# Підсекція загального та прикладного мовознавства

УДК 81’23

## МЕТОД ГЛИБИННОГО НАВЧАННЯ В МОДЕЛЮВАННІ ЛІНГВОПЕРСОНИ

***І. Г. Данилюк***

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.

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 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.

**Відомі обмеження глибинного навчання. 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. **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. **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. **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.

**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 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. 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 Г. Маркусом – 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.