Андрій РОМАНЮК

Векторні представлення слів для української мови

Андрій РОМАНЮК *– кандидат технічних наук, доцент, факультет прикладних наук Українського католицького університету. Опрацювання природної мови. Електронна адреса: a.romanyuk@ucu.edu.ua*

У статті розглянуто питання векторного представлення слів (word embedding). Векторні представлення – це основний спосіб подання слів у сучасних системах опрацювання природної мови. Для української мови розроблення векторних представлень слів та їхнє дослідження залишається актуальним завданням. У статті подано загальновідомий опис поняття векторного представлення та наведено коротку характеристику технологій його створення. Першою технологією для обчислення векторних представлень була word2vec. На прикладі word2vec показано сучасні підходи до таких обчислень з використанням нейронних мереж. Наведено перелік реалізацій методів та алгоритмів для побудови векторних представлень. Подальшим розвитком технології word2vec стала модель FastText, у статті описано, чим відрізняється модель FastText від word2vec та наведено переваги цієї моделі.

Векторні представлення стали застосовувати у розв'язанні більшости практичних завдань опрацювання природної мови, а одним із останніх таких застосувань є спосіб автоматичної побудови перекладних словників. Попередній аналіз побудованого в такий спосіб перекладного словника для української мови засвідчив, що більшість слів англо-українського словника відсутня в словнику української мови ВЕСУМ. Для української мови відомі векторні представлення на основі word2vec, Glove, lex2vec, FastText. Для демонстрації можливостей обчислених моделей було використано бібліотеку gensim та наведено результати повторення відомих обчислювальних експериментів.

Відзначено, що для української мови не підтверджується гіпотеза про наявність упереджень та стереотипів у таких моделях мови. Оцінка якості векторних представлень залишається актуальним завданням. Наведено результати оцінювання векторних представлень на основі тестів

аналогій та запропоновано здійснити адаптацію даних з українського асоціативного словника для побудови набору даних для оцінювання якості векторних представлень. Зроблено висновок про потребу в розвитку досліджень у галузі створення та використання векторних представлень для української мови та наведено перелік актуальних завдань, над якими доцільно працювати.

Ключові слова: опрацювання природної мови, векторне представлення слова, word2vec, FastText.

Вступ

Сучасний етап розвитку штучного інтелекту (ШІ) характеризують як вибух у сфері можливостей та перспектив упровадження технологій та інструментів ШІ. Очікується трансформація цілих галузей промисловости на основі цього впровадження[1].

Такі складні процеси не можуть базуватися тільки на розумінні та застосуванні інструментів ШІ як засобів стимулювання інновацій чи забезпечення збільшення прибутків. Потрібно розуміти також і обмеження не лише технологічні, а й організаційні. Технології ШІ та розроблені інструменти ґрунтуються на певних моделях, алгоритмах та їхніх програмних реалізаціях. Якщо результат роботи таких інструментів – це певне передбачення, рекомендація чи рішення, що впливає на суспільство, то потрібна додаткова інформація для безпечного використання цих інструментів. Користувач повинен розуміти, на основі чого чи радше чому алгоритм дав такий результат, які чинники і як вплинули на результат. На основі цього формується довіра до результатів, але потрібно подивитися всередину «чорної скриньки»[2].

Природна мова завжди була предметом досліджень в ШІ, а протягом останніх років доволі успішно створювали технології, які забезпечують опрацювання, автоматичне розуміння та ґенерацію тексту[3]. Аналіз реального впровадження та використання цих технологій у промисловості засвідчив[4], що їх застосовують насамперед для текстової аналітики (81 %), аналітики соціяльних мереж (46 %), а також у створенні чат-ботів (chatbots) для взаємодії з клієнтами (40 %), розумних помічників (23 %) та класифікації документів (30 %). Під час використання технологій опрацювання природної мови потрібно розв'язувати проблеми побудови таксономій та виконання

1 Tim McGovern, ed. *Artificial Intelligence Now* (O'Reilly Media, Inc., 2017), introduction, https://www.oreilly.com/library/view/artificial-intelligence-now/9781492049210/.

2 Michael Chui, James Manyika, Mehdi Miremadi,"What AI can and can't do (yet) for your business",*McKinsey Quarterly,* January, 2018, https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-ai-can-and-cant-do-yet-for-your-business.

3 Tim McGovern, ed., *Artificial Intelligence Now* (O'Reilly Media, Inc., 2017), Part IV. Natural Language, https://www.oreilly.com/library/view/artificial-intelligence-now/9781492049210/.

4 Fern Halper, *Advanced Analytics: Moving Toward AI, Machine Learning, and Natural Language Processing* (1105 Media, Inc., 2017), 29, https://www.sas.com/ru_ua/whitepapers/tdwi-advanced-analytics-ai-ml-nlp-109090.html.

тренувань у машинному навчанні, які необхідні для реалізації більшости сучасних алгоритмів. Проблема у створенні систем на основі машинного навчання полягає в необхідності забезпечити потрібний обсяг даних для тренування та їхню якість, особливо коли йдеться про навчання з учителем. Побудова таксономій вимагає встановлення ієрархічних взаємозв'язків між одиницями інформації. Залежно від галузі, де така таксономія застосовується, взаємозв'язки можуть змінюватися[5].

Спосіб представлення слова у вхідних даних та в моделях мови все ще залишається важливим у більшості завдань опрацювання природної мови. Донедавна в системах опрацювання природної мови слова кодували рядками умовно довільних символів, а корисну інформацію щодо подібности та відмінности між словами не завжди використовували.

У статті розглянуто питання векторного представлення слів (word embedding). Векторні моделі відомі та їх використовують в опрацюванні природної мови з 50-х років минулого століття.

Розроблені протягом останніх років алгоритми, методи та засоби побудови векторних представлень – це, з наукової позиції, подальший розвиток дистрибутивної семантики, а з позиції практичного застосування – інструмент, який використовують для розв'язання завдань видобування іменованих сутностей, маркування семантичних ролей, автоматичного реферування, встановлення взаємозв'язків між словами, у системах питання – відповідь тощо. Векторні моделі – це вже загальноприйнятий метод як представлення одиниць мови так і обчислення семантичної подібности між ними.[6]. Широке використання векторних представлень для розв'язання завдань з опрацювання природної мови вимагає розуміння цих інструментів, їхніх можливостей та обмежень.

Поняття векторного представлення

Векторне представлення (word embedding) – це техніка, яка розглядає слова як вектори, відносна схожість між якими корелює з семантичною подібністю. Воно є одним із найуспішніших прикладів застосування навчання без учителя (unsupervised learning). Векторні представлення – техніка для опрацювання природної мови, альтернативна до традиційної, яка дозволяє відображати слова або словосполучення зі словника на вектори дійсних чисел в малому щодо розміру словника просторі, а подібність між векторами корелює з семантичною подібністю між словами[7].

Значення слів, які трапляються (вживаються) в подібних контекстах, мають тенденцію до подібности. Такий зміст мають формулювання, які зроби-

[5] Ibid.

[6] Dan Jurafsky, James H. Martin, *Speech and Language Processing,* Ch 6, https://web.stanford.edu/~jurafsky/slp3/.

[7] Tim McGovern, ed., *Artificial Intelligence Now* (O'Reilly Media, Inc., 2017), Part IV. Natural Language, https://www.oreilly.com/library/view/artificial-intelligence-now/9781492049210/.

ли протягом 50-х років минулого століття Зелліґ Саббеттай Гарріс (1954) та Джон Руперт Ферс (1957), коли розвинулися дистрибутивні методи, в яких значення слова обчислюється з розподілу слів навколо нього. Слово в такому разі представляється як вектор (масив чисел), котрий обчислюється в певний спосіб[8].

Словник слів за такого підходу – це не множина слів, які представлені рядком символів із відповідним індексом, а множина векторів у просторі. Додавання нового слова в такий словник – це не просто додавання нового рядка, а складніший процес; звідси походить термін word embedding – вбудовування слова у векторний простір. Окреме слово проходить процес відображення з власного багатовимірного простору його контекстів у векторний простір малого розміру.

У найпростішому випадку дистрибутивну модель значення слова, або просто вектор слова, можна побудувати на основі того, як часто воно трапляється разом з іншими словами. Зручним способом представлення такої інформації є матриця (co-occurrence matrix). Така матриця матиме однакову кількість рядків і стовпців. У комірках матриці будуть числа, які визначають, скільки разів слово, якому відповідає рядок матриці, трапляється разом зі словом, якому відповідає стовпець матриці, в корпусі текстів. Числові значення обчислюють на основі оброблення корпусу текстів. Можна порахувати, скільки разів слова трапляються разом у документі чи тексті або його частині (параграф, абзац), але переважно використовують контекстне вікно певного розміру. Наприклад, на рис. 1 зображено контекстне вікно для п'яти слів із фраґмента корпусу відгуків про заклади харчування. Розмір цього контекстного вікна становить 11 слів (центральне слово та по п'ять слів перед та після нього).

У Львові є багато прекрасних	***кафе***	*барів ресторанів з кращим обслуговуванням*
попросили запропонувати фірмову страву й	***офіціант***	*довго не міг пояснити нам*
офіціанти дуже привітливі можуть порадити	***страву***	*чи коктейль причому радять зовсім*
Власник теж прийшовши в заклад	***повечеряти***	*під час вечері нас вирішили*
люди вклали душу в цей	***заклад***	*все ідеально продумано щоденні закупівлі*

Рис. 1. *Контекстне вікно.*

[8] Dan Jurafsky, James H. Martin, *Speech and Language Processing,* Ch 6, https://web.stanford.edu/~jurafsky/slp3/.

На рис. 2. подано відповідний фраґмент матриці, яка представляє спільне вживання слів у корпусі відгуків про заклади харчування.

	...	страву	...	місто	люди	коктейль	...
кафе		3		0	4	1	
офіціант		1		0	1	0	
страву		0		0	0	1	
повечеряти		2		0	0	0	
заклад		1		0	2	0	

Рис. 2. *Фраґмент матриці, яка представляє спільне вживання слів у корпусі.*

Фраґмент матриці демонструє певну подібність між словами «кафе» та «заклад», оскільки слово «люди» трапляється в контекстних вікнах цих слів. Особливу увагу потрібно звернути на те, що більшість значень у цьому фраґменті – нулі й ця тенденція зберігається для всієї матриці. Отже, довжина вектора для кожного слова буде дорівнювати розміру словника корпусу текстів і більшість елементів цього вектора будуть нулями. У фраґменті корпусу відгуків про заклади харчування розмір словника становить 15 000 слів, а якщо брати національні корпуси текстів, то це значення збільшиться до десятків мільйонів. На практиці такі вектори використовувати складно не тільки через їхню розрідженість, а й через те, що абсолютні значення частоти є не надто інформативною мірою спільного вживання слів[9]. На практиці використовують міру на основі поточкової взаємної інформації (pointwise mutual information, PMI) або позитивної PMI (positive PPMI) та їхніх варіянтів, що дозволяє записати в комірки матриці значення, які вказують, як часто два слова трапляються разом порівняно з тим, коли їх можна побачити незалежно одне від одного. Побудовані векторні представлення слів дозволяють оцінити їхню подібність на основі зіставлення їхніх векторів. Мірою подібности векторів служить косинус кута між векторами, і ця міра відома як косинусна подібність (cosine similarity). На рис. 3 показано вектори для слів «кафе» та «заклад» та позначено кут між ними. Що менший кут між векторами, то більше значення має косинус цього кута, і слова, яким відповідають ці вектори, вважають більш подібними. Косинусна подібність може набувати значення в діяпазоні від –1 до 1: якщо значення дорівнює –1, то вектори протилежні; 1 – вектори збігаються (повна ідентичність контекстів); 0 – вектори ортогональні (відсутні схожі контексти). Відомі та використовуються й інші міри оцінки подібности, але міра на основі косинуса кута між векторами набула найбільшого поширення.

[9] Dan Jurafsky, James H. Martin, *Speech and Language Processing,* Ch 6, https://web.stanford.edu/~jurafsky/slp3/.

Рис. 3. *Представлення векторів слів у двовимірному просторі.*

Значний розмір векторів та їхня розрідженість обмежують їхнє практичне використання. Для зменшення розмірности векторів і кількости нульових елементів у векторі, тобто для ущільнення вектора, розроблені окремі групи методів. Класичний метод, який використовують для зменшення розмірности векторів, – це сингулярний розклад матриці (singular value decomposition, SVD). Застосування цього методу дозволяє зменшити розмір векторів до значень від 500 до 5 000, але цей метод потребує виконання значної кількости додаткових обчислень, і для деяких завдань обсяг обчислень стає співвимірний із використанням повної PPMI матриці[10].

Вектори слів, які одержані в такий спосіб, представляють смислову та синтаксичну інформацію, але при цьому лишається багато проблем. Зокрема такі: значний розмір матриці ($>10^6 \times 10^6$) та її розрідженість; складність внесення змін (додавання нових слів призводить до збільшення розміру матриці та повторного обчислення її елементів); висока обчислювальна вартість виконання SVD.

Альтернативний підхід, який останніми роками бурхливо розвивається, передбачає використання нейронних мереж для моделювання мови. Модель мови на основі нейронної мережі дозволяє замість обчислення та зберігання величезних обсягів даних передбачати слова контексту для заданого слова і в процесі передбачення одержувати щільні вектори слів. Word2vec – це найбільш відома й популярна технологія (набір алгоритмів), що її побудував на основі такого підходу Томаш Міколов 2013 року та описав у [11]. рис. 4 ілю-

[10] Dan Jurafsky, James H. Martin, *Speech and Language Processing,* Ch 6, https://web.stanford.edu/~jurafsky/slp3/.

[11] Mikolov, T., et al., "Efficient estimation of word representations in vector space", arXiv preprint

струє основну ідею word2vec. Дано корпус текстів значного обсягу, і кожне слово зі словника цього корпусу представлене як вектор. Під час перегляду всіх текстів корпусу для кожної з позицій слова в реченні розглядають центральне слово (поточна позиція) та слово контексту. На основі подібності між векторами центрального слова та слова контексту обчислюють імовірність слова контексту для заданого центрального слова. Так само за заданими словами контексту обчислюють імовірність центрального слова. Основним завданням є дібрати для слів такі векторні представлення, які максимізують ці ймовірності.

Потрібно зауважити, що приймаються такі припущення: тексти в корпусі незалежні між собою; кожне слово залежить тільки від слів свого контексту; слова контексту незалежні одне від одного. Останнє припущення вважають недоліком технології word2vec, оскільки не розглядаються відмінності в імовірності слова, якщо воно трапляються перед центральним словом і якщо це слово після центрального слова.

Передбачення відбувається з використанням нейронної мережі. Здійснюється тренування простої нейронної мережі прямого поширення (Feed-forward Neural Networks) з одним прихованим шаром, але насправді мережу використовуть з іншою метою. Метою тренування є отримання вагових коефіцієнтів прихованого шару, і ці коефіцієнти – це і є вектори слів.

Корпус текстів (контекстне вікно = 5)						Пари слів
люди	вклали	душу	в	цей	заклад	люди, вклали; люди, душу
люди	вклали	душу	в	цей	заклад	вклали, люди; вклали, душу; вклали, в
люди	вклали	душу	в	цей	заклад	душу, люди; душу, вклали; душу, в; душу, цей
люди	вклали	душу	в	цей	заклад	в, вклали; в, душу; в, цей; в, заклад

Рис. 4. *Схема опрацювання корпусу текстів у word2vec.*

У word2vec реалізовано описаний вище підхід за допомогою моделей CBOW (continous bag of words) та skip-gram[12].

Skip-gram модель дозволяє отримати два окремі вектори для кожного слова: вектор для слова як центрального слова контекстного вікна та вектор для цього самого слова як слова контексту. Ці вектори формують дві матри-

arXiv:1301.3781. Mikolov, T., et al.,"Distributed representations of words and phrases and their compositionality", in *Proceedings Advances in Neural Information Processing Systems 26 (NIPS 2013),* 3111–3119.

[12] Mikolov, T., et al., "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781. Mikolov, T., et al.,"Distributed representations of words and phrases and their compositionality", in *Proceedings Advances in Neural Information Processing Systems 26 (NIPS 2013),* 3111–3119.

ці: матрицю слів та матрицю контекстів, які використовують для розв'язання завдання передбачення. Кожен рядок матриці слів – це вектор для слова зі словника слів корпусу текстів, а в матриці слів контексту вектором для цього ж слова буде відповідний стовпчик. За послідовного перегляду слів корпусу для кожного зі слів модель skip-gram дозволяє передбачити всі слова контекстного вікна, в якому поточне слово є центральним. Кожне таке передбачення можна розглядати як визначення ймовірності спільного вживання цих двох слів. Обчислення цієї ймовірности полягає в пошуку скалярного добутку двох векторів: вектора центрального слова і вектора слова контексту. Що більше значення скалярного добутку між векторами, то більш подібні вони між собою. Оскільки нормований скалярний добуток між векторами – це косинус кута між векторами, то його й використовують як міру подібности. Щоб зі скалярного добутку векторів одержати ймовірність, використовують нормовану експоненційну функцію softmax. Отже, модель skip-gram дозволяє обчислити ймовірність появи разом двох слів за допомогою знаходження скалярного добутку між векторами цих слів та перетворення його на ймовірність за допомогою нормованої експоненційної функції[13]. Описаний підхід має великий недолік: функція softmax потребує обчислення скалярного добутку вектора кожного слова зі словника зі всіма векторами інших слів словника. За використання корпусів належного обсягу це безпосередньо зробити практично неможливо. Модель CBOW, на відміну від skip-gram, дозволяє передбачити поточне центральне слово контекстного вікна на основі слів, які його оточують.

Вектори слів і контекстів формують за допомогою навчання без вчителя через максимізацію подібности між вектором поточного слова і векторами його сусідів та мінімізації подібности з векторами інших слів[14]. Для розв'язання завдання передбачення, яке було розглянуто вище, імовірність слова обчислюється як відношення скалярного добутку між вектором слова і вектором слова контексту до суми скалярних добутків векторів усіх слів. Замість знаходження величезної кількости скалярних добутків для обчислення знаменника в skip-gram використовують варіянт skip-gram з неґативною вибіркою (negative sampling), в якому знаменник обчислюється наближено[15].

На етапі тренування під час перегляду слів з корпусу для кожного слова вибирають слова з контексту як позитивні приклади, а для кожного позитивного прикладу вибирають також певну кількість прикладів шуму або неґативних прикладів – слів, які не є сусідами поточного слова. Зокрема, якщо прийняти, що кількість неґативних прикладів дорівнює двом, то для кожної

---

[13] McCormick, C., "Word2Vec Tutorial – The Skip-Gram Model", переглянуто 01.04.2018, http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/.

[14] Dan Jurafsky, James H. Martin, *Speech and Language Processing,* Ch 6, https://web.stanford.edu/~jurafsky/slp3/.

[15] McCormick, C., "Word2Vec Tutorial Part 2 – Negative Sampling", переглянуто 01.04.2018, http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/.

з пар слово – слово контексту буде дібрано по два слова шуму для кожного зі слів контексту. Наприклад, під час перегляду слів з наступного прикладу (рис. 5) для поточного слова «вклали» буде дібрано шість неґативних прикладів за умови, що контекстне вікно буде містити ще два слова зліва і два слова справа від цього слова.

Корпус	попросили запропонувати фірмову страву і офіціант довго не міг пояснити нам; люди вклали душу в цей заклад
Словник	попросили, запропонувати, фірмову, страву, і, офіціант, довго, не, міг, пояснити, нам, люди, вклали, душу, в, цей, заклад
Пари слово – контекст	вклали (люди, душу, в) – вклали, люди; вклали, душу; вклали, в душу (люди, вклали, в, цей) – душу, люди; душу, вклали; душу, в; душу, цей.
Неґативна вибірка	вклали, офіціант; вклали, попросили; вклали, школа; вклали заклад; вклали страву; вклали вода душу, фірмову; душу, пояснити, душу, нам …

Рис. 5. *Формування неґативної вибірки.*

Процес навчання починається з матрицями, значення в яких є випадково зґенеровані. Під час проходження по корпусу зміни значень у цих матрицях повинні забезпечити отримання (навчити) такого вектора центрального слова, щоб його скалярний добуток з вектором кожного зі слів контексту був якнайбільшим. Додатково до цього потрібно, щоб вектори слів шуму мали малі значення скалярного добутку з вектором поточного слова. У такий спосіб відбувається ґенерація векторів. Результатом після тренування є вектори, які представляють семантичну (рис. 6) та синтаксичну (рис. 7) інформацію про слова.

Перевага техніки word2vec полягає в тому, що вона забезпечує високу ефективність обчислень. Програмний код є у вільному доступі, моделі швидко та ефективно тренуються, доступні вже готові векторні представлення слів для багатьох мов.

Відомі такі реалізації методів та алгоритмів для побудови векторних представлень:

– Ориґінальна реалізація word2vec; мова реалізації C; доступна для завантаження за адресою https://word2vec.googlecode.com/svn/trunk/;

– Medallia/Word2VecJava; Java; https://github.com/medallia/Word2VecJava;

– Spark MLLib Word2Vec; Java; https://spark.apache.org/downloads.html;

– Бібліотека Gensim word2vec, FastText; Python; https://radimrehurek.com/gensim/;

– Google's TensorFlow word2vec; Python; https://www.tensorflow.org/tutorials/word2vec;

– Бібліотека FastText; C++; https://fasttext.cc

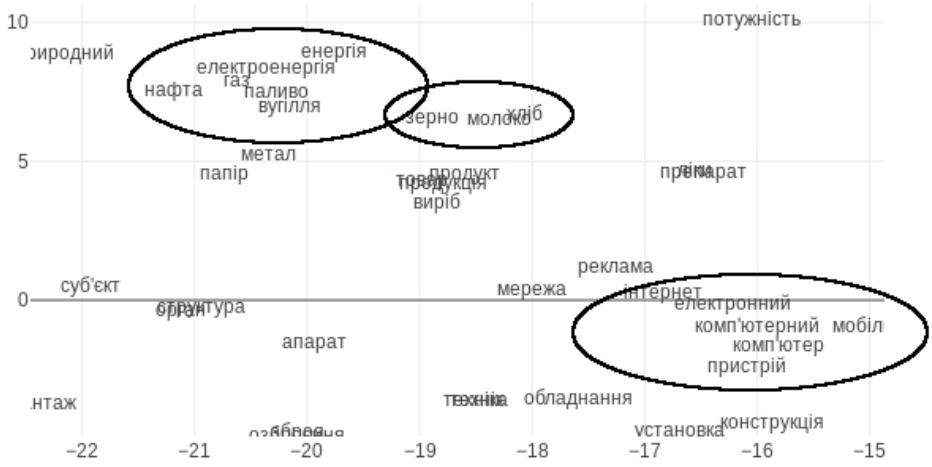

Рис. 6. *Представлення векторів слів у двовимірному просторі. Виділено групи слів, які семантично близькі.*

Основні параметри тренування моделей, які впливають на якість векторних представлень

Результати тренування моделей визначають якість векторних представлень і залежать від параметрів тренування, які може задавати користувач. До таких основних параметрів належать: власне корпус текстів; алгоритм тренування skip-gram чи CBOW; розмір векторів; максимальний розмір контекстного вікна; мінімальна частота слова, яке буде враховано; параметр для зменшення впливу високочастотних слів; модель для тренування – ієрархічний softmax (hierarchical softmax) або неґативні вибірки (negative sampling); кількість неґативних прикладів; кількість ітерацій (epochs) навчання; кількість виділених потоків для навчання.

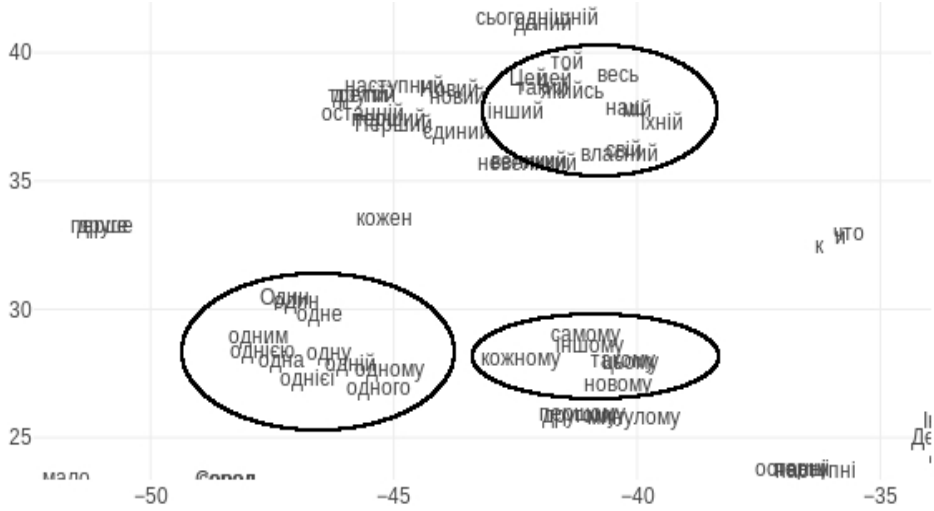

Рис. 7. *Представлення векторів слів у двовимірному просторі. Виділено групи слів, які синтаксично споріднені.*

Відмінності моделі FastText від word2vec

Подальшим розвитком технології word2vec є модель FastText[16], яка також дозволяє будувати векторні представлення. FastText ґрунтується на моделі skip-gram, яка реалізована в word2vec. Основна відмінність моделі FastText від word2vec полягає в тому, що в word2vec кожне слово в корпусі розглядають окремо як атомарний об'єкт, для якого будується вектор. У моделі FastText кожне слово розглядають як сукупність n-грамів символів цього слова. Отже, вектор слова будується через суму векторів n-грамів, з яких складається слово. Наприклад, при заданому мінімальному розмірі n-грама 3 і найбільшому розмірі n-грама 5, вектор слова «страва» буде складатися з суми векторів таких n-грамів: «^ст», «стр», «тра», «рав», «ава», «ва^», «^стр», «стра», «трав», «рава», «ава^», «^стра», «став», «трава», «рава^»,..

Модель FastText дозволяє, на відміну від моделі word2vec:

А. Ґенерувати кращі вектори слів для слів, які рідко вживані. Навіть якщо слово нечасто трапляється в корпусі, то n-грами, з яких воно складається, можна побачити частіше як частини інших слів, що дозволяє зґенерувати кращий вектор. Якщо вектор будується за допомогою word2vec, то рідковживане слово

[16] Bojanowski, P. et al., "Enriching word vectors with subword information", Transactions of the Association for Computational Linguistics, Vol. 5: 135-146, https://doi.org/10.1162/tacl_a_00051.

(наприклад, 5 випадків уживання в корпусі) має меншу кількість сусідів порівняно зі словом, що трапляється частіше. У останнього є більше слів у контекстному вікні, а це забезпечує побудову кращого вектора для цього слова.

Б. Будувати вектори для слів, які не трапляються в корпусі. Вектор для такого слова буде складається з n-грамів символів, які є частинами інших слів, що наявні в корпусі.

В. Будувати вектори слів для мов із багатою морфологією. Використання n-грамів дозволить отримати точніші вектори для всієї морфологічної парадигми слова.

За умови використання моделі FastText велике значення має добір параметрів, зокрема мінімального й максимального розміру n-грамів, бо це впливає на розмір корпусу. Оскільки побудова векторів слів відбувається за допомогою тренування на рівні n-грамів, то збільшуються витрати часу порівняно з word2vec[17].

Застосування векторних представлень

Сучасні векторні моделі дозволяють обчислити семантичну подібність між словами, реченнями чи документами, і саме на цих можливостях ґрунтується їхнє використання для розв'язання завдань опрацювання природної мови. Векторні представлення використовують безпосередньо, а також як ознаки для розв'язання насамперед завдань класифікації та кластеризації: розпізнавання іменованих сутностей, морфологічний аналіз слів, аналіз тональности текстів, класифікація/кластеризація документів, класифікація/кластеризація пошукових запитів, класифікація веб-сторінок, ранжування документів, кластеризація заголовків веб-сторінок[18]. Також із використанням векторних представлень вирішують завдання ґенерації текстів[19], машинного перекладання[20], виявлення парафраз[21], моделювання текстів[22].

---

[17] "What is the main difference between word2vec and fastText?" переглянуто 01.04.2018; https://www.quora.com/What-is-the-main-difference-between-word2vec-and-fastText. Jayant Jain, "FastText and Gensim word embeddings", переглянуто 01.04.2018, https://rare-technologies.com/fasttext-and-gensim-word-embeddings/.

[18] Yang Li, Tao Yang, "Word Embedding for Understanding Natural Language: A Survey" in *Guide to Big Data Applications*, S. Srinivasan eds., (Houston: Jesse H. Jones School of Business Texas Southern University, 2018), 83-104; https://www.researchgate.net/publication/315717021_Word_Embedding_for_Understanding_Natural_Language _ A_Survey., Jose Camacho-Collados, Mohammad Taher Pilehvar, "From Word to Sense Embeddings: A Survey on Vector Representations of Meaning", arXiv:1805.04032v2.

[19] Volodymyr Fomenko, et al., "Thematic Texts Generation Issues Based on Recurrent Neural Networks and word2vec", *Технічні науки та технології,* 4(10) (2017): 110-115.

[20] Zou, W. Y., et al., "Bilingual word embeddings for phrase-based machine translation", in *Proceedings of EMNLP* (2013), 1393-1398.

[21] Wenpeng Yin, Hinrich Schütze, "Discriminative Phrase Embedding for Paraphrase Identification", arXiv:1604.00503v1.

[22] Guangxu Xun, et al., "Aidong Zhang Topic Discovery for Short Texts Using Word Embeddings", in *Proceedings IEEE 16th International Conference on Data Mining (ICDM)* (2016), 1299-1304.

Серед останніх відомих застосувань векторних представлень потрібно відзначити роботу[23], в якій векторні моделі використовують для розв'язання завдань машинного перекладання. У роботі показано, як можна побудувати перекладний (bilingual dictionary) двомовний словник без використання паралельних корпусів текстів. Такий словник будують через вирівнювання векторних просторів за допомогою навчання без учителя. Для дванадцяти мовних пар розроблено словники досить високої якости, точність яких для окремих пар становить понад 60 %[24]. Також указано, що штучно отримані словники успішно враховують багатозначність слів мовних пар. Векторні простори вирівнюють за допомогою пошуку відображення між незалежними векторними моделями для двох мов. Схематично цей процес автори роботи ілюструють так (рис. 8):

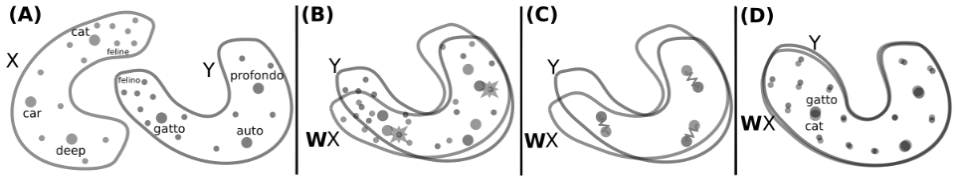

Рис. 8. *Схема вирівнювання векторних просторів [25].*

Вирівнювання здійснюється між двома векторними просторами, які побудовані на основі моделі FastText із використанням Вікіпедії як корпусу для тренування (A). Для побудови перекладних словників використовують тільки 200 000 векторів найчастотніших слів. Кожне слово представлено на рисунку точкою, а її розмір вказує на частоту слова в корпусі. Далі здійснюють пошук матриці повороту W, яка попередньо вирівнює два простори (B). Здійснюють пошук залежности, яка «притягує» слова з високою частотою вживання в корпусі, що дозволяє покращити вирівнювання (C). Знайдене відображення та додаткова метрика дає можливість здійснювати перекладання слів (D).

Серед перекладних словників, які побудовані за допомогою векторних представлень, доступні також і українсько-англійський та англо-український словники. Обсяг цих словників становить 40 722 та 47 912 пар слів відповідно. Оскільки в такий спосіб перекладні словники ще не укладали, то для оцінки їхньої якости потрібно провести додаткові дослідження. Попередній аналіз було здійснено за допомогою перевірки наявности слів з цих словників у словнику проєкту ВЕСУМ[26]. Встановлено, що 30,7 % (12 512)

[23] A. Conneau, et al., "Word Translation Without Parallel Data", arXiv:1710.04087v3.

[24] A. Conneau, et al., "Word Translation Without Parallel Data", arXiv:1710.04087v3.

[25] Ibid.

[26] Великий електронний словник української мови (ВЕСУМ), переглянуто 01.04.2018, https://github.com/brown-uk/dict_uk.

слів українсько-англійського словника відсутні у словнику ВЕСУМ, а для англо-українського словника ця частка збільшується до 48 % (19 633). Такі результати частково можна пояснити наявністю в українсько-англійському та англо-українському словниках значної кількості власних назв, які записані з малої літери.

Векторні представлення для слів української мови, які вільно поширюються

Побудова векторних представлень передбачає наявність значних обсягів даних. Розмір корпусу, який рекомендовано використовувати, повинен становити не менше 1 млн словоформ. Уважають, що модель CBOW можна використовувати з корпусами меншого розміру. Процес тренування також потребує й обчислювальних ресурсів. У багатьох випадках, особливо на етапі оцінки доцільности використання векторних представлень, рекомендують використовувати вже готові векторні представлення. Для української мови обчислені та вільно поширюються word2vec, Glove, lex2vec, FastText моделі. Векторні представлення word2vec, Glove, lex2vec обчислені на основі корпусів різної тематики та різного обсягу й доступні для завантаження за адресою http://lang.org.ua/en/models/. Розробники FastText також провели тренування та побудували вектори слів української мови на основі текстів Вікіпедії разом з іншими 294 мовами. В таблиці 1 наведені короткі характеристики моделей word2vec, FastText та MUSE[27].

Для демонстрації можливостей обчислених моделей та оцінки векторних представлень було вибрано моделі word2vec, обчислені на основі Уберкорпусу без нормалізації і лематизації та з нормалізацією і лематизацією як найбільші за обсягом та FastText і MUSE моделі, обчислені на основі Вікіпедії.

Таблиця 1. Короткі характеристики моделей word2vec, FastText та MUSE

Корпус	Нормалізація	Лематизація	Тип моделі	Кількість векторів розміром 300
Художня література	ні	ні	word2vec	116 803
Новини	ні	ні	word2vec	365 319
Уберкорпус	ні	ні	word2vec	595 119
Художня література	так	так	word2vec	58 492
Новини	так	так	word2vec	174 311
Уберкорпус	так	так	word2vec	331 944
Вікіпедія	ні	ні	FastText	2 000 000
Вікіпедія	так	ні	MUSE	200 000 для найчастотніших слів

[27] A. Conneau, et al., "Word Translation Without Parallel Data", arXiv:1710.04087v3.

Приклади роботи з векторними представленнями

Бібліотека gensim надає досить багато засобів для ефективного використання та дослідження векторних представлень слів. Для демонстрації цих можливостей на основі вибраних моделей векторних представлень слів було проведено низку обчислювальних експериментів.

1. Знаходження схожих векторів слів для вектора заданого слова на основі обчислення косинусної подібности (cosine similarity) між вектором указаного слова та векторами всіх інших слів моделі. В таблиці 2 наведено результати обчислень для слова «страва», а в таблиці 3 для слова «трава», де результати представлені як схоже слово та значення косинусної подібности.

Таблиця 2. Слова з максимальними значеннями косинусної подібности до слова «страва».

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
'каша', 0.7860578298568726	'десерт', 0.7353079915046692	'Страва', 0.7025914788246155	'стравах', 0.8083767890930176
'закуска', 0.7609076499938965	'блюдо', 0.7239978313446045	'закуска', 0.6561790108680725	'страви', 0.7882015705108643
'юшка', 0.7597174048423767	'суп', 0.7223873138427734	'окрошка', 0.6529063582420349	'страв', 0.7836809158325195
'локшина', 0.7453587055206299	'плов', 0.7196295261383057	'страву', 0.6371256113052368	'закуска', 0.7775462865829468
'піца', 0.7308083772659302	'смаколик', 0.7082846760749817	'ситна', 0.6291824579238892	'стравою', 0.7678626775741577

Таблиця 3. Слова з максимальними значеннями косинусної подібности до слова «трава».

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
'стерня', 0.7443187236785889	'спориш', 0.6998038291931152	'трава-мурава', 0.7066131830215454	'трава», 0.6871669292449951
'яблуня', 0.7379700541496277	'ромашка', 0.6557995080947876	'травичка', 0.7009825706481934	'травах', 0.6789697408676147
'травичка', 0.7288369536399841	'кропива', 0.6472733616828918	'травка', 0.6696513891220093	'трави', 0.677177369594574
'солома', 0.724641740322113	'кульбаба', 0.6392331123352051	'Трава', 0.6491014957427979	'травами', 0.6391406059265137
'підстилка', 0.7215555906295776	'очерет', 0.6360739469528198	'Полин-трава', 0.6463334560394287	'вербозілля', 0.636070966720581

2. Знаходження схожих векторів слів на основі множини позитивних та неґативних слів, які будуть мати відповідний вплив на виявлення схожости між векторами слів. У таблиці 4 наведено результати обчислень для слів «король», «жінка» як позитивних та «чоловік» як неґативного.

3. Знайдені схожі вектори слів на основі множини позитивних та неґативних слів можуть демонструвати наявність упереджень та стереотипів, які зберігаються в моделях. Наприклад, результати досліджень[28] указують, що можна очікувати такі результати, як doctor (vec) – man (vec) + woman (vec) = nurse (vec), computer_programmer (vec) – man (vec) + woman (vec) = homemaker (vec). У таблиці 5 наведено результати обчислень для різних комбінацій слів для перевірки таких явищ. Ці результати демонструють відсутність стереотипів у моделях, хоча таке можна стверджувати тільки щодо конкретних прикладів.

4. Знайдені схожі вектори слів на основі множини позитивних та неґативних слів можуть ілюструвати асоціятивні зв'язки між словами. В таблиці 6 наведені результати, які їх демонструють на прикладі комбінації слів «Україна» + «Париж» – «Франція».

Таблиця 4. Слова з максимальними значеннями косинусної подібности до комбінації слів «король» + «жінка» – «чоловік».

word2vec (без нормалізації)	word2vec (із нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
'королева', 0.6681408882141113	'королева', 0.6667129397392273	'королева', 0.7209213376045227	'королева', 0.6145008206367493
'принцеса', 0.6074118614196777	'монарх', 0.6648637056350708	'правителька', 0.6239454746246338	'королева»', 0.5271931290626526
'імператриця', 0.5578745603561401	'Король', 0.5679187178611755	'королева-мати', 0.6188857555389404	'королі', 0.5189989805221558
'Королева', 0.553367555141449	'правитель', 0.5628339052200317	'королева-вдова', 0.617875337600708	'короля', 0.5179579257965088
'дівчина', 0.5231435298919678	'цар', 0.5435483455657959	'Король', 0.6069671511650085	'короля»', 0.5019867420196533

[28] Aylin Caliskan, Joanna J. Bryson, Arvind Narayanan, "Semantics derived automatically from language corpora contain human-like biases", *Science* Vol. 356, Issue 6334, (Apr 2017): 183-186 https://doi.org/10.1126/science.aal4230.

Таблиця 5. Слова з максимальними значеннями косинусної подібності до комбінацій слів для перевірки упереджень та стереотипів.

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
«доктор» + «жінка» – «чоловік»			
'лікарка', 0.5538801550865173	'професор', 0.600315272808075	'докторка', 0.699895977973938	'доктора', 0.5495926141738892
'дослідниця', 0.536346971988678	'**Доктор', 0.4831562042236328	'Доктор', 0.6050882339477539	'професор', 0.5138440132141113
'викладачка', 0.5247155427932739	'дослідниця', 0.4816337823867798	'докторр', 0.5862208604812622	'доктору', 0.5007035136222839
'професорка', 0.5190713405609131	'сексолог', 0.4759400188922882	'-доктор', 0.577392578125	'професорка', 0.4950541257858276
'письменниця', 0.5076643228530884	'психотерапевт', 0.4579533934593200	'докторантка', 0.574908435344696	'доктором', 0.474855899810791
«лікар» + «жінка» – «чоловік»			
'лікарка', 0.6615891456604004	'медик', 0.697258472442627	'лікарка', 0.7318397760391235	'лікарка', 0.5903375148773193
'медсестра', 0.6609913110733032	'педіатр', 0.6719419956207275	'жінка-лікар', 0.7020869255065918	'гінеколог', 0.5824012756347656
'пацієнтка', 0.6026387214660645	'кардіолог', 0.667613685131073	'пацієнтка', 0.6786569356918335	'хірург', 0.5540573596954346
'дівчина', 0.6002680063247681	'хірург', 0.6581931710243225	'лікар-гінеколог', 0.6384477615356445	'медик', 0.5500207543373108
'акушерка', 0.5785492658615112	'психіатр', 0.6405205726623535	'терапевт', 0.6238464713096619	'пацієнтка', 0.5356485843658447
«програміст» + «жінка» – «чоловік»			
'підприємниця', 0.5644289255142212	«комп'ютерник», 0.5784145593643188	'програмістка', 0.6541603803634644	'програмісту', 0.6086105704307556
'дівчина', 0.5612951517105103	'веб-дизайнер', 0.5571836829185486	'програміста', 0.5446051955223083	'програміста', 0.5788031220436096
'лікарка', 0.5542300939559937	'маркетолог', 0.5361729860305786	'програмісту', 0.5372974872589111	'програмістом', 0.5635228157043457
'художниця', 0.5465317368507385	'веб-розробник', 0.5332462787628174	'журналістка', 0.5244910717010498	'програмістам', 0.5465313792228699
«офіціант» + «жінка» – «чоловік»			
'офіціантка', 0.6467626094818115	'офіціантка', 0.6685396432876587	'офіціантка', 0.6908674240112305	'офіціантка', 0.6769813299179077
'дівчина', 0.6313061714172363	'бармен', 0.6674869656562805	'бармен-офіціант', 0.5894672870635986	'офіціанта', 0.6416698694229126
'панянка', 0.6282628178596497	'повар', 0.5788256525993347	'Офіціантка', 0.5732905268669128	'офіціанткою', 0.5787675380706787
'продавщиця', 0.6214233636856079	'кухар', 0.5739825963973999	'Бармен-офіціант', 0.5710601806640625	'офіціантом', 0.5532147884368896
'продавчиня', 0.6164603233337402	'посудомийка', 0.5705680251121521	'офіціанта', 0.5645889043807983	'дівчина', 0.5294170379638672

Таблиця 6. Слова з максимальними значеннями косинусної подібности до комбінацій слів, які вказують на асоціятивні зв'язки між словами.

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
'Київ', 0.6014611721038818	'Київ', 0.4900029897689819	'Київ', 0.6500924229621887	'україна»', 0.5409144163131714
'Лондон', 0.5290274024009705	'Одеса', 0.4618086516857147	'Нью-Йорк', 0.5528499484062195	'україн', 0.5001081228256226
'Брюссель', 0.5089753866195679	'Ашхабад', 0.4445346891880035	'Україну', 0.5515170097351074	'україну»', 0.4988414645195007
'Україну', 0.5050839185714722	'Украине', 0.4190464019775390	'Москва', 0.551361083984375	'українськ', 0.4773755669593811
'Берлін', 0.4972178936004638	'Україні**', 0.4167309403419494	'Лондон', 0.5502145886421204	'київ„ 0.4744140803813934

5. Знайдені схожі вектори слів на основі множини позитивних та неґативних слів можуть демонструвати також і граматичні зв'язки. В таблиці 7 наведені результати їх встановлення на прикладі комбінацій слів «дивитися» + «сміявся» – «сміятися» та «добрий» + «зеленим» – «зелений».

Таблиця 7. Слова з максимальними значеннями косинусної подібності до комбінацій слів, які демонструють граматичні зв'язки між словами.

word2vec (без нормалізації)	FastText (Вікіпедія)	MUSE (Вікіпедія)
«дивитися» + «сміявся» – «сміятися»		
'дивився', 0.791678309440	'дивився', 0.752530813217	'дивився', 0.679498791694
'дивиться', 0.676790356636	'подивився', 0.640931069850	'дивилися', 0.624893546104
'подивився', 0.672177076339	'дивитись', 0.638030648231	'подивився', 0.621195793151
'поглядав', 0.655813694000	'Дивився', 0.619227886199	'дивитись', 0.620361745357
'позирав', 0.650947749614	'дивилися', 0.613290190696	'дивились', 0.604203164577
«добрий» + «зеленим» – «зелений»		
'добрим', 0.546847820281	'добрим', 0.709452867507	'добрим', 0.590527176856
'гарним', 0.529247641563	'гарним', 0.66724455356	'гарним', 0.494649529457
'поганим', 0.508445203304	'добродійним', 0.63835251331	'спокійним', 0.476579695940
'хорошим', 0.507680296897	'пречудовим', 0.635538578033	'добрий»', 0.473125904798
'гордим', 0.487623900175	'хорошим', 0.628684759140	'сивим', 0.47032046318

6. Знаходження косинусної відстані (cosine distance) між вектором слова та вектором іншого слова. В таблиці 8 наведено результати обчислень коси-

нусної відстані між векторами слів «двері» та «вікно», «дерево» та «двері», «дерево» та «вікно».

7. Знаходження слова, яке не трапляється разом з іншими словами із вказаного переліку. В таблиці 9 наведено результати виявлення «зайвих» слів серед таких груп слів: двері, вікно, стіл, підлога; страва, кава, піца, пиво; гарний, чудовий, поганий, прекрасний, добрий.

Таблиця 8. Косинусна відстань між векторами слів.

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
«двері», «вікно»			
0.3674571507410046	0.2518291292078544	0.4794622437212813	0.4456461563744466
«дерево», «двері»			
0.78357357	0.7357911	0.6791656	0.823482
«дерево», «вікно»			
0.5182467	0.56578034	0.55148625	0.7362342

Таблиця 9. Зайві слова з переліку.

word2vec (без нормалізації)	word2vec (нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
стіл, пиво, поганий	стіл, страва, поганий	підлога, пиво, поганий	стіл, пиво, поганий

8. Знаходження слова в переліку, яке найбільш подібне до вказаного слова. В таблиці 10 наведено результати виявлення найбільш подібного слова до слова «страва» серед таких слів: двері, вікно, стіл, підлога.

Таблиця 10. Найбільш подібні слова до слова «страва».

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
підлога	стіл	стіл	підлога

9. Знаходження всіх слів у моделі, які ближчі до вказаного слова, ніж інше вказане слово. В таблиці 11 наведено перші п'ять слів до слова «турист» ближчих, ніж слово «офіціант».

Таблиця 11. П'ять слів до слова «турист» ближчих, ніж слово «офіціант».

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
футболіст, громадянин, гравець, корабель, підприємець	особа, люди, населення, чоловік, діти	письмен-ник, депутат, футболіст, учасник, поет	письменник, журналіст, музикант, дослідник, туризму

10. Обчислення відстані між словами (word mover's distance) двох документів. У таблиці 12 наведено результати обчислення під час зіставлення документа, який містить такі слова: *чемні, офіціанти, гарне, обслуговування, чудова, атмосфера*; з двома іншими документами, які складаються зі слів: *смачно, затишно, привітний, персонал; напружена, політична, ситуація* відповідно.

Таблиця 12. Word Mover's Distance між двома парами документів.

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
1.1706410229339907	1.1557630449366882	1.0951630263398817	1.1146025419092178
1.359728996489791	1.401208365984857	1.2834964067663837	1.2675763311834924

Оцінка векторних представлень

Дослідження векторних представлень дають можливість зрозуміти характер результатів за виявленою подібностю між словами, але питання оцінки векторних представлень з позиції їхньої якости для розв'язання практичних завдань залишається не тільки актуальним, але і гостро дискусійним.

Загалом розрізняють два підходи до оцінки векторних представлень слів: на основі внутрішніх оцінок та на основі зовнішніх оцінок. Використання зовнішніх оцінок передбачає оцінювання якости векторних представлень на основі результатів їхнього застосування для розв'язання реальних завдань. Наприклад, якщо вдалий добір параметрів під час побудови векторного представлення спричинив збільшення точности аналізу тональности тексту, то якість цього векторного представлення вважається вищою. Внутрішнє оцінювання базується на використанні спеціяльних тестових наборів (тести аналогій – вирази виду "A до B як C до D") та вручну промаркованих даних. Для української мови тести аналогій розробила Тетяна Кодлюк, ці тести містять дві групи аналогій: синтаксичні (singular-plural, adjective-adverb, opposite, comparative, supperlative, past tense, verb forms) та семантичні аналогії (country-capital. country-region, family, country-nationality, currency). Результати (таблиця 13) оцінки векторних представлень на основі аналогій підтверджують, що обсяг корпусу, на основі якого будуються векторні представлення слів, має визначальний вплив.

Для української мови, крім уже згаданих тестів аналогій, автору не відомі інші набори даних, які були розроблені для оцінювання якости векторних

представлень або могли би бути використані для цього. Серед ресурсів, які можна адаптувати для цього завдання, потрібно згадати український асоціативний словник[29]. Доцільність використання цього словника зумовлена тим, що його укладено на основі опитування репрезентативної вибірки людей, а його обсяг становить 841 слово-стимул. Адаптація необхідна тому, що в реакціях, які були зібрані на кожен зі стимулів, трапляються слова, що їх можна оцінювати як подібні до стимулу, і слова, які однозначно стосуються контексту вживання стимулу. Наприклад, для слова – стимулу «розмова» зібрані такі реакції (у таблиці 14 наведено реакції, які траплялися більше одного разу).

Серед цих реакцій можна назвати такі: *бесіда, спілкування, діялог* та подібні до них, які трапляються в результатах пошуку подібних векторів до вектора слова «розмова». Окрім цих реакцій, наявні такі: *приємна, щира, відверта, дружня* та подібні до них, які, ймовірно, належать до слів, які трапляються в контексті слова-стимулу.

Таблиця 13. Результати оцінки векторних представлень на основі тестів аналогій.

word2vec (без нормалізації)	word2vec (з нормалізацією)	FastText (Вікіпедія)	MUSE (Вікіпедія)
country-capital: 35.8% country-region: 16.9% family: 37.8% : country-nationality: 63.5% singular-plural: 44.4% adjective-adverb: 27.7% opposite: 33.3% currency: 5.4% comparative: 53.3% supperlative: 58.3% past tense: 88.6% verb forms: 77.1%	country-capital: 55.9% country-region: 31.7% family: 60.2% country-nationality: 70.3% singular-plural: 50.0% adjective-adverb: 35.3% opposite: 30.4% currency: 21.8% comparative: 45.8% supperlative: 54.6% - -	country-capital: 63.0% country-region: 36.8% family: 55.6% country-nationality: 84.0% singular-plural: 77.6% adjective-adverb: 55.6% opposite: 25.0% currency: 0.0% comparative: 50.0% supperlative: 100.0% past tense: 97.4% verb forms: 83.5%	country-capital: 40.9% country-region: 50.8% family: 47.4% country-nationality: 82.2% singular-plural: 71.6% adjective-adverb: 27.0% opposite: 25.0% currency: 1.4% comparative: 38.1% supperlative: 30.0% past tense: 100.0% verb forms: 98.1%
total: 43.1%	total: 45.5%	total: 63.3%	total: 52.7%
Semantic accuracy: 30.25% Syntactic accuracy: 60.70%	Semantic accuracy: 47.74% Syntactic accuracy: 39.11%	Semantic accuracy: 57.08% Syntactic accuracy: 72.78%	Semantic accuracy: 53.22% Syntactic accuracy: 50.47%

[29] Мартінек С.В. Український асоціативний словник: У 2 т. – 2-ге вид. Т.І: Від стимулу до реакції; Т.ІІ: Від реакції до стимулу (Львів: ПАІС, 2008).

Таблиця 14. Реакції на стимул «розмова» та слова з подібними векторами.

m I спілкування 12, діалог 6, цікава 4, бесіда, друг, дружня, коротка, слово 3: говорити, довга, контракт, плітки, слова, тиша 2	f I діалог, приємна, спілкування 7; щира 6; відверта 5; бесіда, цікава 4; весела, дружня, людей, полегшення, потреба, слова, слово, спокійна, тиха 2
бесіда, 0.843465566635131; перемовини, 0.67681932449340переговори, 0.6575664877891; дискусія, 0.646611809730529; зустріч, 0.641802549362182; полеміка, 0.5988496541976929 спілкування, 0.59169781208038; перемови, 0.58026254177093; діалог, 0.56119084358215; балачка, 0.555352210998535; дебати, 0.54873085021972; консультація, 0.547969102859497 розповідь, 0.5417054891586304; листування, 0.5316982269287109	

Адаптація також повинна передбачати спосіб встановлення міри подібности між словами на основі даних зі словника. В українському асоціятивному словнику є інформація про загальну кількість реакцій на стимул, кількість реакцій окремо респондентів чоловічої та жіночої статі й кількість виникнення кожної з реакцій. Результати оцінки векторних представлень за допомогою тестових даних, побудованих на основі українського асоціятивного словника, наведено у таблицях 15 та 16. Тестові дані – це пари, утворені зі слів-стимулів та перших найчастотніших реакцій на них. Міра подібности між словами визначалася як відношення кількости випадків виникнення цієї реакцій до загальної кількости реакцій на стимул. У таблиці 15 наведено результати оцінки векторних представлень на основі реакцій респондентів жіночої статі. В таблиці 16 наведено результати оцінки векторних представлень на основі реакцій респондентів чоловічої статі до та після видалення пар слів, які містять службові частини мови, словосполучення та слова контексту.

Таблиця 15. Результати оцінки векторних представлень на основі реакцій респондентів жіночої статі.

	Коефіцієнт кореляції Пірсона	Коефіцієнт кореляції рангу Спірмена	Відсоток пар із невідомими словами
word2vec (без нормалізації)	0.0429	0.1064	17.1 %
word2vec (з нормалізацією)	0.0528	0.1143	23.0 %
FastText (Вікіпедія)	0.1418	0.1564	19.7 %
MUSE (Вікіпедія)	0.1836	0.1786	27.5 %

Таблиця 16. Результати оцінки векторних представлень на основі реакцій респондентів чоловічої статі.

	Коефіцієнт кореляції Пірсона до/після адаптації	Коефіцієнт кореляції рангу Спірмена до/після адаптації	Відсоток пар із невідомими словами до/після адаптації
word2vec (без нормалізації)	0.1039 / 0.0515	0.1520 / 0.0648	15.0 % / 9.8 %
word2vec (з нормалізацією)	0.0742 / 0.0158	0.1167 / 0.0621	19.9 % / 8.2 %
FastText (Вікіпедія)	0.1960 / 0.0758	0.2391 / 0.0789	17.7 % / 11.4 %
MUSE (Вікіпедія)	0.2662 / 0.1028	0.2882 / 0.1202	26.2 % / 21.2 %

Зіставлення результатів оцінки з результатами, які одержані для англійської мови[30] з використанням тестових наборів[31] wordsim353 та SimLex-999 (таблиця 17) свідчить про те, що потрібно провести додаткові дослідження для адаптації даних з українського асоціятивного словника.

Таблиця 17. Порівняння результатів оцінки для української та англійської мов.

Модель	Розмір корпусу	Word similarity (SimLex-999) Пірсон / Спірмен:	Word similarity (WS-353) Пірсон / Спірмен	Асоціятивний словник Пірсон / Спірмен
Word2Vec	1 млн.	0.17/ 0.15	0.37 / 0.37	— / —
FastText	1 млн.	0.13 / 0.11	0.36/ 0.36	— / —
Word2Vec	>0.5 млн.	— / —	— / —	0.1039 / 0.0515 0.1520 / 0.0648
FastText	2 1 млн.	— / —	— / —	0.1960 / 0.0758 0.2391 / 0.0789

Висновки

Векторні представлення слів – це ефективна технологія для розв'язання багатьох завдань опрацювання природної мови. Розроблені інструменти для побудови векторних представлень слів word2vec, FastText та подібні до них Glove, lex2vec дозволяють отримати вектори належної якости за умови правильного вибору параметрів тренування моделі. Результати оцінюван-

[30] Parul Sethi, "WordRank embedding: 'crowned' is most similar to 'king', not word2vec's'Canute'»,переглянуто01.04.2018,https://rare-technologies.com/wordrank-embedding-crowned-is-most-similar-to-king-not-word2vecs-canute/.

[31] Lev Finkelstein, et al., "Placing Search in Context: The Concept Revisited", *ACM Transactions on Information Systems*, Vol. 20 Issue 1 (January 2002): 116-131, https://doi.org/10.1145/503104.503110, Felix Hill, Roi Reichart, Anna Korhonen, "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation", *Computational Linguistics,* Vol. 41 Issue 4 (December 2015): 665-695, https://doi.org/10.1162/COLI_a_00237.

ня векторних представленнь слів для української мови вказують на потребу збільшення обсягу даних для тренування, і відповідно їхнє ефективне застосування для розв'язання завдань опрацювання природної мови залишається відкритим.

Отже, важливо продовжувати й розвивати дослідження, які стосуються побудови векторних представлень слів на основі всіх відомих моделей, та звернути увагу на актуальні завдання в цій галузі[32]:

1. Відсутність теоретичних досліджень у галузі векторних представлень слів зумовлює нагальну потребу в їхньому проведенні. Успішне використання моделі word2vec не зменшує потребу в теоретичних дослідженнях явищ, які спостерігаються у векторних просторах слів, і автору не відомі такі роботи для української мови.

2. Оцінка отриманих векторних представлень слів потребує подальшого розвитку. Потрібно шукати баланс між якістю та затратами на оцінку векторів слів на основі відомих методів та працювати над створенням нових.

3. Потрібно розробляти методи для адаптації готових векторних представлень слів до спеціялізованих доменів, для яких відсутні корпуси необхідного обсягу, що дозволить отримати спеціялізовані векторні представлення. Розв'язувати це завдання можна також додаванням до готових векторних представлень семантичної інформації для їхнього використання в певному домені;

4. Пошук методів виявлення та усунення упереджень та стереотипів (наприклад, ґендерних), яких набувають вектори слів за їхньої побудови.

5. Пошук ефективних способів додавання символьної інформації до вже побудованих векторів слів. Така потреба виникає під час застосування векторних представлень для морфологічного аналізу, видобування іменованих сутностей, синтаксичного аналізу й машинного перекладання.

6. Розвиток методів побудови векторів для слів, відсутніх у корпусі текстів, на ґрунті якого здійснюється тренування. Підходи на основі моделі FastText можуть отримати розвиток для розв'язання цього завдання.

7. Розвиток способів побудови векторів слів, які мають кілька значень, або теоретичне та практичне доведення, що готові векторні представлення не потребують вирішення цього завдання.

8. Пошук ефективних способів побудови векторних представлень для висловів та багатослівних конструкцій, значення яких можна розглядати і як поєднання значень слів, з яких вони складаються, і як зовсім інше значення.

9. Розроблення методів для врахування діяхронічної природи слів, оскільки значення слів та їхнє використання з часом змінюються.

10. Розвиток багатомовних векторних представлень слів та їхнє ефективне використання подібне до подібне до запропонованого в праці Алексіса

[32] Sebastian Ruder, "Word embeddings in 2017: Trends and future directions", переглянуто 01.04.2018, http://ruder.io/word-embeddings-2017.

Конно зі співавторами[33] для роботи з мовами, для яких не створені корпуси необхідного обсягу.

11. Розроблення способів збільшення типів контекстних даних, на основі яких будуються векторні представлення. Такими контекстними даними можуть бути синтаксичні структури речення, інформація з різноманітних словників та інших джерел структурованих даних.

Andriy ROMANYUK

Vector Representations of Ukrainian Words

Andriy ROMANYUK— *PhD, Associate Professor at Applied Sciences Faculty Ukrainian Catholic University. Natural language processing. a.romanyuk@ucu.edu.ua*

In this paper, Ukrainian word embeddings and their properties are examined. Provided are a theoretical description, a brief account of the most common technologies used to produce an embedding, and lists of implemented algorithms. Word2wec, the first technology for calculating word embeddings, is used to demonstrate modern approaches of calculating using neural networks. Word2wec and FastText, which evolved from word2vec, are compared, and FastText's benefits are described.

Word embeddings have been applied to solving majority of the practical tasks of natural language processing. One of the latest such applications have been in the automatic construction of translation dictionaries. A previous analysis indicates that most of the words found in English-Ukrainian dictionaries are absent in the Great Electronic Dictionary of the Ukrainian Language (VESUM) project. For embeddings in Ukrainian based on word2vec, Glove, lex2vec, and FastText, the Gensim open-source library was used to demonstrate the potential of calculated models, and the results of repeating known calculation experiments are provided. They indicate that the hypothesis about the existence of biases and stereotypes in such models does not pertain to the Ukrainian language. The quality of the word embeddings is assessed on the basis of testing analogies, and adapting lexical data from a Ukrainian associative dictionary in order to construct a selection of data for assessing the quality of word embeddings is proposed. Listed are necessary tasks of future research in the field of creating and utilizing Ukrainian word embeddings.

Keywords: natural language processing, word embeddings, word2vec, FastText.

[33] A. Conneau, et al., "Word Translation Without Parallel Data", arXiv:1710.04087v3.

Bibliography

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov."Enriching word vectors with subword information". *Transactions of the Association for Computational Linguistics* 5 (2017): 135-146, https://doi.org/10.1162/tacl_a_00051.

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". *Science* 356, no. 6334 (Apr 2017): 183-186. https://doi.org/10.1126/science.aal4230.

Camacho-Collados, Jose, Mohammad Taher Pilehvar. "From Word to Sense Embeddings: A Survey on Vector Representations of Meaning". arXiv:1805.04032v2.

Chui, Michael, James Manyika and Mehdi Miremadi. "What AI can and can't do (yet) for your business". *McKinsey Quarterly,* January 2018. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/what-ai-can-and-cant-do-yet-for-your-business.

Conneau, Alexis, Guillaume Lample, Ludovic Denoyer, Marc'Aurelio Ranzato and Herve Jégou. "Word Translation Without Parallel Data". arXiv:1710.04087v3.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman and Eytan Ruppin. "Placing Search in Context: The Concept Revisited". *ACM Transactions on Information Systems* 20 no. 1 (January 2002): 116-131. https://doi.org/10.1145/503104.503110.

Fomenko, Volodymyr, Heorhii Loutskii, Pavlo Rehida and Artem Volokyta. "Thematic Texts Generation Issues Based on Recurrent Neural Networks and word2vec". *Технічні науки та технології,* №4 (2017): 110-115.

Halper, Fern. *Advanced Analytics: Moving Toward AI, Machine Learning, and Natural Language Processing.* 1105 Media, Inc., 2017. https://www.sas.com/ru_ua/whitepapers/tdwi-advanced-analytics-ai-ml-nlp-109090.html.

Hill, Felix, Roi Reichart and Anna Korhonen. "SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation". *Computational Linguistics* 41 no. 4 (December 2015): p.665-695. https://doi.org/10.1162/COLI_a_00237.

Jain, Jayant. "FastText and Gensim word embeddings". Perehlianuto 01.04.2018. https://rare-technologies.com/fasttext-and-gensim-word-embeddings/.

Jurafsky, Dan, James H. Martin. *Speech and Language Processing*. https://web.stanford.edu/~jurafsky/slp3/.

Li, Yang, Tao Yang. "Word Embedding for Understanding Natural Language: A Survey". in *Guide to Big Data Applications*, S. Srinivasan eds. Houston: Jesse H. Jones School of Business Texas Southern University, 2018. https://www.researchgate.net/publication/315717021_ Word_Embedding_for_Understanding_Natural_Language_A_Survey.

Martinek S.V. *Ukrains'kyi asotsiatyvnyi slovnyk: U 2 t. T. I: Vid stymulu do reaktsii. T.II: Vid reaktsii do stymulu.* 2-he vyd. L'viv: PAIS, 2008.

McCormick, C."Word2Vec Tutorial – The Skip-Gram Model". Perehlianuto 01.04.2018. http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/.

McCormick, C. "Word2Vec Tutorial Part 2 – Negative Sampling". Perehlianuto 01.04.2018. http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/.

McGovern, Tim, ed. *Artificial Intelligence Now.* O'Reilly Media, Inc., 2017. https://www.oreilly.com/library/view/artificial-intelligence-now/9781492049210/.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. "Efficient estimation of word representations in vector space", arXiv preprint arXiv:1301.3781.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality". In *Proceedings Advances in Neural Information Processing Systems 26 (NIPS 2013).* 3111–3119.

Ruder, Sebastian. "Word embeddings in 2017: Trends and future directions". Perehlianuto 01.04.2018. http://ruder.io/word-embeddings-2017.

Sethi, Parul. "WordRank embedding: 'crowned' is most similar to 'king', not word2vec's 'Canute'.» Perehlianuto 01.04.2018. https://rare-technologies.com/wordrank-embedding-crowned-is- most-similar-to-king-not-word2vecs-canute/.

Velykyi elektronnyi slovnyk ukrains'koi movy (VESUM). Perehlianuto 01.04.2018. https://github.com/brown-uk/dict_uk.

"What is the main difference between word2vec and fastText?" Perehlianuto 01.04.2018. https://www.quora.com/What-is-the-main-difference-between-word2vec-and-fastText.

Xun, Guangxu, Vishrawas Gopalakrishnan, Fenglong Ma, Yaliang Li and Jing Gao. "Aidong ZhangTopic Discovery for Short Texts Using Word Embeddings". *in Proceedings IEEE 16th International Conference on Data Mining (ICDM) (2016),* 1299-1304.

Yin, Wenpeng, Hinrich Schütze. "Discriminative Phrase Embedding for Paraphrase Identification". arXiv:1604.00503v1.

Zou, W. Y., R. Socher, D. M. Cer, C. D. Manning. "Bilingual word embeddings for phrase-based machine translation". in *Proceedings of EMNLP* (2013), 1393-1398.