DOI: 10.25140/2411-5363-2025-1(39)-151-160
УДК 004.855.2

Дмитро Олегович Журко[1], Ірина Володимирівна Білоус[2]

[1] аспірант кафедри інформаційних технологій та програмної інженерії
Національний університет «Чернігівська політехніка» (Чернігів, Україна)
E-mail: dm.zhurko@gmail.com. **ORCID:** https://orcid.org/0009-0001-4192-7780

[2] кандидат технічних наук, доцент, завідувач кафедри інформаційних технологій та програмної інженерії
Національний університет «Чернігівська політехніка» (Чернігів, Україна)
E-mail: iryna.bilous.it@gmail.com. **ORCID:** https://orcid.org/0000-0003-3092-678X. **ResearcherID:** G-3887-2014

ВИКОРИСТАННЯ МОДЕЛЕЙ ВКЛАДАННЯ СЛІВ В ОБРОБЦІ ПРИРОДНОЇ МОВИ

У статті представлено результати науково-методичного дослідження, присвяченого застосуванню вкладання слів в обробці природної мови. Зокрема, розглянуто основні моделі, такі як Word2Vec, GloVe, FastText. ELMo і BERT та проаналізовано вплив різних параметрів на точність і ефективність цих моделей. Описано наявні корпуси текстів українською мовою, які вже зібрані спільнотою, та можуть застосовуватися для навчання власних моделей. Досліджено використання зібраних корпусів текстів для тренування моделей, дана оцінка поточному прогресу та виявлені пріоритетні напрями для подальших досліджень. Результати роботи можуть бути застосовані для побудови власних моделей у різних предметних областях.

Ключові слова: *обробка природної мови; вкладання слів; векторна модель.*
Табл.: 1. Рис.: 2. Бібл.: 19.

Актуальність теми дослідження. Стрімке збільшення обчислювальних потужностей за останній час стало ключовою причиною значного прогресу у сфері обробки природної мови. Зокрема, це дозволило обробляти великі масиви текстів та вирішувати задачі пошуку, класифікації та ранжування текстів у різних предметних областях.

Одним зі способів представлення текстів, зручних для подальшої обробки алгоритмами, є представлення слів у вигляді векторів дійсних чисел. Такий підхід у мовному моделюванні, коли слова або фрази зі словника відображаються у вектори дійсних чисел, називається вкладання слів. Векторні представлення слів, отримані такими моделями, як Word2Vec, GloVe, FastText, ELMo і BERT, можна застосувати як вхідні дані для подальшої обробки та аналізу текстів.

Важливість цієї теми також зумовлена зростаючою потребою у високоякісних мовних моделях для текстів українською мовою, які можуть використовуватися в різних предметних областях, таких як освіта, охорона здоров'я, юридична справа чи медіа. При цьому першим кроком є збір необхідних корпусів текстів та їх якісна підготовка для навчання власних моделей, що дозволить підвищити ефективність вже існуючих алгоритмів саме для україномовних текстів.

Постановка проблеми. Існуючі моделі вкладання слів, такі як Word2Vec, GloVe, FastText, ELMo і BERT потребують адаптації перед застосуванням відповідно до наявного корпусу текстів. Це вимагає аналізу параметрів моделей та їхніх архітектур. Додатковим завданням є необхідність оцінки ефективності застосування різних моделей векторного представлення слів у конкретних задачах. Порівняння різних моделей та їхніх параметрів потребує проведення численних експериментів та аналізу результатів.

Крім того, важливою проблемою є відсутність достатньої кількості високоякісних корпусів текстів українською мовою з різних предметних областей. Їх збір потребує додаткових зусиль, що впливає на кількість досліджень у цій сфері.

Аналіз досліджень і публікацій. Ідея представлення слів у вигляді вектора не нова і широко досліджена [1-2]. Сама концепція семантичного простору, де лексичні одиниці (слова або фрази) представлені у вигляді векторів, заснована на вирішенні обчислювальних задач щодо отримання характеристик розподілу слів у тексті. Це дозволило використовувати ці представлення для практичних застосувань, таких як вимірювання схожості

між словами, фразами або цілими документами. Подальший розвиток цього підходу насамперед був зумовлений необхідністю зменшення високої розмірності отриманого векторного простору, що ускладнювало обчислення [3].

Новий поштовх у розвитку даної моделі був отриманий у 2013 році, коли команда розробників із Google представила Word2vec – модель векторного представлення слів на основі штучних нейронних мереж [4]. У 2014 році було розроблено GloVe (Global Vectors) — модель, яка враховує глобальну статистику спільної появи слів [5], а у 2015 році з'явився FastText, що працює з морфологічними компонентами слів — N-грамами, що виявилося особливо корисним для мов зі складною морфологією, таких як українська [6].

Важливим кроком у розвитку обробки природної мови стала поява моделі ELMo (Embeddings from Language Models), яку у 2018 році представила команда з Інституту Аллена науки про мозок [7]. Модель створює контекстуалізовані векторні представлення слів залежно від їх оточення у реченні. На відміну від попередніх моделей, ELMo використовує двонаправлені рекурентні нейронні мережі (BiLSTM), що дозволяє враховувати навколишній контекст слова при кожному його вживанні. Це суттєво покращило якість представлення багатозначних слів і складних синтаксичних зв'язків.

Проте справжнім проривом у подальшому розвитку архітектур нейронних мереж для глибокого навчання стала поява у 2018 році трансформерів. Зокрема, представлена дослідниками Google модель BERT (Bidirectional Encoder Representations from Transformers) використовує двонаправлений контекст, тобто розуміє зв'язки між словами не лише зліва направо чи справа наліво, а й з обох напрямків одночасно. Це дозволило їй досягти визначних результатів у задачах розпізнавання іменованих сутностей, класифікації текстів та пошуку. Саме з її появою почалося значне зростання якості NLP-моделей у багатьох мовах [8].

Для багатомовного застосування, включаючи українську мову, було розроблено Multilingual BERT (mBERT) та XLM-R (Cross-lingual Language Model – RoBERTa). Остання підтримує понад 100 мов і демонструє високу ефективність навіть для низькоресурсних мов завдяки багатоетапному попередньому тренуванню та перенесенню знань між мовами [9].

Розглянуті моделі використовуються для вирішення різноманітних задач, у тому числі задачі ранжування. Зокрема, було представлено підхід до покращення ранжування документів за допомогою подвійних векторних представлень слів. Поєднання двох різних моделей векторного представлення дало змогу суттєво покращити точність ранжування документів [10].

Окрему увагу варто звернути на зусилля дослідників, спрямовані на збір необхідних для навчання моделей корпусів текстів. Зокрема, важливим досягненням є зібраний Корпус української мови. Крім нього, українська спільнота зібрала корпус публіцистичних текстів та корпус законів та правових актів. Отримані корпуси застосовуються зокрема й для обчислення векторного представлення слів [11].

Зібрані корпуси текстів українською мовою дослідники використовують для побудови моделей векторного представлення слів з їх подальшим практичним застосуванням. Так, у 2016 році група дослідників побудувала модель вкладання слів для кластеризації україномовних академічних текстів за темами. Автори представили документи як усереднені вектори слів і досягли лише 5 % помилкових класифікацій без використання паралельних корпусів [12].

В іншому дослідженні 2023 року була застосована модель RoBERTa для задачі аналізу тональності україномовних відгуків у електронній комерції. Дослідники зібрали корпус відгуків, що включає як літературну мову, так і елементи сленгу, суржику та

іншомовних запозичень, після чого здійснили бінарну класифікацію (позитивна/негативна тональність), досягнувши точності у 92 %. Запропонований підхід може бути використаний у маркетингових аналітичних системах для моніторингу клієнтських відгуків та оцінюванні якості обслуговування [13].

У 2024 році група дослідників використала векторні представлення частин слів для створення зворотного словника термінів у галузі фізики українською мовою. Автори сегментували фізичні терміни та їх визначення з «Пояснювального словника з фізики» на морфеми, після чого побудували модель векторного представлення отриманих частин слів. Ці представлення використовувались для знаходження найбільш відповідного терміна за його визначенням, таким чином довівши можливість застосування моделей вкладання слів у задачах термінологічної інженерії українською мовою [14].

Виділення недосліджених частин загальної проблеми. Очікувано, що більшість досліджень зосереджено саме на англомовних текстах, що створює труднощі в застосуванні моделей векторного представлення слів до українських текстів через відмінності у структурі та лексичному складі мов. Це, зокрема, включає й налаштування параметрів моделей для оптимальної роботи з україномовними даними.

Крім того, для досліджень необхідні великі корпуси текстів українською мовою з різних предметних областей. При цьому тексти для різних предметних областей можуть мати різний вплив на ефективність моделей та їх застосування в задачах обробки природної мови.

Мета дослідження. Проаналізувати існуючі методи векторного представлення слів для задач обробки природної мови українською мовою, враховуючи наявні корпуси текстів, та запропонувати їх вдосконалення.

Виклад основного матеріалу. Векторне представлення слів — це низка методів, які використовуються в обробці природної мови, і полягають у відображенні слів у відповідні їм вектори дійсних чисел у багатовимірному просторі. При цьому семантично схожі слова матимуть відповідні їм вектори, що знаходитимуться ближче один до одного у векторному просторі, аніж слова, що є семантично несхожими. Такий підхід є гарним способом представити людську мову у вигляді зрозумілих математичних об’єктів (векторів), алгоритми обробки яких відомі. У подальшому ці дані можна використовувати для пошуку, порівняння або сортування текстів.

Word2Vec. Одним із найбільш відомих і широко використовуваних методів векторного представлення слів залишається Word2Vec. Word2Vec – це модель векторного представлення слів, розроблена командою дослідників із Google у 2013 році [4]. Вона працює на основі штучних нейронних мереж і може використовувати один із двох підходів: модель «неперервної торби слів» (continuous bag-of-words, CBOW) або skip-grams. У той час як перша архітектура прогнозує поточне слово на основі контексту (оточуючих слів), то друга архітектура навпаки прогнозує контекст (оточуючі слова) на основі поточного слова (рис. 1).

Word2Vec є досить швидкою та ефективною у навчанні, що дозволило використовувати її для великих корпусів текстів, що й стало причиною її популярності. Але при цьому для досягнення високої точності при навчанні моделі їй необхідний великий обсяг текстових даних, який не завжди досяжний. До того ж модель має проблеми з обробкою рідкісних слів або слів, які не траплялися під час навчання.

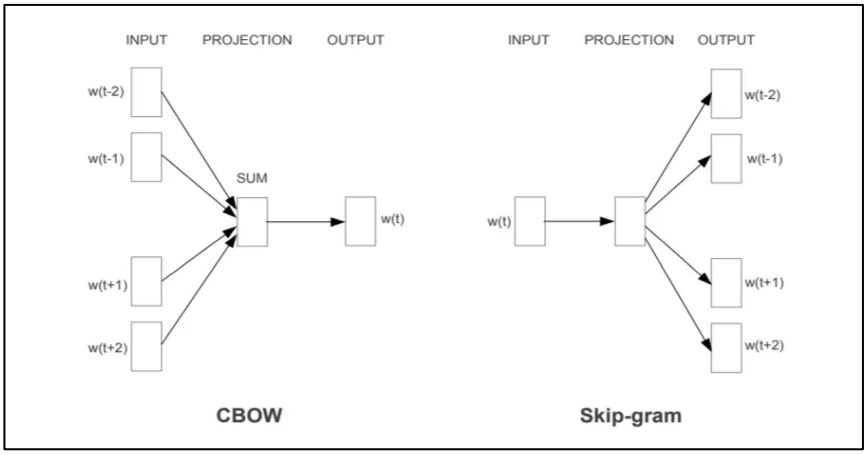

Рис. 1. Графічне представлення моделі CBOW (continuous bag-of-words) і моделі skip-gram

Джерело: [15].

У моделі CBOW навколишні слова (w(t±i)) перетворюються у векторні представлення та сумуються, щоб передбачити слово посередині (w(t)). Модель skip-gram навпаки використовує векторне представлення вхідного слова (w(t)) для отримання набору ймовірностей оточуючого контексту (w(t±i)).

GloVe. GloVe (Global Vectors) – це модель векторного представлення слів, розроблена в Стенфордському університеті у 2014 році [5]. Вона використовує статистику попарної спільної появи слів у корпусі текстів. При цьому GloVe намагається врахувати глобальну статистику тексту на відміну від Word2Vec, який фокусується лише на локальному контексті.

Усі слова в корпусі текстів аналізуються, і підраховується частота спільної появи кожної пари слів у певному контекстному вікні. На основі цього формується матриця спільної появи, де кожен елемент показує, скільки разів пара слів з'являється разом у заданому контексті (рис. 2).

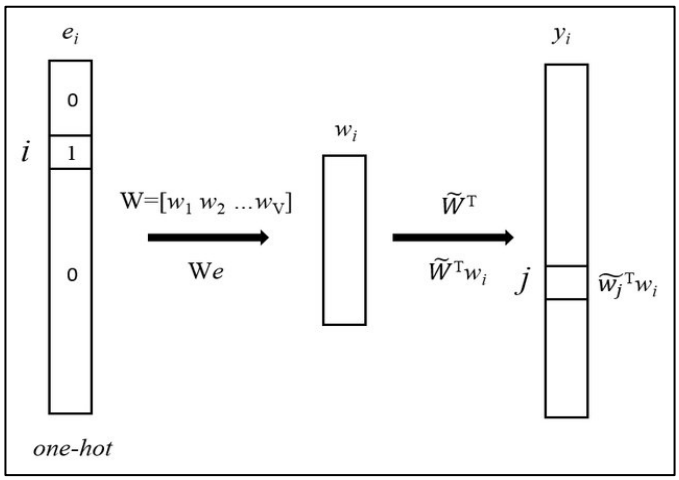

Рис. 2. Архітектура моделі GloVe

Джерело: [16].

Унітарний код (one-hot) слова у вигляді вектора e_i перемножується з матрицею векторів $W = [w_1\ w_2\ \ldots\ w_V]$ усіх слів словника, що дає вектор w_i. Потім обчислюються скалярні добутки $W^T w_i$ вектора w_i з усіма іншими векторами слів словника для отримання вектора y_i. Кожен елемент y_j у цьому векторі відповідає ймовірності слова j у контексті слова i.

Головною перевагою цієї моделі є використання глобальної статистики усього тексту, що дозволяє краще захоплювати його загальну семантику. Але це потребує значних обчислювальних ресурсів для побудови матриці спільної появи та її обробки. До того ж неправильне налаштування моделі, зокрема вибір розміру контекстного вікна, може обмежувати можливість захоплення довгострокових залежностей.

FastText. FastText — це метод векторного представлення слів, розроблений дослідниками з Facebook AI Research у 2015 році [6]. Він розширює можливості Word2Vec, враховуючи морфологічні особливості слів, що дозволяє створювати більш точні векторні представлення, особливо для мов з багатою морфологією.

Головною особливістю цієї моделі є те, що кожне слово розбивається на підпослідовності символів, так звані N-грами. Далі використовуються ті самі підходи, що і в Word2Vec, але вектори створюються не тільки для самих слів, а і для N-грам. Підсумковий вектор слова буде сумою векторів усіх його N-грам.

Таким чином ця модель враховує внутрішню структуру слів, що дозволяє краще працювати з мовами з багатою морфологією. А завдяки використанню N-грам, FastText може створювати вектори для слів, які не зустрічалися в тренувальній вибірці, що особливо корисно для рідкісних або нових слів. Проте використання N-грам потребує більше обчислювальних ресурсів та пам'яті для збереження векторів, а саме налаштування моделі може бути більш складним через додаткові параметри, пов'язані з обробкою N-грам.

ELMo. ELMo (Embeddings from Language Models) – це модель векторного представлення слів, запропонована у 2018 році командою з Інституту Аллена науки про мозок [7]. Вона формує вектори з урахуванням того, в оточенні яких саме інших слів вживається цільове слово.

Основою ELMo є двонаправлені рекурентні нейронні мережі (BiLSTM). Для кожного слова модель генерує вектор, враховуючи як попередні, так і наступні слова. Таким чином кожен вектор є результатом поєднання кількох рівнів абстракції — від лексичного до синтаксичного та семантичного.

Модель ELMo стала важливим досягненням у розвитку моделей вкладання тексту. На відміну від раніше популярних методів, таких як Word2vec чи GloVe, ELMo створює окремі векторні представлення слова для кожного окремого випадку його вживання, беручи до уваги контекст. Word2Vec при цьому генерує єдиний вектор для слова, незалежно від різних випадків його вживання.

BERT. BERT (Bidirectional Encoder Representations from Transformers) – модель на основі нової архітектури нейронних мереж для глибокого навчання – трансформерів. Вона була представлена дослідниками із Google у 2018 році [8]. Трансформер – це архітектура глибокого навчання, яка обробляє послідовності даних за допомогою механізму уваги. Оригінальні трансформери використовують кодувально–декодувальну архітектуру, тобто складаються з шарів кодування, які перетворюють вхідний текст на набір внутрішніх представлень, та декодування, які перетворюють внутрішні представлення від кодувальника назад у слова.

Проте BERT має свої особливості у порівнянні зі звичайними трансформерами. Ця модель використовує лише кодувальники, оскільки її задача – отримати значення тексту. При цьому вона обробляє контекст двонаправлено, тобто враховує як попередні, так і наступні слова. Це дає їй глибше розуміння контексту, ніж у класичних трансформерів, ідеально підходячи для задач, де важлива семантика всього тексту.

Порівняння методів векторного представлення слів. У табл. 1 наведено порівняння розглянутих методів векторного представлення слів. Зокрема, можна зробити висновок, що різні підходи у побудові моделей, зокрема й у виборі архітектури, вимагають різних ресурсів, при цьому збільшуючи ефективність розуміння контексту та багатозначних слів.

Таблиця 1 – Порівняння методів векторного представлення слів

Критерій	**Word2Vec**	**GloVe**	**FastText**	**ELMo**	**BERT**
Архітектура	Нейронна мережа (CBOW, Skip-gram)	Матриця спільної появи слів у всьому корпусі	Нейронна мережа (CBOW, Skip-gram з N-грамами)	Двонаправлені рекурентні нейронні мережі	Трансформер (лише кодувальник)
Контекст	Обмежене симетричне вікно	Глобальний (на основі статистики)	Обмежене вікно з урахуванням частин слів	Повне речення (двостороннє)	Повне речення (двостороннє)
Морфологія	Працює з цілими словами	Працює з цілими словами	Працює з частинами слів (N-грамами)	Використовує посимвольну обробку (CharCNN)	Працює з частинами слів (WordPiece)
Багатозначність	Не підтримується	Не підтримується	Не підтримується	Підтримується	Підтримується
Вимоги до ресурсів	Низькі	Низькі	Середні	Високі	Високі

Джерело: розроблено автором.

Аналіз параметрів моделей. Налаштування дослідником різних параметрів навчання моделей векторного представлення слів впливає на точність та ефективність моделей. Зокрема, важливим параметром при навчанні є розмірність вектора. При цьому збільшення розмірності векторів зазвичай покращує точність, але також збільшує обчислювальні витрати.

Іншим параметром, що налаштовується в процесі навчання моделей, є розмір контекстного вікна. Менші контекстні вікна краще захоплюють локальні залежності, тоді як більші – глобальні залежності. Звісно, свій вплив на точність моделі має і частота слів. При цьому слова, що трапляються частіше, можуть отримувати занадто високу вагу, тому застосовується down-sampling для зменшення впливу дуже частих слів.

Вибір параметрів навчання моделей векторного представлення слів суттєво вплине на ефективність їх роботи з текстами українською мовою, враховуючи її морфологічну складність, багатослівність та наявність діалектизмів.

Оптимальна розмірність вектора для невеликих корпусів (1–10 млн слововживань) має бути меншою, ніж для корпусів на 100+ млн слововживань (як-от Корпус української мови). При цьому варто пам’ятати, що занадто великі розмірності можуть призвести до перенавчання або просто зайвих витрат ресурсів.

Через гнучкий порядок слів в українській мові доцільніше використовувати ширше контекстне вікно, ніж в інших мовах — 5–10 слів. Це дозволить краще захопити синтаксичні зв’язки, які можуть розриватися через вставні конструкції або зміну порядку.

Для очищення словника від шуму для великих корпусів загального призначення мінімальна частота слова для входження має бути встановлена 5. Для менших корпусів із вузькою спеціалізацією (медицина, право тощо) її доцільно зменшити до 2–3.

Варто пам’ятати, що для динамічних предметних областей (наприклад, новини, соцмережі) доцільно регулярно оновлювати модель або донавчати її на нових даних, щоб зберігати актуальність векторного простору.

Таким чином, збільшення розміру векторів підвищує точність, причому для всіх розглянутих моделей, але вимагає більше ресурсів. Збільшення контекстного вікна може покращити розуміння глобальних залежностей, але може і збільшити небажаний шум.

Огляд існуючих корпусів текстів українською мовою. Важливим етапом, з якого починається будь-яка обробка природної мови, є збір корпусу текстів, необхідних для навчання моделей. Очікувано, що найбільш представленими в дослідженнях є моделі, натреновані з використанням текстів англійською мовою. Разом з тим, кожна мова має свої лексичні та морфологічні особливості, тому важливо навчати моделі на корпусі текстів тією мовою, для якої ця модель призначена.

Головним надбанням для дослідників україномовних текстів є Корпус української мови, що містить 100 млн слововживань. Він складається з різних підкорпусів, таких як законодавчі, наукові, фольклорні тексти, публіцистика, художня проза та поетична мова. Найбільший розділ – публіцистика (47 млн слововживань). Корпус розміщено на лінгвістичному порталі MOVA.info, який забезпечує пошук у ньому [11].

Іншим вагомим досягненням є Браунський корпус української мови, що містить 1 млн слововживань. Він побудований на принципах, що були покладені в основу корпусу англійської мови Brown, зокрема використання текстів із різних галузей людської діяльності для якнайбільш рівномірного представлення лексичного багатства української мови. Дані корпусу доступні для використання згідно з умовами ліцензії «Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License», він вільно поширюється за допомогою вебсервісу для спільної розробки програмного забезпечення GitHub [17].

Провідною ініціативою у розвитку української обробки природної мови можна назвати спільноту lang.org.ua. Вона збирає різноманітні корпуси текстів українською мовою, словники, газетири (географічні довідники) та самостійно створює сервіси, бібліотеки й моделі для дослідників та розробників, зокрема й векторні моделі слів, моделі для аналізу тональності текстів та розпізнавання іменованих сутностей, які вільно поширюються на їх сайті [18].

Так, команда проєкту створила анотацію розпізнавання іменованих сутностей для частини Браунського корпусу української мови. Крім цього, вони зібрали корпус текстів української періодики та законів та юридичних актів. На основі цього вони створили моделі векторного представлення слів за допомогою Word2Vec, LexVec та GloVe.

Адаптація моделей до текстів українською мовою. Оскільки більшість існуючих моделей векторного представлення слів спочатку були створені для англійської мови, їх застосування до українських текстів потребує певної адаптації.

Одним із найбільш ефективних методів адаптації мовних моделей до іншомовних корпусів текстів є донавчання (fine-tuning, дослівно: тонке настроювання) — процес повторного навчання вже попередньо натренованої моделі на новому корпусі, що краще відповідає специфіці мови або предметної галузі [19]. Замість повного навчання з нуля, fine-tuning дозволяє використати вже наявні знання моделі про мову, набуті під час попереднього навчання на великому загальному корпусі, і адаптувати їх до нових умов за допомогою додаткового тренування на більш вузькому корпусі текстів.

У випадку роботи з україномовними текстами цей підхід дає змогу досягти високої точності, навіть якщо модель спочатку була навчена переважно на англомовних або багатомовних даних. Наприклад, багатомовна версія BERT (mBERT або XLM-R) може бути донавчена на корпусах текстів українською мовою, в результаті чого модель краще відображатиме лексичні та граматичні особливості саме української мови. Донавчання є стандартною технікою при створенні мовних моделей, орієнтованих на специфічні мови або сфери застосування.

Висновки. Існуючі методи векторного представлення слів можуть застосовуватися для різноманітних задач з природної обробки мови текстів українською мовою, враховуючи наявні корпуси текстів. Для основних моделей, таких як Word2Vec, GloVe, FastText, ELMo та BERT, визначені особливості їх реалізації, а також досліджено вплив різних параметрів та архітектур на точність і ефективність моделей.

Окрему увагу приділено огляду існуючих корпусів текстів українською мовою. При цьому виявлено, що більшість досліджень зосереджені на англомовних текстах, що створює виклики для застосування моделей векторного представлення слів до українських текстів через відмінності у структурі та лексичному складі мов.

Таким чином, існує необхідність подальшого збору та підготовки корпусів текстів, а також адаптації існуючих моделей до специфіки української мови для покращення їх точності та ефективності.

Список використаних джерел

1. Salton, G. (1962). Some experiments in the generation of word and document associations. In *The December 4-6, 1962, fall joint computer conference*. ACM Press. https://doi.org/10.1145/1461518.1461544.

2. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220.

3. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). Neural probabilistic language models. *Journal of Machine Learning Research*, (3), 1137–1155. https://www.jmlr.org/papers/ volume3/bengio03a/bengio03a.pdf.

4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/pdf/1301.3781.

5. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics. https://doi.org/10.3115/v1/d14-1162.

6. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. https://arxiv.org/abs/1612.03651v1.

7. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* https://arxiv.org/abs/1802.05365.

8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423.

9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

10. Nalisnick, E., Mitra, B., Craswell, N., & Caruana, R. (2016). Improving document ranking with dual word embeddings. In *The 25th international conference companion*. ACM Press. https://doi.org/10.1145/2872518.2889361.

11. Дарчук, Н. (2010). Дослідницький корпус української мови: Основні засади і перспективи. *Вісник Київського національного університету імені Тараса Шевченка. Літературознавство. Мовознавство. Фольклористика*, *21*, 45-49.

12. Kutuzov, A., Kopotev, M., Sviridenko, T., & Ivanova, L. (2016). Clustering comparable corpora of Russian and Ukrainian academic texts: Word embeddings and semantic fingerprints. https://doi.org/10.48550/arXiv.1604.05372.

13. Zalutska, O., Molchanova, M., Sobko, O., Mazurets, O., Pasichnyk, O., Barmak, O.V., & Krak, I. (2023). Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. *International Conference on Computational Linguistics and Intelligent Systems*. https://ceur-ws.org/Vol-3387/paper26.pdf.

14. Vakulenko M., Slyusar V. (2024). Automatic smart subword segmentation for the reverse Ukrainian physical dictionary task. *CEUR Workshop Proceedings*. https://ceur-ws.org/Vol-3723/paper4.pdf.

15. Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. https://doi.org/10.48550/arXiv.1309.4168.

16. Liu, C., Zhang, P., Li, T., & Yan, Y. (2019). Semantic Features Based N-Best Rescoring Methods for Automatic Speech Recognition. *Applied Sciences*, *9*(23), 5053. https://doi.org/10.3390/app9235053.

17. *GitHub - brown-uk/corpus: Браунський корпус української мови*. (n.d.). GitHub. https://github.com/brown-uk/corpus.

18. *Про нас: lang-uk*. (n.d.). Головна: lang-uk. https://lang.org.ua/uk/about.

19. Liu, Z., Winata, G. I., Madotto, A., & Fung, P. (2020). Exploring fine-tuning techniques for pretrained cross-lingual models via continual learning. https://arxiv.org/abs/2004.14218.

References

1. Salton, G. (1962). Some experiments in the generation of word and document associations. In *The December 4-6, 1962, fall joint computer conference*. ACM Press. https://doi.org/10.1145/1461518.1461544.

2. Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. https://doi.org/10.1145/361219.361220.

3. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). Neural probabilistic language models. *Journal of Machine Learning Research*, (3), 1137–1155. https://www.jmlr.org/papers/ volume3/bengio03a/bengio03a.pdf.

4. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. https://arxiv.org/pdf/1301.3781.

5. Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics. https://doi.org/10.3115/v1/d14-1162.

6. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. https://arxiv.org/abs/1612.03651v1.

7. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).* https://arxiv.org/abs/1802.05365.

8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423.

9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747.

10. Nalisnick, E., Mitra, B., Craswell, N., & Caruana, R. (2016). Improving document ranking with dual word embeddings. In *The 25th international conference companion*. ACM Press. https://doi.org/10.1145/2872518.2889361.

11. Darchuk, N. (2010). Doslidnytskyi korpus ukrainskoi movy: osnovni zasady i perspektyvy [Research corpus of the Ukrainian language: Basic principles and prospects] *Visnuk Kyyivskoho natsionalnoho universitetu imeni T. Shevchenka. Seriya: Literaturoznavstvo. Movoznavstvo. Folklorystyka – Bulletin of Taras Shevchenko National University of Kyiv. Series: Literary studies. Linguistics. Folklore studies*, (21), 45-49.

12. Kutuzov, A., Kopotev, M., Sviridenko, T., & Ivanova, L. (2016). Clustering comparable corpora of Russian and Ukrainian academic texts: Word embeddings and semantic fingerprints. https://doi.org/10.48550/arXiv.1604.05372.

13. Zalutska, O., Molchanova, M., Sobko, O., Mazurets, O., Pasichnyk, O., Barmak, O.V., & Krak, I. (2023). Method for Sentiment Analysis of Ukrainian-Language Reviews in E-Commerce Using RoBERTa Neural Network. *International Conference on Computational Linguistics and Intelligent Systems*. https://ceur-ws.org/Vol-3387/paper26.pdf.

14. Vakulenko M., Slyusar V. (2024). Automatic smart subword segmentation for the reverse Ukrainian physical dictionary task. *CEUR Workshop Proceedings*. https://ceur-ws.org/Vol-3723/paper4.pdf.

15. Mikolov, T., Le, Q. V., & Sutskever, I. (2013). Exploiting similarities among languages for machine translation. https://doi.org/10.48550/arXiv.1309.4168.

16. Liu, C., Zhang, P., Li, T., & Yan, Y. (2019). Semantic Features Based N-Best Rescoring Methods for Automatic Speech Recognition. *Applied Sciences*, *9*(23), 5053. https://doi.org/10.3390/app9235053.

17. *GitHub – brown-uk/corpus: Braunskyi korpus ukrainskoi movy [GitHub - brown-uk/corpus: The Brown Corps of the Ukrainian language]*. (n.d.). GitHub. https://github.com/brown-uk/corpus

18. *Pro nas: lang-uk*. (n.d.). *Holovna: lang-uk.* [About us*: lang-uk*. (n.d.). *The Head: lang-uk.* https://lang.org.ua/uk/about.

19. Liu, Z., Winata, G. I., Madotto, A., & Fung, P. (2020). Exploring fine-tuning techniques for pretrained cross-lingual models via continual learning. https://arxiv.org/abs/2004.14218.

Отримано 21.03.2025

UDC 004.855.2

Dmytro Zhurko[1], Iryna Bilous[2]

[1] PhD student, recipient of the Doctor of Philosophy degree in specialty 122
Chernihiv Polytechnic National University (Chernihiv, Ukraine)
E-mail: dm.zhurko@gmail.com. **ORCID:** https://orcid.org/0009-0001-4192-7780

[2] PhD in Technical Sciences, Associate Professor of Information Technology and Software Engineering Department,
Chernihiv Polytechnic National University (Chernihiv, Ukraine)
E-mail: iryna.bilous.it@gmail.com. **ORCID:** https://orcid.org/0000-0003-3092-678X. **ResearcherID:** G-3887-2014

USING WORD EMBEDDING MODELS IN NATURAL LANGUAGE PROCESSING

The paper presents findings of the scientific and methodological investigation of word embedding models for application in natural language processing (NLP). The research is timely due to rapid advancements in computational power, enabling large-scale text analysis.

The study is focused on adapting existing word embedding models — Word2Vec, GloVe, FastText, ELMo and BERT — to the under-represented Ukrainian language. While effective for English texts, these models are difficult to use for Ukrainian due to linguistic specifics and lack of quality resources.

The objective is to compare these models for Ukrainian text processing and suggest improvements based on existing corpora. The paper reviews their architectures, training parameters, and highlights their pros and cons. FastText is effective with infrequent and morphologically rich words and is hence a good candidate for Ukrainian; Word2Vec is effective and simple to use; GloVe is effective for capturing global co-occurrence.

Compared to these static embedding models, ELMo generates contextualized word representations that had better handle polysemy and syntactic variation, although it requires more computational resources. BERT further improves contextual understanding through its transformer-based bidirectional architecture, outperforming previous models in NLP tasks; however, it is more demanding in terms of memory and training time.

Comparative assessment is given that illustrates how word frequency thresholds, context window size, and vector size affect accuracy in models. The study stresses relevence of tuning parameters to match task-specific and linguistic needs.

The research also encompasses Ukrainian-language resources like Ukrainian Language Corpus (100M tokens), Brown Ukrainian Corpus (1M tokens), and community resources like lang.org.ua. However, there is a lack of available, annotated resources that limit larger-scale experiments.

In conclusion, this work highlights the need for ongoing development of the corpus and better embedding model adaptation to Ukrainian. It makes recommendations to enhance accuracy and efficiency of models for educational, health care, legal, and mass media applications.

Keywords: *natural language processing; word embedding; vector model.*

Tables: 1. Figures: 2. References: 19.

Журко Д. О., Білоус І. В. Використання моделей вкладання слів в обробці природної мови. *Технічні науки та технології.* 2025. № 1(39). С. 151-160.