Issue 13, 2023

https://doi.org/10.23939/sisn2023.13.189

УДК 004.8, 004.032.26

МЕТОД ПОБУДОВИ ЕМБЕДИНГІВ ОЗНАК У ЗАДАЧАХ ГЛИБИННОГО НАВЧАННЯ НА ОСНОВІ ОНТОЛОГІЙ

Василь Литвин[1], Соломія Мушаста[2]

Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж, Львів, Україна
[1] E-mail: vasyl.v.lytvyn@lpnu.ua, ORCID: 0000-0002-9676-0180
[2] E-mail: solomiyanytrebych@gmail.com, ORCID: 0000-0003-4932-4113

У роботі досліджено проблему ембедингу ознак, які використовують у датасетах для навчання нейронних мереж. Використання ембедингів підвищує продуктивність нейронних мереж, а отже, є важливою ланкою підготовки даних для методів глибинного навчання. Такий процес ґрунтується на семантичній метриці. Запропоновано для ембедингу використовувати онтології предметних областей, до яких належить відповідна ознака. У цій роботі розроблено такий метод й досліджено його використання для завдання рубрикування текстових документів. Результати досліджень підтвердили перевагу розробленого методу.

Ключові слова: онтологія; нейронна мережа; ембединг; семантична метрика; природномовний текст.

Вступ

У сфері опрацювання природних мов (Natural Language Processing, NLP) зазвичай користуються словниками, що складаються із тисячі слів. Ці словники вводяться у модель з використанням методики швидкого кодування (One-Hot Encoding). Із математичного погляду це рівносильно наявності окремого стовпця для кожного слова. Коли слово передається в модель, у відповідному стовпці ставлять одиницю, тоді як у всіх інших – нулі. Це зумовлює появу дуже сильно розріджених наборів даних. Вирішення цієї проблеми полягає в створенні ембедингу. Ембединг – загальна назва низки методик мовного моделювання та навчання ознак під час опрацювання природної мови, в яких слова або фрази зі словника відображають у вектори дійсних чисел. По суті, йдеться про те, що ембединг, на основі навчального тексту, групує слова з подібним значенням і повертає їхнє місце розташування. Наприклад: значення ембедингу слова "fun" може бути подібне до значень ембедингів слів "humor" і "dancing" або словосполучення "machine learning". Тобто для отримання ембедингу використовують деяку семантичну метрику. Пропонуємо будувати такі ембединги на основі онтологій.

Постановка проблеми

Створення ембедингів ознак (*feature embeddings*) – один із найважливіших етапів підготовки табличних даних, який використовують, наприклад, для навчання нейромережевих моделей. На практиці нейронні мережі демонструють істотне підвищення продуктивності під час застосування подібних репрезентативних властивостей. Якщо його не використовувати з такими даними, то це призводить до значного погіршення точності моделей. Вважається, що алгоритми градієнтного

бустингу є найкращим вибором для вирішення завдань, що передбачають роботу зі структурованими наборами даних. Насправді нейромережеві методи моделювання, поліпшені за рахунок ембедингів, часто дають кращі результати, ніж методи, основані на градієнтному бустингу. Ускладнення виникають переважно, якщо нейромережевим моделям доводиться працювати з розрідженими категоріальними ознаками. Ембединги – це можливість зменшення розмірності таких ознак. Це сприяє підвищенню продуктивності моделі. Структуровані набори даних теж часто містять розріджені категоріальні ознаки. Наприклад, у даних про продажі можуть бути стовпці з поштовим індексом та ідентифікатором магазину. Оскільки у цих стовпцях можуть міститись сотні або тисячі унікальних значень, їх використання призведе до появи проблем із продуктивністю нейромереж. У цьому випадку теж можна скористатися ембедингами.

Сьогодні вирішення завдання побудови ембедингів основане на деякій семантичній метриці, що, як правило, ґрунтується на статистичному аналізі. Однак такі метрики не повністю враховують семантики понять, які є значеннями певних ознак. Таку проблему можна вирішити, насамперед, застосуванням онтологій.

Аналіз останніх досліджень та публікацій

Розглянемо підходи до побудови семантичних метрик. Всі відомі роботи стосуються функціонування систем, які вирішують проблеми, пов'язані із опрацюванням природомовних текстів. До них належать: проблема семантичної близькості слів, семантичної багатозначності та лексико-семантичної неоднозначності, що вирішуються за допомогою використання метрик для вимірювання семантичної відстані. Однак для кожної конкретної ПО, залежно від розв'язуваної задачі, евристично вибирають певну метрику для підрахунку семантичної відстані, тому що ця метрика залежить від топології ПО. Загалом на вибір метрики впливають [1, 3–5]:

– наявність перегинів;

– довжина шляху між поняттями;

– різні типи перегинів;

– різні типи зв'язків, які існують в онтології;

– максимальна довжина шляху між поняттями;

– урахування локального і глобального контексту.

Дуже важливо, визначаючи семантичну близькість понять, враховувати особливість шляху між поняттями [2]:

– чим довший шлях між поняттями, тим менша семантична близькість;

– наявність перегину на шляху знижує семантичну близькість;

– різні типи перегинів на шляху можуть по-різному впливати на семантичну близькість;

– перегин шляху на високому рівні ієрархії гірший, ніж на нижчому рівні.

Під час вибору відповідної метрики також необхідно враховувати топологію зв'язків між поняттями. За звичайної класифікації, тобто за наявності в основному зв'язку "IS-A" набагато вигідніше взяти простішу метрику, таку як, наприклад, метрика Leacock і Chodorov, що у дослідженнях теж враховували тільки таксономію понять. Якщо є багато типів зв'язків, то необхідно визначитися з вагою кожного типу зв'язку. Оцінки підбирають евристично залежно від топології мережі [1, 7–10].

Виділяють кілька способів визначення подібності текстових документів. Вважають, що документи і запити подаються за допомогою індексних термінів або ключових слів. Позначимо за допомогою символу |.| – розмір множини ключових слів, що задають ТД. Простий коефіцієнт відповідності $|X \cap Y|$ відображає кількість загальних індексних термінів. Обчислюючи коефіцієнт, не беруть до уваги потужності множин X і Y. Коефіцієнти подібності для документів на основі ключових слів наведено у табл. 1.

Таблиця 1

Коефіцієнти подібності для документів на основі ключових слів

Формула	Назва
$\frac{\lvert X \cap Y\rvert}{\lvert X\rvert+\lvert Y\rvert}$	Коефіцієнт Дайса (dice)
$\frac{\lvert X \cap Y\rvert}{\lvert X \cup Y\rvert}$	Коефіцієнт Джаккарда (jaccard)
$\frac{\lvert X \cap Y\rvert}{\lvert X\rvert^{\frac{1}{2}}\times\lvert Y\rvert^{\frac{1}{2}}}$	Косинусний коефіцієнт
$\frac{\lvert X \cap Y\rvert}{\min\left(\lvert X\rvert,\lvert Y\rvert\right)}$	Коефіцієнт перекриття

У роботі [6] детально проаналізовано наявні семантичні метрики, які можна використати для визначення схожості текстових документів (ТД), які ґрунтуються на:

1) залежності від частоти слів у ТД;

2) введенні відстані в заданій таксономії понять;

3) одночасно на основі першого та другого підходів.

Метрики, які ґрунтуються на подібності таксономії, використовують для обчислення подібності концептів WordNet [11], GermaNet [12], Вікіпедій.

К. Леачок і М. Холоров [5] запропонували обчислювати близькість концептів як відстань між концептами в таксономії, нормалізовану за допомогою урахування глибини таксономії. У формулі $lch(C_1,C_2) = -\log\frac{length(C_1,C_2)}{2D}$, функція $length(C_1,C_2)$ – кількість вершин вздовж найкоротшого шляху між вершинами C_1 і C_2; D – максимальна глибина таксономії. У роботі автори розглянули тільки одне відношення IS-A і тільки між іменниками.

У роботі [12] використано відстань, що задає як глибину концептів у ієрархії, так і глибину найближчого спільного батька *lcs* (least common subsumer): $wup(C_1,C_2) = \frac{lcs(C_1,C_2)}{depth(C_1)+depth(C_2)}$.

П. Рєзник [7, 13] запропонував вважати, що відстань між термінами тим менша, чим інформативніший термін, розміщений на вищому щаблі в ієрархії.

У роботі [14] метрику П. Рєзника *res* адаптовано до вікіпедій:

$$res_{hypo}(C_1,C_2) = 1 - \frac{\log\left(hypo\left(lcs(C_1,C_2)\right)+1\right)}{\log(C)},$$

де *lcs* – найближчий спільний батько концептів C_1 і C_2; *hypo* – кількість гіпонімів цього батька; C – загальна кількість концептів у ієрархії.

У статті [9] *lin* визначає подібність об'єктів *A* і *B* як відношення кількості інформації, необхідної для опису подібності *A* і *B*, до кількості інформації, що цілком описує *A* і *B*. Для вимірювання подібності між словами *lin* враховує частотний розподіл слів у тексті (аналогічно до міри П. Рєзника):

$lin(C_1,C_2) = \frac{2\times\log\left(P(C_0)\right)}{\log\left(P(C_1)\right)+\log\left(P(C_2)\right)}$, де C_0 – найближчий загальний суперклас в ієрархії для обох концептів C_1 і C_2.

Проаналізувавши наведені вище метрики, можна зробити висновок, що жодна з них не враховує онтологію предметної області, а онтологія задає семантичну специфіку та таксономію понять як предметної області, так і задач, які у ній виникають. Тому актуальне дослідження розв'язування прикладних семантичних задач, зокрема побудови ембедингів ознак із використанням онтологічного підходу

Формулювання цілі статті

У цій роботі запропоновано ввести семантичну метрику на основі онтологій для побудови ембедингів ознак, які використовують у нейромережевих моделях для розв'язування прикладних задач методами глибинного навчання.

Виклад основного матеріалу

Нехай побудовано онтологію *O* деякої предметної області (ПО):

$$O = \langle C, R, F \rangle,$$

де *C* – задає множину термінів предметної області, множину понять (концептів, термінів) предметної області, яку задає онтологія *O*; $R: C \to C$ – множина взаємозв'язків між термінами заданої предметної області, тобто це відображення *C* у *C*; *F* – множина інтерпретацій з використанням дескриптивної логіки**.**

Множину взаємозв'язків між термінами *R* можна поділити на групи (кореляти, синонімія, рід – вид, частина – ціле тощо) – $R = \{R_1, R_2, ..., R_k\}$. Позначимо через n_i кількість відношень типу R_i в онтології. Тип взаємозв'язку матиме більшу вагу, якщо цей тип частіше трапляється в онтології. Вагу типу відношення визначимо як $L_i = \frac{n_i}{N}$.

Зважимо нашу семантичну мережу, яка задає онтологію:

$$l_i = \frac{K}{L_i} = \frac{K \times N}{n_i},$$

де K – деяка константа, яка відображає онтологію [1].

Зауважимо, що у роботі [16] типам зв'язків присвоєно такі ваги важливості: для зв'язку типу спеціалізації ("IS-A") – $s = 0{,}9$; для узагальнення ("KIND-OF") – $g = 0{,}4$; для причинного зв'язку ("CAUSED-BY") – $r_{CBY} = 0{,}3$, для характеризуючого зв'язку ("CHARACTERIZED-BY") – $r_{WRT} = 0{,}2$.

Визначимо множину значень певної ознаки (терміна в онтології) $C = \{C_1, C_2, ...C_n\}$, які, на думку експерта, найкраще відображають ознаку в деякому датасеті. Метод знаходить множину документів, в яких є подібні терміни. Для кожного документа T_s побудуємо множину потужністю m, яка складається з термінів, які входять в онтологію ПО й частіше трапляються у документі T_s: $\hat{C}^s = \{\hat{C}_1^s, \hat{C}_2^s, ..., \hat{C}_m^s\}$. Методом Флойда – Уоршалла або Дейкстри [5], використовуючи семантичну метрику, яку розробив В. В. Литвин і яка описана в роботах [1, 16], знайдемо $n \times m$ найкоротших відстаней $d_{ij}^s = d(C_i, C_j^s)$ між термінами з множини C та $\hat{C}^s$. Якщо множина $\hat{C}^s$ містить менше ніж m елементів, то вважаємо, що решта відстаней $d_{ij}^s = d(C_i, C_j^s)$, яких не вистачає, дорівнюють найдовшій відстані від елемента C_i в межах зваженого концептуального графа. Тоді відстань до знайденого документа T_s обчислюють відповідно до формули: $d^s = \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij}^s$. Ранжуємо знайдені терміни в онтології за зростанням значення d^s.

Для прикладу розглянемо онтологію матеріалознавства, фрагмент якої наведено на рис. 1. Візьмемо ключове слово corrosion (корозія). Дослідимо три текстові документи, для яких застосуємо

ембединги. Тобто замінимо ці тексти на числові величини, залежно від відстані до терміна corrosion (корозія), використовуючи онтологію матеріалознавства.

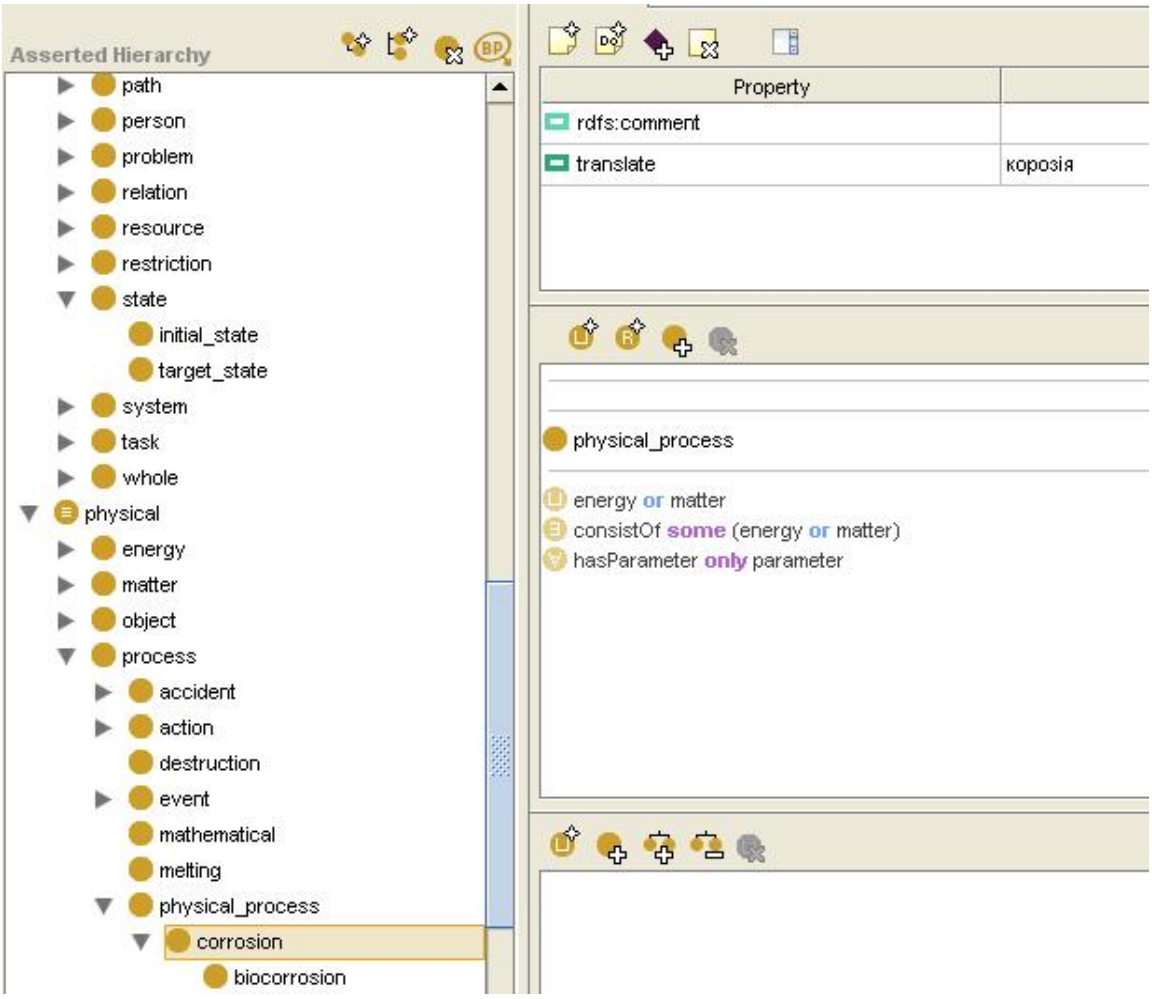

Рис. 1. Фрагмент онтології матеріалознавства

1. *The correlation between difractometric inverstigations and calculations, based on the model of rigid spheres, allowed us to make prediction of the change of the surface tension and to evaluate the steel wettability by extremum of a continuous function of structural melt factor. The influence of stainless steel elements laser doped into the surface on structural factors of melts Pb and Li Pb was investigated*).

2. *The damaging of power plant equipment, made of stainless austenitic steels is considered. It has been found that initiation of intergranular stress corrosion cracks in the weld region of the welded joints made of this steel is caused by interaction of 3 factors – the determined degree of basic metal sensitization, high service stress, that is higher than the material yield strength and the increased oxygen concentration in the heat carrier*).

3. *The expressions for the change of initial stresses, caused by small disturbance of the basic state of any nature (thermal, electrical, magnetic) are constructed. Two types of expressions, differing in kinetics, were obtained. In the first case the change is caused by disturbance of the deformed state. In the second one – by a small additional deformation and also by disturbance of other basic state parameters in both cases. The physical sense of various components of expressions is analyzed*).

Концептуальні графи цих анотацій з вагами важливості понять та відношень наведено на рис. 2. Ваги важливості понять задано частотним методом, тобто частотою вживання цих понять у наукових

текстах із матеріалознавства. Зважені концептуальні графи цих текстів наведено на рис. 3. Над поняттями наведено їх індекси.

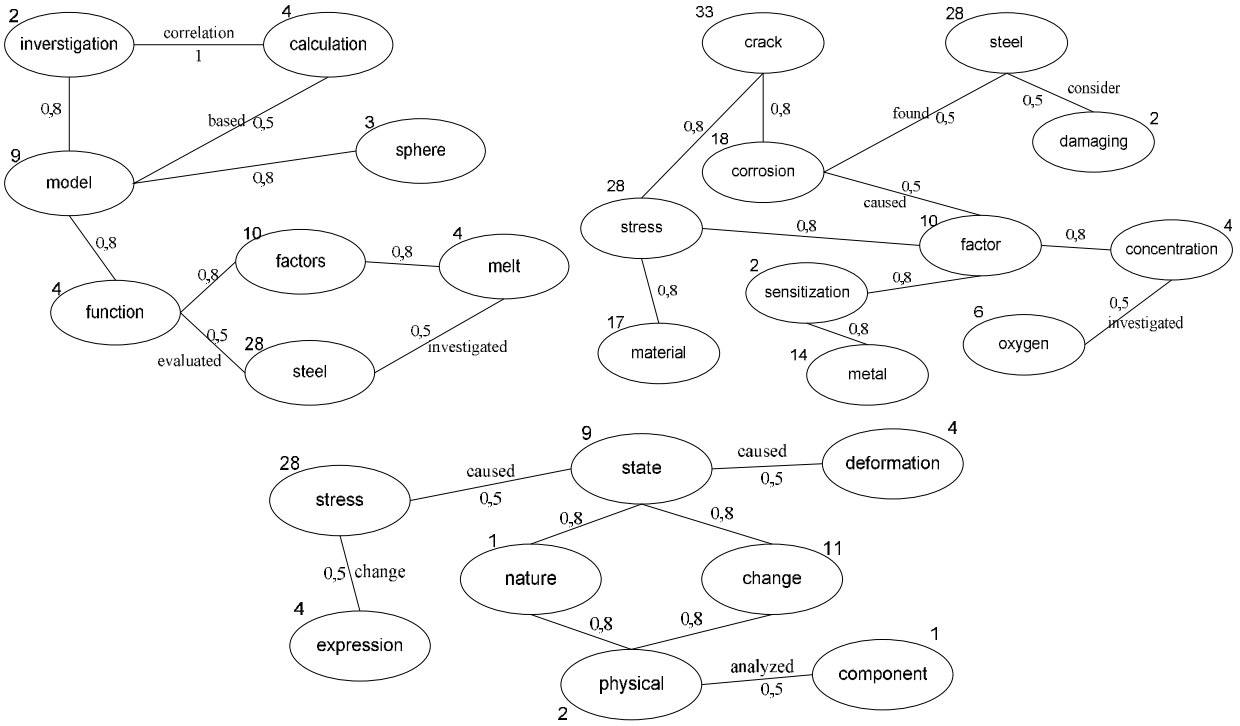

Рис. 2. Концептуальні графи трьох текстових документів

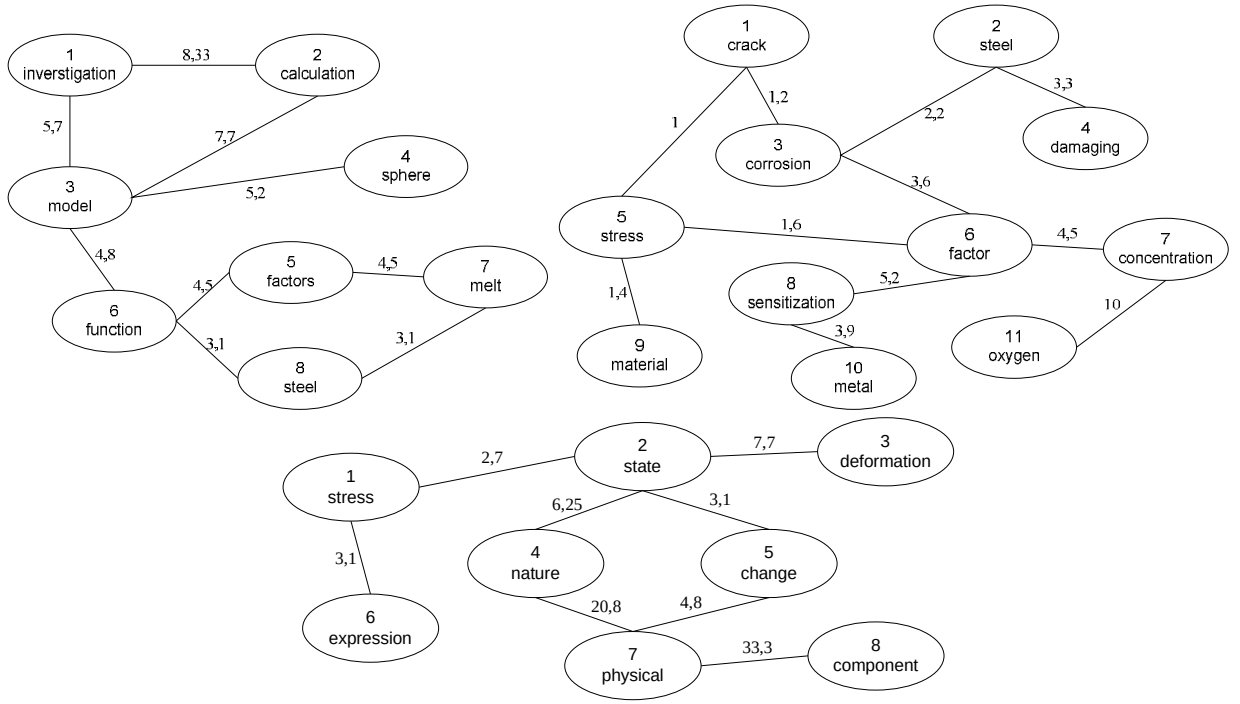

Рис. 3. Зважені концептуальні графи трьох текстових документів

Наприклад, бачимо, що для першого документа, відповідно до онтології матеріалознавства, шлях від “corrosion” до “model” такий: “corrosion” – “physical process” – “process” – “model”. Враховуючи ваги понять та ваги зв’язків (перші два ієрархічні, третій функціональний), отримаємо: $d = 4,6$. Для третього документа: “corrosion” – “physical process” – “process” – “state”. Відстань дорівнюватиме $d = 4,9$, оскільки ваги понять “model” та “state” становлять: $W = 9$. Аналогічно можна знайти відстані до інших термінів від ключового слова “corrosion”.

Подивимося, як змінилися результати роботи нейронної мережі після ембедингу на прикладі задачі класифікації текстових документів [17]. Класифікація текстових документів (ТД) розглядається як один з можливих варіантів вирішення проблеми використання інформаційних ресурсів. Коротко її характеризують так. Різні сховища знань (зокрема і бібліотеки) нагромадили величезні інформаційні масиви. Проблема полягає в складності орієнтування у цих масивах через неадекватність їх розмірів. Класифікацію природомовних текстів називають рубрикуванням. Використання рубрикаторів дає змогу зменшити витрати на пошук потрібної інформації, поданої електронними текстами. Застосування семантичного підходу (онтологій) дає змогу підвищити релевантність такого пошуку, тоді як використання методів самонавчання (штучних нейронних мереж, генетичних алгоритмів, байєсівських мереж) спрощує процедуру побудови класифікатора [18, 19]. Задача класифікації тестових документів (ТД) визначається так. Існує множина текстів $T = \{T_1, T_2, ..., T_M\}$, множина N рубрик, які розглядатимемо як класи Class={Class$_1$, Class$_2$,…, Class$_N$}. Процедура класифікації f текстів T_i Î T полягає у виконанні певних процедур, на основі яких роблять висновок про відповідність T_i одній зі структур Class$_j$, що означає зарахування T_i до класу Class$_j$.

Результати досліджень наведено у табл. 2. Порівнювали результати рубрикування на основі тієї самої структури нейронної мережі: 1) з ембедингом, використовуючи метрику Жакара; 2) з ембедингом на основі запропонованої метрики; 3) без ембедингу.

Таблиця 2

Результати досліджень

Метод	Правильність класифікації, %
Нейронна мережа з ембедингом, використано метрику Жакара	90,4
Нейронна мережа з ембедингом, використано запропоновану метрику	93,5
Нейронна мережа без ембедингу	84,7

Тестування показало, що запропонований метод дає найкращий результат класифікації текстових документів на основі нейронних мереж.

Висновки

Розроблено метод ембедингу ознак датасетів на основі онтологій. Метод ґрунтується на семантичній метриці, визначеній у межах онтології ПО, до якої належать ознаки. Враховано відношення між термінами онтології та їх близькість для заміни текстових ознак числовими значеннями. Наведено приклад використання такої заміни. Досліджено використання розробленого методу для задачі рубрикування на основі нейронної мережі. Результати досліджень свідчать про перевагу розробленого методу.

Список літератури

1. Литвин В. В. Бази знань інтелектуальних систем підтримки прийняття рішень: монографія. Львів: Видавництво Львівської політехніки, 2011. 240 с.

2. Вдовіченко А. В. Інтелектуалізовані пошукові системи. Класифікація та порівняння. Искусственный интеллект, ІПШІ “Наука і освіта”. 2002. № 3. С. 61–70.

3. Strube M., Ponzetto S. WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence. (AAAI 06). Boston, Mass., July 16–20, 2022. URL: http://www.eml-research.de/english/research/nlp/public

4. Jarmasz M., Szpakowicz S. (2020). Roget’s Thesaurus and semantic similarity. In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003). Borovets, Bulgaria, September, 212–219.

5. Fellbaum C. (1998). WordNet: an electronic lexical database. MIT Press, Cambridge, Massachusetts, 423 p.

6. Литвин В. В., Мороз О. В. (2013). Метод контекстного пошуку на основі тезаурусу предметної області. *Східно-Європейський журнал передових технологій*, № 6/2(66), С. 22–27.

7. Resnik P. (1995). Disambiguating noun groupings with respect to WordNet senses. In Proceedings of the 3rd Workshop on Very Large Corpora. MIT, June. URL: http://xxx.lanl.gov/abs/cmp-lg/9511006.

8. Resnik P. (2019). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* (JAIR), Vol. 11, 95–130.

10. Lin D. (2018). An information-theoretic definition of similarity. In Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July. URL: http://www.cs.ualberta.ca/~lindek/papers.htm

11. WordNet: a lexical database for the English language. Cognitive Science Laboratory Princeton University, 2006. Режим доступу: http://wordnet.princeton.edu/.

12. Gruninger M., Fox M. (1995). Methodology for the Design and Evaluation of Ontologies. Proceedings of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, 231–238.

13. WordNet: a lexical database for the English language. Cognitive Science Laboratory Princeton University, 2006. Режим доступу: http://wordnet.princeton.edu/.

14. Дубинский А. Г. Розробка моделей і вдосконалення структури систем інформаційного пошуку в глобальній комп’ютерній мережі: автореф. дис... канд. техн. наук: 05.13.06 / НАН України; Національна бібліотека України ім. В. І. Вернадського. К.: 2001. 17 с.

15. Bulskov H., Knappe R., Andreasen R. (2004). On Querying Ontologies and Databases. FQAS, 191–202.

16. Кравець П. О., Литвин В. В., Висоцька В. А. Моделювання ігрової задачі призначення персоналу для виконання ІТ-проектів на основі онтологій. *Радіоелектроніка, інформатика, управління*. 2022. № 1. С. 130–145.

17. Bublyk M., Kowalska-Styczeń A., Lytvyn V., Vysotska V. (2021). The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis. *Energies*, 14(18), 5951. URL: https://www.mdpi.com/1996-1073/14/18/5951/htm.

18. Kravets P., Lytvyn V., Vysotska V. (2020). Game Model of Ontological Project Support. *Radio Electronics, Computer Science,* Control, Vol. 1(1), 172–183. URL: http://ric.zntu.edu.ua/article/view/228160/227318

19. Карпов І. А., Буров Є. В. Використання онтологічних мереж у системах підтримки прийняття рішень в умовах неоднозначності. *Вісник Нац. ун-ту “Львівська політехніка”. Серія: Інформаційні системи та мережі*. 2020. Вип. 7. С. 8–15. URL: https://science.lpnu.ua/uk/sisn/vsi-vypusky/vypusk-7-2020/vykorystannya-ontologichnyh-merezh-u-systemah-pidtrymky-pryynyattya

References

1. Lytvyn V. V. (2011). Knowledge bases of intelligent decision support systems: monograph. Lviv: Publishing House of Lviv Polytechnic, 240 p.

2. Vdovichenko A. V. (2002). Intelligent search systems. Classification and comparison. *Artificial intelligence, IPSI “Science and education”,* No. 3, 61–70.

3. Strube M., Ponzetto S. (2022). WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence. (AAAI 06). Boston, Mass., July 16–20, 2022. Access mode: http://www.eml-research.de/english/research/nlp/public

4. Jarmasz M., Szpakowicz S. (2020). Roget’s Thesaurus and semantic similarity. In Proceedings of Conference on Recent Advances in Natural Language Processing (RANLP 2003). Borovets, Bulgaria, September, 212–219.

5. Fellbaum C. (1998). WordNet: an electronic lexical database. MIT Press, Cambridge, Massachusetts, 423 p.

6. Wu Z., Palmer M. (1994). Verb semantics and lexical selection. In Proc. of ACL-94, 133–138.

7. Resnik P. (1995). Disambiguating noun groupings with respect to WordNet senses. In Proceedings of the 3rd Workshop on Very Large Corpora. MIT, June. Access mode: http://xxx.lanl.gov/abs/cmp-lg/9511006

8. Resnik P. (2019). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research* (JAIR), Vol. 11, 95–130.

10. Lin D. (2018). An information-theoretic definition of similarity. In Proceedings of International Conference on Machine Learning, Madison, Wisconsin, July. Access mode: http://www.cs .ualberta.ca/~lindek/papers.htm

11. WordNet: a lexical database for the English language. Cognitive Science Laboratory Princeton University, 2006. Access mode: http://wordnet.princeton.edu/.

12. Gruninger M., Fox M. (1995). Methodology for the Design and Evaluation of Ontologies. Proceedings of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, 231–238.

13. WordNet: a lexical database for the English language. Cognitive Science Laboratory Princeton University, 2006. Access mode: http://wordnet.princeton.edu/.

14. Dubinsky A. G. (2001). Development of models and improvement of the structure of information search systems in the global computer network: abstract. dis... cand. technical sciences: 05.13.06 / NAS of Ukraine; National Library of Ukraine named after V. I. Vernadskyi. K., 17 p.

15. Bulskov H., Knappe R., Andreasen R. (2004). On Querying Ontologies and Databases. FQAS, 191–202.

16. Kravets P. O., Lytvyn V. V., Vysotska V. A. (2022). Simulation of the game task of assigning personnel for the execution of IT projects based on ontologies. *Radio electronics, informatics, management,* No. 1, 130–145.

17. Bublyk M., Kowalska-Styczeń A., Lytvyn V., Vysotska V. (2021). The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis. *Energies*, 14(18), 5951. Access mode: https://www.mdpi.com/1996-1073/14/18/5951/htm

18. Kravets P., Lytvyn V., Vysotska V. (2020). Game Model of Ontological Project Support. *Radio Electronics, Computer Science, Control*, Vol. 1(1), 172–183. Access mode: http://ric.zntu.edu.ua/article/view/228160/227318.

19. Karpov I. A., Burov E. V. (2020). The use of ontological networks in decision support systems under conditions of ambiguity. *Bulletin of the Lviv Polytechnic National University. Series: Information systems and networks*, is. 7, 8–15. Access mode: https://science.lpnu.ua/uk/sisn/vsi-vypusky/vypusk-7-2020/vykorystannya-ontologichnyh-merezh-u-systemah-pidtrymky-pryynyattya.

METHOD OF BUILDING EMBEDDINGS OF SIGNS IN DEEP LEARNING PROBLEMS BASED ON ONTOLOGIES

Vasyl Lytvyn[1], Solomiya Mushasta[2]

Lviv Polytechnic National University,
Information Systems and Networks Department Lviv, Ukraine
[1] E-mail: vasyl.v.lytvyn@lpnu.ua, ORCID: 0000-0002-9676-0180
[2] E-mail: solomiyanytrebych@gmail.com, ORCID: 0000-0003-4932-4113

This paper investigates the problem of embedding features used in datasets for training neural networks. The use of embeddings increases the performance of neural networks, and therefore is an important part of data preparation for deep learning methods. Such a process is based on semantic metrics. It is proposed to use ontologies of the subject areas to which the corresponding feature belongs for embedding. This work developed such a method and investigated its use for the task of categorizing text documents. The research results showed the advantage of the developed method.

Key words: ontology; neural network; embedding; semantic metrics; natural language text.