



МГУ имени М.В. Ломоносова

Научно-исследовательский вычислительный центр

Лаборатория анализа информационных ресурсов

Отчет за 2021 год

- **Тема:**
Методы структурирования трудноформализуемых предметных областей на основе автоматизированного формирования больших графов знаний и онтологий по разнородным потокам текстовых данных
- Этап 2. (2021 г.) Разработка методов автоматизированного формирования больших лингвистических онтологий предметной области, методов извлечения сложных типов именованных сущностей, исследование методов автоматизированного формирования графов знаний

Кадровый состав

- Всего 32 чел., из них 9 (3,6 ставки = $2*1,0+1*0,5+5*0,2+1*0,1$) науч. сотрудников, 4 докторов наук, 12 канд. наук, 1 науч. сотр. без степени, 23 (11,15 = $9*1,0+5*0,25+9*0,1$) ИТР, 4 аспирантов, 4 студентов
- **ЛАИР (лаб. 404):**
 - Добров Б.В., Лукашевич Н.В., Мячина А.В., Сидоров А.В., Тихомиров М.М. [1/2], Чернышев Д.И., Штернов С.В.
 - + Рубцова Ю.В. (РНФ Швядас В.К.)
 - + Кремнева М.Д. (420-ЛКБ)
 - + Болдырев И.А. [1/10], Волков Е.В. (407) [1/4], Архипенко К.В. [1/10], Майоров В.Д. [1/10]
- **РНФ (Лукашевич Н.В.) [1/10] :**
 - Артемова Е.Л., Батура Т.В., Браславский П.И., Иванов В.В., Тутубалина Е.В., Шелманов А.О.
- **РНФ (Швядас В.К., лаб. 407) [1/10 – 1/4] :**
 - Швядас В.К. , Балдин С.М., Кирилин Е.М., Коннов С.И., Копылов К.Е., Нилов Д.К., Пушкарев С.В., Смирнов В.А., Чухарев Ф.С.
- **ЛКБ (лаб. 420):**
 - Быстрицкий Д.К., Быстрицкий Н.Д., Макаров-Землянский Н.В.

Планы на 2021 год

Исследование методов интеграции в лингвистические онтологии новых текстовых объектов, именованных сущностей

- **исследовать методы выявления сложных текстовых образов именованных сущностей** – вложенных, разрывных, неполных, в том числе разработать методики разметки таких сущностей для возможности применения методов машинного обучения;
- **разработать методы автоматического пополнения больших лингвистических онтологий** (с небольшим количеством фиксированных отношений) предметной области, на примере естественно-научных и технических текстов;
- **разработка методов глубокого машинного обучения для интегрирования большой номенклатуры типов именованных сущностей с понятиями онтологии;**
- **исследовать методы извлечения значимых текстовых фрагментов в текстах для задач абстрактного аннотирования**, выявления аргументации;
- **исследовать методы формирования графов значимых для содержания текста текстовых объектов**, включающих понятия, описанные в онтологии, а также терминоподобные слова и словосочетания, в том числе из внешних источников энциклопедического типа.

Выявление сложных текстовых образов именованных сущностей – вложенных, разрывных, неполных



- Сформирован новый датасет на русском языке NEREL с размеченными именованными сущностями и отношениями между ними. Особенностью NEREL является разметка вложенных именованных сущностей и их отношений. Отношения между сущностями размечаются в рамках связного текста и не ограничиваются уровнем предложения
- В NEREL поддерживается 29 типов именованных сущностей и 49 типов отношений между сущностями. Размечено более 56 тысяч сущностей и 39 тысяч отношений в более чем 900 документов Russian Wikinews
- Подготовлено научное соревнование в рамках Dialogue Evaluation 2022

Методы автоматического пополнения больших лингвистических онтологий (с небольшим количеством фиксированных отношений) предметной области

- Получен результат, что комбинации векторных представлений, обученных на общей предметной области, рассчитанные на больших текстовых коллекциях из сети Интернет, оказывают существенное влияние на качество пополнения таксономий, таких как WordNet, RuWordNet, Онтологии Естественных Наук и Технологий (ОЕНТ).
- Мета-векторные представления строятся с использованием автокодировщиков в комбинации с дополнительной функцией потерь, добавляющей информацию об онтологии (triplet loss).
- Предсказание гиперонимов реализовано через алгоритм машинного обучения, где входные признаки формируются на основании метрик близости между сущностями в онтологии и целевым словом, используя векторную модель.
В качестве алгоритма машинного обучения использовалась логистическая регрессия, которая обучалась на задаче классификации, то есть, предсказания, является ли обрабатываемый концепт гиперонимом целевого слова.

Методы глубокого машинного обучения для интегрирования большой номенклатуры типов именованных сущностей с понятиями онтологии

Предсказание гиперонимии

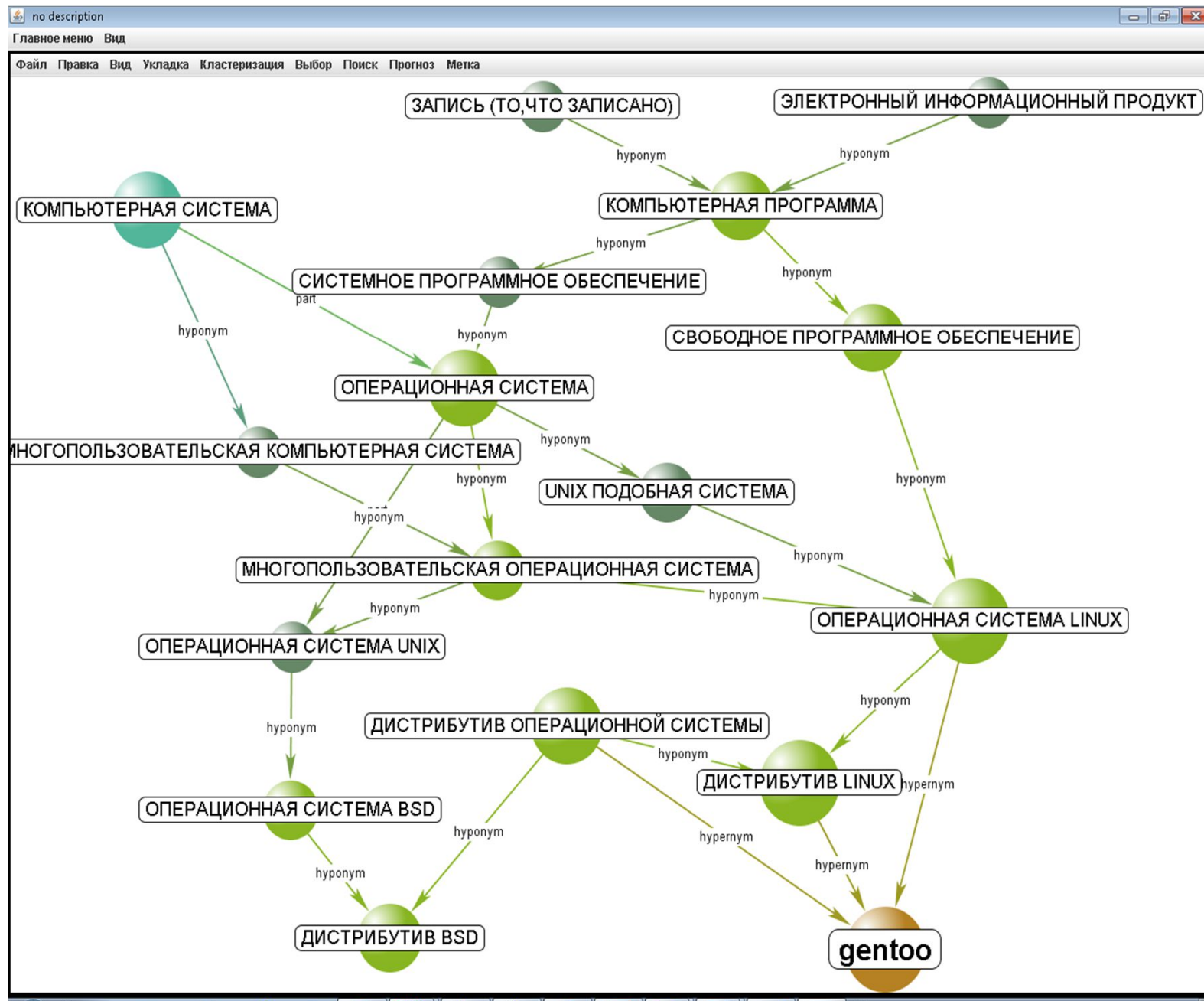
Слово

Количество

ДИСТРИБУТИВ LINUX	0.974 hypernym
ОПЕРАЦИОННАЯ СИСТЕМА LINUX	0.964 hypernym
ДИСТРИБУТИВ ОПЕРАЦИОННОЙ СИСТЕМЫ	0.943 hypernym
ОПЕРАЦИОННАЯ СИСТЕМА	0.943 hypernym
СВОБОДНОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ	0.863 hypernym
КОМПЬЮТЕРНАЯ ПРОГРАММА	0.854 hypernym
ДИСТРИБУТИВ BSD	0.753 hypernym
ОПЕРАЦИОННАЯ СИСТЕМА BSD	0.713 hypernym
МНОГОПОЛЬЗОВАТЕЛЬСКАЯ ОПЕРАЦИОННАЯ СИСТЕМА	0.661 hypernym
UNIX ПОДОБНАЯ СИСТЕМА	0.644 hypernym

КОМПЬЮТЕРНАЯ СИСТЕМА	0.925 part_whole
МНОГОПОЛЬЗОВАТЕЛЬСКАЯ КОМПЬЮТЕРНАЯ СИСТЕМА	0.474 part_whole
ВЕРСИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ	0.332 part_whole
ИНСТАЛЛЯЦИЯ ПРОГРАММЫ	0.309 part_whole

Предсказание гиперонимов и меронимов



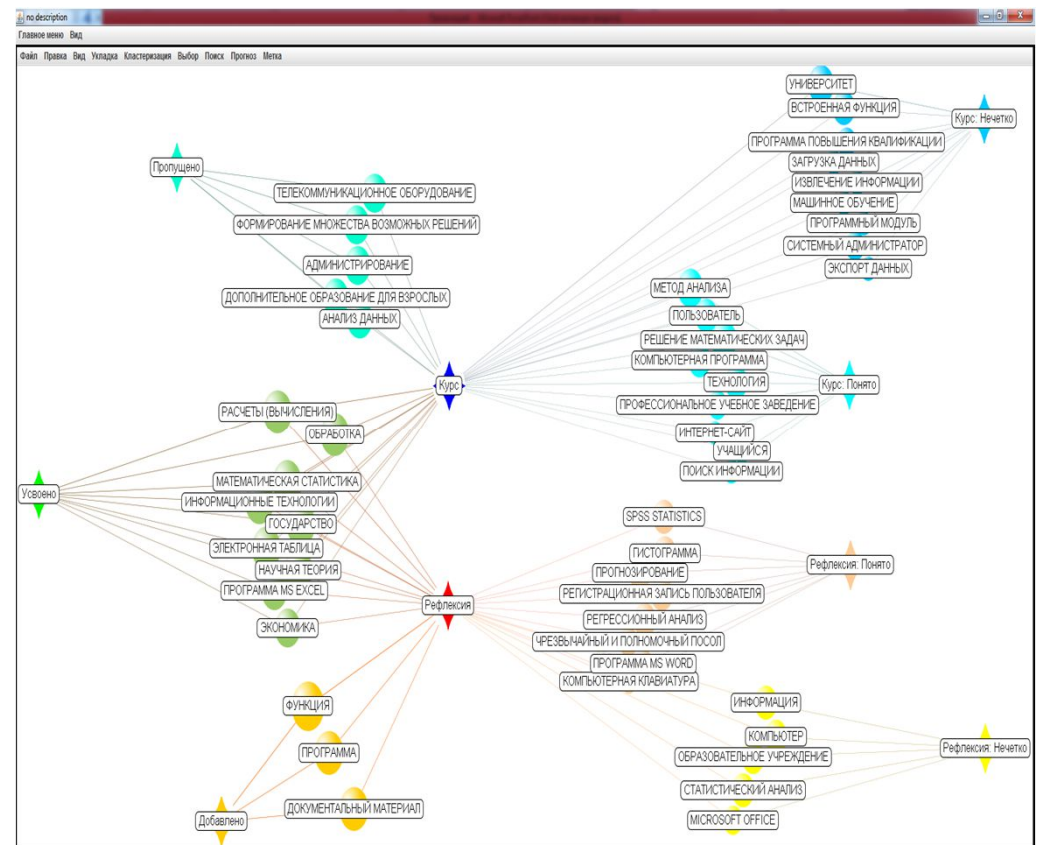
Методы наполнения «текстовых вершин» графа знаний, (фрагмент текста, содержащий неструктурированное знание по заданной теме)

- Цель – автоматизированное формирование «Википедии» большого размера по предметной области
- Исследовались методы абстрактного аннотирования извлечения значимых текстовых фрагментов с использованием современных нейросетевых подходов
- Реализован метод абстрактного реферирования новостей на основе автоматически сформированной обучающей коллекции, где в качестве абстрактной аннотации рассматривалась близкая по смыслу новость, опубликованная в качественном СМИ

Методы анализа текстовых материалов вида «резюме и вакансии», учебные курсы

- Для оценки компетенции в областях знаний важно понимать иерархию «пререквизитов» - понятий, которые необходимо усвоить, чтобы понимать некоторое новое понятие (Для понимания концепта «машинное обучение» необходимо иметь понимание о концептах «методы оптимизации», «линейная алгебра»)

- Для онтологии ОЕНТ получен результат, что отношения «пререквизит» могут автоматически выводиться по иерархии существующих отношений лингвистических онтологий типа RuTez, возможно, с добавлением небольшого количества отношений вручную



НИР (1)

- 28 мая 2020 г. - 30 декабря 2022 г. *РНФ 20-11-20166*
Автоматические методы построения и пополнения баз знаний на основе кросс-языковых технологий
Руководитель: Лукашевич Н.В. Участники НИР: ..., Тихомиров М.М.
- 1 октября 2019 г. - 1 октября 2022 г. *РФФИ 19-37-90119\19*
Автоматические методы улучшения качества векторных представлений лексических единиц за счет распространения контекстов по лексико-семантическим отношениям
Руководитель: Лукашевич Н.В. Отв. исполнитель: Тихомиров М.М.
- 19 февраля 2020 г. - 30 декабря 2022 г. *РФФИ 20-07-01059\20*
Автоматический анализ тональности текстов с множественными оценками на основе оценочных фреймов
Руководитель: Лукашевич Н.В. Участники НИР: ..., Тихомиров М.М., Штернов С.В.
- Чернышёв Д.И. получил грантовую поддержку Фонда «Интеллект» о **Улучшение качества автоматических методов порождения сложноструктурированных повествовательных текстов на основе интеграции в нейросетевые модели абстрактного реферирования общих и специальных знаний**

НИР (2)

- 17 марта 2021 г. - 31 декабря 2024 г. **РНФ 21-71-30003**
Разработка и реализация принципов суперкомпьютерного кодизайна для развития сквозных цифровых технологий и создания высокопроизводительных вычислительных платформ для предсказательного моделирования и применения искусственного интеллекта: от микромира, до живых организмов и Земной системы
Руководитель: Швядас В.К. Участники НИР: ... Волков Е.В. ... Добров Б.В., ... Лукашевич Н.В., ... Тихомиров М.М. ... Чернышев Д.И.
- 25 декабря 2020 г. – май 2022 г.
СЧ НИР «Задел-СООИ-1-М»
 - ***1й этап: декабрь 2020 г. – октябрь 2021 г.***
- 01 сентября 2021 г. – 31 мая 2023 г.
НИР «Задел-Запрос-1»
 - ***1й этап: сентябрь 2021 г. – декабрь 2021 г.***

Публикации, конференции, дипломники 2021

- Опубликовано
 - ЛАИР:
 - 7 статей в журналах
 - 16 работ в рецензируемых сборниках
- ПК (ОК) 22 конференции
- Дипломные работы: 12

Планы на 2022 год

Разработка методов автоматизированного формирования больших графов знаний предметной области

- исследовать **методы автоматического извлечения неизвестных отношений** (Open Information Extraction) с использованием методов машинного обучения и ранее созданных онтологических ресурсов;
- разработать методы автоматизированного формирования больших онтологий предметной области **с развитым набором отношений**;
- разработать **методы разрешения многозначности текстового выражения именованных сущностей** в разных документах;
- разработать **методы связывания различных текстовых вариантов извлеченных именованных сущностей** на основе результатов обработки больших текстовых коллекций.

Публикации ЛАИР (1)

- 2021 Multi-Step Transfer Learning for Sentiment Analysis
Golubev A., Loukachevitch N.
в журнале Lecture Notes in Computer Science, с. 209-217 DOI
- 2021 NEREL: набор данных на русском языке с вложенными именованными сущностями, отношениями и событиями
Денисов И.В., Рожков И.С., Лукашевич Н.В.
в журнале Интеллектуальные системы. Теория и приложения, издательство [б.и.] (М.), том 25, № 4, с. 243-249
- 2021 Use of Bert Neural Network Models for Sentiment Analysis in Russian
Golubev A.A., Loukachevitch N.V.
в журнале Automatic Documentation and Mathematical Linguistics, издательство Allerton Press Inc. (United States), том 55, № 1, с. 17-25 DOI
- 2021 Using Embedding-Based Similarities to Improve Lexical Resources
Loukachevitch N.V., Tikhomirov M.M., Parkhomenko E.A.
в журнале Lobachevskii Journal of Mathematics, издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation), том 42, № 7, с. 1532-1546 DOI
- 2021 Автоматическое извлечение мнений по вопросам репродуктивного поведения
Калабихина И.Е., Лукашевич Н.В., Банин Е.П., Алибаева К.В., Ребрей С.М.
в журнале Программные системы: теория и приложения, издательство Федеральное государственное бюджетное учреждение науки Институт программных систем им. А.К. Айламазяна Российской академии наук (Веськово), том 52, № 5
- 2021 Исследование моделей нейронных сетей типа BERT для анализа тональности текстов на русском языке
Голубев А.А., Лукашевич Н.В.
в журнале Научно-техническая информация. Серия 2: Информационные процессы и системы, издательство ВИНТИ (М.), том 1, с. 32-4
- 2021 Разработка автоматизированной системы пополнения таксономии на конкретную предметную область
Тихомиров Михаил
в журнале Интеллектуальные системы. Теория и приложения, издательство [б.и.] (М.), том 25, № 4, с. 250-254

Публикации ЛАИР (2)

- 2021 Automatic Generation of Annotated Collection for Recognition of Sentiment Frames
Solomatina Yuliana, Loukachevitch Natalia
// Lecture Notes in Computer Science, издательство Springer International Publishing AG (Cham, Switzerland), том 12602, с. 162-174 DOI
- 2021 Automatic Sentiment Analysis of Texts: The Case of Russian
Loukachevitch Natalia
// The Palgrave Handbook of Digital Russia Studies, место издания Palgrave Macmillan Ltd Cham, Switzerland, с. 501-516
- 2021 Comparing similarity of words based on psychosemantic experiment and RuWordNet
Solovyev V., Loukachevitch N.
// GWC 2021 - Proceedings of the 11th Global Wordnet Conference, с. 199-206
- 2021 Domain-specific Taxonomy Enrichment based on Meta-Embeddings
Tikhomirov Mikhail, Loukachevitch Natalia V.
// CEUR Workshop Proceedings, серия Proceedings of the XXIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021), том 3036, с. 285-298
- 2021 Exploring graph-based representations for taxonomy enrichment
Nikishina I., Loukachevitch N., Logacheva V., Panchenko A.
// GWC 2021 - Proceedings of the 11th Global Wordnet Conference, с. 126-136
- 2021 Linking Russian Words to Semantic Frames of FrameNet
Zlochevskaya D., Loukachevitch N., Nevzorova O.
// Proceedings of OSTIS-2021 conference, с. 293-298
- 2021 Meta-Embeddings in Taxonomy Enrichment Task
Tikhomirov M.M., Loukachevitch N.V.
// Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2021), место издания Москва, том 20 DOI
- 2021 Methods for Verification of Sentiment Frames
Matueva Irina, Loukachevitch Natalia
// Communications in Computer and Information Science, издательство Springer International Publishing AG (Cham, Switzerland), том 1427, с. 52-64 DOI

- 2021 NEREL: A Russian Dataset with Nested Named Entities, Relations and Events
Loukachevitch Natalia, Artemova Ekaterina, Batura Tatiana, Braslavski Pavel, Denisov Ilia, Ivanov Vladimir, Manandhar Suresh, Pugachev Alexander, Tutubalina Elena
// Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications, место издания INCOMA Ltd. Shoumen, BULGARIA
- 2021 RuThes Thesaurus for Natural Language Processing
Loukachevitch Natalia, Dobrov Boris
// The Palgrave Handbook of Digital Russia Studies, место издания Palgrave Macmillan Ltd Cham, Switzerland, с. 319-334
- 2021 Semantic Similarity of Words in RuWordNet Thesaurus and in Psychosemantic Experiment
Solovyev Valery, Loukachevitch Natalia
// Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics, серия Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics. Intercognsci 2020. Advances in Intelligent Systems and Computing, издательство Springer (New York), том 1358, с. 395-402
- 2021 TatWordNet: A Linguistic Linked Open Data-Integrated WordNet Resource for Tatar
Kirillovich Alexander, Shaekhov Marat, Galieva Alfiya, Nevzorova Olga, Ilvovsky Dmitry, Loukachevitch Natalia
// 3rd Conference on Language, Data and Knowledge (LDK 2021), серия Open Access Series in Informatics (OASIs), Schloss Dagstuhl – Leibniz-Zentrum für Informatik Dagstuhl, Germany, том 93
- 2021 Transfer Learning for Improving Results on Russian Sentiment Datasets
Golubev Anton, Loukachevitch Natalia
// Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2021), Москва, том 20, с. 268-277
- 2021 Use of Augmentation and Distant Supervision for Sentiment Analysis in Russian
Golubev Anton, Loukachevitch Natalia
// Text, Speech, and Dialogue, издательство Springer International Publishing (New York), с. 184-196
- 2021 Using ontology for natural sciences and technologies for vacancies analysis
Loukachevitch N.V., Komissarov A.A., Dobrov B.V., Shternov S.V.
// CEUR Workshop Proceedings, серия Proceedings of the International Workshop on Digital Technologies for Teaching and Learning (DTTL-2021), том 2090, с. 30-38
- 2021 Abstractive Summarization of Russian News Learning on Quality Media
Чернышев Даниил Иванович, Добров Борис Викторович
// Lecture Notes in Computer Science, издательство Springer International Publishing AG (Cham, Switzerland),

Публикации по теме (1)

- Голубев А.** Исследование моделей нейронных сетей типа BERT для анализа тональности текстов на русском языке / А.А. Голубев, Н.В. Лукашевич // Научно-техническая информация. Серия 2: Информационные процессы и системы, издательство ВИНТИ (М.). – 2021. - том 1, С. 32-41.
- Денисов И.** NEREL: набор данных на русском языке с вложенными именованными сущностями, отношениями и событиями / И.В. Денисов, И.С. Рожков, Н.В. Лукашевич // Интеллектуальные системы. Теория и приложения (ранее: Интеллектуальные системы по 2014, № 2, ISSN 2075-9460). - М. – 2021.- Том 25, № 4, С. 243-249.
- Тихомиров М.** Разработка автоматизированной системы пополнения таксономии на конкретную предметную область / Михаил Тихомиров // Интеллектуальные системы. Теория и приложения (ранее: Интеллектуальные системы по 2014, № 2, ISSN 2075-9460), М. – 2021. - том 25, № 4, С. 250-254.
- Chernyshev D.** Abstractive Summarization of Russian News Learning on Quality Media / Daniil Chernyshev, Boris Dobrov // Lecture Notes in Computer Science, издательство Springer International Publishing AG (Cham, Switzerland). – 2021. - том 12602, с. 96-104.
- Golubev A.** Use of Bert Neural Network Models for Sentiment Analysis in Russian / A.A. Golubev, N.V. Loukachevitch // Automatic Documentation and Mathematical Linguistics, Allerton Press Inc. (United States). – 2021. - Том 55, № 1, С. 17-25.
- Loukachevitch N.** Automatic Sentiment Analysis of Texts: The Case of Russian / Natalia Loukachevitch // The Palgrave Handbook of Digital Russia Studies, место издания Palgrave Macmillan Ltd Cham, Switzerland. – 2021. - С. 501-516.
- Loukachevitch N.** NEREL: A Russian Dataset with Nested Named Entities, Relations and Events / Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Iliia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, Elena Tutubalina // Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications, место издания INCOMA Ltd. Shoumen, BULGARIA. – 2021.
- Loukachevitch N.** RuThes Thesaurus for Natural Language Processing / Natalia Loukachevitch, Boris Dobrov // The Palgrave Handbook of Digital Russia Studies, Palgrave Macmillan Ltd Cham, Switzerland. – 2021. - С. 319-334.

Публикации по теме (2)

- Loukachevitch N.** Using Embedding-Based Similarities to Improve Lexical Resources / N.V. Loukachevitch, M.M. Tikhomirov, E.A. Parkhomenko // Lobachevskii Journal of Mathematics, издательство Kazanskii Gosudarstvennyi Universitet/Kazan State University (Russian Federation). – 2021. - том 42, № 7, С. 1532-1546.
- Loukachevitch N.** Using ontology for natural sciences and technologies for vacancies analysis / N.V. Loukachevitch, A.A. Komissarov, B.V.Dobrov, S.V. Shternov // CEUR Workshop Proceedings, серия Proceedings of the International Workshop on Digital Technologies for Teaching and Learning (DTTL-2021). – 2021. - том 2090, С. 30-38.
- Solovyev V.** Semantic Similarity of Words in RuWordNet Thesaurus and in Psychosemantic Experiment / Valery Solovyev, Natalia Loukachevitch // Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics, серия Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics. Intercognsci 2020. Advances in Intelligent Systems and Computing, Springer (New York) . – 2021. - том 1358, С. 395-402.
- Tikhomirov M.** Domain-specific Taxonomy Enrichment based on Meta-Embeddings / Mikhail Tikhomirov, Natalia V. Loukachevitch // CEUR Workshop Proceedings, серия Proceedings of the XXIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021). – 2021. - том 3036, С. 285-298.
- Tikhomirov M.** Meta-Embeddings in Taxonomy Enrichment Task / M.M. Tikhomirov, N.V. Loukachevitch // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2021) – М. – 2021. - том 20.
- Nikulin M.** Prospects of Using Biocatalysis for the Synthesis and Modification of Polymers / Nikulin, M.; V. Švedas // *Molecules* **2021**, 26, 2750. <https://doi.org/10.3390/molecules26092750>
- Drobot V.V.** PLUMED plugin integration into high performance pmemd program for enhanced molecular dynamics simulations. / V.V. Drobot, E.M. Kirilin, K.E. Kopylov, I.V. Pochinok, V.K. Švedas // Supercomputing Frontiers and Innovations, 2021, 8 N4.

Учебная деятельность

Лукашевич Н.В.

- ВМиК
 - Анализ текстовых данных и информационный поиск
 - Методы и средства информационного поиска
 - Машинное обучение в автоматической обработке текстов
- Филологический факультет
 - Лингвистические аспекты новых информационных технологий
 - Искусственный интеллект
 - Введение в информационный поиск для лингвистов
- МГТУ имени Н.Э.Баумана
 - Информационный поиск и извлечение информации из текстов)
 - Автоматическая обработка текстов