения количества начавшихся составляющих и уменьшается с количеством завершившихся при линейном расстоянии больше 2. Получается, что в случае Word2Vec слова, непосредственно связанные с данным, относятся к немного другой области (что логично, так как они описывают характеристики и связи данного слова). С ростом расстояния по дереву при спуске вниз вновь начинают появляться слова из той же предметной области. Окончание составляющей означает смену описания одной области и начало следующей.

4. Заключение

В данной статье мы провели анализ векторного представления Bert для слов, находящихся в предложении рядом. Было выяснено, что для произвольной пары слов в одном предложении косинусное сходство не падает ниже 0,6 при среднем 0,9586. Для соседних слов среднее значение колеблется от 0,955 до 0,975, чем больше синтаксических составляющих заканчивается между этими словами, тем ниже значение косинусного сходства. С ростом расстояния между словами косинусное сходство падает, причем наблюдается зависимость от количества закончившихся (в большей мере) и начавшихся (в меньшей мере) синтаксических составляющих. Это особенно хорошо видно при визуализации данных в виде матрицы, в которой по строкам идет количество переходов на один уровень выше по дереву зависимостей, а по столбцам — количество переходов по дереву зависимостей вниз. Поведение косинусного сходства для векторов Word2Vec ведет себя совершенно иначе, показывая рост сходства слов по мере увеличения линейного расстояния между словами от 2 до 4.

Таким образом, из результатов наших экспериментов следует, что модель Bert расставляет векторы словам исходя из переданного ей контекста, причем близким словам присваиваются более близкие векторы. Степень близости векторов коррелирует с границами синтаксических составляющих, пролегающих между выбранными словами.

Подобное свойство можно использовать для построения классификаторов границ составляющих, предложений или абзацев, но точность таких классификаторов не будет слишком высокой, так как интервалы, в которых принимают значение косинусная мера сходства, в значительной мере пересекаются. Низкое значение косинусной меры может использоваться как универсальный признак границы составляющих.

Заметим, что в отличие от Word2Vec, модель Bert присваивает векторы всем словам или токенам в составе слов, в том числе, служебным частям речи: предлогам, союзам и т. д. Таким образом, Bert векторизует слова, определяющие синтаксические связи между словами и типы

этих связей. Векторы, получаемые для слов, некоторым образом оптимизируются с тем, чтобы они оказали рядом. За счет этого, Bert относит к некоторой области в векторном пространстве не отдельные слова, а группы стоящих рядом слов. Так, для предложений «*Мама мыла раму*» и «*папа мыл раму*» векторные представления для двух вхождений слов «*рама*» будут значительно отличаться (косинусное сходство ниже 0.8). А в паре «*Папа мыл раму*» и «*Папа пил виски*» аналогичная ситуация наблюдается для слова «*папа*».

Следовательно, векторизация Bert обращает больше внимания на область всего текста, чем на значения отдельных слов. Исходя из этого можно сказать, что модель Bert заслуживает названия контекстуализированной не за то, что она подбирает вектор для слова исходя из контекста, а за то, что она подгоняет все вектора слов к некоторой единой тематике выданного ей текста. Однако, эксперименты с нейронными сетями демонстрируют, что в этой новой области векторного пространства, соответствующей тематике текста, Bert вполне успешно различает значения отдельных слов.

5. Список источников

[1] BERT: Pre-training of deep bidirectional transformers for language understanding. / Devlin J., Chang M.-W., Lee K., Toutanova K. // In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019. P. 4171–4186.

[2] Pethe C., Kim A., Skiena S. Chapter Captor: Text Segmentation in Novels // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2019. P. 8373–8383.

[3] Mohebbi H., Modarressi A., Pilehvar M.T. Exploring the Role of BERT Token Representations to Explain Sentence Probing Results. // In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021. P. 792–806.

[4] Тестелец Я. Г. Введение в общий синтаксис. М.: РГГУ, 2001. 800 с.

[5] Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. М.: Наука, 1985. 144 с.

[6] Distributed Representations of Words and Phrases and their Compositionality / Mikolov T., Chen K., Corrado G., Dean J. // NIPS, 1–9 Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems, 2013. P. 3111-3119