Вестник Московского университета. Серия 9. Филология. 2026. № 2. С. 47–59
Lomonosov Philology Journal. Series 9. Philology, 2026, no. 2, pp. 47–59

АВТОМАТИЧЕСКОЕ РАЗРЕШЕНИЕ МНОГОЗНАЧНОСТИ АББРЕВИАТУР В БИОМЕДИЦИНСКИХ ТЕКСТАХ НА РУССКОМ ЯЗЫКЕ

О.А. Митрофанова

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия; o.mitrofanova@spbu.ru

П.А. Гусяцкая

Московский государственный университет имени М.В. Ломоносова, Москва, Россия; polinagousyatskaya@gmail.com

Аннотация: В данной статье описывается исследование, посвященное разрешению неоднозначности (омонимии) инициальных аббревиатур в русскоязычных биомедицинских текстах. Неоднозначность является неотъемлемым свойством естественного языка, распространяющимся даже на кодифицированные сегменты, в частности на терминосистемы. Содержание биомедицинского домена представлено совокупностью терминологически насыщенных текстов, относящихся к разным направлениям медицины и смежных наук, вследствие чего аббревиатуры в таких текстах склонны к межсистемной неоднозначности. Процедура разграничения омонимичных аббревиатур опирается на методологию дистрибутивной семантики, предполагающую использование корпус-ориентированного подхода и контекстного анализа в разрешении неоднозначности. В рамках исследования был собран уникальный корпус неоднозначных медицинских аббревиатур на русском языке и контекстов их употребления. В корпус вошли 76 аббревиатур, имеющих от 2 до 11 значений; для каждого значения был проведен отбор контекстов методом случайной выборки, объем контекстов составляет примерно 150–160 словоупотреблений. На корпусе были проведены эксперименты по классификации с использованием классического алгоритма — машины опорных векторов (SVM), и предобученной нейросетевой модели RuBioBERT. Модель SVM, демонстрирующая самые высокие результаты среди классических методов машинного обучения, подтвердила свой статус и на терминологических текстах, показав аккуратность и F-меру в 0.94. Модель RuBioBERT достигла наивысших результатов для данной задачи — 0.98 по обеим метрикам. На нейросетевой модели были проведены дополнительные эксперименты с привлечением широкого и узкого контекста: результаты классификации показали, что более информативным для модели является широкий контекст. Также были сделаны предварительные выводы о влиянии абсолютного количества контекстов и их семантической близости на качество классификации.

Ключевые слова: разрешение неоднозначности; аббревиатура; биомедицинские тексты; классификация контекстов; дистрибутивно-семантические модели

Финансирование: Исследование выполнено при поддержке СПбГУ, шифр проекта 123042000068-8.

doi: 10.55959/MSU0130-0075-9-2026-49-02-4

Для цитирования: *Митрофанова О.А., Гусяцкая П.А.* Автоматическое разрешение неоднозначности аббревиатур в русскоязычных биомедицинских текстах // Вестн. Моск. ун-та. Серия 9. Филология. 2026. № 2. С. 47–59.

AUTOMATIC ACRONYM DISAMBIGUATION IN RUSSIAN BIOMEDICAL TEXTS

Olga A. Mitrofanova

Saint Petersburg State University, Saint Petersburg, Russia; o.mitrofanova@spbu.ru

Polina A. Gousyatskaya

Lomonosov Moscow State University, Moscow, Russia; polinagousyatskaya@gmail.com

Abstract: This article is dedicated to automatic disambiguation of acronyms in Russian-language biomedical texts. Ambiguity is an inherent property of natural language, prominent even in its codified segments, namely, terminological systems. The biomedical domain is represented by a set of terminologically rich texts related to different fields of medicine and related sciences, therefore abbreviations in such texts are prone to intersystem ambiguity. The procedure for distinguishing homonymous abbreviations is based on the methodology of distributive semantics, which implies corpus-oriented approach and contextual analysis in resolving ambiguity. The problem in question called for compiling a unique corpus of contexts containing acronyms in different senses. The corpus includes 76 abbreviations exhibiting from 2 to 11 meanings; through random sampling we performed the context selection for each meaning, the sample size is approximately 150–160 tokens. Classification experiments were carried out using the classical algorithm — support vector machine (SVM), and the pre-trained RuBioBERT neural network model. The SVM model, which offers a reliable solution to the problem of resolving lexical ambiguity in texts of general semantics, confirmed its status in terminological texts, showing accuracy and F-measure of 0.94. The RuBioBERT model achieved the highest results for this task — 0.98 for both metrics. Additional experiments were performed with the neural network model involving a wide and a narrow context: classification results showed that a wide context is more informative. Preliminary conclusions were also drawn about the influence of the absolute number of contexts and their semantic proximity on the quality of classification.

Keywords: Word Sense Disambiguation; classification; acronyms; biomedical domain; context classification; distributional semantic models

Funding: This research was supported by Saint-Petersburg State University, project No. 123042000068-8.

For citation: Mitrofanova O.A., Gousyatskaya P.A. (2026) Automatic Acronym Disambiguation in Russian. *Lomonosov Philology Journal*, no. 2, pp. 47–59.

Введение

Исследовательская проблема, решение которой представлено в статье, связана с разрешением неоднозначности особого разряда лексических единиц, а именно аббревиатур, используемых в русскоязычных биомедицинских текстах. Инициальные аббревиатуры являются самостоятельными единицами терминосистемы и, благодаря своей компрессионной природе, обладают отличной от полной формы дистрибуцией и выполняют иные языковые функции [Алексеев 2019]. Сокращение плана выражения, происходящее при формировании подобных единиц, значительно увеличивает потенциал их неоднозначности.

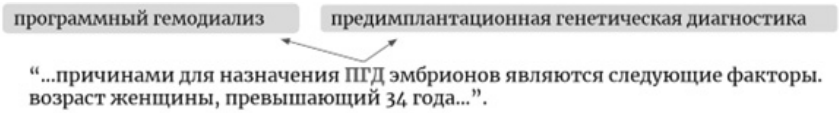


Рис. 1. Пример неоднозначной аббревиатуры в контексте

Неоднозначность — это неотъемлемое свойство естественного языка, распространяющееся в том числе на более кодифицированные его сферы — языки для специальных целей и терминосистемы. Биомедицинская терминосистема представляет собой совокупность нескольких поддоменов, соответствующих ветвям медицины и смежным наукам, что обуславливает то, что единицы ее терминосистемы проявляют межсистемную неоднозначность, другими словами, являются омонимами.

Предыдущие исследования

Проблема автоматического разрешения лексической неоднозначности (Word Sense Disambiguaton, WSD) на материале русскоязычных биомедицинских текстов в литературе не представлена — настоящее исследование предлагает первое решение данной задачи в текстах обсуждаемого домена. Куда более широко проблема представлена на материале англоязычных текстов как общей, так и биомедицин-

ской тематики. В частности, благодаря тому, что англоязычный биомедицинский домен располагает несколькими крупными лексическими ресурсами для различных задач обработки естественного языка, широко распространены подходы, основанные на знаниях. Наиболее часто встречаются попытки расширить векторное представление терминов с помощью лексических ресурсов (прежде всего, wordnet-подобных словарей), а также эксперименты на выявление наиболее информативных для WSD признаков [Skreta et al. 2021; Li et al. 2019; Wu et al. 2015].

Подходы, основанные на знаниях, для русскоязычного биомедицинского домена не так актуальны из-за отсутствия высокоструктурированных лексических ресурсов достаточного объема. Исследования в этой области фокусируются на различных способах выделения информативных морфосинтаксических и семантических контекстных параметров: наиболее важными для настоящего исследования являются работы [Азарова, Марина 2006], где описан опыт классификации контекстов в тезаурусе RussNet на основе рамок валентностей, и [Митрофанова и др. 2012], посвященная разрешению лексико-семантической неоднозначности с помощью выделения конструкций с целевым словом.

Сбор данных

В рамках исследования был собран корпус контекстов, содержащих неоднозначные аббревиатуры в разных значениях. Первым шагом стало формирование списка неоднозначных аббревиатур и их значений: это было осуществлено вручную путем анализа словарных ресурсов. Каждое значение аббревиатуры представлялось как полная форма терминологического словосочетания, или развертка.

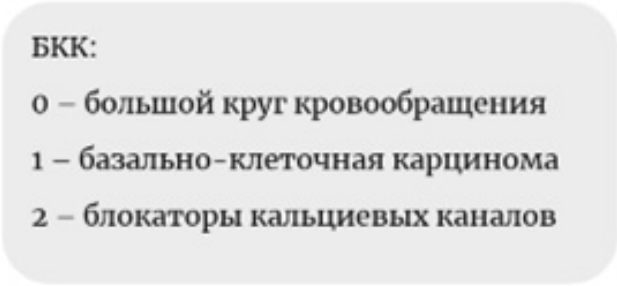
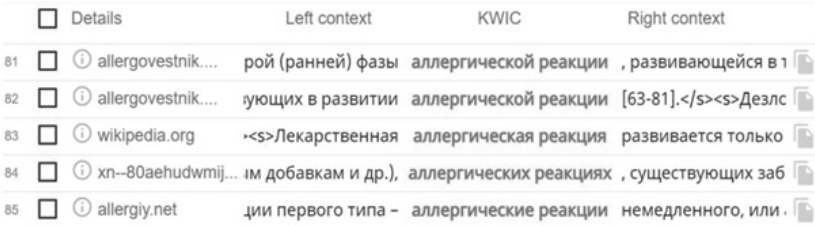


Рис. 2. Представление значений аббревиатуры в виде разверток

Для каждого значения при помощи ресурса Sketch Engine было собрано некоторое количество контекстов: данный корпусной ресурс дает возможность выгрузить из интернета конкорданс контекстов, в которых встречается заданное ключевое слово. Контекст было решено оставить максимально широким: 500 символов, или 70–80 токенов (слов) слева и справа от целевой развертки. Для обеспечения однозначности контекстов, выгружаемых для одного значения аббревиатуры, поиск в Sketch Engine велся по разверткам, а при добавлении в корпус развертка заменялась на аббревиатуру. Жанровый состав получившегося корпуса является смешанным: в основном он состоит из научно-популярных статей, посвященных медицинским феноменам, инструкций к лекарствам, обсуждений на форумах и т. д.



	Details	Left context	KWIC	Right context
81	allergovestnik....	рой (ранней) фазы	аллергической реакции	, развивающейся в т
82	allergovestnik....	ующих в развитии	аллергической реакции	[63-81].</s><s>Дезлс
83	wikipedia.org	><s>Лекарственная	аллергическая реакция	развивается только
84	xn--80aehudwmij...	им добавкам и др.),	аллергических реакциях	, существующих заб
85	allergiy.net	ции первого типа –	аллергические реакции	немедленного, или

Рис. 3. Конкорданс, формируемый
при помощи инструмента Sketch Engine

Нашей целью было выгрузить для каждого значения 80–200 контекстов: именно такое количество было признано, во-первых, минимальной адекватной выборочной совокупностью большого корпуса и, во-вторых, минимальным количеством данных, обуславливающим успешную автоматическую классификацию [Азарова, Марина 2006; Кашкин, Ляшевская 2015]. Однако частотность некоторых аббревиатур, а также отдельных значений оказалась неравномерной, в результате чего набор данных распался на три условные группы: сбалансированные аббревиатуры, все значения которых представлены 80–200 контекстами, аббревиатуры с частотными и периферийными значениями, а также редкие аббревиатуры, ни одно из значений которых не достигло отметки в 80 контекстов.

В окончательный набор данных вошло 76 аббревиатур. Количество значений варьируется от аббревиатуры к аббревиатуре: большинство из них — бинарные, остальные 30 % корпуса представлено аббревиатурами с 3–11 значениями. Средняя неоднозначность по корпусу насчитывает 3.61 значения.

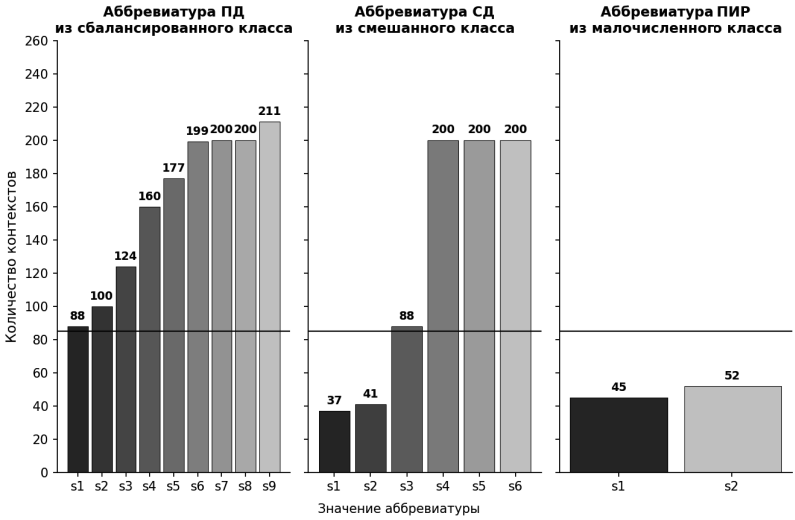


Рис. 4. Структура значений аббревиатур из трех классов

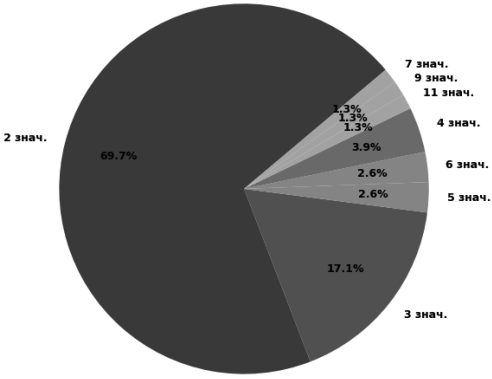


Рис. 5. Распределение аббревиатур в корпусе по количеству значений

Таблица 1

Фрагмент корпуса (контексты сокращены)

аббр.	развертка	№ знач.
ПГ	…в соединенном королевстве великобритании и северной ирландии, а симптомы заболевания у пациента возникли в декабре 2021 г. каждый случай циркуляции вирусов ПГ среди домашней птицы создает риск спорадического заражения человека…	0
ПГ	…в типичном развитии ПГ можно выделить 4 стадии, которым соответствуют местные симптомы заболевания: 1 стадия. на губах, языке, уголках рта, в других областях появляются зудящие, покалывающие ощущения, …	1

Принцип работы алгоритма

В терминах машинного обучения разрешение лексической неоднозначности соответствует задаче классификации: модели, обученной на нашем наборе данных, предлагалось предсказать метку класса по подаваемому на вход контексту. Техническая реализация классификации осуществлялась средствами языка Python: библиотеками ресурса Scikit-learn, а также моделями платформы HuggingFace.

Рис. 6. Схема работы классификатора

В рамках исследования мы протестировали на наборе данных два классификатора: классический алгоритм машинного обучения — машину опорных векторов и модель архитектуры Transformer — RuBioBERT [Yalunin 2022], предобученную на русскоязычных медицинских текстах.

Выбор SVM для эксперимента был обусловлен тем, что этот алгоритм был признан алгоритмом выбора среди классических методов машинного обучения в задаче разрешения неоднозначности на текстах общей тематики [Zhong, Ng 2010]. Более того, SVM дает более интерпретируемые результаты, чем RuBioBERT: линейный классификатор SVM является, по сути, двухклассовым — классификация производится по методу «один против всех», что дает нам ценные сведения об успешности отделения каждого класса от всех конкурирующих. Более того, в цели данного исследования входило сравнение результатов классического и нейросетевого алгоритмов.

Результаты

Машина опорных векторов на корпусе достигает средней аккуратности и F1-меры в 0.93. Такие показатели подтверждают, что SVM можно признать алгоритмом выбора для задачи разрешения неоднозначности не только на текстах общей тематики, но и на исследуемом нами материале. Стоит отметить, что рассматриваемый в нашем исследовании корпус представляет малые данные, что обуславли-

вает некоторые неточности в работе классификатора, которые, впрочем, не оказывают существенного влияния на усредненные показатели. Так, на отдельных аббревиатурах SVM показывает предельно возможный результат, что можно интерпретировать именно с точки зрения недостатка контекстов для того или иного значения.

Модель RuBioBERT на корпусе достигает аккуратности и F1-меры в 0.98, что на 5 % превышает условный baseline, заданный линейным классификатором. Таким образом, при всех достоинствах классического алгоритма, нейросетевая модель все-таки показывает свое преимущество.

Таблица 2

Результаты классификации

метрика	*F1*	*ACC*
SVM	0.93	0.93
RuBioBERT	0.98	0.976

Выводы и наблюдения

Машина опорных векторов (SVM)

Эксперименты с SVM, помимо всего прочего, позволили нам сделать несколько выводов о специфике текстов биомедицинского домена и факторах, влияющих на успешность классификации.

Так, по всей видимости, одним из таких факторов является абсолютное количество контекстов, доступных для каждого из значений неоднозначной аббревиатуры. В [Азарова, Марина 2006] утверждается, что адекватной выборочной совокупностью для неоднозначной лексемы является 100 контекстов. Эксперименты на небольшой тренировочной выборке — в 25 и 50 контекстов с широким контекстным окном — описаны в [Leacock, Chodorow 1998], однако такие алгоритмы достигали аккуратности в 40–50 %. При расширении выборки до 100 контекстов и более аккуратность поднималась до 80 %, что позволило рассматривать порог в 100 контекстов как минимально достаточный для успешной классификации значений.

По всей видимости, пороговое число контекстов, обеспечивающих успешное выделение контекстуальных маркеров, на наших данных располагается ниже указанного порога. Так, на аббревиатурах малочисленного подкорпуса (в него входили, например, аббревиатуры с соотношением контекстов по значениям 64/35 (АЕ), 64/26 (ЛТГ), 23/59 (МПА), 26/42 (ПГБ), 78/45 (ПДП) и т. д.) классификатор показывает адекватные результаты: аккуратность и F1-меру в 0.89.

По всей видимости, необходимым условием адекватного выделения класса является наличие у аббревиатуры хотя бы одного "сильного" значения, представленного достаточным количеством контекстов.

Вторым фактором успешной классификации стоит признать лексическое разнообразие контекстов, представляющих каждое из значений неоднозначной аббревиатуры. Анализ текстов корпуса показал, что знаменательные слова, составляющие контексты того или иного значения, можно поделить на две группы — в настоящей работе мы будем называть их тематическим и фактографическим контекстом. К первому будут относиться лексемы, принадлежащие к тематическим полям, соответствующим подсферам медицины, отраженным в нашем корпусе: кардиологии, фармакологии, иммунологии и т. д. Ко второму мы будем относить лексемы, соответствующие именованным сущностям, которые принято выделять из биомедицинских текстов: болезнь, лекарство, химическое соединение, орган и т. д.

Пример лексем **тематического** и фактографического контекстов:

(1) …однако **почечная недостаточность** при **ГБ**, если нет злокачественного течения, развивается редко. факторы риска **артериальной гипертонии…**

Полученные результаты указывают на то, что в рамках задачи разрешения неоднозначности эти контексты неравнозначны — а именно, на широком контекстном окне ключевое значение для успешного выделения класса будет иметь тематический контекст, в то время как фактографический по этому признаку будет нейтрален. Кроме того, эксперимент показал, что близость тематических контекстов значений неоднозначной аббревиатуры может приводить к ухудшению результатов классификации даже при условии достаточного количества контекстов. Приведем в пример результаты работы SVM на аббревиатуре «АР», где классификатор выделяет малочисленный, но лексически специфичный класс «авторентген» более успешно, чем частотные, но лексически близкие классы «аллергическая реакция» и «аллергический ринит».

Таблица 3

Пример выделения лексически уникального, но малочисленного класса

аббр.	*развертка*	*кол-во контекстов*	*F1-мера*
АР	аллергическая реакция	200	0.89
	аллергический ринит	200	0.86
	авторентген	15	0.91

Наблюдения, касающиеся тематического и фактографического контекстов, побудили нас провести отдельные эксперименты на широком и узком контекстах. Идея сравнения широкого и узкого контекстов основывается на двух подходах к лексическому значению слова в контексте, описанных в работах "One Sense per Discourse" [Gale, Church, Yarowsky 1992] и "One Sense per Collocation" [Yarowsky 1993], — первая постулирует, что неоднозначная единица с высокой вероятностью будет встречаться в одном и том же значении в рамках одного текста (дискурса), вторая — что то или иное значение многозначного слова реализуется в локальном контексте внутри коллокации.

При расчете узкого контекстного окна мы пользуемся выводами [Митрофанова и др. 2012] о том, что локальный контекст неоднозначной лексемы может содержать сочетаемостную информацию, способную улучшить результаты разрешения лексико-семантической неоднозначности. Большинство полных форм аббревиатур, входящих в наш корпус, являются именными группами, поэтому в настоящем эксперименте контексты были ограничены с помощью окна [-3; +4], соответствующего синтаксическим группам, появляющимся в локальном контексте существительного.

Аккуратность и F-мера классификатора на локальном контексте понизились до средних показателей в 0.83. Это позволяет предположить, что для разрешения неоднозначности лексем терминологического узкоспециального текста, каковыми являются тексты биомедицинского домена, широкий тематический контекст оказывается более полезным. Это предположение в целом соответствует выводам [Martinez, Agirre 2020] о том, что принцип "One Sense per Collocation" не в полной мере применим к корпусам, для которых характерна тематическая или жанровая вариативность, поскольку коллокации, выученные моделью на одной теме, не всегда возможно обобщить в отношении других предметных областей.

Не противоречат полученные результаты и нашей гипотезе о тематической и фактографической лексике, составляющей каждый из контекстов. Экспертная оценка контекстов, ограниченных окном [-3; +4], позволяет предположить, что в ближайшем окружении аббревиатуры чаще встречаются именно фактографические лексемы, ранее показавшие себя как менее качественные контекстные маркеры, помогающие алгоритму правильно совершить классификацию. Характер распределения лексем по контекстам двух видов отражен в примерах (**тематический** контекст — полужирным, фактографический — подчеркнутым):

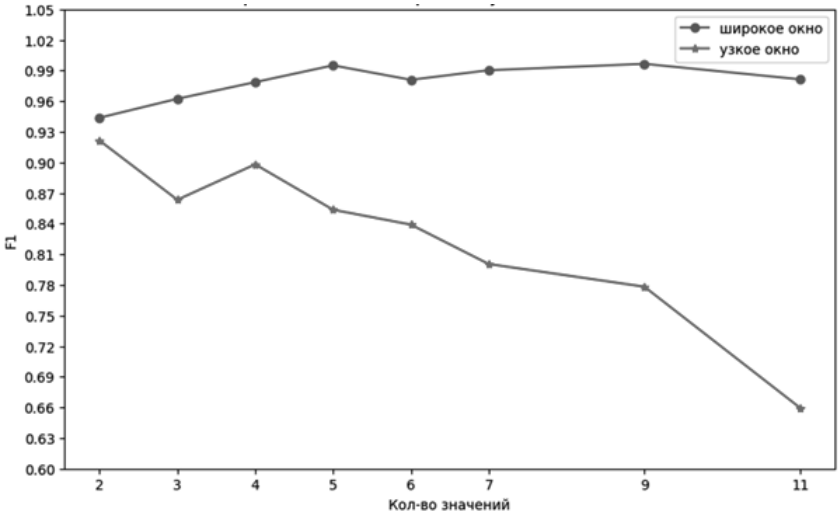


Рис. 7. F1-мера RuBioBERT на широком и узком контекстном окне

(1) при этом виде АС **коронарные сосуды** расположены **проксимально**

(2) заболеваниях методах диагностики АС **порок сердца** сопровождавшийся деформацией

(3) **генетическая** консультация лечение СД не может быть вылечен

Заключение

В статье описан опыт автоматического разрешения неоднозначности аббревиатур в русскоязычных биомедицинских текстах. В рамках исследования собран корпус контекстов употребления неоднозначных аббревиатур в разных значениях и проведены классификационные тесты на классическом алгоритме и нейросетевой модели, показавшие конкурентные по отношению к существующим на данный момент результаты. Анализ результатов классификации позволил сделать ряд выводов о специфике решения задачи WSD в терминологических текстах, а также факторов, потенциально влияющих на результаты: абсолютного количества контекстов, широты контекста и лексической близости контекстов значений одной аббревиатуры.

СПИСОК ЛИТЕРАТУРЫ

1. *Азарова И.В., Марина А.С.* Автоматизированная классификация контекстов при подготовке данных для компьютерного тезауруса RussNet // Компьютерная

лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог». 2006. С. 13–17.
2. *Алексеев Д.И.* Сокращенные слова в русском языке. М., 2019.
3. *Кашкин Е.В., Ляшевская О.Н.* Типы информации о лексических конструкциях в системе ФреймБанк // Труды института русского языка им. В.В. Виноградова. 2015. №. 6. С. 464–556.
4. *Митрофанова О.А. и др.* Автоматическое разрешение лексико-семантической неоднозначности и выделение конструкций (на материале Национального корпуса русского языка) // Лексикология. Лексикография и Корпусная лингвистика. 2013. С. 122–143.
5. *Gale W.A., Church K., Yarowsky D.* One sense per discourse // Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, 1992.
6. *Leacock C., Chodorow M., Miller G.A.* Using corpus statistics and WordNet relations for sense identification // Computational Linguistics. 1998. V. 24. № 1. P. 147–165.
7. *Li I. et al.* A neural topic-attention model for medical term abbreviation disambiguation // arXiv preprint arXiv:1910.14076. 2019.
8. *Martinez D., Agirre E.* One sense per collocation and genre/topic variations //arXiv preprint cs/0010027. 2000.
9. *Skreta, M., Arbabi, A., Wang, J. et al.* Automatically disambiguating medical acronyms with ontology-aware deep learning // Nat Commun. 2021. V. 12, P. 5319.
10. *Wu Y. et al.* Clinical abbreviation disambiguation using neural word embeddings // Proceedings of BioNLP 15. 2015. P. 171–176.
11. *Yalunin A., Nesterov A., Umerenkov D.* RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining // arXiv preprint arXiv:2204.03951. 2022.
12. *Yarowsky D.* One sense per collocation // Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, 1993.
13. *Zhong Z., Ng H. T.* It makes sense: A wide-coverage word sense disambiguation system for free text // Proceedings of the ACL 2010 system demonstrations. 2010. P. 78–83.

REFERENCES

1. Azarova I.V., Marina A.S. Avtomatizirovannaya klassifikaciya kontekstov pri podgotovke danny`x dlya komp`yuternogo tezaurusa RussNet [Automatic context classification as part of preparing the data for a thesaurus RussNet]. *Komp`yuternaya lingvistika i intellektual`ny`e texnologii: Trudy` mezhdunarodnoj konferencii «Dialog»,* 2006, pp. 13–17. (In Russ.)
2. Alekseev D.I. *Sokrashhenny`e slova v russkom yazy`ke.* [Abbreviated words in the Russian language]. Moscow., Knizhny`j dom «LIBROKOM» Publ., 2019. 346 p.
3. Kashkin E.V., Lyashevskaya O.N. Tipy` informacii o leksicheskix konstrukciyax v sisteme FrejmBank [Types of information about lexical collocations in FrameBank]. *Trudy` instituta russkogo yazy`ka im. VV Vinogradova. 2015, № 6, pp. 464–556.* (In Russ.)
4. Mitrofanova O.A. i dr. Avtomaticheskoe razreshenie leksiko-semanticheskoj neodnoznachnosti i vy`delenie konstrukcij (na materiale Nacional`nogo korpusa russkogo yazy`ka) [Automatic word sense disambiguation and collocation extraction (based on the Russian National Corpus)]. *Leksikologiya. Leksikografiya i Korpusnaya lingvistika,* 2013, pp. 122–143. (In Russ.)
5. Gale W. A., Church K., Yarowsky D. One sense per discourse. Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, 1992.

6. Leacock C., Chodorow M., Miller G.A. Using corpus statistics and WordNet relations for sense identification // Computational Linguistics. 1998. V. 24. №. 1. P. 147–165.
7. Li I. et al. A neural topic-attention model for medical term abbreviation disambiguation // arXiv preprint arXiv:1910.14076. 2019.
8. Martinez D., Agirre E. One sense per collocation and genre/topic variations // arXiv preprint cs/0010027. 2000.
9. Skreta M., Arbabi A., Wang J. et al. Automatically disambiguating medical acronyms with ontology-aware deep learning // Nat Commun. 2021. V. 12, P. 5319.
10. Wu Y. et al. Clinical abbreviation disambiguation using neural word embeddings // Proceedings of BioNLP 15. 2015. P. 171-176.
11. Yalunin A., Nesterov A., Umerenkov D. RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining // arXiv preprint arXiv:2204.03951. 2022.
12. Yarowsky D. One sense per collocation // Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, 1993.
13. Zhong Z., Ng H. T. It makes sense: A wide-coverage word sense disambiguation system for free text //Proceedings of the ACL 2010 system demonstrations. 2010. P. 78–83.

Поступила в редакцию 22.10.2025
Принята к публикации 18.02.2026
Отредактирована 20.03.2026

Received 22.10.2025
Accepted 18.02.2026
Revised 20.03.2026

ОБ АВТОРАХ

Ольга Александровна Митрофанова — к.ф.н., доцент кафедры математической лингвистики филологического факультета Санкт-Петербургского государственного университета; o.mitrofanova@spbu.ru

Полина Андреевна Гусяцкая — аспирант кафедры теоретической и прикладной лингвистики филологического факультета МГУ имени М.В. Ломоносова; polinagousyatskaya@gmail.com

ABOUT THE AUTHORS

Olga A. Mitrofanova — PhD, Associate Professor, Department of Computational Linguistics, Faculty of Philology, Saint Petersburg State University; o.mitrofanova@spbu.ru

Polina A. Gousyatskaya — PhD Student, Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University; polinagousyatskaya@gmail.com