Вестник Московского университета. Серия 9. Филология. 2023. № 6. С. 40–51
Lomonosov Philology Journal. Series 9. Philology, 2023, no. 6, pp. 40–51

ПОДХОДЫ К АВТОМАТИЧЕСКОМУ РАЗРЕШЕНИЮ МНОГОЗНАЧНОСТИ НА ОСНОВЕ НЕРАВНОМЕРНОСТИ РАСПРЕДЕЛЕНИЯ ЗНАЧЕНИЙ СЛОВ В КОРПУСЕ

Д.А. Зарипова

Московский государственный университет имени М.В. Ломоносова, Москва, Россия; diana.ser.sar96@gmail.com

Н.В. Лукашевич

Московский государственный университет имени М.В. Ломоносова, Москва, Россия; louk_nat@mail.ru

Аннотация: Задача автоматического разрешения многозначности (Word Sense Disambiguation, WSD) является ключевой задачей семантической обработки текста, решение которой влияет на качество более сложных семантических задач. Однако задача выбора одного из значений многозначного слова в контексте вызывает трудности даже у носителей языка. Тем более непростой она оказывается для систем автоматического разрешения многозначности. Поэтому так важны любые наблюдения и эвристики, способные упростить задачу либо повысить качество работы алгоритмов WSD.

Исследователями был выявлен ряд закономерностей распределения значений слов в корпусе. В статье будут рассмотрены три из них: 1) наиболее частотное значение (Most Frequent Sense, MFS); 2) гипотеза «одно значение на документ» (One Sense per Discourse) и 3) гипотеза «одно значение на словосочетание» (One Sense per Collocation).

По результатам экспериментов на материале корпуса русских текстов, метод, основанный на выборе самого частотного значения по корпусу во всех контекстах, достиг относительно высокого значения точности как на обучающей выборке, так и на тестовой (85,7 % и 71,1 % соответственно). Гипотеза «одно значение на документ» подтвердилась в 93 % текстов. Гипотеза «одно значение на словосочетание» подтверждается в 84,46 % отобранных по определенным правилам пар из текстов. Исключения связаны с трудностями при семантической разметке слов в корпусе.

Эвристики, основанные на неравномерности распределения значений многозначных слов, позволяют упростить задачу автоматического разрешения многозначности, а также могут применяться при создании тренировочных данных для обучения моделей WSD.

Ключевые слова: автоматическое разрешение многозначности; наиболее частотное значение; «одно значение на документ»; «одно значение на словосочетание»

doi: 10.55959/MSU0130-0075-9-2023-47-06-4

Для цитирования: *Зарипова Д.А., Лукашевич Н.В.* Подходы к автоматическому разрешению многозначности на основе неравномерности распределения значений слов в корпусе // Вестн. Моск. ун-та. Серия 9. Филология. 2023. № 6. С. 40–51.

APPROACHES TO AUTOMATIC WORD SENSE DISAMBIGUATION BASED ON UNEVEN DISTRIBUTION OF WORD SENSES IN CORPUS

D.A. Zaripova

Lomonosov Moscow State University, Moscow, Russia; diana.ser.sar96@gmail.com

N.V. Loukachevitch

Lomonosov Moscow State University, Moscow, Russia; louk_nat@mail.ru

Abstract: Word Sense Disambiguation (WSD) is a key task of automatic semantic analysis that affects other upstream tasks. Nevertheless, the selection of appropriate sense of ambiguous word in context is a complicated task even for human native speakers. It is even more relevant for automatic disambiguation models. That is why we need any observations and heuristics able to make the WSD task simpler or performance higher.

Researchers have noticed that the distribution of ambiguous word senses follow certain laws. In the paper we discuss three hypotheses about word senses distribution in corpus: 1) Most Frequent Sense, MFS; 2) One Sense per Discourse and 3) One Sense per Collocation.

The following results were obtained on the material of a corpus of Russian texts. Most Frequent Sense based algorithm demonstrates relatively high precision on both training and test set (85.7% and 71.1% respectively). The One Sense per Discourse hypothesis has been confirmed in 93% of texts. The One Sense per Collocation hypothesis has been confirmed in 84.46% word pairs from texts. The exceptions are related to difficulties and errors by manual word sense labeling.

Heuristics based on uneven distribution of ambiguous words in corpus allow to make WSD task simpler and can be used by collecting training data sets for WSD models.

Keywords: word sense disambiguation; most frequent sense; one sense per discourse; one sense per collocation

For citation: Zaripova D.A., Loukachevitch N.V. (2023) Approaches to Automatic Word Sense Disambiguation Based on Uneven Distribution of Word Senses in Corpus. *Lomonosov Philology Journal. Series 9. Philology,* no. 6, pp. 40–51.

Введение

Задача ***автоматического разрешения многозначности*** (*Word Sense Disambiguation*, *WSD*) является ключевой задачей семантической обработки текста, от качества решения которой зависит каче-

ство решения более высокоуровневых задач. Так, результаты WSD могут быть использованы в области ***информационного поиска*** [Hristea, Colhon 2020], при ***машинном переводе*** [Tang et al. 2018]. Поэтому очень важны любые закономерности, эвристики и наблюдения, которые могли бы упростить задачу автоматического разрешения многозначности, повысить его качество.

В статье будут описаны три закономерности распределения значений слов в корпусе, которые подходят для этой задачи. Статья состоит из трех разделов. В разделе 1 освещаются типы неравномерностей распределения значений: подраздел 1.1 посвящен самому частотному значению, раздел 1.2 — гипотезе «одно значение на дискурс», раздел 1.3 — гипотезе «одно значение на словосочетание». Раздел 2 содержит результаты экспериментов на материале корпуса русских текстов с целью проверить указанные выше закономерности. В разделе 3 приведены краткие выводы.

1. Типы неравномерностей распределения значений

Несмотря на многообразие моделей распределения значений по контекстам, исследователями были выделены несколько наиболее ярких типов неравномерностей распределения значений многозначных слов, которые будут рассмотрены в следующих подразделах.

1.1. Самое частотное значение (Most Frequent Sense)

Исследователи давно заметили, что значения многозначных слов распределены по корпусу неравномерно, то есть вероятности разных значений одного и того же слова не равны между собой. В частности, можно выделить наиболее частотное значение слова, которое встретилось в семантически размеченном корпусе чаще всех остальных.

Поэтому в задаче автоматического разрешения многозначности в качестве ***базового алгоритма для сравнения*** (*baseline*) зачастую принимается простой алгоритм, выбирающий для каждого многозначного слова наиболее частотное из его значений. Принцип работы такого алгоритма предельно прост: для каждого употребления многозначного слова приписать ему метку того значения, которое встретилось чаще всего в рамках корпуса с семантической разметкой.

Из-за важности наиболее частотного значения предложены различные алгоритмы автоматического определения наиболее частотного значения на неразмеченном корпусе. Так, в статье [Preiss et al. 2009] предлагается алгоритм поиска наиболее частотного значения слова с использованием алгоритма ранжирования и меры семантической близости слов и концептов Wikipedia. В [Calvo, Gelbukh 2015] проверяется гипотеза скоррелированности наиболее частотного значения и числа отношений, которые имеет соответствующий

синсет в онлайн-тезаурусе WordNet[1] [Miller 1998]. Авторы работы [Bhingardive et al. 2015] ставят своей целью разработать метод получения наиболее частотных значений без использования больших объемов размеченных данных и применяют для этого вектора сокращенной размерности. Работа [Loukachevitch, Mischenko 2018] посвящена обзору разных подходов к определению наиболее частотного значения для русского языка. В [Hauer et al. 2019] предложен метод определения наиболее частотных значений на основе соседей многозначного слова, а также наиболее частотных переводов.

1.2. Гипотеза One Sense per Discourse

Гипотеза «одно значение на дискурс» (в рамках данной статьи сочетания «одно значение на дискурс» и «одно значение на документ» понимаются как синонимы) была изначально сформулирована в статье [Gale et al. 1992]. В ходе исследований на материале параллельных корпусов авторы обнаружили, что ***многозначные слова редко употребляются в пределах одного дискурса (например газетной статьи) в двух и более своих значениях***. В рамках экспериментов многозначным признавалось то слово английского языка, которое переводилось несколькими способами на французский язык. В результате проверки гипотезы на статьях энциклопедии было выявлено, что только 6 статей из 300 размеченных содержали многозначные слова в разных значениях.

Авторы предлагают применять гипотезу «одно значение на дискурс», во-первых, для повышения качества работы алгоритмов по автоматическому разрешению многозначности в качестве дополнительного ограничительного фактора: выбирается тот результат, при котором все вхождения многозначного слова в текст размечены одним значением. Во-вторых, данная эвристика сильно упрощает разметку данных: теперь достаточно снять многозначность с одного вхождения многозначного слова в текст, а все остальные его вхождения разметить тем же семантическим тегом.

Эвристика довольно быстро нашла применение в реальных алгоритмах автоматического разрешения многозначности [Yarowsky 1995; Rapp 2004]. В исследовании [Carpuat 2009] данная эвристика применяется в машинном переводе. Результаты эксперимента показали, что более чем в 80 % случаев эвристика выполнялась. Показано, что принятие гипотезы «один перевод на дискурс» может помочь увеличить качество работы алгоритмов машинного перевода.

Однако в статье [Krovetz 1998] гипотеза подвергается критическому изучению и сомнению, ее заголовок говорит сам за себя — *More*

[1] https://wordnet.princeton.edu/

than One Sense Per Discourse. В работе указывается на то, что применимость гипотезы «одного значения на документ» зависит от подхода к выделению значений слов. Если рассматривать только не связанные друг с другом никакими семантическими связями значения (***bank-модель***; ***омонимия***), гипотеза [Gale et al. 1992] в большинстве случаев окажется верной. Однако при втором подходе к выделению значений слов, при котором значения могут быть связаны между собой различными отношениями (***полисемия***), картина усложняется.

В работе [Krovetz 1998] на материале двух корпусов с ручной семантической разметкой — *SemCor*[2] и *DSO Corpus*[3] — были проведены эксперименты с целью выяснить, как часто слова встречаются в более чем одном значении в рамках одного дискурса. Были получены следующие результаты: почти 33 % многозначных слов из корпуса SemCor имели более одного значения на дискурс, в корпусе DSO все слова встречались с более чем одной семантической меткой в рамках одного дискурса. В среднем в 39 % текстовых файлов, содержащих размеченное целевое многозначное слово, данное слово было размечено двумя и более тегами. По результатам экспериментов было установлено, что совместная встречаемость значений слов разных частей речи связана с разными типами полисемии.

1.3. Гипотеза One Sense per Collocation

Исследователями выделяется еще одна известная гипотеза — "One Sense per Collocation". Она заключается в том, что в рамках одного словосочетания многозначное слово тяготеет к какому-то одному конкретному значению.

В классической работе [Yarowsky 1993] обращается внимание на то, что использование локального контекста (ближайших линейных соседей слов) может быть полезным в решении задачи автоматического разрешения многозначности. В работе многозначными признаются «значения, которые обычно не переводятся с помощью одного слова на иностранный язык» [Yarowsky 1993: 266]. Рассматривается несколько типов словосочетаний: непосредственное линейное соседство слов в тексте, первое слово слева или справа, относящееся к определенной части речи, пары слов, связанные определенными синтаксическими отношениями. Доля словосочетаний с одними и теми же значениями в разных вхождениях в среднем составляет 95 %.

[2] https://www.sketchengine.eu/semcor-annotated-corpus/
[3] https://catalog.ldc.upenn.edu/LDC97T12

Гипотеза «одно значение на словосочетание» может помочь в разметке и создании алгоритмов автоматического разрешения многозначности. В статье [Martinez, Agirre 2000] авторы проверяют данную гипотезу для более сложной классификации значений на материале двух корпусов. Авторы утверждают, что для более тонких различий в значениях (то есть при наличии у слова более двух значений) гипотеза проявляется слабее (70 %). Кроме того, авторы отмечают важную особенность гипотезы: в рамках одного корпуса, тексты которого принадлежат, как правило, одному жанру, слова используются в рамках словосочетаний в одном значении. Однако при смене корпуса необходимо учитывать вариативность жанров и тем.

Отдельно следует упомянуть такой известный ресурс, как SyntagNet[4] [Maru et al. 2019], который содержит более 80 тысяч словосочетаний, размеченных по значениям. Создание такого рода ресурса подразумевает принятие гипотезы одного значения на словосочетание.

2. Анализ неравномерности распределения значений на семантическом корпусе русского языка

Авторами данной статьи были проведены собственные исследования на материале корпуса русскоязычных текстов с разметкой по значениям с целью проверить особенности распределения значений многозначных слов.

2.1. Описание корпуса

Процесс создания корпуса подробно описан в работе [Kirillovich et al. 2022]. Авторы использовали для разметки и помещения в корпус тексты средней длины, собранные в рамках проекта OpenCorpora[5]. Для разметки текстов по значениям были выбраны инвентари значений из тезауруса RuWordNet[6], аналога WordNet для русского языка [Loukachevitch et al. 2016].

Отобранные тексты были поделены на предложения, слова прошли процесс лемматизации, были сопоставлены с лексическими единицами RuWordNet. Корпус состоит из 807 текстов, число лемм составляет 109 893 (136 лемм на документ в среднем), число синсетов RuWordNet равно 8619.

[4] http://syntagnet.org/
[5] http://opencorpora.org/
[6] https://ruwordnet.ru/ru

2.2. Исследование неравномерности распределения значений на материале корпуса

2.2.1. Самое частотное значение (MFS)

Метод на основе самого частотного значения был протестирован в оригинальной статье [Kirillovich et al. 2022]. На обучающей части корпуса такой алгоритм показал значение точности **85,7 %**. Если не считать частотность значения на обучающих данных, а применять к тестовой выборке, то точность определения значения составляет **71,1 %**, т. е. в целом гипотеза подтверждается.

2.2.2. Гипотеза One Sense per Discourse

Эксперименты для проверки гипотезы «одно значение на дискурс» проводились в несколько этапов. На первом этапе были отобраны релевантные тексты, то есть такие тексты, которые содержат как минимум одно слово (рассматриваются леммы, а не словоформы), встретившееся в рамках документа-текста более одного раза. По каждому такому тексту собирались релевантные слова (леммы), а именно такие слова, которые встретились в рамках данного текста не менее двух раз. Далее отбирались многозначные слова, которые входят в состав RuWordNet. Процент случаев соблюдения гипотезы «одно значение на документ» подсчитывался следующим образом: в рамках выборки релевантных текстов, отобранных описанным выше способом, для каждого текста получали процент релевантных слов, встретившихся в данном тексте ровно в одном значении. Затем высчитывалось среднее арифметическое всех таких значений относительно длины выборки релевантных текстов. Статистика по данному эксперименту приводится в таблице 1:

Таблица 1

Процент релевантных текстов	≈ 94 %
Средний процент релевантных слов в тексте	3,63 %
Одно значение на документ	93 %

2.2.3. Гипотеза One Sense per Collocation

С помощью библиотеки **spaCy**[7] для языка программирования Python из корпуса с семантической разметкой для русского языка были выделены словосочетания — пары слов, связанные синтаксическими отношениями. Рассматриваются слова трех морфологических категорий — глагол, существительное и прилагательное.

[7] https://spacy.io/

Затем были отобраны те словосочетания, которые содержат хотя бы одно многозначное слово из RuWordNet и которые в лемматизированной форме встретились в корпусе более одного раза. Всего было собрано 2066 таких словосочетаний.

По подсчетам получается, что в 84,46 % отобранных пар либо оба слова многозначны и встретились в корпусе каждое в одном значении в рамках словосочетания, либо одно слово однозначное, а второе многозначное и встречалось только в одном значении в пределах словосочетания. Еще в 7,07 % пар закономерность одного значения на словосочетание не соблюдена только для одного из многозначных слов пары.

2.2.4. Анализ результатов

Большинство случаев нарушения гипотезы «одного значения на документ» можно отнести к ошибкам и сложностям в ручной разметке употреблений многозначных слов по значениям. Например, в следующих двух предложениях лексема *вычисления* употребляется, вероятнее всего, в одном значении — 'компьютерные вычисления', однако в корпусе им приписаны разные семантические теги:

(1) *В 2010 году видеокарты nVidia и ATI будут обеспечивать аппаратную реализацию* ***вычислений****, связанных с искусственным интеллектом в играх.* **Значение**: 140919-N КОМПЬЮТЕРНЫЕ ВЫЧИСЛЕНИЯ.

(2) *По статистике, до 90%* ***вычислений*** *в играх связаны с такими «рутинными интеллектуальными операциями», как определение прямой видимости противника, поиск кратчайшего пути до цели и т. д.* **Значение**: 111735-N ВЫЧИСЛИТЬ, ИСЧИСЛИТЬ.

Однако были случаи, когда одно из употреблений многозначного слова в тексте относилось к устойчивому словосочетанию, а другое — нет, и в таких случаях разметка разными синсетами оправдана. Пример:

(1) ***В связи*** *со стихийным бедствием произошли перебои в электроснабжении, нарушилась телефонная* ***связь****.*

Также встретились комбинации значений многозначных слов в рамках одного текста вне устойчивых сочетаний, например: 'момент времени' vs. 'промежуток времени' для лексемы *время*, 'федеративное государство' vs. 'общественное объединение' для *федерации*, 'бразды правления' и 'орган власти' для лексемы *власть*.

Если рассматривать случаи несоблюдения гипотезы «одного значения на словосочетание», то можно выделить следующие категории: 1) устойчивые словосочетания, в которых затруднительно приписать значения словам-компонентам (*Академия наук*, *Академия музыки*, *ценная бумага*, *войти/входить в состав*, *вступить в брак*)

[Телия 1996; Баранов, Добровольский 2022]; 2) словосочетания с прилагательными *новый*, *старый*, *великий*, *крупный*; 3) словосочетания с существительными, обозначающими абстрактные понятия (*вид*, *власть*, *время*, *высота*, *место*); 4) словосочетания, состоящие из двух глаголов (например {*вестись, быть*}); 5) словосочетания, многозначные по своей сути: *выборы президента* — речь может идти как о выборах президента страны, так и о выборах президента организации; 6) географические обозначения: *Новая Гвинея*, *Южная Корея*; 7) словосочетания с глаголами типа *быть* (выделено 6 значений в RuWordNet), *жить* (выделены значения 'обитать' и 'существовать', которые зачастую трудно разделить), а также *мочь*, *находиться*.

Таким образом, стоит отметить, что для успешного решения задачи автоматического разрешения лексической многозначности недостаточно применять алгоритм пословного снятия многозначности, лежащий в основе большинства предлагаемых методов, необходимы также словари устойчивых словосочетаний с приписанной словосочетанию единой меткой значения.

3. Заключение

В статье были рассмотрены три типа неравномерностей распределения значений слов в корпусе: 1) самое частотное значение; 2) «одно значение на документ» и 3) «одно значение на словосочетание», приведена краткая история изучения вопроса.

Также описаны эксперименты на материале корпуса текстов на русском языке объемом более 209 000 лемм. Результаты экспериментов показали, что гипотезы «одного значения на документ» и «одного значения на словосочетание» действительно подтверждаются для большинства употреблений многозначных слов. Исключения носят точечный характер и не в последнюю очередь связаны с ошибками/трудностями задачи ручной разметки многозначных слов по значениям в рамках связного текста.

СПИСОК ЛИТЕРАТУРЫ

1. *Баранов А., Добровольский Д.* Аспекты теории фразеологии. Litres, 2022.
2. *Телия В.Н.* Русская фразеология. Семантический, прагматический и лингвокультурологический аспекты. М., 1996.
3. *Bhingardive S., Singh D., Rudramurthy V., Redkar H., Bhattacharyya P.* Unsupervised most frequent sense detection using word embeddings // Proceedings of the 2015 conference of the North American Chapter of the Association for Computational Linguistics. 2015. P. 1238–1243.
4. *Calvo H., Gelbukh A.* Is the most frequent sense of a word better connected in a semantic network? // Advanced Intelligent Computing Theories and Applications: 11th International Conference, ICIC 2015. Springer, 2015. P. 491–499.

5. *Carpuat M.* One translation per discourse // Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. 2009. P. 19–27.
6. *Gale W.A., Church K., Yarowsky D.* One sense per discourse // Speech and Natural Language: Proceedings of a Workshop. 1992. P. 233–237.
7. *Hauer B., Luan Y., Kondrak G.* You shall know the most frequent sense by the company it keeps // 13th International Conference on Semantic Computing (ICSC). 2019. P. 208–215.
8. *Hristea F., Colhon M.* The long road from performing word sense disambiguation to successfully using it in information retrieval: An overview of the unsupervised approach // Computational Intelligence. 2020. V. 36. № 3. P. 1026–1062.
9. *Kirillovich A., Loukachevitch N., Kulaev M., Bolshina A., Ilvovsky D.* Sense-Annotated Corpus for Russian // Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria. 2022. P. 130–136.
10. *Krovetz R.* More than one sense per discourse // NEC Princeton NJ Labs., Research Memorandum. 1998. V. 23.
11. *Loukachevitch N., Mischenko N.* Evaluation of approaches for most frequent sense identification in Russian // 7th International Conference, AIST 2018, Springer, 2018. P. 99–110.
12. *Loukachevitch N., Lashevich G., Gerasimova A., Ivanov V., Dobrov V.* Creating Russian WordNet by conversion // Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue". 2016. P. 405–415.
13. *Martinez D., Agirre E.* One Sense per Collocation and Genre/Topic Variations // 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. 2000. P. 207–215.
14. *Maru M., Scozzafava F., Martelli F., Navigli R.* SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019. P. 3534–3540.
15. *Miller G.A.* WordNet: An electronic lexical database. MIT press, 1998.
16. *Preiss J., Dehdari J., King J., Mehay D.* Refining the most frequent sense baseline // Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. 2009. P. 10–18.
17. *Rapp R.* Utilizing the One-Sense-per-Discourse Constraint for Fully Unsupervised Word Sense Induction and Disambiguation // LREC. 2004.
18. *Tang G., Sennrich R., Nivre J.* An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation // Proceedings of the Third Conference on Machine Translation. 2018. P. 26–35.
19. *Yarowsky D.* One sense per collocation // Proceedings of the workshop on Human Language Technology. 1993. P. 266–271.
20. *Yarowsky D.* Unsupervised word sense disambiguation rivaling supervised methods // 33rd annual meeting of the association for computational linguistics. 1995. P. 189–196.

REFERENCES

1. Baranov A., Dobrovol'skii D. Aspekty teorii fraseologii [Aspects of phraseological theory]. *Litres*, 2022.
2. Teliya V.N. Russkaya frazeologiya. Semanticheskii, pragmaticheskii i lingvokul'turologicheskii aspekty [Russian Phraseology. Semantic, pragmatic and linguistic-cultural aspects]. Moscow, *Yazyki russkoi kul'tury*, 1996. 191 p.
3. Bhingardive S., Singh D., Rudramurthy V., Redkar H., Bhattacharyya P. Unsupervised most frequent sense detection using word embeddings. *Proceedings of the 2015 con-*

ference of the North American Chapter of the Association for Computational Linguistics, 2015, pp. 1238–1243.
4. Calvo H., Gelbukh A. Is the most frequent sense of a word better connected in a semantic network? *Advanced Intelligent Computing Theories and Applications: 11th International Conference, ICIC 2015. Proceedings, Cham: Springer*, 2015, pp. 491–499.
5. Carpuat M. One translation per discourse. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 2009, pp. 19–27.
6. Gale W.A., Church K., Yarowsky D. One sense per discourse. *Speech and Natural Language: Proceedings of a Workshop Held at Harriman*, 1992, pp. 233–237.
7. Hauer B., Luan Y., Kondrak G. You shall know the most frequent sense by the company it keeps. *13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 208–215.
8. Hristea F., Colhon M. The long road from performing word sense disambiguation to successfully using it in information retrieval: An overview of the unsupervised approach. *Computational Intelligence*, 2020, v. 36, № 3, pp. 1026–1062.
9. Kirillovich A., Loukachevitch N., Kulaev M., Bolshina A., Ilvovsky D. Sense-Annotated Corpus for Russian. *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria*, 2022, pp. 130–136.
10. Krovetz R. More than one sense per discourse // NEC Princeton NJ Labs., Research Memorandum. 1998. V. 23.
11. Loukachevitch N., Mischenko N. Evaluation of approaches for most frequent sense identification in Russian. *7th International Conference, AIST 2018, Springer* 2018, pp. 99–110.
12. Loukachevitch N., Lashevich G., Gerasimova A., Ivanov V., Dobrov V. Creating Russian WordNet by conversion. *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*, 2016, pp. 405–415.
13. Martinez D., Agirre E. One Sense per Collocation and Genre/Topic Variations. *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 207–215.
14. Maru M., Scozzafava F., Martelli F., Navigli R. SyntagNet: Challenging supervised word sense disambiguation with lexical-semantic combinations. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3534–3540.
15. Miller G.A. WordNet: An electronic lexical database. *MIT press*, 1998.
16. Preiss J., Dehdari J., King J., Mehay D. Refining the most frequent sense baseline. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 2009, pp. 10–18.
17. Rapp R. Utilizing the One-Sense-per-Discourse Constraint for Fully Unsupervised Word Sense Induction and Disambiguation. *LREC*, 2004.
18. Tang G., Sennrich R., Nivre J. An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 26–35.
19. Yarowsky D. One sense per collocation. *Proceedings of the workshop on Human Language Technology*, 1993, pp. 266–271.
20. Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. *33rd annual meeting of the association for computational linguistics*, 1995, pp. 189–196.

Поступила в редакцию 10.05.2023
Принята к публикации 17.10.2023
Отредактирована 07.11.2023

Received 10.05.2023
Accepted 17.10.2023
Revised 07.11.2023

ОБ АВТОРАХ

Зарипова Диана Александровна — аспирант кафедры теоретической и прикладной лингвистики филологического факультета Московского государственного университета имени М.В. Ломоносова; diana.ser.sar96@gmail.com

Лукашевич Наталья Валентиновна — доктор технических наук, кандидат физико-математических наук, профессор кафедры теоретической и прикладной лингвистики филологического факультета Московского государственного университета имени М.В. Ломоносова, ведущий научный сотрудник НИВЦ МГУ; louk_nat@mail.ru

ABOUT THE AUTHORS

Diana Zaripova — PhD student, Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University; diana.ser.sar96@gmail.com

Natalia Loukachevitch — Prof. Dr., Department of Theoretical and Applied Linguistics, Faculty of Philology, Lomonosov Moscow State University; louk_nat@mail.ru