И

Проблемы искусственного интеллекта. 2026. N 3 (42). С. 200-212
Problems of Artificial Intelligence. 2026;3(42):200-212.
Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей
Научная статья

УДК 004.827 **DOI** 10.24412/2413-7383-2026-3-42-200-212

Э. С. Израилова
Комплексный научно-исследовательский институт им. акад. Х.И. Ибрагимова РАН, г. Грозный, Россия
uelisa@yandex.ru, ORCID: https://orcid.org/0000-0003-1655-4414/

МЕТОД АВТОМАТИЧЕСКОГО РАЗРЕШЕНИЯ ОМОГРАФИИ НА ОСНОВЕ ЕВКЛИДОВОЙ МЕТРИКИ WEN

E. S. Izrailova
Kh. Ibragimov Complex Institute of the Russian Academy of Sciences (CI RAS), Grozny, Russia
uelisa@yandex.ru, ORCID: https://orcid.org/0000-0003-1655-4414/

THE METHOD FOR AUTOMATIC DISAMBIGUATION OF HOMOGRAPHY BASED ON THE EUCLIDEAN METRIC WEN

В статье проведен обзор классических методов разрешения лексической многозначности, использующих векторное представление слов и евклидово расстояние. Представлен новый метод WSD – метод автоматического разрешения омографии на основе евклидовой метрики. Приведено описание работы метода WEN на примере распознавания омографа «замок» в предложении, показан процесс «оцифровки» слов и предложений из баз, вычисление расстояний от тестового предложения до баз предложений с различными значениями омографа. Представлены данные об экспериментальном исследовании метода WEN на базах предложений с омографами для малоресурсного чеченского языка. Проведено сравнение полученных метрик точности с аналогичными методами для различных языков.

Ключевые слова: разрешение лексической многозначности (WSD), метод WEN, евклидово расстояние, малоресурсные языки, омографы, эмбеддинг.

The article provides an overview of classical methods for resolving lexical ambiguity using vector representation of words and Euclidean distance. A new WSD method is presented, a method for automatic homography resolution based on the Euclidean metric. The description of the WEN method is given using the example of recognizing the "lock" homograph in a sentence, the process of "digitizing" words and sentences from databases, calculating distances from a test sentence to sentence databases with different homograph values. The data on the experimental study of the WEN method on sentence databases with homographs for the low-resource Chechen language are presented. The obtained accuracy metrics are compared with similar methods for different languages.

Keywords: word sense disambiguation (WSD), WEN method, Euclidean distance, low-resource languages, homographs, embedding.

Введение

Разрешение лексической многозначности (Word Sense Disambiguation, WSD) является одной из классических и наиболее сложных проблем в области компьютерной лингвистики и обработки естественного языка. WSD – это задача автоматического определения того, в каком значении (смысле) употреблено слово. Для решения проблемы устранения многозначности слов необходимо наличие больших текстовых корпусов, различных словарей и баз данных для реализации методов WSD, как алгоритмических моделей, так и методов машинного обучения.Основные научные исследования в области разрешения лексической многозначности (Word Sense Disambiguation, WSD) и созданные на их основе методы получены для английского, русского, китайского и других языков, достаточно обеспеченных электронными ресурсами.

В мире насчитывается более 7000 живых разговорных языков, и только для небольшого числа из них существуют необходимые информационные ресурсы и программное обеспечение для реализации естественно-языковых технологий. То есть большинство языков мира являются малоресурсными, они мало изучены и не обеспечены корпусами и специализированными базами данных, вследствие чего затрудняется создание для таких языков различных автоматизированных систем языкового моделирования и обработки [1]. В России по международной классификации все языки кроме русского относятся к малоресурсным языкам, в их числе и чеченский язык, для которого не существует описания всех необходимых сведений в достаточном объеме и необходимом формате. Основные проблемы, возникающие при разработке прикладных программ для чеченского языка – нехватка электронных ресурсов для обработки языка и речи, в том числе одноязычных корпусов, двуязычных электронных словарей, орфоэпических и грамматических словарей, тезаурусов, толковых словарей, словарей омографов и других различных лингвистических ресурсов. Таким образом, в данном исследовании решение проблемы WSD необходимо проводить в условиях ограниченных языковых информационных ресурсов.

Цель данной работы – разработка метода, обеспечивающего автоматическое разрешение омографии в текстах на чеченском языке. Для достижения цели работы были поставлены и решены следующие задачи:

1. Исследование современных подходов к разрешению омографии и устранению лексической многозначности.
2. Разработка баз данных, алгоритма и метода автоматического разрешения омографии в текстах на чеченском языке.

Научная новизна исследования состоит в том, что предложен новый метод разрешения омографии для малоресурсного чеченского языка, отличающийся формированием числового вектора контекстных значений слов во входном предложении и обеспечивающий формирование последовательности слов с оптимальными контекстными значениями омографов.

Обзор методов разрешения лексической многозначности WSD

Все методы WSD можно разделить на три большие категории в зависимости от того, какой ресурс они используют:

1. Методы, основанные на знаниях. Эти методы используют готовые лингвистические ресурсы (словари, тезаурусы, онтологии) и не требуют размеченных данных для обучения [2-6].

2. Методы, основанные на машинном обучении (Supervised Learning). Это наиболее точные на сегодняшний день методы, но они требуют больших размеченных вручную корпусов текстов (SemCor — семантический корпус) [7-12].

3. Методы, основанные на неразмеченных данных (Unsupervised Learning) [13], [14].

В данной работе рассматривается вторая категория методов – методы, основанные на машинном обучении. Данные модели довольно часто используют векторные представления слов – эмбеддинги для решения задачи разрешения лексической многозначности WSD. Приведем несколько наиболее известных методов, использующих эмбеддинги для снятия лексической многозначности.

Авторы в статье [15] показали, что слова граф со скрытой иерархической структурой, которую должно выявить следующее поколение самообучающихся векторных представлений слов. В этой статье, опираясь на древовидности пространства, авторы предлагают встраивать слова в декартово произведение гиперболических пространств, которое теоретически связано с гауссовыми векторными представлениями слов и их геометрией Фишера.

Метод формирующий эмбеддинги синсетов (Sense Embeddings) представлен в работе [16] Для каждого синсета в WordNet собираются все входящие в него леммы и слова из глосс (определений). Задача WSD решается как поиск ближайшего соседа: вектор контекста сравнивается с векторами всех возможных значений через косинусное сходство (эквивалент евклидова расстояния для нормализованных векторов).

Авторы статьи [17] описали новый метод EWISE (Embedding Wordnet Ideas for Sense Embeddings), который на основе машинного обучения создаёт эмбеддинги значений, комбинируя информацию из WordNet и больших корпусов. Также метод использует нейросетевую архитектуру для объединения информации из глосс, примеров употребления и семантических отношений.

Графовые методы с евклидовой проекцией описываются в работе [18]. Граф WordNet встраивается в евклидово пространство. Контекст целевого слова векторизуется усреднением эмбеддингов слов. Выбирается синсет, чей вектор минимально удален от вектора контекста.

В трудах [19], [20] авторами представлены современные нейросетевые подходы BERT и контекстуальные эмбеддинги. Эти трансформеры генерируют контекстуализированные представления слов, которые затем сравниваются с эмбеддингами значений. Вычисляется косинусное сходство (линейно связанное с евклидовым расстоянием) между контекстуальным вектором и векторами-прототипами. Выбирается значение с максимальным сходством.

В статье [20] описываются модели SenseBERT / SemBERT, которые специально дообучаются с задачей предсказания значений слов. Данные модели обучается не только предсказывать замаскированные слова, но и их значения (синсеты WordNet). SenseBERT значительно улучшает понимание лексики, что продемонстрировано в ходе экспериментов по устранению неоднозначности значений слов в рамках конкурса SemEval.

Для сравнения векторных эмбеддингов используются метрики, такие как косинусное расстояние или евклидова метрика. В задачах компьютерной лингвистики евклидова метрика используется для количественного определения семантической близости между словами или контекстами. Обзор научной литературы по проблеме снятия лексической многозначности (Word Sense Disambiguation, WSD) показал, что существуют несколько методов, использующих евклидову метрику для вычисления расстояний между векторами слов. Перечислим их:

1. Метод ближайших контекстов с евклидовым расстоянием. Авторы статьи [21] предлагают метод автоматического снятия лексической многозначности для всех слов в документе на японском языке. Основная гипотеза заключается в том, что значения слов можно определить, вычисляя евклидовы расстояния между векторами окружения целевого слова и векторами его гипотетических синонимов. Чем меньше евклидово расстояние между вектором контекста и вектором синонима, тем выше вероятность, что значение слова соответствует этому синониму.

2. Метод ε-фильтрации на основе расстояния Хаусдорфа. В работе [22] предлагается новый метод WSD, который использует ε-фильтрацию для вычисления близости контекстов. Вместо простого усреднения векторов слов, метод рассматривает контексты как множества векторов. Для сравнения этих множеств используется метрика Хаусдорфа, которая, в свою очередь, опирается на базовую метрику в пространстве векторов (например, евклидову). Для достижения более высокой точности выполняется предварительная є-фильтрация слов как в предложении, так и в наборе синонимов.

3. Кластеризация на основе евклидовой метрики – метод, которые преобразует граф семантических отношений (например, из WordNet) в евклидово векторное пространство. В этом пространстве местоположение каждого понятия или набора синонимов задается вектором, что позволяет применять алгоритмы кластеризации (например, K-средних), использующие евклидову метрику для группировки слов со схожим значением.

Представленное в данной статье исследование показывает использование евклидовой метрики для сравнения расстояний между тестовым предложением, в котором находится целевое многозначное слово, и базами предложений с различными значениям многозначного слова.

Предлагаемый метод автоматического разрешения омографии на основе евклидовой метрики WEN

Идея рассматриваемого метода заключается в том, чтобы определить к какому из n множеств B_i «ближе» содержащее данный омограф тестовое предложение. Поэтому в векторном пространстве необходимо ввести понятие расстояния между предложениями [23].

Приведем детальное описание схемы алгоритма WEN в соответствии с рисунком 1.

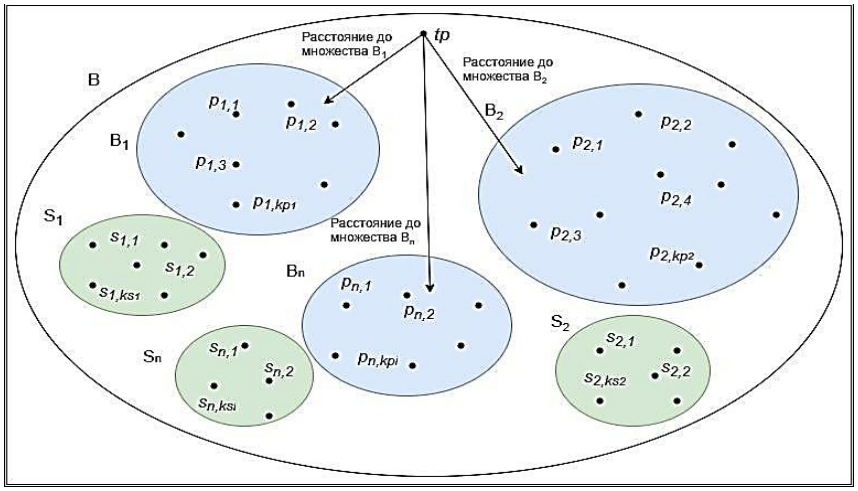


Рисунок 1 – Схема оцифровки предложений в методе WEN

Где: B – база предложений p, содержащих омограф om;

$B_1, B_2, \dots, B_n$ – базы предложений $p_{i,j}$, содержащих омограф om со значениями om_i, $i = 1,2, \dots, n$; $S_1, S_2, \dots, S_n$– базы различных слов-тегов $s_{i,k}$, $k = 1, 2, \dots, ks_i$, содержащихся хотя бы в одном предложении из баз B_i; tp – тестовое предложение.

Схема разделена на несколько ключевых блоков и этапов:

1. Исходные данные и оцифровка.

– Тестовое предложение (tp): Предложение, содержащее омограф, значение которого необходимо определить.

– Базы предложений (B_1, B_2, ..., B_n): Набор эталонных баз. Каждая база B предложения с одним из возможных значений омографа.

– Базы слов (S_1, S_2, ..., S_n): Набор словарей или баз слов, связанных с соответствующими базами предложений, использующихся для формирования контекстных векторов.

Процесс оцифровки состоит в следующем: как тестовое предложение (tp), так и все предложения из эталонных баз (B) преобразуются в числовые вектора (tv для тестового и свои вектора для каждого предложения в базах). Это необходимо для математического сравнения.

2. Вычислительное ядро алгоритма AWEN, в котором происходит вычисление метрики евклидова расстояния между числовым вектором тестового предложения (tp/tv) и каждой из эталонных баз (B_1, B_2, ..., B_n). База представлена как совокупность векторов своих предложений, и расстояние вычисляется до этой совокупности (например, среднее расстояние до всех векторов в базе).

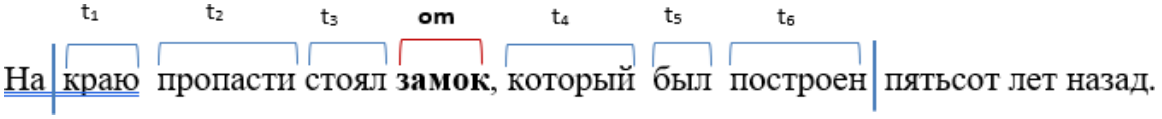
3. Выбор минимального расстояния: Алгоритм выбирает ту эталонную базу B_i, расстояние от которой до тестового предложения оказалось минимальным.

Результат: Значение омографа, связанное с выбранной базой B_i, признается тем значением, в котором омограф используется в тестовом предложении. Происходит распознавание значения омографа.

Этот подход основан на гипотезе, что предложения, в которых омограф используется в одном значении, будут семантически ближе (иметь меньшее векторное расстояние) друг к другу, чем к предложениям, где омограф используется в других значениях.

Описание работы метода WEN на примере распознавания омографа «замок» в предложении

Рассмотрим работу метода распознавания омографов AWEN на примере распознавания омографа зáмок или замóк. Имеются две базы предложений B_1 и B_2. В базе B_1 собраны предложения с омографом замóк, в базе B_2 – с омографом зáмок. Будем распознавать омограф по шести тегам – соседним словам. Число 6 определено экспериментально.



Третий тег t3 «стоял» имеет шесть координат, которые вначале равны нулю (0, 0, 0, 0, 0, 0). Ниже в таблице 1 показан процесс формирования координат выбранного тега по всем базам Bn с предложениями.

Таблица 1– Формирование координат выбранного тега по базе B_1 с предложениями B_1

1. …, …, … замок …, …, … .	1. стоял (0, 0, 0, 0, 0, 0)
Стоял, …, … замок …, …, … .	стоял (1, 0, 0, 0, 0, 0)
…, … замок …, стоял, … .	стоял (1, 0, 0, 0, 1, 0)
Стоял, …, … замок.	стоял (2, 0, 0, 0, 1, 0)
…, … замок …, стоял.	стоял (2, 0, 0, 0, 2, 0)
…, …, … замок …, стоял, … .	стоял (2, 0, 0, 0, 3, 0)
…, …, стоял замок.	стоял (2, 0, 1, 0, 3, 0)
…, …, … замок …, … .	стоял (2, 0, 1, 0, 3, 0)

Таблица 2 – Формирование координат выбранного тега по базе B_2 с предложениями. B_2

…, …, … замок …, …, … .	стоял (2, 0, 1, 0, 3, 0)
…, …, … замок …, … .	….
….	….
….	….
….	… .
100. …, …, … замок …, …, … .	100. стоял (10, 5, 6, 0, 5, 11)

Таблица 3 – Сложение соответствующих координат тегов t_N

краю	(20, 10, 0, 5, 5, 15) +
пропасти	(0, 11, 2, 3, 4, 7) +
стоял	(10, 5, 6, 0, 5, 11) +
который	(1, 1, 0, 75, 8, 0) +
был	(5, 3, 1, 38, 7, 9) +
построен	(4, 3, 0, 25, 15, 12) +
ТР	(40, 33, 9, 146, 44, 54)

В результате суммирования соответствующих координат тегов t_N получаются координаты вектора тестового предложения ТР (40, 33, 9, 146, 44, 54). Для каждого предложения из баз Bn вычисляются координаты, аналогично тому, как вычислялись координаты тестового предложения. Все слова теги из баз B$_n$ собираются в соответствующие им базы S_n. Для каждого слова-тега формируются числовые вектора с координатами, вычисляемыми в зависимости от количества повторов данного слова-тега в базе с предложениями и его позиции относительно омографа.

Формирование числового вектора для слова-тегов в базах S на основе предложений из баз B производится аналогично. Затем также вычисляются числовые вектора для остальных предложений $sp1,j$ из первой базы B1 и для других баз. Теперь каждому предложению pi,j, j = 1, 2, … , kpi, из базы Bi, i = 1,2, … , n, соответствует числовой вектор $vp_{i,j} = (np_{i,j,1}, np_{i,j,2}, \dots, np_{i,j,N})$,

Каждому предложению $p_{i,j}$, j = 1, 2, … , kp_i , из базы B_i, i = 1,2, … , n, соответствует числовой вектор $vp_{i,j} = (np_{i,j,1}, np_{i,j,2}, \dots, np_{i,j,N})$. Пусть $p_{1,1}$(27, 20, 57, 155, 51, 27) – числовой вектор первого предложения из первой базы. Экспериментально определен весовой вектор w = (0,9; 0,95; 1; 1; 0,95; 0,9), смысл которого: чем дальше тег от омографа, тем меньше его влияние на значение омографа.

Найдем расстояние от вектора первого предложения $p_{1,1}$ из первой базы до вектора тестового предложения по формуле:

$$(tp, p_{i,j}) = \sqrt{\sum_{m=1}^{N} w_m (tw_m - np_{i,j,m})^2}, j = 1,2, \dots, kp_i, i = 1,2, \dots, n.$$

Подставляем значения:

ρ (tp, $p_{1,1}$) = $\sqrt{0,9(40-27)^2 + 0,95(33-20)^2 + 1(9-57)^2 + 1(146-155)^2 + 0,95(44-51)^2 + 0,9(54-27)^2}$ = 3400

Пусть $p_{1,2}$ = (9, 15, 10, 7, 26, 13) – числовой вектор второго предложения из первой базы. Вычислим расстояние от него до тестового предложения:

ρ (tp, $p_{1,2}$) = $\sqrt{0,9(40-9)^2 + 0,95(33-15)^2 + 1(9-10)^2 + 1(146-7)^2 + 0,95(44-26)^2 + 0,9(54-13)^2}$ = 150

Затем находим среднее арифметическое расстояние r_1 от тестового предложения до предложений из базы B_1: допустим r_1 = 1005. Аналогично находим среднее арифметическое расстояние r_2 от тестового предложения до предложений из базы B_2: r_2 = 993. Сравниваем r_1 и r_2 . Имеем $r_2 < r_1$. Выбираем номер (индекс) меньшего. Это 2. Значит омограф в тестовом предложении тот же, что и в базе B_2, т.е. $om = om_2$ = зáмок. В случае $r_1 = r_2$ выбираем тот омограф, который в жизни встречается чаще, т.е. из базы, в которой больше предложений.

Результаты работы программы на основе метода WEN

Для исходных параметров алгоритма экспериментальным путем установлены оптимальные значения: количество тегов N = 6, т.е. значение омографа определяется по шести словам в предложении: по трём – слева и по трём – справа от омографа; весовой вектор w = (0,9; 0,95; 1; 1; 0,95; 0,9). Для тестирования результатов работы алгоритма [23] создан модуль самотестирования программы распознавания омографов. Для модуля выделяются случайным образом выбранные части баз B_1 и B_2. На оставшихся частях проводится обучение. Экспериментально определено оптимальное соотношение между обучающей и тестовой частями баз – 80%/20%. В итоге программа вывела средние значения для разных метрик, которые отображены в таблице 4.

Таблица 4 – Точность распознавания омографов по методу WEN, %

Омограф	**N**	**WEN**	**Омограф**	**N**	**WEN**
бала	578	36,7	йаха	235	49,3
белира	960	70,3	йолу	2556	20,7
бен	1813	35,6	лар	1405	41,8
дакъа	588	41,5	цIе	2400	32,6
дала	6500	87,2	ча	346	42,4
дан	3783	41,3	шун	652	32,4
де	1168	31,5	**среднее значение**	**1768**	**43,3**

где N -количество предложений в базе.

Таблица 5 – Сравнение результатов точности распознавания многозначных слов различными методами, %

Язык	**Корпуса и базы**	**Метод**	**Accuracy (%)**
Карельский	WCorpus 333 000 предложений/215000 слов	Метод ε- фильтрации на основе расстояния Хаусдорфа	65,9
Ассамский	SeAnDa (20000 предложений)	Naive Bayes Classi- fier	50,2
Японский	4 688 762 слов	all-words WSD w2v	51,3
		all-words WSD c2v	52,8
		all-words WSD w2v+c2v	53,9
Чеченский	CheWSData (15000 предложений)	Naive Bayes	69,1
		Logistic Regression	56,3
		Random Forest	63,0
		WEN	**43,3**

Из таблицы 5 видно, что наиболее высокие метрики точности для метода ε-фильтрации на основе расстояния Хаусдорфа (65,9%) для карельского языка, Наивный Байесовский классификатор для чеченского языка показал результаты точности 69,1%. Точность предлагаемого метода WEN на самой нижней позиции 43,3%.

После эксперимента возник вопрос: приведет ли увеличение баз данных с размеченными предложениями к повышению показателей точности распознавания многозначных слов? Для ответа на возникший вопрос был проведен эксперимент – увеличены базы предложений для наиболее распространенного омографа «Дала» (таблица 6). Изначально в базе было 1720 размеченных предложений и для этого количества предложений максимальные метрики точности метода WEN составили: F1- 65.10 и Accuracy – 65.87. После увеличения количества размеченных предложений в базе с1720 до 6500, метрики точности также увеличились до следующих показателей: F1- 87.52 и Accuracy – 87.18.

Таблица 6 – Результаты вычисления метрик точности по методу WEN для омографа «Дала» (F1, Accuracy %)

Омограф "Дала"					
S	**N**	**L**	**W**	**F1**	**Accuracy**
1720	4	2	[0.9, 1, 1, 0.9]	60.59	61.05
	6	3	[0.8, 0.9, 1, 1, 0.9, 0.8]	**65.10**	**65.87**
	6	2	[0.9, 1, 1, 0.9, 0.8, 0.7]	64.21	65.23
	6	4	[0.7, 0.8, 0.9, 1, 1, 0.9]	63.51	63.77
	8	4	[0.7, 0.8, 0.9, 1, 1, 0.9, 0.8, 0.7]	63.42	64.87
	8	5	[0.6, 0.7, 0.8, 0.9, 1, 1, 0.9, 0.8]	63.95	65.17
	8	3	[0.8, 0.9, 1, 1, 0.9, 0.8, 0.7, 0.6]	64.19	**65.84**
6500	4	2	[0.9, 1, 1, 0.9]	83.63	83.05
	6	3	[0.8, 0.9, 1, 1, 0.9, 0.8]	86.69	86.01
	6	2	[0.9, 1, 1, 0.9, 0.8, 0.7]	84.85	85.09
	6	4	[0.7, 0.8, 0.9, 1, 1, 0.9]	85.55	85.58
	8	4	[0.7, 0.8, 0.9, 1, 1, 0.9, 0.8, 0.7]	82.4	83.62
	8	5	[0.6, 0.7, 0.8, 0.9, 1, 1, 0.9, 0.8]	85.91	86.71
	8	3	[0.8, 0.9, 1, 1, 0.9, 0.8, 0.7, 0.6]	**87.52**	**87.18**

В таблице 6 приведены в качестве примера результаты вычисления точности распознавания омографа «Дала» по методу WEN, где: S – количество предложений в базе, N – количество слов рядом с омографом, L – количество слов слева от омографа, W – весовой вектор, показывающий влияние слова в зависимости от его позиции относительно омографа.

Заключение

Методы, использующие векторные представления слов (эмбеддинги) и геометрические расстояния, в частности евклидово расстояние, являются классическими и основополагающими методами WSD. Хотя современные методы, основанные на нейронных сетях, дают более высокие оценки по метрикам точности WSD, классические векторные методы для малоресурсных языков могут быть приемлемыми, так как менее требовательны к размеру исходных баз данных.

Однако современные системы WSD редко используют какой-то один метод. Наиболее эффективным на сегодняшний день является гибридный подход: для массовых, частотных слов используется обучение с учителем, что даёт максимальную точность; для редких слов, для которых нет размеченных примеров, применяются методы, основанные на знаниях (упрощенный алгоритм Леска или поиск по семантической близости в WordNet).

Таким образом, WSD остается задачей, где комбинация "классической лингвистики" (словари) и решений на основе нейронных сетей (глубокое обучение) дает наилучший результат. Исходя из сказанного, подведем итог: предложенный метод WEN может использоваться как часть гибридной программной системы разрешения лексической многозначности.

Список источников

1. Карпов А.А. Речевые технологии для малоресурсных языков мира / А.А. Карпов, В.О. Верходанова // Вопросы языкознания. 2015. № 2. С. 117–135.
2. Lesk M. Automatic sense disambiguation using machine-readable dictionaries: how to tell a pinecone from an ice cream cone / M. Lesk // In Proceedings of the 5th annual international conference on Systems documentation. 1986. P. 24–26.
3. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art / Lastra-Díaz J.J., Goikoetxea J., Taieb M.A.H. et al. // Engineering Applications of Artificial Intelligence. 2019. № 85. P.645–665.
4. Zero-shot word sense disambiguation using sense definition embeddings / Kumar S., Jat S., Saxena K., Talukdar P. // In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. P. 5670–5681.
5. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation / Scozzafava F., Maru M., Brignone F. et al. // In Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations. 2020. P. 37–46.
6. Ниценко, А. В. Об использовании семантической информации для снятия омонимии именительного и винительного падежа (как элемента создания онтологии) / А. В. Ниценко, В. Ю. Шелепов // Проблемы искусственного интеллекта. 2024. № 4(35). С. 16-24. DOI 10.24412/2413-7383-2024-4-16-24. EDN FKGYSQ.
7. Conia S., Navigli R. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration // In Proceedings of the 16th Conference of the European Chapter of the Asso- ciation for Computational Linguistics: Main Volume. – 2021. – P. 3269–3275.
8. Amrami A., Goldberg Y. Towards better substitution-based word sense induction // arXiv. preprint arXiv:1905.12598. 2019.Arefyev N., Sheludko B., Panchenko A. Combining Lexical Substitutes in Neural Word Sense Induc- tion // Proceedings of the International Conference on Recent Advances in Natural Language Pro- cessing (RANLP 2019). Varna, Bulgaria: INCOMA Ltd. 2019. P. 62–70.
9. Vasilescu, F. Evaluating Variants of the Lesk Approach for Disambiguating Words / F. Vasilescu, P. Langlais, G. Lapalme // In Lrec. 2004.

10. 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. 2019. P. 4171–4186.
11. El-Razzaz M. Arabic gloss WSD using BERT / M. El-Razzaz, M.W. Fakhr, F.A. Maghraby // Applied Sciences. 2021. 11(6). P. 2567.
12. Haveliwala T.H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search / T.H. Haveliwala // IEEE transactions on knowledge data engineering. 2003. V. 15, N 4. P. 784–796.
13. Deep Contextualized Word Representations / Peters M., Neumann M., Iyyer M. [et al.] // In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguis- tics: Human Language Technologies. 2018. P. 2227–2237.
14. Tifrea, A. Poincar'e glove: Hyperbolic word embeddings / A. Tifrea, G. Bécigneul, O. E. Ganea // arXiv preprint arXiv:1810.06546. 2018.
15. Loureiro D. Language modelling makes sense: Propagating representations through Word- Net for full-coverage word sense disambiguation / D. Loureiro, A. Jorge // Proceedings of the 57th annual meeting of the as- sociation for computational linguistics. 2019. C. 5682-5691.
16. Zero-shot word sense disambiguation using sense definition embeddings / Kumar S. et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019. C. 5670-5681.
17. Moradi, B. Unsupervised Word Sense Disambiguation Using Word Embeddings / B. Moradi, E. Ansari and Z. Žabokrtský // 25th Conference of Open Innovations Association (FRUCT), Helsinki, Finland, 2019, pp. 228-233, doi: 10.23919/FRUCT48121.2019.8981526.
18. Hadiwinoto C. Improved word sense disambiguation using pre-trained contextualized word representations / C. Hadiwinoto, H. T. Ng, W. C. Gan // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. C. 5297-5306.
19. Wiedemann G. et al. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings //arXiv preprint arXiv:1909.10430. – 2019.
20. SenseBERT: Driving some sense into BERT / Levine Y. [et al.] //Proceedings of the 58th annual meeting of the association for computational linguistics. 2020. C. 4656-4667.
21. Unsupervised all-words wsd using synonyms and embeddings / Suzuki R. [et al.] // Journal of Natural Language Processing. 2019. Т. 26. №. 2. С. 361-379.
22. Kirillov, A. WSD algorithm based on a new method of vector-word contexts proximity calculation via ε-filtration / A. Kirillov, N. Krizhanovskaya, A. Krizhanovsky // Труды Карельского научного центра Российской академии наук. 2018. No. 7. P. 149-163. DOI 10.17076/mat829.
23. Алгоритм распознавания омографов на основе евклидовой метрики / Израилова Э.С., Астемиров А.В., Бадаева А.С., Султанов З.А., Умархаджиев С.М., Хехаев М.-С.Л., Ясаева М.Л. // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 1. С. 41–50. doi: 10.17586/2226-1494-2024-24-1-41-50.

References

1. Karpov A. A. Speech technologies for under-resourced languages of the world / A. A. Karpov, V. O. Verkhodanova // Voprosy Jazykoznanija. – 2015. – No. 2. – Pp. 117–135.
2. Lesk M. Automatic sense disambiguation using machine-readable dictionaries: how to tell a pinecone from an ice cream cone / M. Lesk // In Proceedings of the 5th annual international conference on Systems documentation. –1986. – P. 24–26.
3. A reproducible survey on word embeddings and ontology-based methods for word similarity: Linear combinations outperform the state of the art / Lastra-Díaz J.J., Goikoetxea J., Taieb M.A.H. et al. // Engineering Applications of Artificial Intelligence. – 2019. – № 85. – P. 645–665.
4. Zero-shot word sense disambiguation using sense definition embeddings / Kumar S., Jat S., Saxena K., Talukdar P. // In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – 2019. – P. 5670–5681.
5. Personalized PageRank with syntagmatic information for multilingual word sense disambiguation / Scozzafava F., Maru M., Brignone F. et al. // In Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations. – 2020. – P. 37–46.
6. Nicenko A.V. The use of semantic information to disambiguate the nominative/ accusative homonyms: an element of creating ontology / A.V. Nicenko, V.Ju. Shelepov // Problems of artificial intelligence. – 2024. – № 4(35). – P. 16-24.
7. Conia S. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration / S. Conia, R. Navigli // In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. – 2021. – P. 3269–3275.

8. Amrami A. Towards better substitution-based word sense induction // arXiv. preprint arXiv:1905.12598. – 2019.Arefyev N., Sheludko B., Panchenko A. Combining Lexical Substitutes in Neural Word Sense Induction / A. Amrami, Y. Goldberg // Proceedings of the International Conference on Recent Advances in Natural Language Pro- cessing (RANLP 2019). – Varna, Bulgaria: INCOMA Ltd. – 2019. – P. 62–70.
9. Vasilescu F. Evaluating Variants of the Lesk Approach for Disambiguating Words / F. Vasilescu, P. Langlais, Lapalme G. // In Lrec. – 2004.
10. 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. – 2019. – P. 4171–4186.
11. El-Razzaz M. Arabic gloss WSD using BERT / M. El-Razzaz, M.W. Fakhr, F.A. Maghraby // Applied Sciences. – 2021. – № 11(6). – P. 2567.
12. Haveliwala T.H. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search / T.H. Haveliwala // IEEE transactions on knowledge data engineering. – 2003. – V. 15, N4. – P. 784–796.
13. Deep Contextualized Word Representations / Peters M., Neumann M., Iyyer M. et al. // In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – 2018. – P. 2227–2237.
14. Tifrea A. Poincar'e glove: Hyperbolic word embeddings / A. Tifrea, G. Bécigneul, O. E. Ganea // arXiv preprint arXiv:1810.06546. – 2018.
15. Loureiro D. Language modelling makes sense: Propagating representations through Word- Net for full-coverage word sense disambiguation / D. Loureiro, A. Jorge // Proceedings of the 57th annual meeting of the as- sociation for computational linguistics. – 2019. – C. 5682-5691.
16. Zero-shot word sense disambiguation using sense definition embeddings / Kumar S. et al. // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. – 2019. – C. 5670-5681.
17. Moradi, B. Unsupervised Word Sense Disambiguation Using Word Embeddings, / B. Moradi, E. Ansari and Z. Žabokrtský // 25th Conference of Open Innovations Association (FRUCT). – Helsinki, Finland, 2019. – P. 228-233. – doi: 10.23919/FRUCT48121.2019.8981526.
18. Hadiwinoto C. Improved word sense disambiguation using pre-trained contextualized word representations / C. Hadiwinoto, H. T. Ng, W. C. Gan // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – 2019. – C. 5297-5306.
19. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings / Wiedemann G. et al. // arXiv preprint arXiv:1909.10430. – 2019.
20. SenseBERT: Driving some sense into BERT / Levine Y. et al. // Proceedings of the 58th annual meeting of the association for computational linguistics. – 2020. – C. 4656-4667.
21. Unsupervised all-words wsd using synonyms and embeddings / Suzuki R. et al. // Journal of Natural Language Processing. – 2019. – Т. 26. – №. 2. – С. 361-379.
22. Kirillov A. NWSD algorithm based on a new method of vector-word contexts proximity calculation via ε-filtration / A. Kirillov, N. B. Krizhanovskaya, A. A. Krizhanovsky // Transactions of Karelian Research Centre RAS. – 2018. – No. 7. – P. 149–163. – DOI 10.17076/mat829.
23. Homograph recognition algorithm based on Euclidean metric / Izrailova E. S., Astemirov A. V., Badaeva A. S., Sultanov Z. A., Umarkhadzhiev S. M., Khekhaev M. L., Yasaeva M. L. // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. – 2024. – Vol. 24, No. 1. – Pp. 41–50. – Doi 10.17586/2226-1494-2024-24-1-41-50

RESUME

E. S. Izrailova

The method for automatic disambiguation of homography based on the euclidean metric WEN

The resolution of lexical ambiguity (Word Sense Disambiguation, WSD) is one of the classic and most difficult problems in the field of computational linguistics and natural language processing. To solve the problem of eliminating the ambiguity of words, it is necessary to have large text corpora, various dictionaries and databases for the implementation of WSD methods, both algorithmic models and machine learning methods.

Methods using vector representations of words (embeddings) and geometric distances, in particular the Euclidean distance, are classical and fundamental methods of WSD. Although modern methods based on neural networks give higher scores in terms of WSD accuracy metrics, classical vector methods for low-resource languages may be acceptable, since they are less demanding on the size of the source databases.

The research presented in this article shows the use of Euclidean metrics to compare the distances between a test sentence containing a target polysemous word and sentence databases with different meanings of a polysemous word. The article provides an overview of classical methods for resolving lexical ambiguity using vector representation of words and Euclidean distance.

A new WSD method is presented, a homograph recognition method based on the Euclidean metric. The description of the WEN method is given using the example of recognizing the "lock" homograph in a sentence, the process of "digitizing" words and sentences from databases, and calculating distances from a test sentence to databases with different homograph values. The data on the experimental study of the WEN method on sentence databases with homographs for the low-resource Chechen language are presented. The obtained accuracy metrics are compared with similar methods for different languages.

Thus, WSD remains a task where a combination of "classical linguistics" (dictionaries) and neural network-based solutions (deep learning) gives the best result. Based on the above, let's summarize: the proposed WEN method can be used as part of a hybrid software system for resolving lexical ambiguity.

РЕЗЮМЕ

Э. С. Израилова

Метод автоматического разрешения омографии на основе евклидовой метрики WEN

Разрешение лексической многозначности (Word Sense Disambiguation, WSD) является одной из классических и наиболее сложных проблем в области компьютерной лингвистики и обработки естественного языка. Для решения проблемы устранения многозначности слов необходимо наличие больших текстовых корпусов, различных словарей и баз данных для реализации методов WSD, как алгоритмических моделей, так и методов машинного обучения.

Методы, использующие векторные представления слов (эмбеддинги) и геометрические расстояния, в частности евклидово расстояние, являются классическими и основополагающими методами WSD. Хотя современные методы, основанные на нейронных сетях, дают более высокие оценки по метрикам точности WSD, классические векторные методы для малоресурсных языков могут быть приемлемыми, так как менее требовательны к размеру исходных баз данных.

Представленное в данной статье исследование показывает использование евклидовой метрики для сравнения расстояний между тестовым предложением, в котором находится целевое многозначное слово, и базами предложений с различными значениям многозначного слова. В статье проведен обзор классических методов разрешения лексической многозначности, использующих векторное представление слов и евклидово расстояние.

Представлен новый метод WSD – метод распознавания омографов на основе евклидовой метрики. Приведено описание работы метода WEN на примере распознавания омографа «замок» в предложении, показан процесс «оцифровки» слов

и предложений из баз, вычисление расстояний от тестового предложения до баз с различными значениями омографа. Представлены данные об экспериментальном исследовании метода WEN на базах предложений с омографами для малоресурсного чеченского языка. Проведено сравнение полученных метрик точности с аналогичными методами для различных языков.

Таким образом, WSD остается задачей, где комбинация "классической лингвистики" (словари) и решений на основе нейронных сетей (глубокое обучение) дает наилучший результат. Исходя из сказанного, подведем итог: предложенный метод WEN может использоваться как часть гибридной программной системы разрешения лексической многозначности.

Израилова Элиса Салаудиновна, научный сотрудник лаборатории прикладной математики, сектора искусственного интеллекта Комплексного научно-исследовательского института им. акад. Х.И. Ибрагимова РАН, г. Грозный, Россия. тел. 8(938)-995-24-68,
e-mail: uelisa@yandex.ru, ORCID: https://orcid.org/0000-0003-1655-4414/.*Область научных интересов:* синтез и распознавание речи, синтез чеченской речи, обработка естественного языка, разрешение лексической многозначности.

Israilova Elisa Salaudinovna, Researcher at the Laboratory of Applied Mathematics, Artificial Intelligence Sector of the Kh. Ibragimov Complex Institute of the Russian Academy of Sciences (Cl RAS), Grozny, Russia. tel. 8(938)-995-24-68, e-mail: uelisa@yandex.ru , ORCID: https://orcid.org/0000-0003-1655-4414 /.*Research interests:* speech synthesis and recognition, synthesis of Chechen speech, natural language processing, resolution of lexical ambiguity.

Статья поступила в редакцию: 13.05.2026
одобрена после рецензирования 22.07.2026; принята к публикации 02.09.2026
The article was submitted 13.05.2026; approved after reviewing 22.07.2026; accepted for publication 02.09.2026.