РОЗДІЛ VIII. ПРИКЛАДНА ЛІНГВІСТИКА: НАПРЯМИ Й АСПЕКТИ ДОСЛІДЖЕННЯ

Ілля Данилюк

УДК 81’23

АВТОМАТИЗОВАНІ МЕТОДИ ОПИСУ ТА РОЗПІЗНАВАННЯ МОВНОЇ ОСОБИСТОСТІ

У статті описано методи комп’ютерної лінгвістики, покладені 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. Подано опис ключових методів, як-от машинне навчання, синтез мовлення, моделювання почерку.

Ключові слова: лінгвоперсонологія, мовна особистість, моделювання, методи, машинне навчання, синтез мовлення.

Лінгвоперсонологія – наука, що має предметом вивчення власне-людську мовну особистість – спирається на вже практично столітні теоретичні засади, закладені у працях В. Гумбольдта, Й. Л. Вайсгербера, І. Бодуена де Куртене, В. Вундта, О. Потебні, В. Виноградова, згодом поглиблені та чіткіше окреслені в роботах В. Карасика, Ю. Караулова, В. Нерознака та ін. (теоретичні засади лінгвоперсонології є також предметом розгляду першої публікації циклу (Danylyuk “Teoretychni Zasady i Metody Linhvopersonolohiyi”)). З іншого боку, розвиток інформаційних технологій, здобутки квантитативної та корпусної лінгвістики, можливості автоматичного опрацювання природного мовлення озброїли лінгвоперсонологію новим інструментарієм і, відповідно, накреслили нові завдання та перспективи.

Метою статті є опис та узагальнення методології фундаментального дослідження з лінгвоперсонології. Завдання, які має бути розв’язано, включають визначення підходів до моделювання 1) формально-змістового рівня діяльності мовної особистості; 2) з’ясування обсягу поняття мовної особистості; 2) моделювання формально-звукового рівня діяльності мовної особистості – за допомогою систем синтезу мовлення; 3) моделювання формально-графічного рівня діяльності мовної особистості – почерку.

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).

У лінгвістиці 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 (Karasik).

Саме лінгвістичний підхід, на нашу думку, набуває нової актуальності через можливість збирати велику кількість мовленнєвих даних у вигляді корпусів текстів та звукових корпусів мовлення, а також через появу інструментарію для їх автоматичного опрацювання.

У межах проекту «Комунікативно-прагматична і дискурсивно-граматична лінгвоперсонологія: структурування мовної особистості та її комп’ютерне моделювання» однією з опорних точок ж теза про те, що мовносоціумна особистість може бути описана, змодельована і клонована у віртуальному вимірі на основі аналізу породжених реальною особою (донором) усних або писемних текстів. Відповідно, лінгвістичний підхід до вивчення такої 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.

І. Моделювання формально-змістового рівня діяльності мовної особистості

Корпусні технології давно вже стали одним із основних методів лінгвістичних досліджень. Так, ще в 1960-і роки створювався Браунівський корпус (США), який містив 1 млн слів. У 1970-і роки минулого століття стартував LOB корпус (Великобританія, Норвегія), у 1980-ті роки почали створюватися такі корпусу, як: Машинний Фонд російської мови, Уппсальський корпус російської мови (Швеція) – обидва по 1 млн слів, The Bank of English, Birmingham, – 20 млн слів. У 1990-і роки було створено British National Corpus, який включав на той час 100 млн слів, а також інші національні корпуси для угорської, італійської, хорватської, чеської, японської мови обсягом по 100 млн слів. На початку XXI ст. створювалися такі корпуси, як American National Corpus і Gigaword corpora (англійська, арабська, китайська) на 1 млрд слів, Національний корпус російської мови, над яким працюють лінгвісти Москви і Санкт-Петербурга, містить 300 млн слововживань. В Україні

© Данилюк І.Г., 2016

проблемою корпусної лінгвістики активно займаються вчені Інституту української мови НАН України, Українського мовно-інформаційного фонду, Інституту філології Київського національного університету ім. Т.Шевченка, Національного університету «Львівська політехніка» та ін.

Корпусний менеджер Manatee/Bonito й розроблені для нього корпуси текстів було докладно описано в (Danylyuk "Korpus Tekstiv dlya Vyvchennya Hramatychnoyi Sluzhbovosti", Danylyuk "Korpus Tekstiv dlya Vyvchennya Hramatychnoyi Sluzhbovosti: Klasyfikatsiya Hramatychnykh Klasiv i Pidklasiv"). Корпус текстів Юрія Шевельова (Шереха) викладено на corpora.donnu.edu.ua.

Під терміном лінгвістичний, або мовний, корпус текстів сьогодні розуміють великий, представлений в електронному вигляді, уніфікований, структурований, розмічений, філологічно компетентний масив мовних даних, призначений для вирішення конкретних лінгвістичних завдань. Такими завданнями у проекті «Комунікативно-прагматична і дискурсивно-граматична лінгвоперсонологія: структурування мовної особистості та її комп'ютерне моделювання» є мовна модель, побудована на основі текстів авторства Юрія Шевельова, а також екстракцію знань з цих текстів та дії, спрямовані на розуміння текстів для наповнення бази знань, формування відповідей на запитання і ведення діалогу з моделлю його мовної особистості.

Концептуально завдання автоматичного опрацювання текстів було розглянуто в роботах Н. Хомського, присвячених граматиці природної мови, в яких було описано ключову парадигму комп'ютерної лінгвістики – контекстно-незалежну граматику (CFG). Перші спроби автоматичної обробки текстів зводилися до розбору із застосуванням такої граматики, побудову дерева розбору і переведення його в певне логічне представлення знань за допомогою правил і лексикону. Після цього логічне представлення можна було додати в базу знань і виконувати з ним різні операції: шукати інформацію потрібного змісту чи типу, відповідати на запитання, перевіряти твердження тощо. Однак практичне застосування цього підходу було обмежено труднощами, пов'язаними з необхідністю враховувати загальноприйняті знання про світ, усталеної моделі для чого так і не було створено.

З 1990-х років у розпорядженні вчених з'явилися методи машинного навчання і статистичної лінгвістики. У машинному навчанні ефективними були алгоритми класифікації для різних завдань, пов'язаних з обробкою текстів: визначення спаму, сортування документів за тематиками, виділення іменованих сутностей (власних назв). У комп'ютерній лінгвістиці визначення частин мови стало високоточним завдяки таким статистичним методам, як приховані ланцюга Маркова і моделі максимальної ентропії. З'явилися парсери на основі імовірнісних контекстно-незалежних граматик, а в корпорації IBM було реалізовано масштабний проект зі статистичного машинного перекладу. Нарешті, було закладено основи *глибинного навчання* – найефективнішого на сьогодні з усіх автоматичних методів.

Глибинне навчання – навчання багаторівневих («глибоких») нейронних мереж на великих обсягах даних, що дозволяють уникнути роботи зі ручного розмічування корпусів текстів для машинного навчання, оскільки система «вчиться» виділяти їх автоматично. До речі, перший загальний робочий алгоритм керованого навчання багаторівневої мережі персептронів було опубліковано нашими співвітчизниками О. Івахненком і В. Лапою у (Ivakhnenko).

2010 року було запропоновано модель лексикалізованої ймовірнісної граматики, яка дозволила підвищити точність граматичного розбору до 93%. Точність розбору – це відсоток правильно побудованих граматичних зв'язків, однак імовірність того, що довге речення буде розібрано правильно, зазвичай дуже низька. Одночасно, завдяки новим алгоритмам і підходам, включно з глибинним навчанням, збільшилася швидкість граматичного розбору. Крім того, практично всі провідні алгоритми і моделі сьогодні у відкритому доступі, зокрема й один з найефективніших алгоритм Томаса Міколова (Mikolov).

Сьогодні для моделювання формально-змістового рівня діяльності мовної особистості маємо інструменти, які можна умовно поділити на три класи: *а) методи роботи зі словами, б) методи роботи з реченнями і в) методи для обробки довільних текстів.*

1. Методи роботи зі словами

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.

2. Методи роботи з реченнями

Перше завдання обробки мовлення на рівні речення – встановлення його синтаксичної структури. Інструменти роботи з синтаксисом помітно прогресували – лексикалізовані імовірнісні граматики значно підвищили якість синтаксичного розбору, а зручні для багатьох випадків граматики залежностей досягли якості, достатньої для розв'язання значного класу завдань обробки текстів. Крім того, за останні кілька років в сотні разів збільшилася швидкість алгоритмів синтаксичного аналізу.

Втім, точність автоматичного синтаксичного аналізу, особливо складних речень, і далі є порівняно невисокою. По-перше, багато чого залежить від якості розпізнавання частин мови, яке має бути дуже високим (97-98%), а в довгих реченнях часто трапляються неправильно визначені граматичні класи, що призводить до помилок розбору. По-друге, сам граматичний розбір дає точність щонайбільше 90-93% (відсоток правильно визначених відношень), а це означає, що в довгому реченні практично завжди будуть помилки розбору. Наприклад, за точності розбору 90% ймовірність розбору речення у 10 слів без жодної помилки складе лише 35%.

Методи глибинного навчання дозволяють інакше підійти до роботи з реченнями – моделювати речення як послідовність векторів, отриманих методом word2vec, і використати його в алгоритмах машинного навчання. Стандартні алгоритми машинного навчання працюють з фіксованим набором атрибутів, і їх не можна пристосувати до такої моделі. Для таких випадків запропоновано скористатися рекурентними нейронними мережами, які на вході приймають одне слово у векторному представленні і мають кілька внутрішніх рівнів, а на виході будують класифікатор. На відміну від звичайних нейромереж, внутрішні рівні *рекурентної мережі* (а іноді і верхній рівень) підключені назад у мережу, тобто стан мережі, у який вона перейшла на попередньому слові, буде передано в мережу як додатковий вхід на наступному слові. Таким чином, у нейромережі з'являється аналог «пам'яті», що дозволяє їй послідовно обробляти слова в реченні та будувати припущення окремо щодо кожного слова або всього речення цілком. Інакше кажучи, мережі послідовно передають одне слово речення за іншим, а мережа використовує свої попередні стани для визначення поточного кроку. Однак на практиці прості рекурентні мережі працюють не вельми ефективно через те, що пам'ять про попередні слова в реченні швидко втрачається під час тренування та експлуатації мережі. Тому зазвичай використовують спеціальні елементи пам'яті – *LTSM (Long Term Short Memory)*, що є множиною нейронів і керівних елементів, які визначають, коли треба записувати, читати й очищати пам'ять. Ці елементи дозволяють пам'яті не змінюватися під час тривалого послідовного обчислення і правильно атрибутувати помилку під час навчання.

Рекурентні нейромережі з LTSM добре себе зарекомендували для розв'язання різних завдань – моделювання мови, машинний переклад, – але у цього класу мереж є істотний недолік, пов'язаний з тим, що вони використовують тільки порядок слів у реченні і не працюють з граматичними структурами, отриманими традиційними інструментами, як-от автоматичним морфологічним аналізом. По суті, рекурентні мережі для кожного завдання з нуля «вчаться» граматики мови. Крім того, рекурентна мережа не будує представлення для проміжних фраз, тому для завдань, у яких потрібні якісні представлення різних фраз у складі речення, використовують *рекурсивні нейронні мережі*.

На відміну від рекурентних, рекурсивні мережі працюють не з ланцюжком слів у реченні, а на основі граматики залежностей – для кожного речення будується бінарне дерево для його розбору. Роботу рекурсивної мережі можна описати ось як. Спочатку вона обробляє листочки дерева розбору (листочки дерева – вказівники на два слова речення і на тип граматичної залежності між ними), заміщаючи листочки отриманим вектором тієї ж розмірності, що і вектори слів. І продовжує працювати далі, але тепер листочки вже об'єднують фрази, а не слова – будуються векторні представлення фраз речення. Отже, маючи дерево розбору, можна побудувати рекурсивну мережу з такою ж топологією, як і дерево, замінивши кожен вузол дерева на нейронну мережу. Природно, всі розмножені у такий спосіб мережі мають спільні параметри, тобто під час навчання та експлуатації робота йде з однією мережею.

Під час навчання рекурсивна мережа може навчитися робити якісні представлення не тільки для повних речень, але і для всіх фраз речення. Водночас нейромережа може послабити ефект помилок граматичного розбору. Таким чином, мережа дозволяє визначити міру семантичної близькості як для слів, так і для всіх фраз у реченні. Якщо в рекурсивну нейронну мережу додати елементи пам'яті LTSM, то можна отримати дуже якісні векторні представлення (Tai).

Інший підхід для отримання векторів речення полягає в тому, що для кожного речення, параграфа або цілого документа тренується окремий вектор, який також бере участь в прогнозі контексту кожного слова речення або параграфа, і в процесі навчання вибираються вектори, які найбільшою мірою поліпшують передбачення. За якістю отриманих векторів цей метод (його зазвичай називають doc2vec) змагається з рекурсивними нейромережами, водночас для навчання не потрібна розмічена навчальна вибірка. Щоправда, у цього методу є суттєві недоліки: йому потрібні великі речення або цілі параграфи – він не працює на рівні коротких фраз; і він вимагає значних обчислювальних потужностей.

Ще один підхід до моделювання слів і речення – нейромережі, що працюють із символьними представленнями слів або змішаними представленнями. Ці мережі успішно використовують для доповнення векторних представлень слів. Наприклад, у навчальній вибірці не було слова «побігати», але слово «бігати» траплялося часто, і було отримано його якісне представлення. Тоді нейромержа зі змішаним представленням зможе отримати вектор слова «побігати», застосувавши префікс «по». Тобто нейромережі із символьними представленнями навчаються використовувати морфологію слів, а не тільки працювати з певними словами як неподільними сутностями.

Отже, за допомогою рекурентних і рекурсивних нейромереж можна ефективно розв’язувати прості завдання, пов'язані з автоматичною обробкою текстів: класифікації, визначення тональності, виділення іменованих сутностей, простих фактів тощо.

3. Методи для обробки довільних текстів

Обробка текстів, що складаються з декількох речень, які потрібно розглядати не як незалежні сутності, а як взаємопов'язаний ряд висловлювань, для всіх наявних технологій є суттєвою проблемою. У цьому разі виникає семантичний контекст, який збагачується і модифікується наступними реченнями, і моделювати його дуже складно. У комп'ютерній лінгвістиці вже багато років не має остаточного прийнятного рішення завдання кореферентності або анафоричних відношень. Тільки значне обмеження прикладної сфери дозволяє будувати прийнятні семантичні моделі для текстів, що складаються з декількох речень, і діалогів.

II. Моделювання формально-звукового рівня діяльності мовної особистості

Цей напрям моделювання мовної особистості спирається на дані у формі цифрового звуку – запису живого мовлення. Зібрані у формі звукового корпусу такі дані є джерелом для виділення параметрів, необхідних для системи синтезу мовлення, яка імітуватиме задану мовну особистість.

Комп’ютерний синтез мовлення – порівняно не нова технологія. Електронний вокодер Г. Дадлі було представлено у 1930-ті роки, у 1950-ті з’явився синтезатор «Pattern playback», а 1968 року – перша система синтезу мовлення з довільного тексту, розроблена Норіко Умеда. За тривалий час розвитку виникло кілька підходів.

1. Формантний синтез

В основі формантного синтезу, вперше розробленого у 1929 році Д.Коллардом, лежать лінгвістичні знання про типи фонем і фізичні знання про наявність у їхньому складі формант. Дані про частоту, ширину і рівень 2-4 формант для кожної фонеми втілюються в генеровану систему спектрограму – цілковито штучну. У процесі синтезу можна керувати усіма параметрами – частотою основного тону, тривалістю звуків, амплітудою – моделюючи мовлення будь-якої типу (дитяче, доросле), будь-якого темпу, довільно наголошуючи будь-які фонеми. Схема системи формантного синтезу включає генератори базових формант, генератор носових формант і генератор шуму для породження приголосних фонем. Цей спосіб також має назву «синтез за правилами», бо використовується низка алгоритмізованих правил про вимову фонем у різних позиціях.

Формантний синтез має низку переваг:

- повний контроль над генерованим сигналом, тобто універсальність;
- можливість використання у малопотужних на сьогодні системах.

Недоліки:

- невисока якість мовлення, його виразна штучність.

2. Артикуляторний синтез

Артикуляторний синтез ґрунтується на геометричній моделі мовленнєвого тракту людини. Перший такий синтезатор був розроблений в Haskins Laboratories в середині 1970-их. Донедавна цей тип синтезу не мав комерційних реалізацій через високу складність. Єдиним помітним винятком є розроблений синтезатор Trillium, який сьогодні перебуває у відкритому доступі й має назву gnuspeech (https://www.gnu.org/software/gnuspeech/).

Артикуляторний синтез має низку переваг:

- повний контроль над генерованим сигналом.

Недоліки:

- посередня якість мовлення, складність реалізації.

3. Конкатенативний (компілятивний) синтез

3.1. Предметно-орієнтований синтез

Предметно-орієнтований синтез компілює слова, записані заздалегідь, а також фрази для створення завершених повідомлень. Він використовується в системах, де різноманіття текстів обмежене певною темою/сферою, наприклад, оголошення про відправлення потягів і прогнози погоди. Ця технологія проста у виконанні й досить довго застосовувалася з комерційною метою: побутові електронні прилади, годинник і калькулятори, здатні говорити, довідка мобільних операторів.

Предметно-орієнтований синтез має низку переваг:

- природність звучання цих систем потенційно може бути високою завдяки тому, що кількість можливих речень обмежена і максимально наближена до оригінальної інтонації.

Недоліки:

- системи обмежені вибором слів і фраз у базі даних,
- немає контролю основних параметрів мовлення
- місця склеювання слів і фраз можна почути.

3.2. Сегментний синтез

Синтез мовлення з використанням попередньо записаних відрізків набув поширення у зв'язку з появою можливостей швидкого маніпулювання даними, відібраними з великих масивів. Залежно від розміру вихідних елементів компілювання виділяються такі види синтезу:

- мікросегментний (мікрохвильовий);
- алофонний;
- дифонний;
- напівскладовий;
- складовий.

Найпопулярнішим серед них свого часу став дифонний, який використовує мінімальний корпус дифонів (напівзвуків). Кількість дифонів залежить від фонетики мови: наприклад, іспанська має приблизно 800 дифонів, а німецька – 2500. У базі даних зберігається тільки один варіант кожного дифона. У процесі синтезу система вибирає із бази дифони методом динамічного програмування, кодування із лінійним передбаченням, PSOLA (метод зсування звукових відрізків у часі для досягнення ефектів зміни тону) або MBROLA (міжнародний відкритий проект корпусів дифонів для різних мов).

Сегментний синтез має низку переваг:

- синтез наперед невідомого тексту
- порівняно малий розмір бази даних

Недоліки:

- невисока якість синтезу через спотворення у місцях склеювання
- неможливість керування інтонацією мовлення, причому навіть велика база даних не вирішує проблему

3.3. Unit Selection (вибір одиниць)

Метод спирається на велику базу даних – корпус – записаного мовлення. Цей корпус має глибоку розмітку: поділений на окремі алофони, дифони, напівфони, склади, морфеми, слова, фрази і речення. Як правило, сегментація виконується спеціальним *розпізнавачем*, результати роботи якого виправляють вручну. Для кожного виділеного сегмента встановлюється відповідник – звукозапис або спектрограма. Потім така база даних індексується – укладається перелік усіх одиниць, відсортований за наявністю певних параметрів (частота основного тону, тривалість, амплітуда, позиція в складі, сусідні фонеми тощо). У процесі синтезу з бази даних вибираються (unit selection) і склеюються найбільш відповідні вихідному тексту одиниці різного розміру – від цілого речення до окремих алофонів.

Багаторівнева розмітка бази даних має, наприклад, такий вигляд:

- рівень періодів основного тону, що містить вказівку про вимову алофона після паузи, після голосного чи приголосного, перед паузою;
- рівень міток модифікації мовленнєвого сигналу: подовження алофонів, фонетичні явища типу оглушення чи одзвінчення, акомодації тощо;
- рівні реальної та ідеальної транскрипції;
- рівень слів;
- рівень складів;
- рівень інтонації та пауз;
- рівень міток емоційного забарвлення.

Алгоритм вибору спирається на математичне визначення двох параметрів:

- **ваги заміни** – зваженої суми різниці в ознаках між синтезованим елементом і одиницею в базі (ці ознаки – частота основного тону, тривалість, амплітуда, позиція в складі, сусідні фонеми тощо);
- **ваги зв’язку** – зваженої суми різниці в ознаках між двома сусідніми одиницями.

Метод забезпечує максимальну схожість синтезованого мовлення на природне через те, що вибрані одиниці з бази зазнають мінімального опрацювання, як правило, тільки згладжування спектрограми у місці

склеювання. Однак висока якість можлива тільки за умови використання великої бази даних – десятків годин мовлення.

Метод Unit Selection має низку переваг:

- синтез довільного тексту
- найвища на сьогодні якість синтезованого мовлення

Недоліки:

- великий обсяг даних, потрібних для роботи
- в алгоритмі є особливість: для коротких ненаголошених слів вибираються не найкращі зразки, навіть якщо такі є в базі, що іноді створює неприродність мовлення
- відсутність контролю над усіма параметрами мовлення, зокрема, інтонаційного оформлення.

4. HMM-синтез (на основі статистичних моделей)

Метод базується на використанні прихованих Марковських моделей, тобто для частотного спектра (модель голосового тракту), основної частоти звучання (модель голосу) і тривалості (модель просодики) створюються ймовірнісні моделі. Надалі на основі цих моделей генерується звукова хвиля – як найбільш імовірна для синтезованого ланцюжка символів. Якість синтезу цілковито визначається якістю і розміром розміченого автоматично і вручну звукового корпусу мовлення.

HMM-синтез має низку переваг:

- синтез довільного тексту;
- висока якість синтезованого мовлення.

Недоліки:

- великий обсяг і складність опрацювання базового корпусу, відсутність у методі власне лінгвістичних даних;
- відсутність контролю над усіма параметрами мовлення, зокрема, оформлення тембру, звідси часткова неприродність звучання.

5. Синусоїдальний синтез

Значною мірою експериментальний вид синтезу, що ґрунтується на заміні відомих формант звуків на чистий синтезований сигнал (тон), тобто слухач інтуїтивно домислює необхідний для вимови приголосних шум. Метод розроблений в Haskins Laboratories у 1970-ті роки.

6. CUTE

Нарешті, як і з глибинним машинним навчанням, що стало певною комбінацією різних методів, було запропоновано конкатенативний синтез на основі Unit selection з попереднім добором одиниць на рівні трифонів і згладжуванням переходів між ними (Jin) (назва CUTE є абревіатурою англійських назв цих окремих методів). Його основна ідея полягає у «склеюванні» відібраних з попередньо записаного масиву мовлення якомога довших сегментів, що відповідають синтезованому тексту, а щоби якомога довші сегменти потрапляли у вибірку одиниць для конкатенації, їх попередньо «просіюють» на рівні трифонів. Параметри для згладжування переходів між з'єднаними сегментами обчислюються на зразках із вже відомого масиву. Нарешті, з цього ж масиву копіюється контур ймовірної інтонації. За твердженням розробників, обсяг масиву для навчання системи може становити до 20 хвилин мовлення.

У роботі наведено результати порівняльних тестів з іншими методами голосової конверсії, у яких CUTE значно перевершує конкурентів. Водночас згадуються деякі його недоліки: він так само страждає від недостатньої кількості фонем у базі для синтезу нових слів, через що генеруються фонетично правильні, але не дуже реалістичні результати. Крім того, він залежить від роботи системи розпізнавання мови для коректної фонетичної сегментації.

Втім, немає сумніву, що саме цей метод може бути ефективно використано для моделювання звукового мовлення мовної особистості.

III. Моделювання формально-графічного рівня діяльності мовної особистості

Моделювання індивідуального почерку – завдання, яке по суті стосується аналізу та синтезу зображень. Це одне з основних практичних завдань, що успішно розв'язують методом машинного навчання. А. Грейвс у роботі (Graves) наводить зразок такої системи із застосування рекурентних нейронних мереж (RNN) з архітектурою Long-Short Term Memory (LSTM) для генерації спеціальних послідовностей даних (у цьому разі графічних), кожен елемент в яких обчислюється на підставі попереднього елемента. Якщо навчити штучну нейронну мережу на зображеннях почерку конкретної мовної особистості, то можливим буде обернене генерування таких зображень з випадковою видозміною накреслення літер за заданим текстом.

Отже, сучасні методи комп'ютерної лінгвістики, які використовують класичні словникові підходи, складні лінгвістичні алгоритми, що ґрунтуються на правилах, машинне навчання, штучні нейронні мережі – якщо використовувати їх у сукупності – дозволяють моделювати мовну й мовносоціумну особистість у різних аспектах її мовленнєвої дісльності. Найперспективнішим видається метод глибинного навчання на корпусах текстів, мовлення чи графічних даних з використанням штучних нейромереж різної типології, що й буде розглянуто в наступних публікаціях.

References

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

Danylyuk, Illya. "Korpus Tekstiv dlya Vyvchennya Hramatychnoyi Sluzhbovosti (Text Corpora to Study of a Grammatical Auxiliarity)." *Linhvistychni Studiyi* (*Linguistic Studies*) 26 (2013): 224-230. Print.

Danylyuk, Illya. "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, Illya. "Teoretychni Zasady i Metody Linhvopersonolohiyi (Theoretical Principles and Methods of Lingvopersonology)." *Linhvistychni Studiyi* (*Linguistic Studies*) 31 (2016): 63-66. Print.

Graves, Alex. "Generating Sequences with Recurrent Neural Networks." arXiv preprint arXiv:1308.0850 (2013). Web. 10 Sep. 2016.

Ivakhnenko, A. H., and V. H. Lapa. *Kybernetycheskye Predskazyvayushchye Ustroystva* (*Cybernetic Predicting Devices*). Kyiv: Naukova dumka, 1965. Print.

Jin, Zeyu, et al. "Cute: A concatenative Method for Voice Conversion Using Exemplar-Based Unit Selection." *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016*. Web. 10 Sep. 2016.

Karasik, Vladimir. *Yazykovoy Kruh: Lychnost', Kontsepty, Dyskurs* (*Linguistic Circle: Personality, Concepts, Discourse*). Volgograd: Peremena, 2002. Print.

Mikolov, Tomas, et al. "Efficient Estimation of Word Representations in Vector Space." arXiv preprint arXiv:1301.3781 (2013). 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.

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.

Надійшла до редакції 20 вересня 2016 року.

AUTOMATED LINGUISTIC PERSONALITY DESCRIPTION AND RECOGNITION METHODS

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 (e-mail, sms, voice, audio and video blogs, social networks, etc.).

Purpose: The purpose of the article is to describe the theoretical and practical framework of the project "Communicative-pragmatic and discourse-grammatical lingvopersonology: structuring linguistic identity and computer modeling". The description of key techniques is given, such as machine learning for language modeling, speech synthesis, handwriting simulation.

Results: Lingvopersonology developed some great theoretical foundations, its methods, tools, and significant achievements let us predict that the newest promising trend is a linguistic identity modeling by means of information technology, including language. We see three aspects of the modeling: 1) modeling the semantic level of linguistic identity – by means of the use of corpus linguistics; 2) sound level formal modeling of linguistic identity – with the help of speech synthesis; 3) formal graphic level modeling of linguistic identity – with the help of image synthesis (handwriting). For the first case, we suppose to use machine learning technics and vector-space (word2vec) algorithm for textual speech modeling. Hybrid CUTE method for personality speech modeling will be applied to the second case. Finally, trained with the person handwriting images neural network can be an instrument for the last case.

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. A text corpus and audio corpus are being built, some samples of scientist's handwriting are collected, different techniques, such as machine learning, speech synthesis, handwriting simulation are going to be applied.

Keywords: lingvopersonology, linguistic personality, modeling techniques, machine learning, speech synthesis.

Vitae

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

Correspondence: i.danyluk@donnu.edu.ua