Секция 2.
ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ РЕШЕНИЯ ЛИНГВИСТИЧЕСКИХ ЗАДАЧ

УДК 81'32, 81'33

Астапенко Елизавета Валентинова
студент
Санкт-Петербургский
государственный университет
г. Могилев, Беларусь

Lizaveta Astapenka
Student
St. Petersburg State University
Mogilev, Belarus
li.astapenka@gmail.com

Митрофанова Ольга Александровна
кандидат филологических наук,
доцент
Санкт-Петербургский
государственный университет
г. Санкт-Петербург, Россия

Olga Mitrofanova
PhD in Philology,
Associate Professor
St. Petersburg State University
Saint Petersburg, Russia
o.mitrofanova@spbu.ru

ЛЕКСИКО-СЕМАНТИЧЕСКИЕ ОТНОШЕНИЯ В КОРПУСАХ И ЯЗЫКОВЫХ МОДЕЛЯХ ДЛЯ БЕЛОРУССКОГО ЯЗЫКА

В данной работе исследуется способность языковых моделей фиксировать семантические связи между словами. Описаны эксперименты по автоматическому предсказанию лексико-семантических отношений в *белорусском* языке с помощью различных моделей, а именно Word2Vec, BERT и генеративных языковых моделей. Качество предсказаний моделей определялось с помощью процедуры псевдодизамбигуации (Word2Vec) и экспертной оценки (BERT, Gemini 2.5 Pro). Полученные результаты могут быть применены для решения различных задач: для создания и обогащения лексических баз данных, исследования сочетаемости слов, улучшения машинного перевода, перефразирования, реферирования и других систем автоматической обработки белорусского языка.

Ключевые слова: лексико-семантические отношения; белорусский язык; Word2Vec; BERT; большие языковые модели.

LEXICAL-SEMANTIC RELATIONS IN CORPORA AND LANGUAGE MODELS FOR THE BELARUSIAN LANGUAGE

Our study examines the ability of language models to capture semantic relations between words. We describe experiments on automatic prediction of lexical-semantic relations in the *Belarusian* language using various models (Word2Vec, BERT, and generative language models). The performance of the models was evaluated by pseudo-disambiguation test (Word2Vec) as well as by expert evaluation method (BERT, Gemini 2.5 Pro). The results can be applied to create and enrich lexical databases, to analyse word co-occurrence, to improve machine translation, paraphrasing, summarization, and other systems related to automatic processing of the Belarusian language.

Key words: lexical-semantic relations; the Belarusian language; Word2Vec; BERT; large language models.

Современные нейросетевые модели позволяют решать разнообразные задачи в области обработки естественного языка и исследовать не только морфологические и синтаксические, но и семантические свойства языковых выражений. Распространение моделей распределенных векторных вложений расширяет возможности автоматического распознавания лексико-семантических отношений в корпусах текстов, что особенно актуально для языков, для которых в открытом доступе еще не представлены лексические базы данных и компьютерные тезаурусы. Для белорусского языка существуют различные онлайн-словари (например, Verbum [1], Slounik [2]), инструменты для морфологической и синтаксической разметки (Stanza [3], UD-Pipe [4]). В 2013 г. в рамках проекта «Экспериментальный корпус белорусского языка» был реализован лемматизатор YABC_Tagger [5]. На платформе Corpus.by [6] представлены разработанные в лаборатории распознавания и синтеза речи ОИПИ НАН Беларуси сервисы для автоматической обработки текстов на белорусском языке. В институте языкознания имени Якуба Коласа проводятся работы по созданию Национального корпуса («Беларускі N-корпус») [7], Грамматической базы белорусского языка и т. д. Самые известные NLP-инструменты для белорусского языка представлены в репозитории [8]. Однако на сегодняшний день в открытом доступе не размещены языковые модели, обученные на белорусских корпусах и предназначенные для решения задач лексической семантики. Таким образом, актуальность нашего исследования определяется необходимостью восполнения лакун в области развития общедоступных автоматизированных систем, позволяющих выявлять семантические отношения в лексике, а также большим диапазоном задач в компьютерной лингвистике, для решения которых такие модели могут быть использованы. Цель нашей работы состоит в обучении и оценке языковых моделей белорусского языка для предсказания регулярных парадигматических и синтагматических отношений, наблюдаемых в корпусах текстов.

Для обучения языковых моделей был сформирован корпус белорусского языка. Текстовая коллекция включает в себя три сегмента:

- корпус белорусского языка в проекте Universal Dependencies (UD) объемом 25 231 предложение (305 417 токенов), содержащий художественные, публицистические тексты, тексты Википедии и социальных медиа [9];
- Belacorpus объемом 246 текстов (1 535 047 токенов), включающий тексты различных жанров за период 1987–2010 гг. [10];
- корпус белорусского языка из коллекции Лейпцигского университета (Web 300K) объемом 300 000 предложений, источником которого служат различные веб-сайты на белорусском языке [11].

Итоговый объем корпуса составил 413 012 предложений (5 662 829 токенов). Объединенный корпус представлен в двух версиях: корпус без разметки (для возможности отбора контекстов) и размеченный корпус (для дальнейшего обучения моделей). Корпус UD имел предварительную морфосинтаксическую разметку, лемматизация и частеречная разметка корпусов Belacorpus [10] и Web-300K-2015 [11] была проведена нами с помощью библиотеки Stanza [3].

Для проведения экспериментов по дифференциации синонимов и антонимов был сформирован набор данных, состоящий из 1939 синонимических рядов, которые были извлечены из онлайн-словаря синонимов [12], а также 597 антонимических пар из словаря «Слоўнік лексічных формаў (сінонімы, амонімы, антонімы, паронімы, амографы, амафоны)» [13]. Для повышения сбалансированности данных набор антонимов был расширен за счет автоматического подбора к ним синонимов, благодаря чему общее количество антонимических пар увеличилось до 1624.

На первом этапе предсказание лексико-семантических отношений проводилось с помощью моделей Word2Vec [14]. В результате серии экспериментов был произведен выбор оптимальных параметров обучения моделей skip-gram (vector_size = 250, window = 5, min_count = 5, epochs = 5) и cbow (vector_size = 300, window = 5, min_count = 5, epochs = 5). Оценка качества предсказаний опирается на процедуру псевдодизамбигуации [15], по результатам которой модели достигают высоких значений точности для списков как с целевым существительным, так и с прилагательным (78 % и 90 % для модели skip-gram, 83 % и 92 % для модели cbow). Для визуализации работы моделей на платформе Hugging Face с помощью библиотеки Streamlit было разработано приложение [16]. Пример предсказаний моделей приведен в таблице.

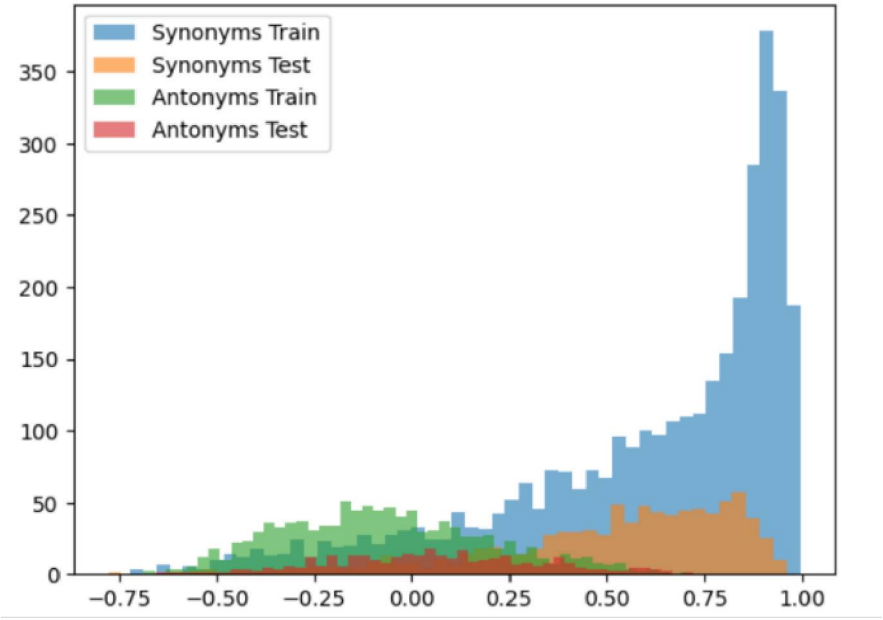
Пример выходных данных для слова *мова* (*язык*)

skip-gram, ADJ	**Cos**	**CBOW, NOUN**	**Cos**
*англійскі*_ADJ	0,593	*правапіс*_NOUN	0,589
*рускі*_ADJ	0,592	*філалогія*_NOUN	0,584
*украінскі*_ADJ	0,578	*літаратура*_NOUN	0,562
*царкоўнаславянскі*_ADJ	0,574	*арфаграфія*_NOUN	0,552
*старабеларускі*_ADJ	0,554	*маўленне*_NOUN	0,54

Сиамские нейронные сети

Для улучшения качества предсказаний использовался подход, объединяющий обращение к дополнительным лексическим ресурсам и применение сиамских нейронных сетей. За основу был взят алгоритм Siamese Semantic Vectors [17] с компонентом Word2Vec skip-gram, более чувствительным к контексту по сравнению с cbow. В качестве обучающих данных выступали списки синонимов и антонимов для белорусского языка, составленные по материалам словарей. Общее число пар синонимов составило 4038, антонимов – 1624. Эти данные были разделены на обучающую и тестовую выборки в процентном соотношении 80:20.

Для оценки моделей была использована метрика, определяющая долю от всех синонимов (антонимов), которую модель смогла рассчитать правильно. Были введены две границы так, что верными считаются все предсказания косинусной меры от 0,6 до 1 для синонимов и от -0,2 до -1 для антонимов. Результаты представлены на рисунке.



Распределение значений косинусного сходства
для синонимических и антонимических пар после обучения нейросети

Приведем в качестве примера изменение значения косинусного сходства для пары синонимов *палахлівы*_ADJ (*пугливый*) – *баязлівы*_ADJ (*трусливый*), а также для пары антонимов *смелы*_ADJ(*смелый*) – *баязлівы*_ADJ (*трусливый*). После обучения нейронной сети оно увеличилось с 0,8713 до 0,9293 в первом случае и уменьшилось с 0,5167 до -0,4429 во втором.

Дообучение модели BERT

Способность моделей BERT к построению контекстуализированных эмбеддингов оказывается особенно полезной при автоматическом подборе синонимических замен слов в предложениях. Задачу предсказания моделью синонимов можно свести к задаче маскированного языкового моделирования, при котором BERT предлагает наиболее вероятные слова вместо специального токена [MASK] в последовательности. Следовательно, необходимо научить модель предсказывать для конкретного контекста не просто возможные варианты заполнения позиции маскированного токена, а именно синонимичные. В этом эксперименте тренировочные данные включили в себя 1203 синонимические пары. Кроме того, были составлены тестовые данные объемом в 100 предложений (по 25 контекстов на 4 основные части речи (NOUN, ADJ, VERB, ADV)).

Для белорусского языка существует несколько многоязычных и одноязычных предобученных моделей, среди которых для дальнейших экспери-

ментов была выбрана модель roberta-small-belarusian (1 язык, 15.7 млн параметров) [18], показавшая лучшие результаты в тестах. При дообучении модели гиперпараметры принимали следующие значения: num_train_epochs = 6 (число эпох), per_device_train_batch_size = 16 (размер батча), learning_rate = 5e-5 (скорость обучения).

Дополнительные методы улучшения качества предсказаний предполагают сравнение эмбеддингов предложений с помощью Sentence Transformer (SBERT) с предварительным добавлением словарных синонимов к списку уже подобранных моделью кандидатов. Поскольку для белорусского языка модели вида SBERT отсутствуют, были выбраны многоязычные модели LaBSE (110 языков, 471 млн параметров) [19] и paraphrase-multilingual-MiniLM-L12-v2 (50 языков, 118 млн параметров) [20]. На тестовой выборке было доказано преимущество первой модели.

Итоговый алгоритм включает в себя нескольких этапов:

- на вход дообученная модель roberta-small-belarusian получает предложение со спецтокеном [MASK] и предсказывает возможные варианты, которые могли бы стоять на его месте;
- к списку также добавляются кандидаты из словаря, которые были найдены в нем по лемме целевого слова, полученной с помощью библиотеки Stanza;
- на место маски в предложение по очереди подставляются варианты замен из списка;
- модель LaBSE сравнивает эмбеддинги предложений и ранжирует кандидатов в зависимости от значения косинусной меры.

В качестве примера выходных данных можно привести предложенные моделью синонимические замены для целевого слова *асаблівасць* в контексте *Была ў Фёдара яшчэ адна рэдкая* [MASK]: *па сваёй уласнай ініцыятыве ён ніколі нікога не падвозіў, не падбіраў*:

- *асаблівасць*: 1,0000
- *адметнасць*: 0,9986
- *характарыстыка*: 0,9974
- *уласцівасць*: 0,9972
- *своеасаблівасць*: 0,9961.

Для оценки качества предсказаний модели до процедуры дообучения и после нами были проведены два опроса. В анкеты были внесены тестовые предложения вместе с полученными для них ответами модели. Все данные были разделены случайным образом на пять выборок. Для оценки того, насколько модель улавливает семантические связи между словами, экспертам нужно было поставить баллы от 0 до 5. В качестве экспертов выступали носители белорусского языка. Общее количество информантов для каждой из анкет составило 25 человек. На основании полученных результатов определялась согласованность экспертов, средняя и относительная оценки. Значение согласованности варьируется от 0,75 до 0,88, что интерпретируется как

значительный уровень согласованности мнений экспертов [21]. По результатам оценки, проведенной для тестовых выборок в целом и всех предложений в отдельности, можно сделать вывод о том, что задача предсказания синонимов в контексте для белорусского языка с помощью модели BERT решается с достаточно высокой точностью.

Генерация синонимов с помощью больших языковых моделей

В связи с распространением больших языковых моделей (LLM) мы решили оценить их способность предсказывать синонимы в контексте для белорусского языка. Работа моделей проверялась на составленном нами тестовом наборе данных, включающем в себя 100 предложений. Среди многоязычных больших языковых моделей были выбраны наиболее известные и доступные в настоящее время, а именно DeepSeek-V3, GPT-4o, Claude 3.5 Haiku, Gemini 2.0 Flash, Gemini 2.5 Pro. Затем был составлен базовый промпт на белорусском языке и получены ответы моделей. На основе анализа ошибок и совпадений предложенных вариантов со словарными данными для дальнейшей экспертной оценки использовалась модель Gemini 2.5 Pro [22]. Тестовые данные были разделены на 5 выборок по 20 предложений. Экспертам предлагалось поставить 0 или 1 для каждого варианта в зависимости от того, может ли он быть использован в качестве замены в предложении. Для всех выборок рассчитывалась средняя оценка, а также согласованность экспертов (с помощью коэффициента каппа Флейса, как и в случае с моделями BERT). В группу экспертов вошли носители белорусского языка (84 человека). Согласованность принимает значения в диапазоне от 0,49 до 0,59, что является достоверным результатом для гуманитарных областей [21]. Средняя оценка варьируется между 50 и 60 баллами для всех выборок, что свидетельствует о том, что модель не в полной мере справляется с задачей генерации синонимов в контекстах для белорусского языка.

В результате исследования было проведено обучение языковых моделей разных архитектур для белорусского языка и их оценка в задаче предсказания парадигматических и синтагматических отношений, наблюдаемых в корпусах текстов. Проделанная работа позволяет заложить основу для последующей разработки более сложных лингвистических систем, охватывающих разнообразные семантические связи в белорусском языке. Обученные и протестированные нами модели могут быть не только интегрированы в подобные системы, но и применены в таких областях, как информационный поиск, машинный перевод, лингводидактика и т. д.

ЛИТЕРАТУРА

1. Verbum [Electronic Resource]. URL: https://verbum.by/ (дата обращения: 15.07.2025).
2. Slounik. URL: https://slounik.org/ (дата обращения: 15.07.2025).
3. Stanza. URL: https://stanfordnlp.github.io/stanza/index.html (дата обращения: 15.07.2025).

4. UD-Pipe. URL: https://ufal.mff.cuni.cz/udpipe/2/models (дата обращения: 15.07.2025).
5. YABC_Tagger. URL: https://github.com/poritski/YABC_Tagger/blob/master/docs/readme_BE.md (дата обращения: 15.07.2025).
6. Corpus.by. URL: https://corpus.by/index.php?lang=be (дата обращения: 15.07.2025).
7. Беларускі N-корпус. URL: https://bnkorpus.info/ (дата обращения: 15.07.2025).
8. be_nlp_speech_resources. URL: https://github.com/navalnica/be_nlp_speech_resources (дата обращения: 15.07.2025).
9. Shishkina Y., Lyashevskaya O. Sculpting enhanced dependencies for Belarusian // Analysis of Images, Social Networks and Texts: 10th International Conference. AIST 2021. Cham: Springer, 2022. Vol 13217. P. 137–147.
10. Belacorpus. URL: https://github.com/Belarusian-Corpus (дата обращения: 15.07.2025).
11. Belarusian Corpora. Wortschatz.uni-leipzig.de. URL: https://wortschatz.uni-leipzig.de/en/download/Belarusian (дата обращения: 15.07.2025).
12. Клышка М. К. Слоўнік сінонімаў і блізказначных слоў / пад рэд. Л. А. Антанюк. 2-е выд., выпр. і дапоўн. Менск: Вышэйшая школа, 1993. 445 с.
13. Хвалей Я. І., Шарпіла У. В. Слоўнік лексічных формаў (сінонімы, амонімы, антонімы, паронімы, амографы, амафоны). Мн.: Парадокс, 2004. 416 с.
14. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. URL: https://arxiv.org/abs/1301.3781 (дата обращения: 15.07.2025).
15. Petrushenko L., Mitrofanova O. Predicting Style-Dependent Collocations in Russian Text Corpora // Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN 2023). 2023. P. 79–89.
16. Lexical-semantic Calculator for the Belarusian Language. https://huggingface.co/spaces/lizaastapenka/lexical-semantic-calculator-for-the-Belarusian-language (дата обращения: 15.07.2025).
17. Siamese Semantic Vectors. URL: https://github.com/maxwelljohn/word-vector-remapping (дата обращения: 15.07.2025).
18. roberta-small-belarusian. URL: https://huggingface.co/KoichiYasuoka/roberta-small-belarusian (дата обращения: 15.07.2025).
19. LaBSE. URL: https://huggingface.co/sentence-transformers/LaBSE (дата обращения: 15.07.2025).
20. Paraphrase-multilingual-MiniLM-L12-v2. URL: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (дата обращения: 15.07.2025).
21. McHugh M. L. Interrater reliability: the kappa statistic // Biochemia medica. 2012. Vol. 22. № 3. P. 276–282.
22. Gemini. URL: https://deepmind.google/technologies/gemini/ (дата обращения: 15.07.2025).