Ілля Данилюк

УДК 81’23

АКТУАЛЬНІ ПРОБЛЕМИ МЕТОДУ ГЛИБИННОГО НАВЧАННЯ В ЛІНГВОПЕРСОНОЛОГІЇ

У статті описано проблеми глибинного навчання як галузі машинного навчання у застосуванні до моделювання мовної особистості. Описано суть методу глибинного навчання, визначено його відомі та потенційні перешкоди в застосуванні до мовних даних, вказано на перспективні напрями подолання цих недоліків.

Ключові слова: лінгвоперсонологія, мовна особистість, глибинне навчання, штучна нейронна мережа.

У межах проекту «Комунікативно-прагматична і дискурсивно-граматична лінгвоперсонологія: структурування мовної особистості та її комп’ютерне моделювання»[1] однією з опорних є теза про те, що мовносоціумна особистість – лінгвоперсона – може бути описана та змодельована у віртуальному просторі на основі аналізу породжених реальною особою (донором) усних або писемних текстів. Відповідно, сучасні інформаційні технології, здатні опрацьовувати велику кількість структурованих і неструктурованих даних, є потужним інструментом і для моделювання мовленнєвої діяльності й, ширше, цілої мовної особистості. Ці завдання покладено на *лінгвоперсонологію*, що має предметом вивчення власне-людську мовну особистість.

Нагадаємо, що підходи до вивчення мовної особистості включають її: 1) психологічний аналіз; 2) соціологічний аналіз; 3) культурологічний аналіз – моделювання лінгвокультурних типажів – узагальнених відомих представників певних груп суспільства, поведінка яких втілює в собі норми лінгвокультури загалом і впливає на поведінку всіх представників суспільства; 4) лінгвістичний аналіз (опис комунікативної поведінки носіїв елітарної або масової мовної культури, характеристика людей з позицій їхньої комунікативної компетенції, аналіз креативної і стандартної мовного свідомості); 5) прагмалінгвістичний аналіз мовної особистості, в основі якого лежить виділення типів комунікативної тональності, характерної для того чи іншого дискурсу (Danylyuk, 2016a).

Власне лінгвістичний підхід до вивчення лінгвоперсони охоплює а) *моделювання формально-змістового рівня* діяльності мовної особистості – за допомогою використання розробок корпусної лінгвістики; б) *моделювання формально-звукового рівня* діяльності мовної особистості – за допомогою систем синтезу мовлення; в) *моделювання формально-графічного рівня* діяльності мовної особистості – почерку (Danylyuk, 2016b).

Для усіх вказаних типів моделювання може бути застосовано метод *машинного навчання*, як було проаналізовано в (Danylyuk, 2017), зокрема глибинного навчання з використанням рекурентних штучних нейронних мереж.

Метою статті є з’ясувати, які потенційні проблеми можуть виникнути в разі використання методу глибинного машинного навчання в моделюванні лінгвоперсони. Завдання, які має бути розв’язано, включають 1) з’ясування суті методу глибинного навчання; 2) опис відомих та потенційних перешкод у застосуванні методу до мовних даних; 3) аналіз наявних і перспективних підходів їх подолання.

Актуальність дослідження зумовлена постійним інтересом до моделювання різних аспектів інтелектуальної діяльності людини, зокрема мовленнєвої, і розширенням інструментарію машинного навчання новими алгоритмами, варіантами архітектури штучних нейронних мереж (ШНМ), а також зростанням потужності апаратного забезпечення для їх тренування (навчання).

І. Особливості методу глибинного машинного навчання

Глибинне навчання (ГН) є галуззю машинного навчання, яка використовує для реалізації своїх завдань *багатошарові* штучні нейронні мережі (ШНМ). Починаючи з 2012 року, зокрема після публікації результатів тренування ШНМ для розпізнавання зображень проекту ImageNet (Krizhevsky та ін.) з використанням *згорткової* багатошарової мережі, де кожен наступний шар є результатом згортання (тобто пропускання через фільтр) даних попереднього. Передумовою стала також можливість навчати мережі з багатьох шарів, що потребує великої кількості обчислень на центральному процесорі або процесорі комп'ютерної відеокарти.

Ефективність методу ГН підтверджена зростанням якості виконання типових завдань машинного навчання – розпізнавання образів і класифікація, кластеризація, прогнозування, – якщо порівнювати з одношаровими ШНМ чи класичними алгоритмами, що ґрунтуються на правилах. До сьогодні зберігається надзвичайно високий інтерес з боку науковців до ГН. Розроблено бібліотеки, як-от TensorFlow чи Keras, що спрощують процес програмування глибинних мереж і дають змогу зосередитися більше на процесі підготування вхідних даних (введення) й аналізу результатів (виведення). Загалом найчастіше мета системи ГН – встановити зіставні зв'язки між вхідними та вихідними даними, і чим більше інформації буде використано для тренування, тим потенційно точнішими будуть результати роботи з новими даними.

[1] Дослідження виконано в межах фундаментального наукового дослідження "Комунікативно-прагматична і дискурсивно-граматична лінгвоперсонологія: структурування мовної особистості та її комп’ютерне моделювання" (0115U000088) і в межах фундаментального наукового дослідження «Об’єктивна і суб’єктивна мовносоціумна граматика : комунікативно-когнітивний та прагматико-лінгвокомп’ютерний виміри» (0118U003137).

Як і традиційне машинне навчання, ГН має характерні слабкі місця, а саме можливість перенавчання та тривалість обчислень. Перенавчання – це побудова такої моделі, яка вірогідно описує тільки тренувальні дані, але не може ефективно використовуватися для нових невідомих даних, тобто не має низьку прогностичну продуктивність. Замість узагальнення, система моделює випадкову похибку, а тому не має практичного застосунку. Тривалість тренування багатошарової ШНМ, хоча і стала меншою з появою багатопотокових процесорів, також не сприяє ефективному використанню ГН у сферах, де даних справді багато.

ІІ. Відомі обмеження глибинного навчання

Ефективність методу ГН прямо залежить від кількості даних. Можливість оперувати справді великою кількістю інформації, власне, спричинила появу багатошарових мереж. Однак даних не може бути безкінечно багато, тому призначення систем ГН – досягти такого рівня узагальнення на наявній тренувальній вибірці, аби ефективно обробляти нові дані з потенційно необмеженого потоку. Узагальнення, отже, є, з одного боку, інтерполяцією між відомими даними з тренувального набору, а з іншого – екстраполяцією, що вимагає виходу за його межі. Тому тренувальні й тестові дані мають бути насправді з однієї вибірки. Розпізнавання звукового мовлення, відповідно, є прикладом завдання для ГН, яке потенційно має ефективний розв’язок, адже зіставлення зацифрованого звукозапису з обмеженим набором звуків мови в типових умовах має незмінну природу. Інша річ – моделювання смислового рівня мовлення чи машинний переклад, де кількість смислів зараз не видається скінченною.

Можна виділити низку відомих проблем ГН, які потребують розв'язання або зміни підходу для ефективного використання методу для моделювання складних об'єктів, яким є, зокрема, і лінгвоперсона.

Кількість тренувальних даних. Системи ГН потребують великої кількості даних. У навчанні без учителя ці дані мають бути структуровані та нормалізовані, а в навчанні з учителем – ще й правильно розмічені. Сьогодні збирання та підготування даних є окремим напрямом дослідження, цілою бізнесовою галуззю, у якій дані продають та купують. Для моделювання мовних та мовленнєвих явищ наявність розмічених даних є неоднаковою для різних мов. Наприклад, для англійської мови чимало готових баз даних є у відкритому доступі (https://deeplearning4j.org/opendata), для української мови, на жаль, навіть прості корпуси текстів є здебільшого закритими.

Принцип навчання системи ГН у цьому плані суттєво відрізняється від того, як засвоює нову, зокрема мовну, інформацію людина. У працях (Lake та ін. 2015, Lake та ін. 2016) продемонстровано, що абстрактне правило, схоже на алгебраїчну формулу, що його люди можуть засвоїти на одному прикладі або як сформульоване визначення, для ГН потребує тисяч однотипних прикладів. Тому в разі, якщо даних небагато, традиційні методи побудови чітко детермінованих алгоритмів можуть бути ефективнішими.

Нові дані. Низький рівень узагальнення для нових даних випливає з попередньої проблеми. Якщо тренувальна вибірка містить небагато прикладів якогось правила, система ГН може не встановити потрібне узагальнення. Експерименти, описані в (Jia and Liang), стосувалися побудови низки ШНМ різної архітектури, натреновані на пошук відповідей на запитання із задачі SQuAD (Stanford Question Answering Database) – у ній метою є виділити кольором слова в певній фразі, які стосуються поставленого користувачем запитання. Наприклад, одна з мереж змогла правильно визначити, проаналізувавши короткий текст, ім’я гравця спортивної команди, якого стосувалося запитання. Однак автори показали, що після додавання до тексту кількох віддалених від теми речень з іменами інших гравців, загальна якість роботи системи впала з умовних 75 % до 36 %. Очевидно, у тренувальній вибірці не було схожих прикладів.

Нечіткі та синкретичні дані. Значною проблемою для аналізу мовлення та моделювання мовної особистості є економія мовних засобів, синкретизм, омонімія та багатозначність. Для умовного розуміння речення *Замóк зáмок на замóк, щоби зáмок не замóк* потрібні знання як лінгвістичні (щодо частин мови чи структури речення), так і екстралінгвістичні. Є корпуси, що мають розмітку таких даних (Bowman та ін.), однак треба усвідомлювати, що побудувати навіть не вичерпний, а мінімально достатній корпус для однієї мови сьогодні є навряд чи реальним завданням.

Ієрархічна структура даних. Системи ГН працюють з текстовими даними, реченнями як з простими ланцюжками слів, не враховуючи ієрархічну будову мови, за якої більші одиниці будуються з менших. Завдяки цій властивості потенційно довжина речення і кількість можливих речень є необмеженими, і водночас побудованими зі скінченного, як видається, набору структур. Текст для системи ГН має вигляд плаского поля або неструктурованого однорангового списку, відповідно, встановлені кореляції між словами чи реченнями будуть неієрархічними. Наприклад, алгоритм word2vec (Mikolov) спирається на ймовірнісну модель мови – кожне слово представлено вектором з дійсних чисел у маленькому (якщо порівняти з розміром повного словника) просторі, наприклад розмірністю в 300 вимірювань. Спочатку векторам присвоюють випадкові значення. Далі в процесі навчання на укладеному корпусі для слова обчислюють вектор, максимально схожий на вектори інших слів, які трапляються у схожих контекстах. За контекст беруть невелике вікно, тобто ланцюжок попередніх і наступних слів, наприклад, у п'ять одиниць. У результаті виявляється, що векторно близькі слова виявляються дійсно семантично близькими. Крім того, виявляється, що багато важливих для обробки природного мовлення відношень закодовано у вектори. Відомий приклад: якщо від вектора слова «Париж» відняти вектор слова «Франція» і додати вектор «Італія», то вийде вектор, дуже близький до вектора «Рим» – відношення «столиця» виявилося закодованим у вектори слів. Однак ієрархічна структура, на кшталт синтаксичних дерев, у цій системі не має ані внутрішнього (навченого), ані зовнішнього (заданого людиною)

представлення. Як вказують вчені у (Lake, Baroni), сьогодні рекурентні мережі добре працюють, узагальнюючи дані, якщо тренувальні та тестові дані – з однієї вибірки, але застосувати композиційні навички мережі не можуть навчитися.

Лінгвістичні знання. Сучасні системи ГН мають самодостатній характер та ізольовані від уже відомих та систематизованих знань. З одного боку, це дає змогу абстрагуватися від надмірного ускладнення системи численними правилами. Зокрема, сучасна якість розпізнавання мовлення з ГН на великій кількості розмічених прикладів вища за всі побудовані раніше системи, що містили моделі фонеми, звука, наголосу, акомодації тощо. З іншого боку, обмеженість тільки тренувальною вибіркою й даними, що можуть бути отримані з неї, унеможливлює для системи ГН встановлення кореляцій для мовних одиниць, яких у ній немає. Натомість навіть простий граматичний словник або схема побудови складного речення з простих містять таку інформацію в готовому вигляді.

ІІІ. Перспективні напрями

Указані проблеми глибинного навчання, як нам видається, містять у собі відповідь на запитання, як їх подолати чи обійти. Зокрема, відмінним до контрольованого глибинного навчання, яке потребує величезної кількості розмічених тренувальних даних, може бути *спонтанне навчання*, або навчання без вчителя. Нагадаємо, що виділяють 3 типи машинного навчання: а) *з учителем:* використовуючи набір об'єктів (прикладів) і правильних реакцій (відповідей) до них навчитися на давати правильну реакцію (відповідь) на заданий об'єкт (приклад). Як-от: на основі розміченого вручну корпусу текстів навчитися визначати частину мови й основні граматичні категорії в інших (не включених до корпусу) текстах; б) *без вчителя:* використовуючи набір об'єктів (прикладів), знайти в них приховані (невідомі наперед) закономірності. Як-от: поділити слова в корпусі текстів на певні класи (групи); в) *з підкріпленням:* використовуючи в певному середовищі контрольованого комп'ютером агента, вчиняти такі дії, щоби досягти максимально можливої кількості позитивних реакцій (відповідей) від середовища. Як-от: у діалоговій системі домогтися подання якнайточнішої відповіді на поставлене природною мовою запитання.

Одним з відомих підходів спонтанного навчання є просте накопичення вхідних даних, які мають схожі властивості, але явно не розмічені, як-от описана в (Le та ін.) система розпізнавання котів від Google. Накопичення персонотекстів, записів спонтанного мовлення лінгвоперсони, зразків почерку, результатів підготовлених експериментів з лінгвоперсоною, на нашу думку, є саме цим перспективним підходом.

Інший напрям пов'язаний із заміною наборів навчальних даних фільмами, які змінюються в часі (Luc та ін.). Суть у тому, що треновані на відеороликах системи можуть використовувати будь-яку пару послідовних кадрів як тренувальні дані в навчанні, метою якого є передбачити наступний кадр. Так, кадр t стає прогнозом для кадру t_1, без жодної потреби необхідності розмічувати ці дані.

Нарешті, третій підхід до спонтанного навчання запропоновано у (Marcus) – розроблення системи, призначеної самостійно ставити завдання, розв'язувати відповідно проблеми високого рівня та інтегрувати абстрактні знання.

Окрім спонтанного навчання, ми вважаємо перспективним також гібридний підхід, що поєднуватиме ГН з класичними директивними системами, у яких абстрактні дані замінено символами, а з ними можливі відповідні логічні, математичні чи подібні операції. Переконливим аргументом на їх користь нам видається те, що вони максимально близькі до того, як функціонує така символьна система, як мова, а також незалежність від того, чи траплялися усі можливі комбінації даних у тренувальній вибірці.

Загалом, у лінгвоперсонології ГН, на нашу думку, має стати ефективним інструментом моделювання мовної особистості, за умови створення репрезентативного корпусу текстів, звукового корпусу та бази даних зразків почерку. Елемент корпусу текстів Юрія Шевельова (Шереха), чию мовну особистість у перспективі має бути змодельовано в межах згаданого проекту, викладено на corpora.donnu.edu.ua.

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. "A large annotated corpus for learning natural language inference". arXiv:1508.05326. (2015). Web. 21 Aug 2015.

Danylyuk, I. "Avtomatyzovani metody opysu ta rozpiznavannya movnoyi osobystosti". (*Automated linguistic personality description and recognition methods*) Linhvistychni studiyi (*Linguistic Studies*) 32 (2016b): 93–99. Print.

Danylyuk, I. "Korpus tekstiv dlya vyvchennya hramatychnoyi sluzhbovosti." (*Text corpora to study of a grammatical auxiliarity*) Linhvistychni studiyi (*Linguistic Studies*) 26 (2013): 224–230. Print.

Danylyuk, I. "Korpus tekstiv dlya vyvchennya hramatychnoyi sluzhbovosti: klasyfikatsiya hramatychnykh klasiv i pidklasiv." (*Text corpora for studying a grammatical auxiliarity: classification of grammatical classes and subclasses*) Linhvistychni studiyi (*Linguistic Studies*) 27 (2013): 221–229. Print.

Danylyuk, I. "Perspektyvy zastosuvannya metodu mashynnoho navchannya dlya modelyuvannya movnoyi osobystosti" (*Prospects of machine learning method for lingual personality modeling*) Linhvistychni studiyi (*Linguistic Studies*) 33 (2017): 159–164. Print.

Danylyuk, I. "Teoretychni zasady i metody linhvopersonolohiyi" (*Theoretical Principles And Methods of Lingvopersonology*) Linhvistychni studiyi (*Linguistic Studies*) 31 (2016a): 63–66. Print.

Jia, R., & Liang, P.. Adversarial Examples for Evaluating Reading Comprehension Systems. arXiv:1707.07328. (2017). Web. 23 Jul. 2017.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. "Imagenet classification with deep convolutional neural networks" Advances in Neural Information Processing Systems 25. pp. 1097–1105. (URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf)

Lake, B. M., & Baroni, M. "Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks". arXiv:1711.00350 (2017). Web. 31 Oct 2017.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. "Human-level concept learning through probabilistic program induction". Science, (2015). 350(6266), pp. 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. "Building Machines That Learn and Think Like People". Behav Brain Sci (2016) pp. 1–101.

Le, Q. V., Ranzato, M.-A., Monga, R., Devin, M., Chen, K., Corrado, G. et al. "Building high-level features using large scale unsupervised learning". Proceedings from International Conference on Machine Learning. arXiv:1112.6209 (2012). Web. 29 Dec 2011.

Luc, P., Neverova, N., Couprie, C., Verbeek, J., & LeCun, Y. "Predicting Deeper into the Future of Semantic Segmentation". International Conference on Computer Vision (ICCV 2017). arXiv:1703.07684 (2017). Web. 22 Mar 2017.

Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013). Web. 10 Sep. 2016.

Marcus, G. "Deep Learning: A Critical Appraisal". arXiv:1801.00631 (2018). Web. 2 Jan 2018.

Надійшла до редакції 15 березня 2018 року.

ACTUAL PROBLEMS OF THE DEEP LEARNING METHOD IN LINGUAL PERSONALITY MODELING

Illya Danyliuk

Department of General and Applied Linguistics and Slavonic Philology, Vasyl Stus Donetsk National University, Vinnytsia, Ukraine

Abstract

Background: The relevance of our research, above all, is theoretically motivated by the development of extraordinary scientific and practical interest in the possibilities of language processing of huge amount of data generated by people in everyday professional and personal life in the electronic forms of communication. Linguopersonology is a new research area for modeling particular linguistic personality.

Purpose: The purpose of the article is to clarify the essence of the method of deep learning, to describe the known and potential obstacles to the applying of that method to the linguistic data; and to analyze existing and perspective approaches to their overcoming.

Results: Amount of training data, dealing with new data, fuzzy and syncretic data, ignoring hierarchical language and speech structure and linguistic knowledge are main DL problems.

Discussion: One way to deal with DL problems is spontaneous unsupervised learning, representing data in the form of movies, and building systems, which are capable to choose and achieve goals. Another way is building hybrid DL systems, which use previous and some linguistic data.

Keywords: linguopersonology, linguistic personality, deep learning, artificial neural network.

Vitae

Illya G. Danyliuk, Candidate of Philology, Doctoral candidate and Associate Professor at Department of General and Applied Linguistics and Slavonic Philology in Donetsk National University. His research areas include applied linguistics, natural language processing, corpus linguistics, and machine grammar.

Correspondence: i.danyluk@donnu.edu.ua

Олена Карпіна

УДК 811.111’374+81’23+616.89-008.454

LINGUO-PSYCHOLOGICAL MODEL OF DEPRESSIVE PERSONALITY

Зроблено спробу лінгвістичної інтерпретації психологічного змісту депресії на лексикографічному рівні з метою з’ясування подібностей і розбіжностей між психологічним трактуванням цього стану та його мовним наповненням у тлумачних словниках сучасної англійської мови. Розмежовано поняття депресивна акцентуація й депресивний стан, а також виявлено відповідні семантичні компоненти в структурі словникових дефініцій.

Ключові слова: афект, депресивний стан, депресивна акцентуація, депресивна особистість, емоційний стан, сема, семантичний компонент, словникова дефініція.

