 © Авторы статьи / Authors of the article

МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ, КОМПЛЕКСОВ И КОМПЬЮТЕРНЫХ СЕТЕЙ/MATHEMATICAL SOFTWARE FOR COMPUTERS, COMPLEXES AND COMPUTER NETWORKS

DOI: https://doi.org/10.60797/IRJ.2026.166.52 EDN: LLOQOW

ПРИМЕНЕНИЕ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ BERT И ELMO В ЗАДАЧЕ ЛЕКСИЧЕСКОЙ ТИПОЛОГИИ

Научная статья

Полозов И.К.[1, *], Волкова И.А.[2]
[1] ORCID : 0000-0003-2679-5465;
[1, 2] Московский государственный университет, Москва, Российская Федерация

* Корреспондирующий автор (ilya-polozov[at]mail.ru)

Аннотация

Работа посвящена использованию векторных представлений ELMo и BERT в задаче определения лексической типологии языков. Также сравнивается способ кластеризация с заданным количеством кластеров K-means и способ кластеризации с автоматическим количеством кластером DBSCAN. Описана задача лексической типологии. Сделан обзор существующих подходов к решению задачи, описаны достоинства и недостатки. Проведены эксперименты для семантической зоны «тянуть-толкать». Выявлено, что на результат влияют как тип векторных представлений, так и способ кластеризации. Способ кластеризации влияет больше, чем векторные представления. Влияние распространяется на точность, полноту и F-меру. BERT показывает результаты лучше, чем ELMo, а K-means лучше, чем DBSCAN. Произведены оценки подходов, найдены оптимальные параметры и сделаны выводы.

Ключевые слова: лексическая типология, ELMo, BERT, K-means, DBSCAN, кластеризация текстов, компьютерная лингвистика.

APPLICATION OF BERT AND ELMO VECTOR REPRESENTATIONS IN LEXICAL TYPOLOGY PROBLEM

Research article

Polozov I.K.[1, *], Volkova I.A.[2]
[1] ORCID : 0000-0003-2679-5465;
[1, 2] Lomonosov Moscow State University, Moscow, Russian Federation

* Corresponding author (ilya-polozov[at]mail.ru)

Abstract

The work is devoted to the use of ELMo and BERT vector representations in the task of determining the lexical typology of languages. It also compares the K-means clustering method with a given number of clusters and the DBSCAN clustering method with an automatic number of clusters. The problem of lexical typology is described. An overview of existing approaches to solving the task is provided, and their pros and cons are described. Experiments were conducted for the semantic zone «pull-push«. It was found that both the type of vector representations and the clustering method affect the result. The clustering method has a greater impact than vector representations. The influence extends to accuracy, completeness, and F-measure. BERT shows better results than ELMo, and K-means is better than DBSCAN. The approaches are evaluated, optimal parameters are found, and conclusions are drawn.

Keywords: lexical typology, ELMo, BERT, K-means, DBSCAN, text clustering, computer linguistics.

Введение

Задача определения лексической типологии языков является актуальной и недостаточно автоматизированной. Данная статья описывает использование векторных представлений ELMo и BERT для помощи лингвистам. Для оценки выбрано семантическое поле «тянуть — толкать». Оно уже исследовано в работе [1], поэтому с ней будет произведено сравнение автоматизированного подхода.

Лексическая типология изучает способы языкового выражения конкретных явлений и сопоставляет соответствующие лексические средства в разных языках. Так, в русском языке один и тот же термин используется для обозначения пальцев руки и ноги, тогда как в английском языке эти понятия различаются лексемами «finger» и «toe». Наряду с межъязыковым сравнением, возможно исследование семантических полей внутри одного языка и анализ средств их выражения. Например, в семантической зоне «чинить — портить» можно выделить такие варианты употребления, как «делать вновь пригодным», «настраивать инструмент», «изменять деятельность», «ухудшать». Подобные типы употреблений представляют собой фреймы [2].

На основе выделенных фреймов может быть построена таблица, в которой строки соответствуют фреймам, столбцы — языкам, а в ячейках указываются лексические единицы, с помощью которых данные фреймы реализуются в разных языках. Статья направлена на автоматизированное формирование подобных фреймов.

Обзор литературы

2.1. Ручные методы

Выделяют четыре основных подхода к исследованию. Первый опирается на использование фреймов [2] и известен как метод Московской лексико-типологической школы. В рамках этого подхода каждая ситуация, входящая в

 © Авторы статьи / Authors of the article

определенное семантическое поле, описывается с помощью фрейма — набора характеристик, выраженных словами. Так, например, фрейм «нажимать предмет вперёд» относится к семантическому полю «тянуть — толкать». В языке фрейм реализуется через конкретные лексемы, при этом количество возможных фреймов может быть значительным. Их отбор осуществляется исследователем на основе словарей, переводных источников и собственных интуитивных представлений. Также возможно использование синхронных переводов текстов.

На следующем этапе составляется таблица: в строках приводятся описания фреймов, в столбцах — лексемы, а в ячейках отмечается, соотносится ли конкретная лексема с данным фреймом. Существует и другой вариант таблицы, где строки соответствуют фреймам, столбцы — языкам, а в ячейках указываются лексемы, с помощью которых в каждом языке выражается данный фрейм. Основным недостатком этого подхода является ручной характер выделения фреймов. Кроме того, при работе со словарями исследователю приходится определять пределы поиска, поскольку в процессе перевода возникают новые фреймы, часто лишь косвенно связанные с изучаемым семантическим полем из-за многозначности слов. В связи с этим также требуется привлечение носителей соответствующих языков.

Второй подход опирается на физическое восприятие человека [3]. Исследователь формирует набор универсальных стимулов, например, объекты с определенным вкусом, запахом, цветом, формой и предъявляет их носителям языка. Задача информанта заключается в максимально точном словесном описании предложенного объекта. Сравнивая ответы носителей разных языков, можно установить, какими лексическими средствами выражаются одни и те же стимулы в разных языках. К недостаткам данного метода относятся невозможность воспроизвести всё разнообразие лексических единиц в виде физических стимулов, а также высокая трудоёмкость и значительные временные затраты. Кроме того, данный подход требует обязательного участия носителей языка.

Третий подход базируется на использовании универсальных семантических примитивов, с помощью которых предполагается возможным описать любую ситуацию [4]. В рамках этого метода применяется система из 64 базовых понятий, из комбинаций которых выводятся все остальные значения. Основными недостатками подхода являются неоднозначность интерпретации выводимых значений и общая методологическая сложность.

Четвёртый подход связан с анализом параллельных корпусов текстов. Исследователь выявляет переводные соответствия для различных реализаций определённой семантической зоны. Существенным ограничением данного метода является отсутствие или недостаточная представленность параллельных корпусов для редких и малораспространённых языков.

Наиболее распространённым в лингвистических исследованиях является фреймовый подход. Так, в работе [5] с его помощью анализируется семантическое поле «мешать». В исследовании [6] рассматривается семантическое поле «домашний скот» на материале германских и славянских языков, основным источником данных служит лексический фонд. В работе [7] при изучении семантических зон «попасть, упасть» и «задеть, попасть» в казымском диалекте хантыйского языка используются корпусные данные, словари и материалы, полученные от носителей языка.

Исследование [8] посвящено семантической зоне «острый» в китайском языке и опирается на данные словарей, текстовых корпусов и материалы, полученные от информантов. Автор работы [9] также обращается к семантическому полю «шахматная игра» в русском языке, применяя модель «центр — периферия», где в центре располагаются наиболее узкоспециализированные семантические признаки, а на периферии — менее специализированные.

В работе [10] для анализа семантической зоны «мягкий», «твёрдый», «жёсткий» используется метод анкетирования носителей языка. В главе «Methodology at work: Semantic fields sharp and blunt» книги [11] описывается семантическая зона «острый — тупой». Показано, что ключевыми параметрами оппозиции выступают тип острого объекта и ощущение, на основе которого определяется степень его остроты. В работе [12] семантическая зона слова «город» исследуется на материале литературных источников.

2.2. Автоматизированные методы

Автоматизация типологических исследований на сегодняшний день остаётся недостаточно хорошо развитой. В работе [13] используются заранее подготовленные анкеты, которые затем автоматически переводятся на другие языки с опорой на словари и параллельные корпуса. Исследование посвящено семантическим зонам «острый — гладкий» и «толстый — тонкий». Основным ограничением данного подхода является потребность в предварительной разработке таких анкет.

В работе [14] в качестве материала используются биграммы Национального корпуса русского языка [15], дополненные различными леммами. Для проведения кластеризации формируются векторные представления: отбираются 10 000 наиболее частотных лексем, после чего для исследуемого слова подсчитывается количество совместных употреблений каждой из этих лексем в окне шириной пять слов. Для анализа применяются алгоритмы иерархической кластеризации, поскольку методы, не требующие заранее заданного числа кластеров, продемонстрировали низкую эффективность. Недостатком данного подхода является отсутствие в векторах семантической информации, а также информации о контексте употребления.

Векторные представления

3.1. ELMo

ELMo являются векторными представлениями, обладающих тем свойством, что для одинаковых слов в разных контекстах будут сгенерированы разные векторы [16]. Например, для слова ключ в следующий предложениях будут сгенерированы разные векторы: «на доске нарисован скрипичный ключ», «в реке бил ключ», «ключ находился в замке». При этом в первом предложении вектор будет ближе к векторам слов, связанных с музыкой, во втором к словам, связанным с реками, а в третьем с замками.

Для получения векторов используются две нейронные сети LSTM (Long Short-Term Memory). Одна читает предложение слева направо и предсказывает следующее слово, а вторая справа налево и предсказывает предыдущее слово. Таким образом, получается учитывать контекст сразу с обеих сторон. Затем внутренние состояния двух сетей

 © Авторы статьи / Authors of the article

суммируются и получается контекстуализированное векторное представление. При этом слова в начале и в конце предложения будут иметь разные векторы.

Однако эволюционно векторы ELMo появились первыми, и сегодня существуют представления, которые лучше захватывают дальний контекст и обучаются эффективнее. Одним из таких представлений являются векторы BERT (Bidirectional Encoder Representations from Transformers).

3.2. BERT

Векторы BERT, как и ELMo, являются контекстуализированными векторными представлениями. Однако контекст учитывается слева и справа не последовательно, как в ELMo, а одновременно [17]. BERT состоит из одинаковых Transformer Encoder-блоков из 12 или 24 слоев. Используются только Encoder блоки. Слова разделяются на подслова. Например, слово «невероятный» может быть разделено на токены «не» и «##вероятный». Это дает возможность обрабатывать редкие слова. В каждое предложение вставляется 2 специальных токена: [CLS] — представляет все предложение целиком, [SEP] — разделяет предложения. При обучении используется механизм внимания. Каждый токен представлен комбинацией 3-х векторных представлений: векторное представление самого слова, представление позиции в предложении и представление принадлежности конкретному предложению.

Для обучения случайно скрываются 15% токенов. Они заменяются на специальный токен [MASK]. BERT должен предсказать это скрытое слово на основе контекста. Например, «Он сидел на [MASK] реки». Вместо слова [MASK] BERT должен научиться предсказывать слово «берегу».

Также его можно дополнительно обучать на задаче предсказания, являются ли 2 предложения последовательными. Например,

1. «Он пошел на рыбалку».
2. «Он поймал много рыбы».
3. «В огороде растет много вкусных овощей».

BERT должен научиться определять, что 2-е предложение с большей вероятностью стоит после 1-го, чем 3-е. Затем BERT дообучается на специфичной для задачи области.

Для лексической типологии могут быть полезны оба представления, поэтому будет произведена оценка их работы и сравнение для данной задачи.

Алгоритм работы

4.1. Векторные представления

Исследование проводится на материале текстов Национального корпуса русского языка. Все слова переводятся в начальную форму, после чего из текста удаляются стоп-слова и специальные символы. Затем в корпусе отбираются все предложения, содержащие слова «тянуть» и «толкать». В файл сохраняется предложение, содержащее найденное слово, все дополнительно обнаруженные слова, номер предложения, а также его исходный вариант.

Затем для каждого слова в каждом предложении вычисляются векторные представления с помощью моделей BERT и ELMo. Они обучены на русской Википедии и новостных корпусах. Лексика схожа с лексикой из корпуса постсоветских текстов Национального корпуса русского языка, поэтому модель должна хорошо работать на данном корпусе. Предложения также сохраняются в структуре. Для сохранения результатов используется библиотека Pickle.

Среди найденных предложений много слов в одинаковой семантической зоне, поэтому необходима фильтрация.

4.2. Фильтрация

Предложения, векторы которых сильно отличаются, считаются новыми фреймами исследуемого поля. Для удаления похожих предложений используется алгоритм K-means [18] и алгоритм кластеризации с автоматическим количеством кластеров DBSCAN [19].

Для выбора количества кластеров в методе в K-means используется метод локтя. Выбирается диапазон значений количества кластеров, например, от 1 до 20. Для каждого количества кластеров запускается работа алгоритма. Затем идет подсчет суммы квадратов внутрикластерных расстояний (Within Cluster Sum of Squares — WCSS) по формуле 1.

$$WCSS(k) = \sum_{i=1}^{k} \sum_{x \in C_i} \left\| x - \mu_i \right\|^2 \quad (1)$$

C_i — i-й кластер

μ_i — центроид кластера

$||x- \mu i||^2$ — квадрат евклидова расстояния

Строится график, на оси X которого обозначается количество кластеров, а на оси Y сумма квадратов внутрикластерных расстояний. График всегда убывает. Он показывает, как уменьшается ошибка кластеризации при увеличении количества кластеров. На нем необходимо найти излом, после которого убывание становится не так значительно по сравнению с предыдущими значениями. Т.е. необходимо найти количество кластеров, после которого улучшения замедляются более медленно, чем до этого. Это количество и будет выбранным k в алгоритме K-means.

Для алгоритма DBSCAN основными параметрами являются:

1) eps — максимальное расстояние между двумя точками, чтобы считать их соседями;

2) min_samples — минимальное количество точек, чтобы сформировать кластер.

min_samples для двумерных данных обычно берется в диапазоне 3-5 и увеличивается для зашумленных данных.

4.3. Псевдокод

Псевдокод представлен в листинге.

```
Вход: корпус НКРЯ
Выход: фреймы
```

 © Авторы статьи / Authors of the article

```
sents ← corpus[тянуть|толкать]
embeddingsBERT ← BERT(sents)
embeddingsELMo ← ELMo(sents)
centroidsBERTKMeans ← KMeans(embeddingsBERT)
centroidsELMoKMeans ← KMeans(embeddingsELMo)
centroidsBERTDBSCAN ← DBSCAN(embeddingsBERT)
centroidsELMoDBSCAN ← DBSCAN(embeddingsELMo)

func FindNearest(centroids, embeddings:
   frames ← []
   FOR centroid in centroids:
      MaxSim ← 0
      MaxSimFrame ← 0
      For emd in embeddings:
         simCetroidSent ← cos(centroid, emd)
         frame ← corpus[number(emd)]
         IF simCetroidSent > MaxSim and NOT(frame in frames):
            MaxSim ← CetroidSent
            MaxSimFrame ← frame
      frames += MaxSimFrame

RETURN frames

framesBERTKMeans ← FindNearest(centroidsBERTKMeans, embeddingsBERT)
framesELMoLMeans ← FindNearest(centroidsELMoKMeans, embeddingsELMo)
framesBERTDBSCAN ← FindNearest(centroidsBERTDBSCAN, embeddingsBERT)
framesELMoDBSCAN ← FindNearest(centroidsELMoDBSCAN, embeddingsELMo)

RETURN (framesBERTKMeans, framesELMoLMeans, framesBERTDBSCAN, framesELMoDBSCAN)
```

Эксперименты

5.1. BERT с K-means

Сначала была исследована зона «толкать». Полученные векторы BERT были кластеризованны методом K-means. В этом методе необходимо выбрать количество кластеров. Для этого был использован метод локтя.

Построим график суммы квадратов внутрикластерных расстояний для большого значения k, например, 360. График представлен на Рисунке 1.

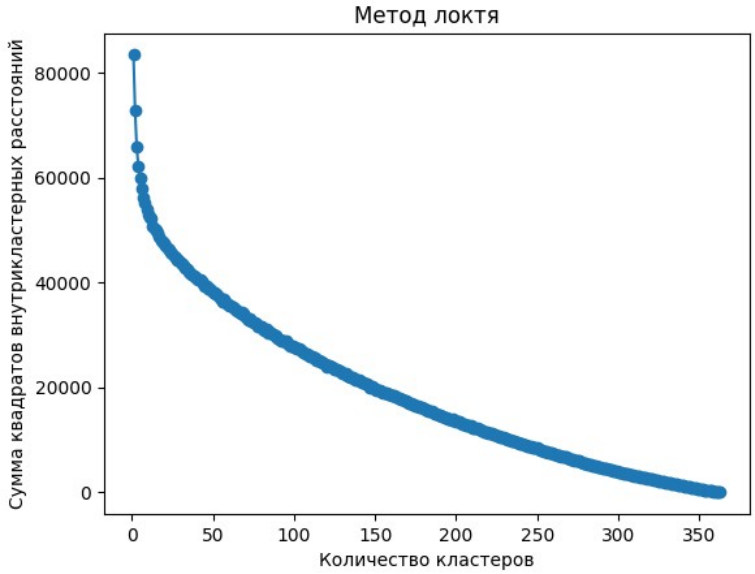

Рисунок 1 - Метод локтя для K-means с векторами BERT зоны «толкать»

DOI: https://doi.org/10.60797/IRJ.2026.166.52.1

Примечание: количество кластеров - 360

Излом в начальной части графика. Построим более точный график для значения k от 1 до 20 на Рисунке 2.

 © Авторы статьи / Authors of the article

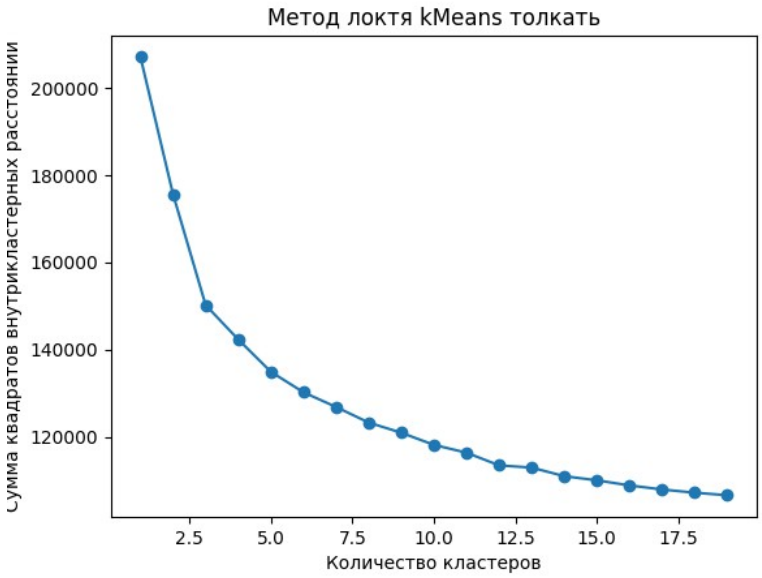

Рисунок 2 - Метод локтя для K-means с векторами BERT зоны «толкать»
DOI: https://doi.org/10.60797/IRJ.2026.166.52.2

Примечание: количество кластеров - 20

По методу локтя было выбрано 12 кластеров, т.к. точка, обозначающая 13 кластеров лежит почти на такой же высоте, как точка 12-и кластеров. При этом все предыдущие изменения (от 1 до 2, от 2 до 3 и т.д.) были значительно больше. Это говорит о том, что среднеквадратичные расстояния стали меняться меньше, чем до этого, и по методу локтя необходимо выбрать 12 кластеров. Затем для каждого кластера найдены центроиды. Далее было найдено 12 наиболее близких векторов предложений к каждому из 12 центроидов. При поиске ближайшего вектора находились иногда одинаковые, поэтому были найдены несколько ближайших векторов для каждого центроида и выбраны уникальные.

Аналогично для зоны «тянуть» был построен график значений k на Рисунке 3.

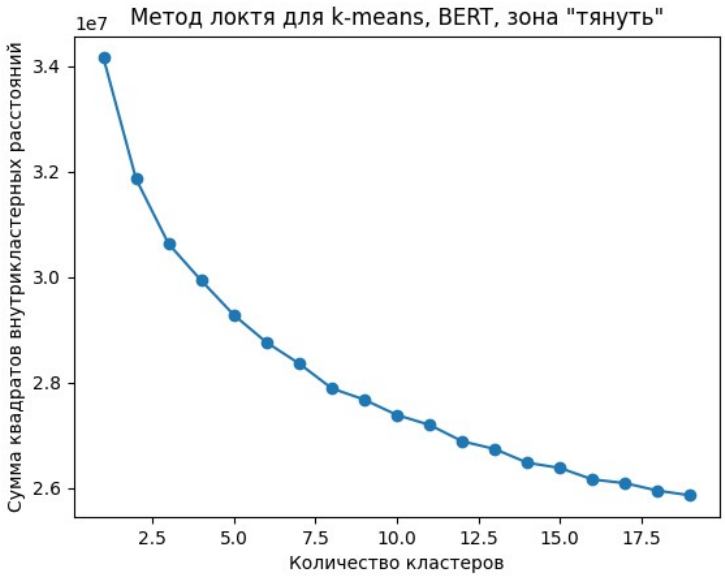

Рисунок 3 - Метод локтя для K-means с векторами BERT зоны «тянуть»
DOI: https://doi.org/10.60797/IRJ.2026.166.52.3

По графику видно, что точка, обозначающая 9 кластеров, значительно меньше отличается по высоте от точки, обозначающей 8 кластеров, чем все предыдущие точки на графике относительно их ближайших. Поэтому по методу локтя было выбрано 8 кластеров.

 © Авторы статьи / Authors of the article

Была произведена оценка. Сначала была подсчитана точность, потом полнота, затем F-мера. Эти метрики применимы к оценке алгоритма, т.к. необходимо понять, сколько найденных фреймов соответствуют фреймам из работы [1], и сколько фреймов из работы [1] было найдено. Для оценки точности среди найденных фреймов были найдены те, которые семантически совпадают с фреймами из работы [1]. Их количество было разделено на общее количество фреймом, найденных алгоритмом. Для оценки полноты было определено, сколько из найденных в работе [1] фреймов совпадает с фреймами, найденными алгоритмом. Их количество было разделено на общее количество фреймов, найденных в работе [1]. F-мера считалась по формуле (2).

$$F = \frac{2PR}{P+R} \quad (2)$$

P — точность.
R — полнота.
Результаты в Таблице 1.

Таблица 1 - Оценка метода BERT + K-means

DOI: https://doi.org/10.60797/IRJ.2026.166.52.4

	Тянуть	Толкать	Тянуть-толкать
Точность	75	100	88
Полнота	70	73	71,4
F-мера	72,4	84,4	79

Подход показал хорошие результаты. Не были найдены только следующие фреймы: «перемещать ногами», «нажимать на кнопку», «сталкивать с высоких объектов», «помещать во внутрь».

5.2. BERT с DBSCAN

Для алгоритма DBSCAN необходимо выбрать значения eps и min_samples. Были найдены следующие оптимальные значения параметров для зоны «тянуть»:

1) eps = 0.27;
2) min_samples = 5.

Оптимальные значения eps лежат в диапазоне от 0.27 до 0.29. Такие значения дают 11 кластеров. Оптимальны именно такие значения, т.к. при выборе значения eps равным 0.26 уменьшается точность, и остается много нерелевантных предложений — 22, а значение 0.3 наоборот оставляет слишком мало предложений — 6, и таким образом уменьшается полнота. Оптимальным значением min_samples будет 5, т.к, если его уменьшить до 4, будет 9 кластеров, и уменьшится полнота, а если повысить до 6, будет 12 кластеров, и уменьшится точность.

Найдены 3 новых фрейма, которых нет в работе [1] (здесь и далее примеры взяты из Национального корпуса русского языка):

1. «Когда Wimm Bill Dann разливал свой сок J7, то это было намного лучше, чем сегодняшние ролики, где показан человек, постоянно тянущийся за стаканом».

2. «Колтунов «раскочегарил» свою «пятерку» и натужно потянулся по остывшему следу».

3. «Глаза Гуревича на секунду затянуло пеленой: он представил себе добрый шмоток сырокопченой свининки».

Для зоны «толкать» такие же параметры, как для зоны «тянуть», дают только один кластер. Поэтому оптимальными для данной зоны являются следующие параметры:

1) eps = 0.2;
2) min_samples = 3.

Диапазон возможных значений min_samples от 1 до 3, а eps от 0.19 до 0.21.

Они дают 9 кластеров. При min_samples = 4 количество найденных кластеров уменьшается до 7.

Оценка работы приведена в Таблице 2.

Таблица 2 - Оценка метода BERT + DBSCAN

DOI: https://doi.org/10.60797/IRJ.2026.166.52.5

	Тянуть	Толкать	Тянуть-толкать
Точность	73	78	75
Полнота	50	64	57
F-мера	59,4	70,3	64,77

Не были найдены фреймы «перемещать ногами»,»помещать во внутрь», «нажимать на кнопку», «сталкивать с высоких объектов»

5.3. ELMo с K-means

 © Авторы статьи / Authors of the article

Получение эмедингов BERT значительно быстрее, чем ELMo. Перевод первых 100 000 предложений в BERT занимает 8,2 секунды, а в ELMo 41.68 секунд на Macbook pro с процессором m1, 16-ю гигабайтами оперативной памяти и 1-м тб постоянной памяти.

Аналогично, как для метода BERT + K-means, построим график суммы квадратов внутрикластерных расстояний в зависимости от значений количества кластеров k для зоны «тянуть» на Рисунке 4.

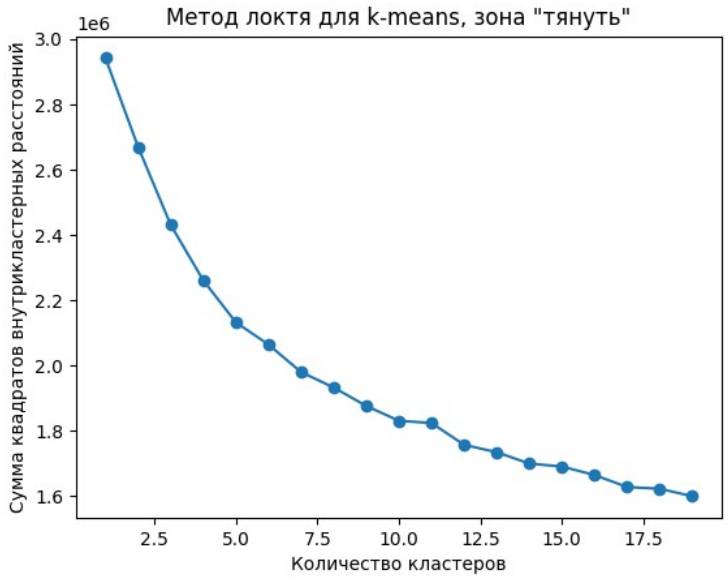

Рисунок 4 - Метод локтя для K-means с векторами ELMo зоны «тянуть»
DOI: https://doi.org/10.60797/IRJ.2026.166.52.6

По методу локтя выбираем 10 кластеров. Примеры найденных предложений:

- *«Правду молвил Ухмыл: такой затянулся узелок, что и не распустишь»*.
- *«Лизавета тянула Сашу к дверям»*.
- *«Дима-а -- протянул мужик»*.
- *«Марик замер и вытянул шею, чтоб ему было все видно»*.
- *«Тут и рекламные щиты, и мигающий неон, и виллы в зелени, а в небе летел самолетик, за которым тянулся какой-то шлейф»*.
- *«Больные, шаркая тапочками, потянулись на процедуры»*.
- *«Но почему-то потянуло прогуляться»*.
- *«Ты не могла бы снять с меня эти браслеты -- спросил Сергей, пошевелив руками, по-прежнему стянутыми за спиной наручниками»*.
- *«Ларт подал ему руку и рывком втянул в круг: -- Время»*.
- *«Ну и тени, конечно, тоже мелькали на дне сознания -- желающий мог дотянуться и до них»*.

Аналогично для зоны «толкать» график на Рисунке 5.

 © Авторы статьи / Authors of the article

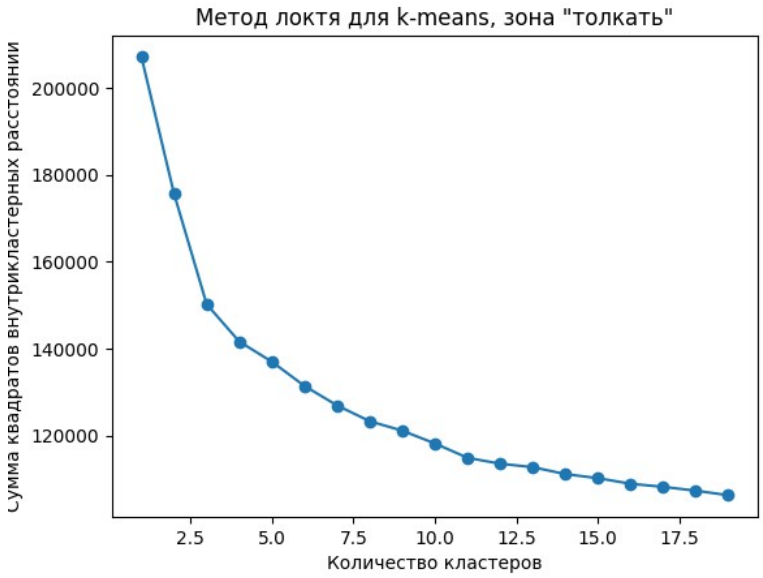

Рисунок 5 - Метод локтя для K-means с векторами ELMo зоны «толкать»
DOI: https://doi.org/10.60797/IRJ.2026.166.52.7

Было выбрано 12 кластеров. Оценка приведена в таблице 3.

Таблица 3 - Оценка метода ELMo + K-means

DOI: https://doi.org/10.60797/IRJ.2026.166.52.8

	Тянуть	Толкать	Тянуть-толкать
Точность	70	83	77
Полнота	60	64	62
F-мера	65	72	68,7

5.4. ELMo с DBSCAN

Для векторов ELMo и зоны «тянуть» были выбраны следующие оптимальные параметры:

1) eps = 0.2;

2) min_samples = 5.

Диапазон оптимальных значений eps от 0.2 до 0.22, min_samples от 4 до 7. Они дают 11 фреймов. При min_samples = 3 находится больше фреймов, и уменьшается точность. При значении eps 0.19 находится 16 фреймов, тоже уменьшается точность. При значении 0.23 находится 6 фреймов, уменьшается полнота.

Для зоны «толкать» подходят такие же оптимальные параметры, как и для зоны «тянуть».

Оценка приведена в Таблице 4.

Таблица 4 - Оценка метода ELMo + DBSCAN

DOI: https://doi.org/10.60797/IRJ.2026.166.52.9

	Тянуть	Толкать	Тянуть-толкать
Точность	55	73	64
Полнота	40	55	48
F-мера	46	63	55

Самые лучшие результаты показал подход BERT с K-means, на втором месте ELMo с K-means, на третьем BERT с DBSCAN, на четвёртом ELMo с DBSCAN. Таким образом, векторы BERT показываются лучше результаты, чем ELMo, а кластеризация K-means лучше, чем DBSCAN, при этом классификатор больше влияет на результат, чем тип векторов.

Результаты

В Таблице 5 приведены сравнительные оценки работы алгоритмов. Жирным выделен самый лучший результат.

 © Авторы статьи / Authors of the article

Таблица 5 - Сравнение работы алгоритмов

DOI: https://doi.org/10.60797/IRJ.2026.166.52.10

	B-K-T1	B-K-T2	B-K-TO	B-D-T1	B-D-T2	B-D-TO	E-K-T1	E-K-T2	E-K-TO	E-D-T1	E-D-T2	E-D-TO
P	75	100	88	73	78	75	70	83	77	55	73	64
R	70	73	71,4	50	64	57	60	64	62	40	55	48
F	72,4	84,4	79	59,4	70,3	65	65	72	68,7	46	63	55

Примечание: B-K-T1 - BERT + K-means для зоны «тянуть»; B-K-T1 - BERT + K-means для зоны «толкать»; B-K-TO - BERT + K-means для зоны «тянуть-толкать»; B-D-T1 - BERT + DBSCAN для зоны «тянуть»; B-D-T2 - BERT + DBSCAN для зоны «толкать»; B-K-TO - BERT + DBSCAN для зоны «тянуть-толкать»; E-K-T1 - ELMO + K-means для зоны «тянуть»; E-K-T2 - ELMO + K-means для зоны «толкать»; E-K-TO - ELMO + K-means для зоны «тянуть-толкать»; E-D-T1 - ELMO + DBSCAN для зоны «тянуть»; E-D-T2 - ELMO + DBSCAN для зоны «толкать»; E-D-TO - ELMO + DBSCAN для зоны «тянуть-толкать»

По всем семантическим зонам («тянуть», «толкать», объединенная «тянуть-толкать») и по всем показателям (точность, полнота, F — мера) самые лучшие результаты показал подход с вектором BERT и фильтрацией K-means. При этом из семантических зон лучших показателей удалось достичь для зоны «толкать».

Второе место для зоны «тянуть» занял подход с векторами ELMo и кластеризацией K-means. По точности он меньше на 5%, по полноте на 10%, по F-мере на 7.4%. На 3 месте подход с векторами BERT и кластеризацией DBSCAN, по точности он больше на 3%, но по полноте меньше на 10% и по F-мере меньше на 5.6%. Четвёртое место занял подход ELMo с кластеризацией DBSCAN. По точности он меньше на 18%, по полноте на 10%, по F-мере на 13.4%.

Второе место для зоны «толкать» занял подход с векторами ELMo и кластеризацией K-means. По точности он меньше на 17%, по полноте на 9%, по F-мере на 12.4%. Третье место занял подход с векторами BERT и кластеризацией DBSCAN. По точности он меньше на 5%, по полноте одинаково, по F-мере меньше на 1.7%. Четвертое место занял подход с векторами ELMo и кластеризацией DBSCAN. По точности он меньше на 5%, по полноте на 9%, по F-мере на 6.7%

Второе место для объединенной зоны «тянуть-толкать» также занял подход с векторами ELMo и кластеризацией K-means. По точности он меньше на 11%, по полноте меньше на 9.6%, по F-мере меньше на 1.3%. Третье место занял подход с векторами BERT и кластеризацией K-means. По точности он меньше на 2%, по полноте на 5%, по F-мере на 3.7%. Четвертое место занял также подход с векторами ELMo и кластеризацией DBSCAN. По точности он меньше на 9%, по полноте на 9%, по F-мере на 10%.

Выводы

Эксперименты показали влияние как векторов, так и способов фильтрации на результат. Векторы BERT показывают результаты лучше, чем векторы ELMo, а кластеризацией с заданным количеством кластеров K-means лучше, чем кластеризация с автоматическим количеством кластеров DBSCAN. При этом на результат больше влияет именно способ фильтрации, чем тип векторов. Влияние распространяется как на точность, так и на полноту на всех подзонах. Исключение составляет только точность подхода BERT с кластеризацией DBSCAN для зоны «тянуть». Он показал улучшение результатов на 3% относительно подхода ELMo с кластеризацией K-means. Также этот подход является исключением для зоны «толкать» по полноте. Он показал такие же результаты, как алгоритм с векторами ELMo и кластеризацией DBSCAN.

Заключение

Для решения задачи лексической типологии применены современные векторные представления и способы кластеризации. Лучшие результаты показал подход с использованием векторов BERT и кластеризацией K-means. Выявлено, что способы фильтрации слов влияют больше, чем векторные представления. Полученные результаты приближены к ручным. Найдены не все фреймы, которые есть в ручном методе, при этом найдены и те, которые не были обнаружены в ручном методе, но они принадлежат зоне «тянуть-толкать». Таким образом, автоматизированные поход не может полностью заменить ручную работу ученых лингвистов, однако может помочь в исследованиях, способствуя увеличению скорости работы и ее полноте.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

 © Авторы статьи / Authors of the article

Список литературы / References

1. Савельева А.Ю. Глаголы семантических зон «ТЯНУТЬ» и «ТОЛКАТЬ» в типологической перспективе / А.Ю. Савельева // Проблемы компьютерной лингвистики и типологии: сб. Всерос. конф. — Воронеж: Издательский дом ВГУ, 2017. — Вып. 6. — С. 142–152.

2. Рахилина Е.В. Фреймовый подход к лексической типологии / Е.В. Рахилина, Т.И. Резникова // Вопросы языкознания. — 2013. — № 2. — С. 3–31.

3. Berlin B. Color Terms: Their Universality and Evolution / B. Berlin. — Berkeley: University of California Press, 1969. — 178 p.

4. Wierzbicka A. Semantic and lexical universals: Theory and emperical findings / A. Wierzbicka // Linguistic Investigations. — 1994. — № 21. — P. 249-261.

5. Дунаева К.О. Семантическое поле «мешать» в типологической перспективе / К.О. Дунаева, В.В. Маринина // XXVI Открытая конференция студентов-филологов в СПбГУ. — Санкт-Петербург: СПбГУ, 2023. — С. 34–37.

6. Шешкина Т.Ф. Германо-славянские параллели семантического поля «Домашний скот» в немецких лексикографических источниках / Т.Ф. Шешкина // Филологические науки. Вопросы теории и практики. — 2020. — № 6. — С. 303–307. — DOI: 10.30853/filnauki.2020.6.57

7. Рыжова Д.А. Фрагмент лексической системы казымского диалекта хантыйского языка: глаголы pitti «упасть, попасть» и χɔjti «задеть, попасть» и их аргументная структура / Д.А. Рыжова // Урало-алтайские исследования. — 2022. — № 2 (45). — С. 123–140. — DOI: 10.37892/2500-2902-2022-45-2-123-140

8. Холкина Л.С. Семантическое поле ОСТРЫЙ в китайском языке: диахроническое развитие и его отражение в современных диалектах / Л.С. Холкина, Л.О. Наний, Ц. Сы // Journal of Language Relationship. — 2023. — № 20 (3-4). — С. 280–298. — DOI: 10.31826/jlr-2023-203-410

9. Влавацкая М.В. Лексико-семантическое поле «шахматная игра» в современном русском языке / М.В. Влавацкая, И.Н. Журавлева // Мир науки, культуры, образования. — 2022. — № 2 (93). — С. 293–297.

10. Кашкин Е.В. Категоризация качественных признаков «мягкий», «твердый», «жесткий» в горномарийском языке / Е.В. Кашкин // Вестник ВГУ. Серия: Лингвистика и межкультурная коммуникация. — 2022. — № 1. — С. 140–150. — DOI: 10.17308/lic.2022.1/9009

11. Rakhilina E. The Typology of Physical Qualities / E. Rakhilina, T. Reznikova, M. Kyuseva et al. // Amsterdam: John Benjamins Publishing Company. — 2022. — Vol. 2. — P. 29–55.

12. Григорьева О.Н. Лексико-семантическая группа «город» в современных российских масс-медиа / О.Н. Григорьева, Н. Цзян // Вестник Московского государственного областного университета. Серия: Русская филология. — 2018. — № 5. — С. 31–37. — DOI: 10.18384/2310-7278-2018-5-31-38

13. Kyuseva M. Automatic data collection in lexical typology / M. Kyuseva, E. Parina, D. Ryzhova // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". — Moscow: ABBYY, 2018. — P. 29–55.

14. Рыжова Д.А. Опыт автоматического построения анкеты для лексико-типологического исследования прилагательных и одноместных глаголов с помощью моделей дистрибутивной семантики / Д.А. Рыжова // ВЕСТНИК РГГУ. Сер.: История. Филология. Культурология. Востоковедение. — 2016. — Т. 18. — С. 140–150.

15. Савчук С.О. Национальный корпус русского языка 2.0: новые возможности и перспективы развития / С.О. Савчук, Т.А. Архангельский, А.А. Бонч-Осмоловская и др. // Вопросы языкознания. — 2024. — № 2. — С. 7–34.

16. Peters M.E. Deep contextualized word representations / M.E. Peters, M. Neumann, M. Gardner et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; — New Orleans: Association for Computational Linguistics, 2018. — P. 2227–2237. doi: 10.18653/v1/N18-1202

17. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee et al. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; — Minneapolis: Association for Computational Linguistics, 2019. — P. 4171–4186. doi: 10.18653/v1/N19-1423

18. Jin X. K-Means Clustering / X. Jin, J. Han // Encyclopedia of Machine Learning. — 2011. — № 1. — P. 563-563. — DOI: 10.1007/978-0-387-30164-8_425

19. Martin E. A density-based algorithm for discovering clusters in large spatial databases with noise / E. Martin, K. Hans-Peter, S. Jörg et al. // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. — 1996. — № 1. — P. 226–231.

Список литературы на английском языке / References in English

1. Saveleva A.Yu. Glagoli semanticheskikh zon «TYaNUT» i «TOLKAT» v tipologicheskoi perspektive [Verbs of the semantic fields "PULL" and "PUSH" from a typological perspective] / A.Yu. Saveleva // Problemi kompyuternoi lingvistiki i tipologii: sb. Vseros. konf [Problems of Computational Linguistics and Typology: Proceedings of the All-Russian Conference]. — Voronezh: VSU Publishing House, 2017. — Iss. 6. — P. 142–152. [in Russian]

2. Raxilina E.V. Frejmovy'j podxod k leksicheskoj tipologii [A frame-based approach to lexical typology] / E.V. Raxilina, T.I. Reznikova // Questions of linguistics. — 2013. — № 2. — P. 3–31. [in Russian]

3. Berlin B. Color Terms: Their Universality and Evolution / B. Berlin. — Berkeley: University of California Press, 1969. — 178 p.

4. Wierzbicka A. Semantic and lexical universals: Theory and emperical findings / A. Wierzbicka // Linguistic Investigations. — 1994. — № 21. — P. 249-261.

 © Авторы статьи / Authors of the article

5. Dunaeva K.O. Semanticheskoe pole 'meshat' v tipologicheskoi perspektive [The semantic field of "to interfere" from a typological perspective] / K.O. Dunaeva, V.V. Marinina // XXVI Otkritaya konferentsiya studentov-filologov v SPbGU [XXVI Open Conference of Philology Students at St. Petersburg State University]. — Saint Petersburg: SPbGU, 2023. — P. 34–37. [in Russian]

6. Sheshkina T.F. Germano-slavyanskie paralleli semanticheskogo polya «Domashnij skot» v nemeczkix leksikograficheskix istochnikax [Germano-Slavic parallels in the semantic field of "domestic cattle" in German lexicographical sources] / T.F. Sheshkina // Philological Sciences. Questions of Theory and Practice. — 2020. — № 6. — P. 303–307. — DOI: 10.30853/filnauki.2020.6.57 [in Russian]

7. Ry'zhova D.A. Fragment leksicheskoj sistemy' kazy'mskogo dialekta xanty'jskogo yazy'ka: glagoly' pitti «upast', popast'» i χɔjti «zadet', popast'» i ix argumentnaya struktura [A fragment of the lexical system of the Kazym dialect of the Khanty language: the verbs pitti "to fall, to hit" and χɔjti "to touch, to hit" and their argument structure] / D.A. Ry'zhova // Ural-Altai studies. — 2022. — № 2 (45). — P. 123–140. — DOI: 10.37892/2500-2902-2022-45-2-123-140 [in Russian]

8. Xolkina L.S. Semanticheskoe pole OSTRY'J v kitajskom yazy'ke: diaxronicheskoe razvitie i ego otrazhenie v sovremenny'x dialektax [The semantic field of SHARP in Chinese: diachronic development and its reflection in modern dialects] / L.S. Xolkina, L.O. Nanij, Cz. Sy' // Journal of Language Relationship. — 2023. — № 20 (3-4). — P. 280–298. — DOI: 10.31826/jlr-2023-203-410 [in Russian]

9. Vlavaczkaya M.V. Leksiko-semanticheskoe pole «shaxmatnaya igra» v sovremennom russkom yazy'ke [The lexical-semantic field of "chess game" in modern Russian] / M.V. Vlavaczkaya, I.N. Zhuravleva // The world of science, culture and education. — 2022. — № 2 (93). — P. 293–297. [in Russian]

10. Kashkin E.V. Kategorizaciya kachestvenny'x priznakov «myagkij», «tverdy'j», «zhestkij» v gornomarijskom yazy'ke [Categorisation of qualitative characteristics "soft", "hard", "rigid" in the Hill Mari language] / E.V. Kashkin // VSU Bulletin. Series: Linguistics and Intercultural Communication. — 2022. — № 1. — P. 140–150. — DOI: 10.17308/lic.2022.1/9009 [in Russian]

11. Rakhilina E. The Typology of Physical Qualities / E. Rakhilina, T. Reznikova, M. Kyuseva et al. // Amsterdam: John Benjamins Publishing Company. — 2022. — Vol. 2. — P. 29–55.

12. Grigor'eva O.N. Leksiko-semanticheskaya gruppa «gorod» v sovremenny'x rossijskix mass-media [The lexical-semantic group «city» in contemporary Russian mass media] / O.N. Grigor'eva, N. Czzyan // Bulletin of Moscow State Regional University. Series: Russian Philology. — 2018. — № 5. — P. 31–37. — DOI: 10.18384/2310-7278-2018-5-31-38 [in Russian]

13. Kyuseva M. Automatic data collection in lexical typology / M. Kyuseva, E. Parina, D. Ryzhova // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". — Moscow: ABBYY, 2018. — P. 29–55.

14. Rizhova D.A. Opit avtomaticheskogo postroeniya anketi dlya leksiko-tipologicheskogo issledovaniya prilagatelnikh i odnomestnikh glagolov s pomoshchyu modelei distributivnoi semantiki [Experience in automatically constructing questionnaires for lexical-typological research of adjectives and monovalent verbs using distributional semantics models] / D.A. Rizhova // VESTNIK RGGU. Ser.: Istoriya. Filologiya. Kulturologiya. Vostokovedenie [RSHU Bulletin. Series: History. Philology. Cultural Studies. Oriental Studies]. — 2016. — Vol. 18. — P. 140–150. [in Russian]

15. Savchuk S.O. Nacional'ny'j korpus russkogo yazy'ka 2.0: novy'e vozmozhnosti i perspektivy' razvitiya [National Corpus of Russian Language 2.0: New Opportunities and Prospects for Development] / S.O. Savchuk, T.A. Arxangel'skij, A.A. Bonch-Osmolovskaya et al. // Issues of Lingusitics. — 2024. — № 2. — P. 7–34. [in Russian]

16. Peters M.E. Deep contextualized word representations / M.E. Peters, M. Neumann, M. Gardner et al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; — New Orleans: Association for Computational Linguistics, 2018. — P. 2227–2237. doi: 10.18653/v1/N18-1202

17. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee et al. // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; — Minneapolis: Association for Computational Linguistics, 2019. — P. 4171–4186. doi: 10.18653/v1/N19-1423

18. Jin X. K-Means Clustering / X. Jin, J. Han // Encyclopedia of Machine Learning. — 2011. — № 1. — P. 563-563. — DOI: 10.1007/978-0-387-30164-8_425

19. Martin E. A density-based algorithm for discovering clusters in large spatial databases with noise / E. Martin, K. Hans-Peter, S. Jörg et al. // KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. — 1996. — № 1. — P. 226–231.