РОЗДІЛ ІХ. ПРИКЛАДНА ЛІНГВІСТИКА: НАПРЯМИ Й АСПЕКТИ ДОСЛІДЖЕННЯ

Ілля Данилюк

УДК 81’23

ПЕРСПЕКТИВИ ЗАСТОСУВАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ МОДЕЛЮВАННЯ МОВНОЇ ОСОБИСТОСТІ

У статті описано підхід до моделювання мовної особистості з використанням машинного навчання, покладений в основу проекту «Комунікативно-прагматична і дискурсивно-граматична лінгвоперсонологія: структурування мовної особистості та її комп’ютерне моделювання», який реалізує кафедра загального та прикладного мовознавства та слов’янської філології ДонНУ імені Василя Стуса.

Ключові слова: лінгвоперсонологія, мовна особистість, машинне 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 точок є теза про те, що мовносоціумна особистість 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.

Метою статті є опис та узагальнення методології фундаментального дослідження з лінгвоперсонології. Завдання, які має бути розв’язано, включають 1) з'ясування сучасного стану розвитку методу машинного навчання; 2) опис моделей, що є результатом машинного навчання; 3) аналіз наявних і перспективних підходів використання штучних нейронних мереж для машинного навчання в межах лінгвоперсонології.

Актуальність нашого дослідження, крім усього, цілком мотивована розвитком надзвичайного теоретично-наукового і практичного інтересу до можливостей опрацювання величезного обсягу мовних даних, які генерує людина у повсякденному професійному та особистому житті в електронних формах комунікації (e-mail, sms, голосовий зв’язок, аудіо- та відеоблоги, соціальні мережі тощо).

Раніше було визначено (Danylyuk, 2016a), 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.

Власне лінгвістичний підхід до вивчення мовної особистості включає, на нашу думку, 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*. Суть цього методу можна окреслити загальним визначенням: "Машинне навчання — процес, у результаті якого машина (комп'ютер) здатна демонструвати поведінку, яку в неї не було явно закладено" (Samuel). Однак значно формальніше визначення, запропоноване у (Mitchell), на нашу думку, описує його точніше: "Кажуть, що комп'ютерна програма навчається на основі досвіду *E* щодо певного класу задач *T* і міри якості *P*, якщо якість розв'язання задач з *T*, виміряна на основі *P*, зростає з набуттям досвіду *E*".

Загалом машинне навчання як окрема галузь інформатики і як метод розвивається із середини минулого століття й сьогодні позиціонується як ефективний механізм опрацювання, зокрема, й лінгвістичних даних.

І. Типи та засади машинного навчання

Машинне навчання (МН), як і будь-яку ІТ-технологію, треба розглядати як інструмент, використання якого передбачає низку передумов: чітке розуміння самого завдання, наявність репрезентативних даних у цифровій формі, можливість виділити у цих даних формальні вимірні ознаки, вибір релевантної моделі для представлення поведінки даних та алгоритму її побудови, а також можливість інтерпретувати отриманий результат. Схематично МН можна представити так:

[1] Дослідження виконано в межах фундаментального наукового дослідження "Комунікативно-прагматична і дискурсивно-граматична лінгвоперсонологія: структурування мовної особистості та її комп’ютерне моделювання " (0115U000088).

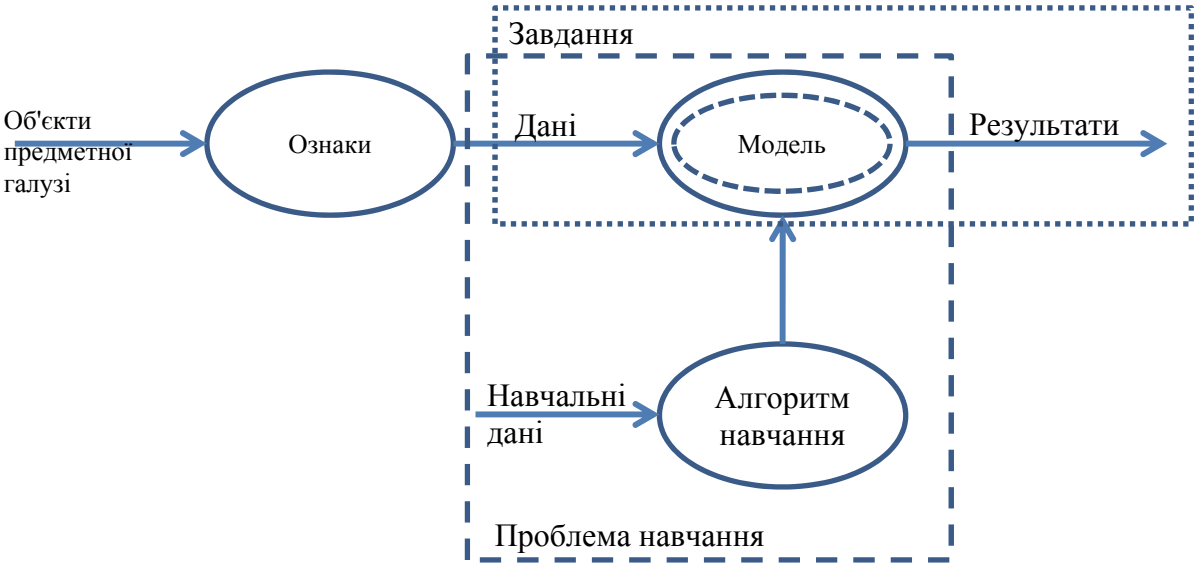


Схема 1. Компоненти машинного навчання

Типове завдання для МН можна сформулювати так: є певний набір об’єктів (прикладів) і певний набір реакцій (відповідей) на них. Між об'єктами і реакціями є певна прихована залежність, яку й треба знайти.

Виділяють 3 типи МН: 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.

Якість МН сильно залежить від того, які *ознаки* буде обрано для опису об’єктів (прикладів). Виділяють декілька способів опису цих ознак. Наприклад, є *бінарні ознаки*: певні характеристики, які можуть бути описаними відповіддю на питання "так" або "ні", є певна характеристика або немає, 1 або 0. Цих ознак може бути декілька, тобто, наприклад, у об'єкта може бути 10 або 100 000 ознак залежно від прикладного завдання. *Номінальні ознаки*: це, координати точки на певній шкалі, якою описано об’єкти (приклади), як-от точка (-4, 10) має дві номінальні ознаки – *x* та *y*. *Порядкові ознаки:* це ознака позиції, наприклад, в списку завдання ранжування. *Кількісні ознаки:* опис певних кількісних параметрів, як-от частота вживання словоформи чи лексеми у корпусі текстів. Системи МН на вхід завжди отримують певний *вектор ознак*, а на виході надають *вектор реакцій (відповідей)*.

Типовими завданнями МН, відповідно, є:

1. Розпізнавання образів і класифікація

Образами можуть виступати різні за своєю природою об'єкти: символи тексту, зображення, зразки звуків тощо. Під час навчання в систему вводять різні зразки образів із зазначенням того, до якого класу вони належать. Зразок зазвичай подається як вектор значень ознак. При цьому сукупність усіх ознак має однозначно визначати клас, до якого належить зразок. У разі, якщо ознак недостатньо, система може співвіднести один і той самий зразок з декількома класами, що є неправильним. Після закінчення навчання системі можна подавати невідомі раніше образи і отримувати відповідь про їхню належність до певного класу.

2. Кластеризація

Кластеризацією називається поділ множини об’єктів (прикладів) на класи, за умови, що ні кількість, ні ознаки класів заздалегідь не відомі. Після навчання така система здатна визначати, до якого з цих класу належить новий невідомий об’єкт (приклад). Система також може сигналізувати про те, що невідомий об'єкт не належить до жодного з виділених класів – це є ознакою нових, відсутніх у навчальній вибірці, даних. Отже, подібна система може виявляти нові, невідомі раніше класи об’єктів. Відповідність між класами, виділеними системою, і класами, що існують у предметній галузі, встановлює людина.

3. Прогнозування

Здатність системи до прогнозування безпосередньо випливає з її здатності до узагальнення і виділення прихованих залежностей між об’єктами (прикладами) й реакціями (відповідями). Після навчання система здатна передбачити майбутнє значення певної змінної на основі декількох попередніх значень і/або якихось

відомих чинників. Треба зазначити, що прогнозування можливе тільки тоді, коли попередні зміни дійсно певною мірою визначають майбутні.

4. Зменшення розмірності й асоціативна пам'ять

Здатність систем з МН до виявлення взаємозв'язків між різними параметрами дає можливість подати об'ємні дані компактно, якщо вони є взаємозалежними. Наприклад, розмічений корпус текстів, у якому кожній словоформі приписано граматичну інформацію, система можна стиснути, об'єднавши в класи однакові сукупності ознак – усі іменники чоловічого роду одними у давальному відмінку. Зворотний процес – відновлення вихідного набору даних із частини інформації – називається (авто)асоціативною пам'яттю. Наприклад, мережа може побудувати вірогідну парадигму закінчення для невідомої словоформи, якщо їй повідомити тільки, що це іменник чоловічого роду.

ІІ. Моделі як результат машинного навчання

Моделі є центральною концепцією МН, оскільки це саме те, що породжується в результаті навчання на даних для розв'язання поставленого завдання.

Простором об'єктів називається множина всіх можливих або описуваних об'єктів (прикладів) незалежно від того, чи були вони в наявному наборі даних. Зазвичай ця множина має якусь геометричну структуру.

Наприклад, якщо всі ознаки числові, то кожну ознаку можна розглядати як точку в декартовій системі координат. *Геометрична модель* будується в просторі прикладів із застосуванням таких геометричних понять, як прямі, площини і відстані. Головна перевага геометричних моделей полягає в простоті їх візуалізації, принаймні, у двомірному і тривимірному просторі. Однак важливо розуміти, що розмірність декартового простору об'єктів дорівнює кількості ознак, а їх можуть бути десятки, сотні, тисячі і навіть більше. Простори настільки високої розмірності наочно уявити складно, але в МН вони трапляються часто.

Є також моделі *імовірнісного характеру*, які ґрунтуються на ідеї, що існує випадковий процес, який встановлює між множиною відомих об'єктів (прикладів), наприклад, значеннями ознак цих об'єктів, і множиною реакцій (відповідей), наприклад, класами об'єктів, певний взаємозв'язок. Встановлення невідомого розподілу ймовірностей і є завданням МН.

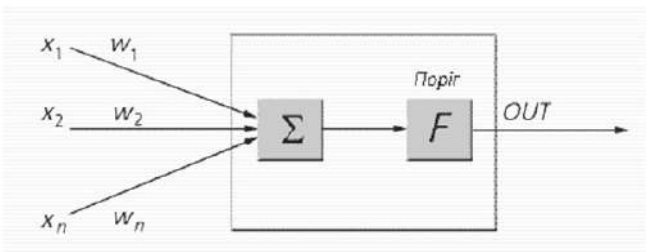
Нарешті, моделі третього типу можна назвати *логічними*, тому що їх легко представити мовою правил, зрозумілих людині, наприклад: [якщо] зліва від слова є прийменник *на* і 3 останні літери слова = *ому* [тоді] клас = місцевий відмінок однини. Такі правила легко організувати у вигляді дерева.

ІІІ. Машинне навчання з використанням штучних нейронних мереж

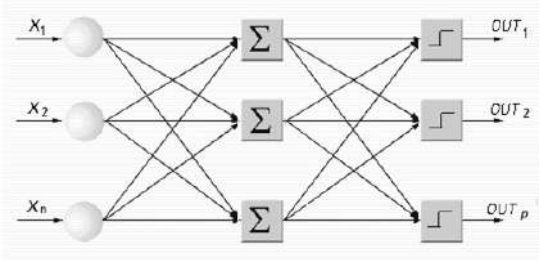
Штучна нейронна мережа (ШНМ) – це математична модель, а також її програмна або апаратна реалізація, побудовані за принципом організації й функціонування біологічних нейронних мереж – мереж нервових клітин живого організму. Це поняття виникло під час вивчення процесів, що відбуваються в мозку, і при спробі змоделювати ці процеси. Першою такою спробою були нейронні мережі Маккалока і Піттса (1943 р.)

ШНМ є системою з'єднаних між собою простих процесорів (штучних нейронів). Такі процесори зазвичай досить прості, особливо якщо порівняти із процесорами, використовуваними в персональних комп'ютерах. Кожний процесор подібної мережі має справу тільки із сигналами, які він періодично одержує, і сигналами, які він періодично надсилає іншим процесорам. З'єднані в досить велику мережу з керованою взаємодією, такі прості процесори разом здатні виконувати досить складні завдання.

У найпростішій ШНМ кожен вузол мережі — це елемент Σ, який множить вхідний сигнал *x* на вагу *w* і підсумовує ці входи (Малюнок 1а). Якщо ця сума більша заданого порогового значення, вихід дорівнює одиниці, в іншому випадку нулю. Такі системи отримали назву *перцептронів*. Вони складаються з одного прошарку штучних нейронів, з'єднаних за допомогою вагових коефіцієнтів з множиною входів (Малюнок 1б).



Малюнок 1а. Окремий вузол мережі

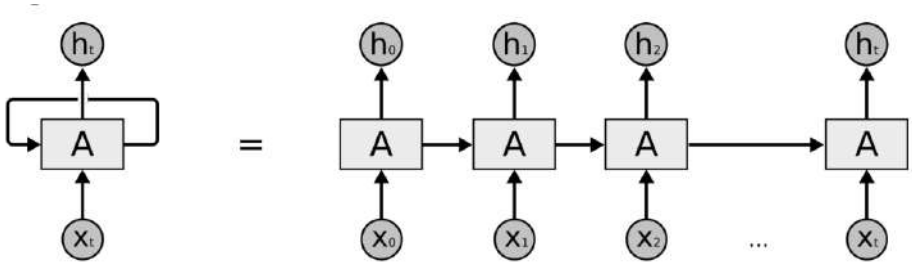


Малюнок 1б. Перцептрон

Окремі перцептрони об'єднуються у блоки, блоки у кластери, а кластери – у шари мережі.

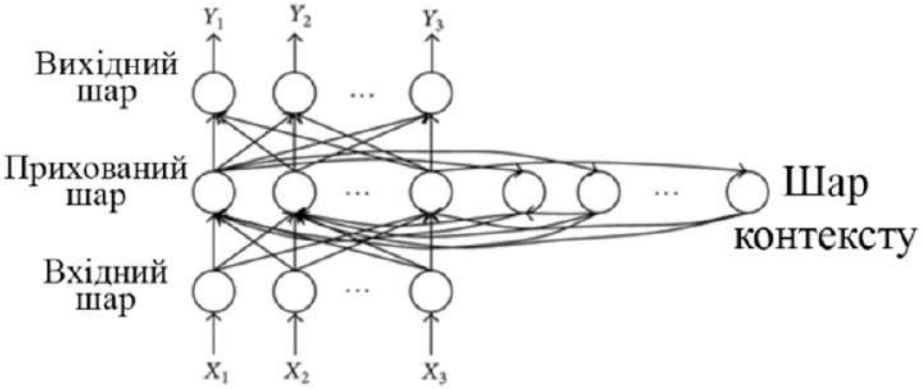
Мережі такого типу під час попереднього навчання, а далі – у процесі роботи опрацьовують об'єкти (приклади) незалежно один від одного, тобто цілковито ігноруючи контекст. Для опрацювання мовних даних, відповідно, потрібні мережі складнішої структури, які використовують *пам'ять* — результати опрацювання попередніх об'єктів.

Першою ідеєю було зациклити одношаровий перцептрон, подаючи на його вхід результати виходу. Відповідно, кожен приклад спочатку проходитиме через мережу і сформує якусь реакцію (відповідь), яка тоді подаватиметься на вхід разом із другим прикладом (Малюнок 3).



Малюнок 2. Зациклений перцептрон

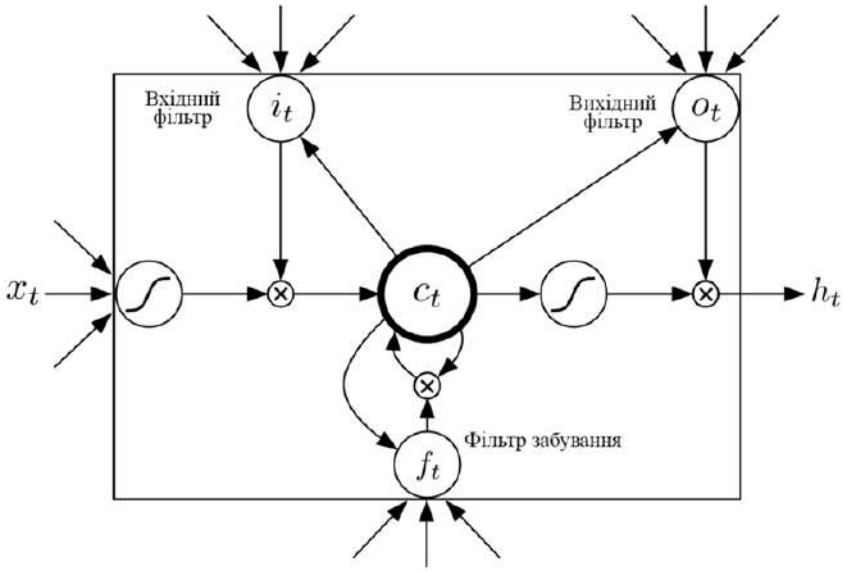
Однак безпосереднє змішування вхідних і вихідних даних призводило до того, що практична цінність таких мереж була невисокою. Подекуди зв'язки в системі були фактично шумом. У 1980-ті роки було запропоновано архітектуру рекурентної (такої, що враховує попередні ітерації) нейронної мережі Елмана. Її характерною особливістю є те, що виходи прихованого шару з попередньої ітерації ми додаються, змішуються або множаться з входами не напряму, а через додатковий вхідний шар, т.зв. шар контексту. Відповідно, починає діяти цикл: прихований шар – шар контексту – прихований шар. У процесі опрацювання послідовності об'єктів (прикладів) інформація прихованого шару постійно буде циркулювати у цьому колі і з кожною ітерацією дещо модифікуватиметься. Нові об'єкти (приклади) змінюватимуть контекст, і контекст, який циркулює всередині мережі та зберігає інформацію про те, які об'єкти (приклади) були перед цим, впливатиме на поточну ітерацію.



Малюнок 3. Штучна нейронна мережа Елмана

Класична ШНМ Елмана є ефективною з одношаровим чи багатошаровим перцептроном, але має недоліки. Основний з них – згасання впливу прикладів з віддаленням контексту. Зазвичай максимальний вплив на результат роботи ШНМ на певній ітерації мають приклади, які були на попередній ітерації, на двох ітераціях тому і т.д., і чим далі, тим цей вплив менший. Відповідно, ШНМ має "досить коротку пам'ять". А з опрацюванням мовлення трапляються ситуації, коли інформація, важлива для правильного аналізу, міститься не у найближчих прикладах, а на 10-20-30 ітерацій назад. Як-от у довгих реченнях, де займенник у кінці абзацу, а кореферентний іменник – на самому початку.

Подолати такий недолік призначена архітектура рекурентних ШНМ із Long Short-Term Memory – LSTM (*довгою короткостроковою пам'яттю*). Ідея полягає в тому, що, оскільки різні об'єкти (приклади) в послідовності мають різний вплив на результат поточної ітерації, запропоновано одним елементам у контексті в попередніх ітераціях надавати більшої ваги, більшого впливу на результат, а іншим – меншої ваги. У мережі з LSTM класична схема рекурентної ШНМ доповнена елементом forget gate (*фільтр забування*), який визначає, з якою імовірністю треба поточний об'єкт (приклад) забути чи запам'ятати для наступних ітерацій. Якщо аналізоване слово в реченні потенційно відіграє важливу роль для майбутньої класифікації слів далі, то йому надається на поточній ітерації висока вага. Якщо ж аналізоване слово відіграє порівняно слабку роль для прогнозування на подальших ітераціях, відсоток його впливу на наступні ітерації зменшується. Наприклад, у реченнях службові слова чи займенники відіграють менш важливу роль для загального розуміння того, про що йдеться, ніж дієслова чи іменники, які досить сильно звужують контекст. Відповідно, перші матимуть менший вплив, ніж другі.



Малюнок 4. LSTM мережа

Традиційно слова обробляли як елементи множини зі словника, обсяг і повнота якого цілковито визначали ефективність такої системи. Однак побудова вичерпного словника – з усіма словоформами чи з включенням професійної і розмовної лексики, жаргону, діалектизмів тощо – надзвичайно важке завдання. На відміну від традиційного підходу, 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 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.

Алгоритм word2vec укладається в парадигму глибокого навчання: він сам знаходить ознаки в режимі «навчання без учителя».

Додаткові відношення (ознаки), встановлені у процесі роботи word2vec, можуть виявитися корисними для завдань обробки текстів, але не зрозуміло, які відношення дійсно містяться в векторах після навчання і наскільки надійно вони закодовані. Є методи, які дозволяють доповнювати векторні представлення слів онтологіями, котрі гарантують, що відношення будуть надійно закодовані у вектори. Наприклад, у (Xu) дослідники запропонували метод навчання векторів, в якому будь-які відношення і таксономії надійно кодуються у векторні представлення.

Втім, стандартний алгоритм word2vec не дозволяє розв’язати проблеми, пов'язані з омонімією. Модифікація алгоритму з елементами автоматичного визначення омонімів і створення окремих векторів для окремих смислів, а також процедури визначення правильного значення омонімів щодо заданого контексту запропоновано у (Bartunov).

Методи обробки природного мовлення здебільшого використовують тільки представлення, ігноруючи синтаксис і семантику, які можна вивести з синтаксичної структури речень. Така модель подання текстів називається «торбою слів» (bag of words) – простий набір слів без урахування їхнього порядку. Наприклад, використовуючи векторне представлення слів можна об'єднати в кластери вектори слів корпусу, на яких тренується модель, і використовувати такі кластери для завдань простої класифікації, як-от – до якого стилю належить текст, визначення авторства тексту тощо. Але якщо завдання полягає в добуванні якісніших семантичних представлень або по суті знань, то потрібні інструменти обробки текстів, які працюють з синтаксичною структурою речення і не ігнорують порядок слів у ньому.

Загалом, у лінгвоперсонології МН, на нашу думку, має стати ефективним інструментом моделювання мовної особистості, за умови створення репрезентативного корпусу текстів. 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

Bartunov, Sergey, et al. "Breaking Sticks and Ambiguities with Adaptive Skip-gram." arXiv preprint arXiv:1502.07257 (2015). Web. 10 Sep. 2016.

Xu, Chang, et al. "Rc-net: A general framework for incorporating knowledge into word representations." Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014. Web. 10 Sep. 2016.

Tai, Kai Sheng, Richard Socher, and Christopher D. Manning. "Improved semantic representations from tree-structured long short-term memory networks." arXiv preprint arXiv:1503.00075 (2015). Web. 10 Sep. 2016.

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

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

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.

Mitchell, T.M. Machine Learning. McGraw-Hill, 1997.

Samuel, A.L. "Some Studies in Machine Learning Using the Game of Checkers". IBM Journal. July 1959. P.210–229.

Надійшла до редакції 20 березня 2017 року.

PROSPECTS OF MACHINE LEARNING METHOD FOR LINGUAL PERSONALITY MODELING

Illya Danyliuk

Department of General and Applied Linguistics and Slavonic Philology, 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 propose Machine Learning method for the project "Communicative-pragmatic and discourse-grammatical linguopersonology: structuring linguistic identity and computer modeling". The description of main ML goals, types and approaches is given.

Results: ML for linguistic personality modeling in linguopersonology seems to be a powerful method. Elmar's and other recurrent artificial networks are used for creating context-dependent language and speech models based on text corpus. Such a corpus for text by Yuriy Shevelyov (Sherekh) is created on corpora.donnu.edu.ua.

Discussion: The project "Communicative-pragmatic, discourse, and grammatical lingvopersonology: structuring linguistic identity and computer modeling", which is implementing by the Department of General and Applied Linguistics and Slavonic philology, selected a task to model Yuriy Shevelyov (Sherekh)' language identity. An architecture for ML network and raw text data are being studied and selected for the main goal.

Keywords: linguopersonology, linguistic personality, machine 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