

Олексій ДОВГАНЬ
Світлана МАТВЄЄВА

ЦИФРОВІ ТЕХНОЛОГІЇ В ЛІНГВІСТИЦІ:

ВІД МОВНИХ КОРПУСІВ ДО ШТУЧНОГО ІНТЕЛЕКТУ

УДК 81’322:004:[004.6+004.8](075.8)
Д 58

Рекомендовано Вченою радою Українського державного університету імені Михайла Драгоманова як підручник для здобувачів освіти спеціальності В11 Філологія *(протокол № 1 від 28 серпня 2025 р.)*

Рецензенти:

Рошко Р., доктор філологічних наук, професор, професор Інституту славістики Польської академії наук;

Дарчук Н., доктор філологічних наук, професор, професор кафедри української мови та прикладної лінгвістики Навчально-наукового інституту філології Київського національного університету імені Тараса Шевченка;

Мазурик Д., кандидат філологічних наук, доцент, доцент кафедри українського прикладного мовознавства Львівського національного університету імені Івана Франка; член Національної комісії зі стандартів державної мови.

Довгань О., Матвєєва С.

Цифрові технології в лінгвістиці : від мовних корпусів до штучного інтелекту : підручник / Міністерство освіти і науки України, Український державний університет імені Михайла Драгоманова. Київ : УДУ імені Михайла Драгоманова, 2025. 226 с. електронне видання

DOI : https://doi.org/10.31392/pidruchnyk.UDU-TTvL.2025

ISBN 978-966-931-336-2

У виданні послідовно розглянуто теоретичні засади та практичні аспекти застосування цифрових методів у мовознавстві. Структура підручника охоплює ключові питання: від принципів створення, структурування та аналізу мовних корпусів до методології автоматичної обробки природної мови й сучасних технологій машинного навчання.

Метою підручника є формування теоретичних знань про можливості цифрових інструментів та їх застосування в лінгвістиці, перекладознавстві та лексикографії, а також розвиток практичних компетенцій, необхідних для проведення емпіричних досліджень і роботи з сучасними лінгвістичними даними.

Видання формує комплекс вмінь і навичок, які дозволять не лише впевнено працювати з цифровими інструментами, а й критично оцінювати результати їх застосування, забезпечуючи баланс між технічними можливостями та гуманітарними цінностями.

Підручник адресовано здобувачам, науково-педагогічним працівникам закладів вищої освіти, а також аспірантам та всім тим, хто прагне інтегрувати цифрові технології в гуманітарну сферу, зберігаючи при цьому аналітичну глибину та академічну доброчесність.

УДК 81’322:004:[004.6+004.8](075.8)

© Ілюстрацію для обкладинки створено за допомогою штучного інтелекту (ChatGPT від OpenAI). Авторська ідея та опис – О. Довгань, 2025.
© Довгань О., Матвєєва С., 2025.

ISBN 978-966-931-336-2

© УДУ імені Михайла Драгоманова, 2025.
Цей матеріал ліцензовано за умовами Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0).
Використання дозволене лише для некомерційних цілей із зазначенням авторства.
Покликання на ліцензію: https://tinyurl.com/2u3yjyn3

ЗМІСТ

СПИСОК УМОВНИХ СКОРОЧЕНЬ……..……..6

ПЕРЕДМОВА…….7

ЛЕКЦІЇ……9

Тема 1. Цифрові технології:

поняття й ключові етапи становлення..……....9

1.1. Поняття цифрових технологій……...……..9

1.2. Історія розвитку цифрових технологій...……...10

1.3. Вплив цифрових технологій на суспільство……...……..37

Запитання для самоконтролю……...38

Тема 2. Текстові дані та робота з ними

в контексті цифрових технологій….……39

2.1. Цифрові технології для роботи з корпусами текстів……39

2.2. Лінгвістична експертиза тексту……40

2.3. Поняття плагіату та його виявлення……..42

2.4. Фактчекінг: методологія та інструменти……...45

2.5. Машинний текст та його аналіз:

автоматизовані та ручні підходи виявлення……...57

Запитання для самоконтролю……...61

Тема 3. Текстові корпуси в цифровій лінгвістиці:

види, можливості, застосування……62

3.1. Поняття текстового корпуса, його основні характеристики……62

3.2. Типологія корпусів……..63

3.3. Структура корпуса: одиниці організації, розмітка, формати……...65

3.4. Застосування корпусів у філології та перекладознавстві……72

3.5. Обмеження та виклики при роботі з текстовими корпусами……...73

Запитання для самоконтролю……...74

Тема 4. Штучний інтелект та штучні нейронні мережі……..……...76
4.1. Основні поняття, принципи роботи та класифікація……76
4.2. Архітектура та специфіка роботи штучних нейронних мереж……...79
4.3. Навчання штучних нейронних мереж (машинне навчання)……88
Запитання для самоконтролю……...96

Тема 5. Методи обробки природної мови…….....98
5.1. Загальне поняття та сфери застосування обробки природної мови……98
5.2. Розпізнавання мовлення……...100
5.3. Попередня обробка звукових сигналів……102
5.4. Синтез мовлення……...106
5.5. Голосовий інтерфейс……114
Запитання для самоконтролю……119

Тема 6. Параметризація обробки природної мови……...120
6.1. Підходи до обробки природної мови: спрощений та розширений……120
6.2. Глибоке навчання в обробці природної мови……...122
6.3. Використання методів обробки природної мови
для вилучення визначень слів із контексту……...135
6.4. Труднощі в роботі систем обробки природної мови
та основні методи їх вирішення……137
Запитання для самоконтролю……144

ЛАБОРАТОРНІ РОБОТИ……145
№ 1/2. Тема 1. Розробка ефективних промптів для штучного інтелекту:
створення, оптимізація, аналіз……...145
№ 3/4. Тема 2. Корпуси текстів, використання цифрових технологій
для їх аналізу (на прикладі AntConc)……152
№ 5/6. Тема 3. Створення бітексту та паралельного корпуса з фрагмента
художнього тексту…...……...161
№ 7/8. Тема 4. Робота антиплагіатних та інших систем……...168

№ 9/10. Тема 5. Методи фактчекінгу та аналізу достовірності інформації у цифрових текстах……173
№ 11. Тема 6. Машинний текст та його виявлення: автоматизовані та лінгвістичні підходи……178

ПІСЛЯМОВА……..182

ОРІЄНТОВНІ ПИТАННЯ НА ІСПИТ……...183
РЕКОМЕНДОВАНА ЛІТЕРАТУРА……..187
СЛОВНИК ОСНОВНИХ ТЕРМІНІВ І ПОНЯТЬ……192
ЦИФРОВІ ІНСТРУМЕНТИ ТА РЕСУРСИ……..209

СПИСОК ДЖЕРЕЛ ТА ЛІТЕРАТУРИ……..212
ІНФОРМАЦІЯ ПРО АВТОРІВ……...225

СПИСОК УМОВНИХ СКОРОЧЕНЬ

БП – багатошаровий перцептрон

ВР – віртуальна реальність

ГА – генетичний алгоритм

ГЗМ – генеративна змагальна мережа

ГНМ – глибока нейронна мережа

ГН – глибоке навчання

ДКП – довготривала короткочасна пам’ять

ДМ – дифузійна модель

ДР – доповнена реальність

ЕОМ – електронна обчислювальна машина

ЗНМ – згорткова нейронна мережа

ЗРБ – закритий рекурентний блок

ІТ – інформаційні технології

КРБ – керований рекурентний блок

МДП – мережа динамічної пам’яті

МН – машинне навчання

НМВ – нейронна машина для відповідей

ОП – одношаровий перцептрон

ОПМ – обробка природної мови

ПЗ – програмне забезпечення

ПК – персональний комп’ютер

РНМ – рекурентна нейронна мережа

СРР – системи розподіленого реєстру

ЦТ – цифрові технології

ШІ – штучний інтелект

ШН – штучний нейрон

ШНМ – штучні нейронні мережі

ПЕРЕДМОВА

Сучасний стан суспільного розвитку позначений інтенсивним впровадженням у всі сфери життя цифрових (інформаційних) технологій: зокрема, у наукову та освітню практику. Природно, що використання таких технологій видозмінює узвичаєні процеси: пошук, аналіз, інтерпретацію та інше даних. Разом із цим, використання вищезазначеного інноваційного інструментарію ставить перед сучасними фахівцями-гуманітаріями цілу низку нових завдань і викликів.

У цьому контексті особливої значущості набувають прикладна лінгвістика, перекладознавство та цифрова гуманітаристика загалом, що функціонують як динамічні міждисциплінарні напрями. Це дозволяє авторам поєднати інструменти обчислювальної обробки даних із методами критичного та іншого аналізу тексту та дискурсу (зокрема, текстовими корпусами).

Підкреслимо, що пропоноване видання не є підручником у традиційному розумінні, оскільки його мета – не перевірка чи оцінювання рівня комп'ютерних знань здобувачів освіти. Навпаки, завдання підручника – сформувати у студентів стійкі навички медіаграмотності і фахової компетентності, розглядаючи останні крізь призму інструментарію прикладної лінгвістики та цифрової гуманітаристики.

Таким чином, медіаграмотність тут розуміється здатністю до критичного сприйняття інформації у цифровому середовищі. Мовиться, зокрема, про підготовку таких даних до обробки цифровими інструментами, власне аналіз та інтерпретацію. Окрім того, представлене видання ставить за мету формування у здобувачів навичок оцінки достовірності інформації, розуміння принципів функціонування цифрових ресурсів і алгоритмів, виявлення маніпулятивних стратегій та усвідомленого використання сучасних технологій для вирішення наукових, професійних та освітніх завдань.

Структура представленого видання поєднує теоретичний і практичний аспекти, охоплюючи питання від побудови та використання мовних корпусів і технологій обробки природної мови до методів автоматизованого виявлення

плагіату, фактчекінгу, аналізу машинного тексту та застосування штучного інтелекту у дослідницькій роботі. Кожну тему акцентовано таким чином, щоб зосередити увагу читачів не лише на технічному описі інструментів, а й на методології їх свідомого використання в умовах інформаційного перенасичення та швидкої зміни технологічних парадигм і культурних практик.

ЛЕКЦІЇ

Тема 1.

ЦИФРОВІ ТЕХНОЛОГІЇ: ПОНЯТТЯ Й КЛЮЧОВІ ЕТАПИ СТАНОВЛЕННЯ

1.1. Поняття цифрових технологій

Цифрові технології (далі – *ЦТ*) – електронні інструменти, пристрої, засоби та ресурси, які генерують, обробляють (аналіз, унаочнення, репрезентація тощо) дані. Відтак, такі технології є інструментарієм, методологією тощо, використовуваними з метою обробки інформації, представленої у дискретній формі, тобто у вигляді нулів та одиниць. Головною відмінністю цифрових від аналогових технологій є природа опрацьовуваних у їх межах сигналів: так, ЦТ оперують згаданими дискретними значеннями, в той час як аналогові – безперервними сигналами.

Вищезазначене дозволяє виокремити такі питомі *риси ЦТ*:

1. Окреслену *дискретність*, яка виявляється у сегментованому представленні даних (бітів). До прикладу, під час використання торрент-клієнта (піринговых (P2P) мереж) обмін даними (файлами) відбувається безпосередньо між користувачами (сідерами – тими, хто роздає дані, та лічерами – тими, хто їх завантажує). Вагомим у цьому прикладі для нас є те, що завантаження відбувається частинами: так, інформація є фрагментованою (дискретною) й «збирається» по комп'ютерам сідерів, а вже на пристроях лічерів поєднується у цілісний файл.

2. *Швидкість*, суть якої у надзвичайно високій інтенсивності опрацювання даних. Так, під час відеодзвінків (Google Meet, Zoom тощо), сториз або відео (Tik-Tok, BeReal та подібні) доступними є численні маски, ефекти та інше, які, здебільшого, використовувані у реальному часі.

3. *Точність*, представлена високою якістю обчислень та передачі даних. Мовиться, першою чергою, про генерацію текстових, графічних та інших даних: до прикладу, створені за допомогою штучного інтелекту (далі – ШІ) колажі,

презентації, аватари (зроблені з низки ваших фотографій, можуть презентувати расову, етнічну й так далі варіативність) та інше.

4. *Адаптивність* або *гнучкість*, локалізована широким спектром (від медицини й виробництва до лінгвістики) виконуваних такими системами завдань та умов. Показовими в цьому плані є штучні (генеративні) нейронні мережі (далі – ШНМ, які ми докладніше розглянемо далі), що демонструють низку «побічних» здатностей. Так, створені для генерації відео або зображень ШНМ мають достатньо потужні «навички» для обробки текстових даних, оскільки без розуміння запиту користувача неможлива успішна генерація відеоряду або графіки (до прикладу, Sora від OpenAI).

1.2. Історія розвитку цифрових технологій

Технології є, першою чергою, соціокультурним феноменом, функціонування й походження якого можна відслідкувати у давнину, коли останній мав форму обчислювальних систем. Однак експоненційного розвитку ЦТ, природно, набули у XX столітті, коли були винайдені електронні обчислювальні машини (далі – ЕОМ). Відповідно, визначимо стрижневі етапи розвитку ЦТ:

1. ***Ручні обчислювальні пристрої*** – етап, позначений використанням пристроїв, які полегшують обчислення, виконання певних дій на них проводилося через фізичне переміщення елементів. Яскравим прикладом такого пристрою є спеціальна дошка – «абак» (у японців – «соробан», у китайців – «суанпан»; див. Рис. 1), обчислення на ній проводилися переміщенням кісток або камінців у спеціальні заглиблення з різних металів (бронзи, каменю, слонової кістки та інших).



Рис. 1. Реконструкція грецького абака.

2. ***Механічні обчислювальні пристрої***, які характеризуються функціонуванням приладів, в основі роботи яких був механічний принцип обчислень. Вищезазначені пристрої будувалися на механічних елементах, та здатні були виконувати більше двох (якщо конкретно – чотирьох) арифметичних дій. Прикладами таких пристроїв були:

– *антикітерський механізм* (дозволяв проводити астрономічні дослідження; див. Рис. 2);



Рис. 2. Антикітерський механізм.

– *лічильна машина Леонардо Да Вінчі* (тринадцятирозрядний десятинний сумуючий пристрій на базі коліщаток із десятьма зубцями, реально збудований за ескізом винахідника лише у XX столітті; див. Рис. 3);

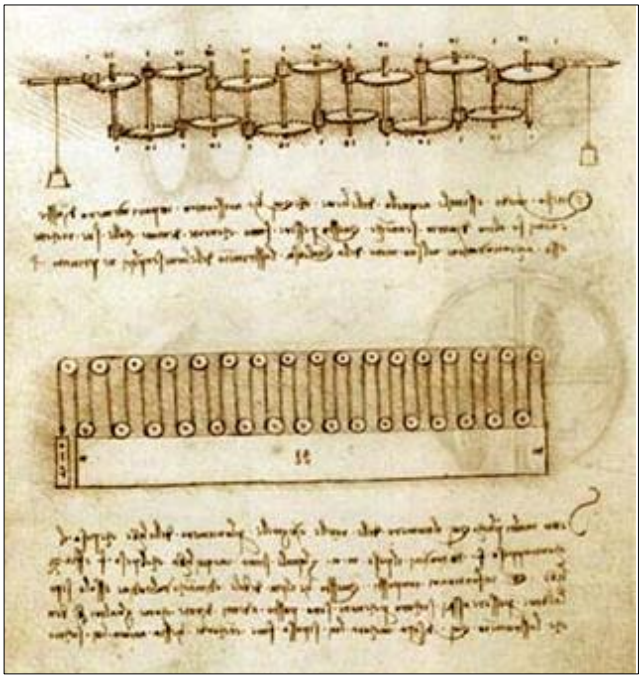


Рис. 3. Креслення Леонардо да Вінчі.

– *лічильний годинник Вільгельма Шиккарда* (по суті, відкрив еру механічної обчислювальної техніки; див. Рис. 4);

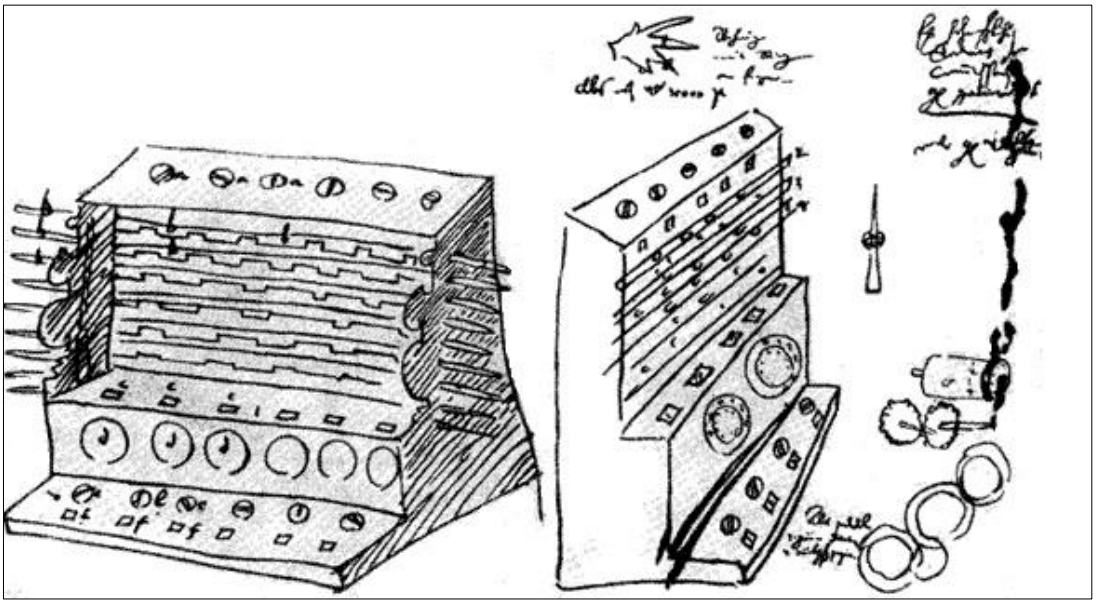


Рис. 4. Ескіз Лічильного годинника Вільгельма Шиккарда.

– *машина Блеза Паскаля (Паскаліна)*: стала основою для створення низки різноманітних за формою й конструкцією арифметичних машин (див. Рис. 5);



Рис. 5. Машина Блеза Паскаля (Паскаліна).

– *машина Готфріда Вільгельма Лейбніца* (творця питомого розділу сучасної математики – диференціального числення, який вдосконалив пристрій Паскаля; див. Рис. 6);



Рис. 6. Машина Готфріда Вільгельма Лейбніца.

– *різниці́ва машина Чарльза Беббіджа* (мала дві тисячі латунних деталей і була здатна виконувати арифметичні операції з точністю до шостого знака після коми й обчислювати похідні другого порядку; була прообразом сучасних комп'ютерів і носила назву диференціальної машини; див. Рис. 7).

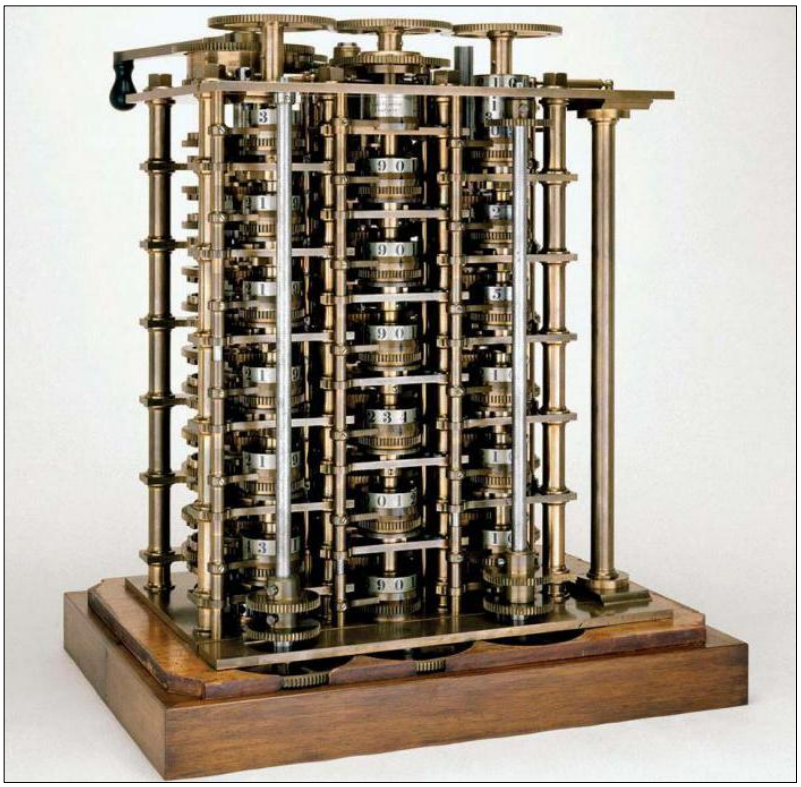


Рис. 7. Різниці́ва машина Чарльза Беббіджа.

3. ***Електронні лампові обчислювальні пристрої*** – етап, що був позначений функціонуванням і створенням вищезазначених машин на базі електромеханічних реле. Разом із тим, у цей період розпочали дослідження нового напряму – вищезазначених ЕОМ. Першим електронним цифровим пристроєм вважається Atanasoff-Berry Computer (див. Рис. 8), названий на честь розробників – Джона Атанасова й Кліффорда Беррі. Значущість цього пристрою в тому, що в ньому не було частин, що рухались: однак це не була програмована машина, проте у ній вперше використали електронні лампи (при цьому вона не була повнофункціональною).

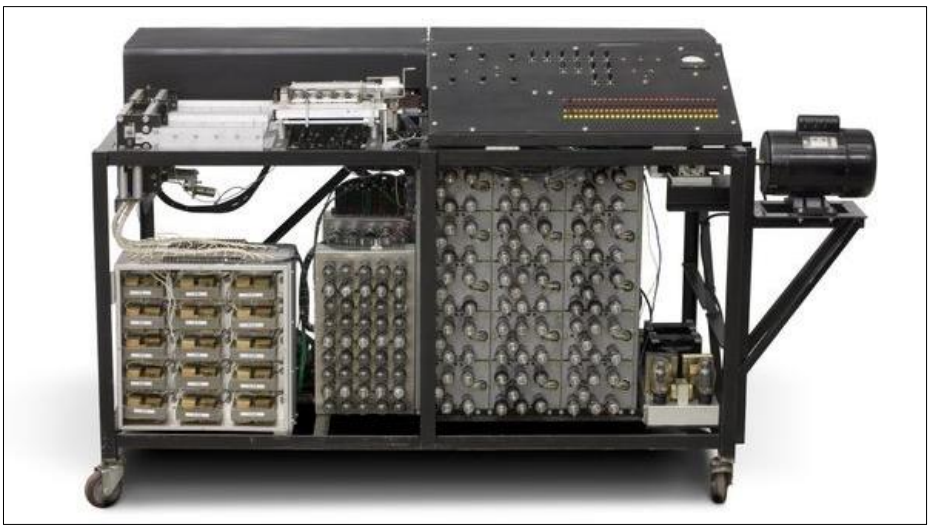


Рис. 8. Atanasoff-Berry Computer.

Натомість першою обчислювальною машиною став Colossus (1943 р.; див. Рис. 9), створений за участю Алана Тьюринга. Машина мала 2000 електронних ламп, а її головним призначенням стало розшифрування радіограм ворога. Через це інформація про пристрій довгий час була засекречена: останнє стало причиною того, що про неї не завжди згадують в історії ЦТ.

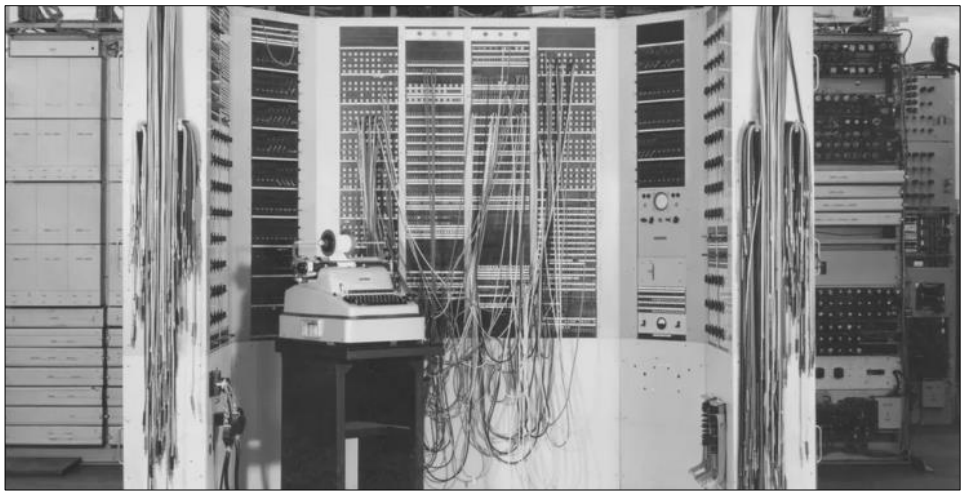


Рис. 9. Colossus.

Згадаємо також Electronic Numerical Integrator, Analyzer and Computer (1945 р.; див. Рис. 10), від якого, здебільшого, починають відлік ери ЕОМ: саме цей пристрій вважається представником їх **першого покоління**. Прикметно, що він так само мав воєнне призначення: розраховував артилерійські таблиці балістики.

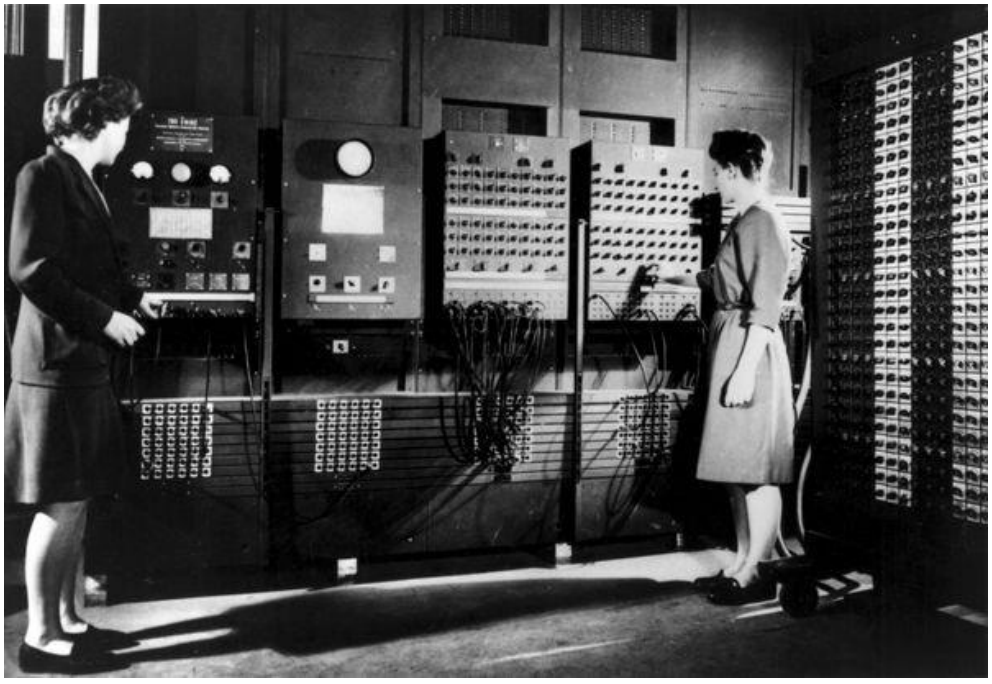


Рис. 10. Програмісти Electronic Numerical Integrator, Analyzer and Computer.

4. ***Транзисторні ЕОМ*** – **друге покоління** вищезазначених пристроїв, яке було позначене зменшенням їх розміру з винаходом транзисторів (напівпровідникових елементів, які дозволяють керувати струмом, споживаючи при цьому суттєво менше електроенергії, виділяючи менше тепла й займаючи менше місця).

В середині 50-х років було винайдено дешевий спосіб виробництва цих деталей, що призвело до того, що вже у другій половині цього періоду з'явилися комп'ютери, повністю базовані на них. Показово, що заміна електронних ламп на транзистори відбувалася паралельно з вдосконаленням методів зберігання інформації: почали використовувати магнітні сердечники та магнітні барабани, а згодом – зберігати дані на дисках.

Одним із перших комп'ютерів на транзисторах став Atlas Guidance Computer (1961 р.; див. Рис. 11), який у подальшому було використано при

управлінні запуском однойменної ракети. Іншим таким пристроєм став комп'ютер 7090, здатний виконувати 229 тисяч операцій завдяки *транзисторному мейнфрейму*. Цікаво, що на базі двох мейнфреймів авіакомпанією SABRE було створено автоматизовану систему продажу й бронювання авіаквитків у 65 країнах світу.



Рис. 11. Atlas Guidance Computer.

Помітним проривом у галузі став міні-комп'ютер компанії DEC – модель Programmed Data Processor – 1 (див. Рис. 12), а також його наступник – Programmed Data Processor – 8. Останній був суттєво зменшений до розміру холодильника, пішов у серійне виробництво й став закупатися лабораторіями, університетами та невеликими підприємствами.



Рис. 12. Programmed Data Processor – 1.

5. ***ЕОМ на інтегральних схемах*** (винайдені незалежно один від одного Джеком Сент Клером Кілбі та Робертом Н. Нойсом) стали представниками **третього покоління**, у побудові яких використали кремнієві кристали з мініатюрними електронними схемами. Розмір кожної з них – не більше горошини, у якій, тим не менш, вміщували тисячі вищезазначених транзисторів, що дозволило зменшити машини настільки, щоб вони могли бути розміщені на письмовому столі.

Відзначимо, що виробництво комп'ютерів на базі мікросхем, до всього, було набагато дешевшим за транзисторне, саме тому ці машини набули поширення. Своєю чергою, це спродукувало збільшення попиту на такі універсальні ЕОМ й експоненційне зростання їх виробництва до промислових масштабів. Також у цей час з'явилася *напівпровідникова пам'ять* (див. Рис. 13), яку досі використовують у персональних комп'ютерах (далі – ПК).

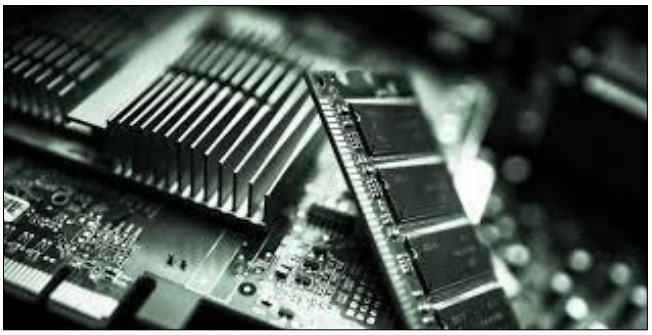


Рис. 13. Напівпровідникова пам'ять.

6. ***Мікропроцесори та ПК*** були представниками **четвертого покоління**, вони позначені використанням у своєму виробництві тисяч інтегральних схем, що знаменувало початок епохи мікрокомп'ютерів. Появу вищезазначених ПК, більшість дослідників пов'язують із процесорами корпорації Intel (див. Рис. 14), яка стала найбільшим виробником мікропроцесорів.



Рис. 14. Процесор корпорації Intel.

Метою компанії було створення напівпровідникової пам’яті, що дозволило їй у подальшому заробляти, беручи замовлення на розробку пристроїв на її базі. Своєю чергою, одне із замовлень для програмованих мікрокалькуляторів спродукувало скорочення номенклатури мікросхем. Це було вирішено шляхом створення «спарки» з вищезазначеної пам’яті й виконавчого пристрою, який працював за командами, які у ній зберігались.

Такий підхід виявився проривом у створенні обчислювальних засобів: універсальний логічний пристрій у вигляді 4-розрядного центрального процесорного пристрою i4004, пізніше назвали *мікропроцесором* (див. Рис. 15). Останній був набором із 4 чіпів, серед яких був такий, що керувався командами, які зберігалися в напівпровідниковій внутрішній пам’яті.

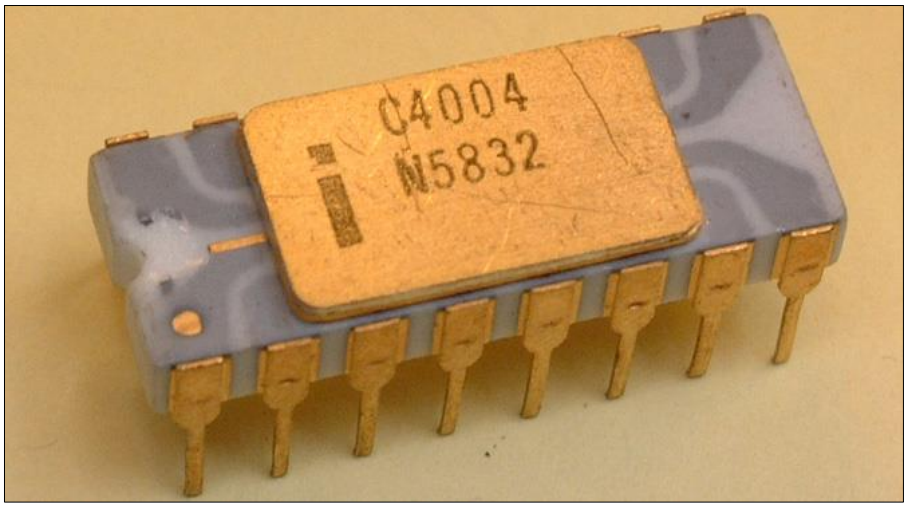


Рис. 15. i4004.

У виробництво такий пристрій був уведений 1971 року під назвою 4004, а вже у 1973 році у Франції з’явився перший комп’ютер на базі мікропроцесора 8008 – Micral. З різних причин цей процесор не мав успіху в Америці (у Радянському Союзі він був скопійований і випускався довгий час під назвою 580ВМ80). Тоді ж група інженерів, що пішла з Intel, організувала фірму Zilog.

Найбільш гучним її продуктом був Z80, що мав розширений набір команд 8080, що забезпечило його комерційний успіх для побутових приладів із напругою живлення 5Ват. На його основі був створений, зокрема, комп’ютер ZX-Spectrum (іноді його називають за ім’ям творця – Sinclair; див. Рис. 16), що став практично прообразом домашнього ПК середини 80-х. Прикметно, що

наявне на початку цього періоду програмне забезпечення (далі – ПЗ) було орієнтоване на обробку текстів і найпростіших електронних таблиць.



Рис. 16. ZX-Spectrum.

12 серпня 1981 року компанія IBM представила ПК, який став, у поєднанні з ПЗ від Microsoft, стандартом для всього парку пристроїв цього типу сучасного світу. Конфігурація першого ПК IBM: процесор Intel 8088 із частотою 4,77 Мгц і 29 тисячами транзисторів, 64 Кб оперативної пам'яті, 1 флоппі-дисковод ємністю 160 Кб, звук – найпростіший вбудований динамік.

У цей час запуск додатків і робота з ними були справжньою мукою: через відсутність жорсткого диска доводилося весь час змінювати гнучкі диски (дискети; див. Рис. 17), не було ні миші, ні графічного віконного користувальницького інтерфейсу, ні точної відповідності між зображенням на екрані й кінцевим результатом. Натомість кольорова графіка була вкрай примітивною, а тому про тривимірну анімацію не було й мови.



Рис. 17. Гнучкі диски (дискети).

У 1984 році відбувся випуск перших комп’ютерів Macintosh 128K (див. Рис. 18) із графічним інтерфейсом, маніпулятором миша і багатьма іншими атрибутами користувацького інтерфейсу, без яких не мисляться сучасні ПК. Сучасників новий інтерфейс не залишив байдужими, однак цей революційний комп’ютер не був сумісний ні з колишніми програмами, ні з апаратними елементами. Разом із тим, тогочасні користувачі вже звикли й пристосувалися до символьного інтерфейсу DOS, тому Macintosh виглядав для них дещо несерйозно й примітивно.

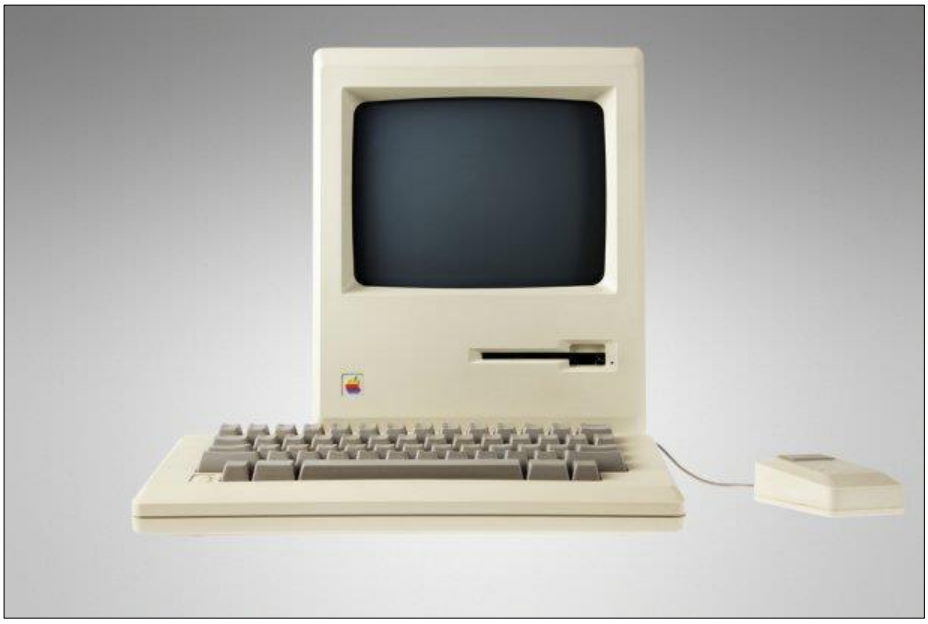


Рис. 18. Macintosh 128K.

7. ***Інтернет*** є найреволюційнішою технологією, що пов’язано з її впливом на комунікацію загалом (зокрема, побутування даних: їх створення, обробку, зберігання й передачу). Саме тому створення й розвиток цієї технології є ключовою подією на шляху до цифровізації (диджиталізації) суспільства, а вплив мережі охопив усі сфери життя: від науки та освіти до промисловості й рекреаційного комплексу.

Розвиток Інтернету розпочався в часи Холодної війни, коли Сполучені Штати Америки (далі – США) шукали способи забезпечення стабільного зв’язку між стратегічними об’єктами в умовах військової загрози. Так, у 1960-х роках за підтримки Агентства перспективних дослідницьких проєктів Міністерства

оборони США було створено ARPANET – першу комп’ютерну мережу, яка стала прототипом сучасного Інтернету (див. Рис. 19).

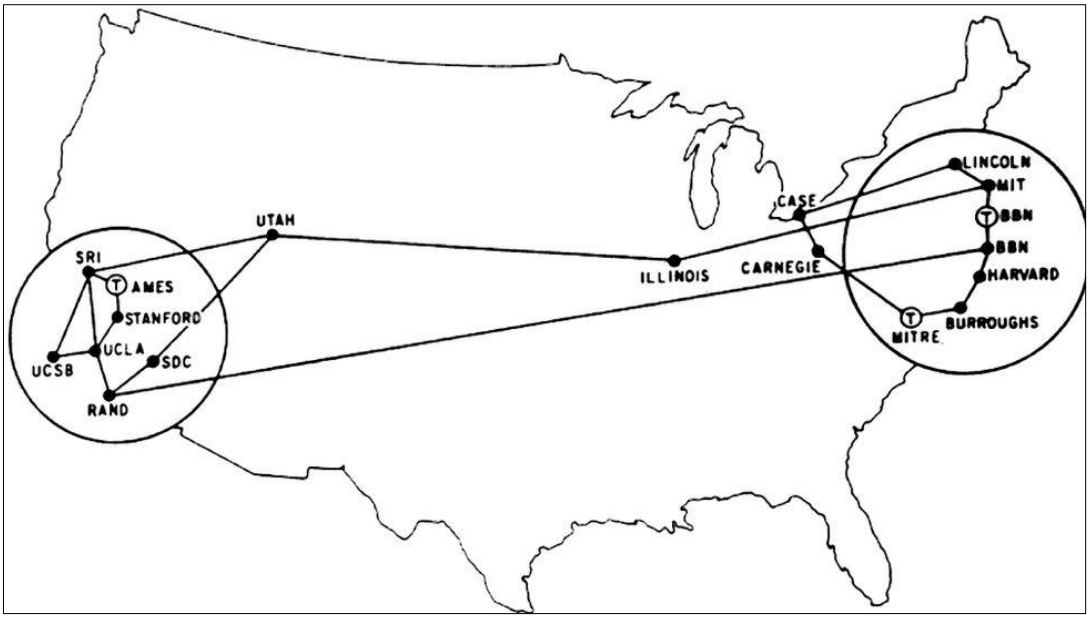


Рис. 19. ARPANET.

Важливі дати:

1969 рік – запуск ARPANET, коли чотири комп’ютери в різних університетах США були об’єднані в мережу.

1970-ті роки – впровадження протоколу TCP/IP, який дозволив стандартизувати передачу даних між різними комп’ютерними мережами.

1983 рік – перехід усіх мереж ARPANET на вищезазначений протокол, що заклало базу для створення єдиної глобальної мережі.

Справжню революцію в історії Інтернету здійснив британський учений Тім Бернерс-Лі, який у 1989 році запропонував концепцію Всесвітньої павутини (World Wide Web). Його винахід базувався на трьох ключових технологіях:

а) *URL* (*Uniform Resource Locator*): універсальній адресації ресурсів;

б) *HTTP* (*Hypertext Transfer Protocol*): протоколі передачі гіпертексту;

в) *HTML* (*Hypertext Markup Language*): мові розмітки для створення вебсторінок.

Прикметно, що перший вебсайт було створено в 1991 році, й вже до кінця десятиліття Інтернет перетворився на масовий інструмент для бізнесу, освіти та

розваг. Головними виявами впливу вищезазначеної технології на суспільство стали:

– *революція в комунікаціях:* Інтернет зробив спілкування миттєвим і глобальним (електронна пошта, форуми, соціальні мережі об'єднали мільйони людей у віртуальному просторі);

– *доступ до інформації:* завдяки Інтернету інформація стала доступною в будь-якому куточку світу (з'явилися пошукові системи, такі як Google, що зробили знання легкодоступними);

– *економічна трансформація:* розвиток електронної комерції (Amazon, Alibaba тощо) та фінансових технологій змінив моделі бізнесу (саме Інтернет став основою для таких інновацій, як блокчейн і криптовалюти);

– *вплив на освіту та науку*: з'явилися онлайн-курси, цифрові бібліотеки, репозитарії, дослідницькі платформи та інші, які зробили освіту інтерактивною й доступною для мільйонів людей.

На сьогодні, Інтернет залишається динамічним середовищем, яке продовжує розвиватися. Серед головних викликів – забезпечення приватності даних, боротьба з кібератаками та мінімізація цифрового розриву між країнами й соціальними групами. У майбутньому очікується подальше зростання впливу Інтернету завдяки таким технологіям, як Інтернет речей (IoT), 5G, ШІ й квантові обчислення.

Отже, поява Інтернету стала однією з найважливіших віх в історії ЦТ, а його розвиток (від першої передачі даних у мережі ARPANET до сучасної Всесвітньої павутини) є шляхом постійного прогресу, який змінив світ і відкрив нові горизонти для людства. Ця технологія продовжує бути рушієм інновацій, впливаючи на всі аспекти життя й забезпечуючи основу для глобальної цифрової епохи.

Прикметно, що ***сучасні технології*** позначені практично досягнутою межею можливостей підвищення швидкості обчислень за рахунок удосконалення елементної бази пристроїв. Відповідно, подолання нових викликів у XXI ст. корелює з випрацюванням інноваційних теоретичних підходів

до обчислень на новітніх елементах і принципах. Саме тому далі розглянемо нові перспективні технології, базовані на новітніх елементах.

8. ***Нейротехнології та ШІ*** є одними з найважливіших досягнень у розвитку ЦТ, які належать до **п'ятого покоління** ЕОМ Вони базуються на імітації роботи людського мозку й мають значний вплив на науку, промисловість та повсякденне життя. Першою концепцією штучних нейронів була запропонована в 1943 році Ворреном МакКаллоком і Волтером Піттсом *математична модель нейрона*, яка стала основою для розвитку вищезазначених ШНМ.

Наступним етапом розвитку цієї технології став створений у 1958 році Френком Розенблаттом *перцептрон* – перша нейронна мережа, здатна до навчання, яка ознаменувала прорив у дослідженнях машинного навчання (далі – МН). У 1980-х роках розвиток *алгоритмів зворотного поширення помилки* дозволив значно підвищити ефективність багатошарових нейронних мереж.

Узвичаєно розвиток ШІ поділяють на такі етапи:

1. *Ранній ШІ (1950-ті – 1970-ті)* – етап, пов'язаний із розробкою вищезазначеним Аланом Тьюрингом тесту, метою якого була оцінка здатності машини імітувати людське мислення. Окрім того, до цього періоду відносять появу перших програм для гри в шахи та рішення задач логіки.

2. *«Зима ШІ» (1970-ті – 2000-ні)* – період, коли ентузіазм щодо цієї технології був знижений через наявні технічні обмеження часу.

3. *Відродження ШІ (2000-ті)* пов'язане з появою великих даних і зростанням обчислювальної потужності, що призвело до нового розквіту ШІ.

За своєю природою *ШІ* – це комплекс технологічних рішень, що дозволяє імітувати когнітивні функції людини (в тому числі самонавчання й пошук рішень без завчасно заготовленого алгоритму) й отримувати під час виконання конкретних завдань результати, які можуть бути порівняні з результатом інтелектуальної діяльності людини.

ШІ можна використовувати в процесі реалізації нових можливостей для:

а) *обробки даних* (генерації, унаочнення, узагальнення тощо);

б) *підтримки в ухваленні рішень* (систематизації даних та іншого, здійснюваного персональним асистентом);

в) *автоматизації поточних процесів* (написанні листів, відповідей у випадку, коли ви не можете відповісти, та іншого).

г) *покращення комунікацій* (удосконалення побутових процесів, захисту даних в процесі спілкування і так далі).

9. Сутнісно ***нейрокомп'ютери*** належать до **шостого покоління** ЕОМ, однак, незважаючи на те, що реальне застосування нейромереж почалося відносно недавно, нейрокомп'ютингу як науковому напрямку пішов сьомий десяток років. Так, перший нейрокомп'ютер був побудований у 1958 році, його розробником був згаданий Френк Розенблатт, який назвав машину Mark I.

Нейрокомп'ютери – це ПК, які складаються з безлічі простих обчислювальних елементів, що працюють паралельно і називаються штучними нейронами. Такі нейрони утворюють так звані нейромережі, висока швидкодія яких досягається саме за рахунок величезної кількості окреслених елементів у їх структурі.

Нейрокомп'ютери побудовані за біологічним принципом: нервова система людини складається з окремих клітин – нейронів, кількість яких у мозку досягає 1012, при тому, що час спрацювання нейрона – 3 мс. Кожен нейрон виконує досить прості функції, при цьому він пов'язаний у середньому з 1–10 тис. інших, які успішно забезпечують роботу людського мозку.

Вищезазначена *теорія нейромереж* уперше була означена в праці В. Маккаллок і В. Піттса (1943 р.): за нею будь-яку арифметичну або логічну функцію можна реалізувати за допомогою простої нейронної мережі. Інтерес до нейрокомп'ютингу знову спалахнув на початку 80-х років і був підігрітий новими працями з багатошаровим перцептроном (про який будемо говорити на інших лекціях) і паралельними обчисленнями.

Вищезазначено, що нейрокомп'ютери використовують принципи роботи біологічних нейронних мереж, реалізовані ними на апаратному чи програмному рівнях:

а) *ШНМ* – моделюють зв’язки між нейронами для вирішення завдань класифікації, прогнозування тощо;

б) *апаратні нейроморфні процесори* (до прикладу, чіпи IBM TrueNorth і Intel Loihi тощо) імітують біологічну структуру мозку.

Отже, нейрокомп’ютери та штучний інтелект відіграють центральну роль у розвитку ЦТ, створюючи нові можливості для науки, бізнесу й суспільства. Їхній вплив на життя зростає, формуючи майбутнє, яке поєднує передові технології з людськими потребами та викликами.

Квантові технології базовані на використанні для зберігання, передачі й обробки інформації істотних квантових властивостей речовини. Історія ЦТ – це послідовність революційних змін, що трансформували суспільство, науку й економіку. Від винаходу перших обчислювальних машин до сучасного ШІ, кожна епоха приносила нові можливості. Одним із найперспективніших напрямів у цій еволюції є квантові технології, базовані на концепції, яка переносить нас у світ, де закони квантової механіки відкривають нові горизонти для обчислень, комунікацій і захисту даних.

Квантова механіка, що виникла на початку XX століття, стала основою сучасного розуміння природи матерії й енергії. Зароджена роботами Вернера Гейзенберга, Альберта Ейнштейна, Макса Планка та інших, ця наука дозволила пояснити явища, які неможливо було зрозуміти за допомогою класичної фізики. Хоча на початкових етапах квантова механіка здавалася абстрактною теорією, у другій половині століття вона дала початок технологіям, що змінили світ: напівпровідниковим пристроям, транзисторам, лазерам тощо.

Розвиток ЦТ протягом XX століття значною мірою був зумовлений досягненнями квантової фізики: так, вищезазначений винахід транзистора став ключовим кроком до мініатюризації електроніки, що згодом привело до появи мікропроцесорів. Однак у 1980-х роках вчені почали замислюватися над тим, чи можливо використовувати квантові ефекти для створення принципово нових обчислювальних систем.

Окреслені ідеї заклали базу для розвитку квантової інформатики: Річард Фейнман (1982 р.) вперше запропонував ідею *квантового комп’ютера* (див.

Рис. 20), який використовував би *суперпозицію* й *заплутаність* – унікальні властивості квантових систем – для виконання обчислень, неможливих на класичних комп’ютерах. На відміну від традиційних бітів, які можуть приймати лише значення 0 або 1, квантові біти (кубіти) здатні перебувати в обох станах одночасно, що дозволяє квантовим комп’ютерам паралельно обробляти величезні обсяги даних.



Рис. 20. Квантовий комп’ютер D-Wave D-Wave 2000Q
з 2000-кубітним процесором.

Перші експериментальні квантові комп’ютери з’явилися наприкінці 1990-х років, однак лише в останні десятиліття технологія почала наближатися до комерційного застосування. До прикладу, компанії Google і IBM продемонстрували прототипи квантових комп’ютерів, що виконують обчислення, недосяжні для сучасних суперкомп’ютерів. Окрім квантових обчислень, ця технологія включає *квантову криптографію* та *квантові комунікації*.

Квантова криптографія, заснована на принципах квантової заплутаності, гарантує абсолютну безпеку передачі даних. Знаковим досягненням у цьому напрямі стала квантова мережа Китаю, яка дозволяє здійснювати захищений

обмін інформацією на великих відстанях. *Квантові сенсори* – ще один напрям, який відкриває нові можливості для медицини, навігації та геофізики. Їхня надзвичайна чутливість дозволяє виявляти найдрібніші зміни у фізичних полях (зокрема, магнітних і гравітаційних).

Хоча квантові технології демонструють величезний потенціал, вони стикаються з низкою викликів. Так, створення стабільних кубітів, боротьба з квантовим шумом і масштабування систем залишаються відкритими проблемами. Разом із тим, у майбутньому квантові технології можуть змінити не лише спосіб, у який ми обчислюємо, а й підхід до науки загалом. Зокрема, вони здатні пришвидшити відкриття нових матеріалів, оптимізувати глобальні логістичні мережі та навіть допомогти розгадати фундаментальні таємниці Всесвіту. Це дозволяє стверджувати, що такі технології посідають особливе місце в історії розвитку ЦТ: вони є не просто черговим етапом їх еволюції, а справжнім проривом, що відкриває нову еру в обчисленнях і комунікаціях.

Натомість в ***оптоелектронних (фотонних) комп'ютерах*** носієм інформації є світловий потік: електричні сигнали перетворюються в оптичні й назад. До прикладу, масштабований і мережевий фотонний квантовий комп'ютер Aurora, який поєднав обидві інноваційні технології. Однак повернемося до оптоелектронних комп'ютерів, у яких використовують оптичне випромінювання в якості носія інформації. Останнє має низку потенційних переваг у порівнянні з електричними сигналами:

а) світлові потоки, на відміну від електричних, можуть перетинатися одне з одним;

б) такі потоки можуть бути локалізовані в поперечному напрямку до нанометрових розмірів і передаватися по вільному простору;

в) швидкість поширення світлового сигналу вище швидкості електричного;

г) взаємодія світлових потоків із нелінійними середовищами розподілена по всьому середовищу (робочій площині), що дає нові ступені свободи в організації зв'язку й створенні паралельних архітектур.

На сьогодні, ведуться розробки зі створення комп’ютерів, повністю базованих на вищезазначеній технології (оптичних пристроїв обробки інформації). Передбачається, що така машина матиме небачену продуктивність і зовсім іншу, ніж електронний комп’ютер, архітектуру: за 1 такт тривалістю менше 1 наносекунди (це відповідає тактовій частоті більше 1000 МГц) у ній буде можлива обробка масиву даних близько 1 мегабайта й більше.

Такий комп’ютер, розміром із ноутбук, може дати користувачеві можливість розмістити в ньому чи не всю інформацію про світ, при цьому він зможе вирішувати завдання будь-якої складності. Прикметно, що вже створено й оптимізовано окремі складники вищезазначених машин, що дозволяє висловити припущення про їх появу в найближчому майбутньому.

Біологічні комп’ютери – новий і перспективний напрям у розвитку ЦТ, у якому інтегровано принципи інформатики, біології та нанотехнологій. До прикладу, CL1 – інноваційний біологічний комп’ютер, у якому інтегровано живі культивовані клітини людського мозку з кремнієвим обладнанням у корпусі. Останній (корпус), до всього, виконує функцію життєзабезпечення вищезазначених клітин (див. Рис. 21).



Рис. 21. CL1.

Такі комп’ютери є важливою частиною історії цифрових інновацій, пропонуючи рішення для задач, які виходять за межі можливостей традиційних кремнієвих процесорів. Сама ідея використання біологічних систем для обчислень з’явилася ще в середині XX століття.

Основні віхи розвитку включають:

– *ДНК як носій інформації:* у 1953 році Джеймс Уотсон і Френсіс Крік відкрили структуру ДНК, що стало базою для ідей використання біологічних молекул у зберіганні даних;

– *біомолекулярні обчислення:* у 1994 році Леонард Адлеман продемонстрував, як можна використовувати молекули ДНК для вирішення задач комбінаторної оптимізації.

Передбачають, що біологічні комп'ютери будуть функціонувати на основі біомолекул (ДНК, РНК або білків) й матимуть унікальні властивості:

а) *паралелізацію обчислень:* завдяки природі біохімічних реакцій, такі комп'ютери зможуть виконувати мільйони операцій одночасно;

б) *енергоефективність:* біологічні системи споживають значно менше енергії порівняно з традиційними електронними пристроями;

в) *мініатюризація:* розміри біологічних компонентів дозволяють створювати обчислювальні системи на нанорівні.

Головною проблемою таких машин очікувано може стати складність контролю, що виявляється у високому рівні *контролю за біохімічними реакціями*, знання про які інтегруватимуть із молекулярної біології. Також згадаємо *обмежену універсальність*, оскільки природа машин не дозволить їм обчислювати різнопланові задачі, натомість концентруючись лише навколо вузькоспеціальних. Разом із тим, біологічні комп'ютери представляють нову еру в історії ЦТ. Хоча їх розвиток перебуває на ранніх стадіях, вже сьогодні вони демонструють величезний потенціал для розв'язання складних задач, змінюючи наше уявлення про можливості обчислювальних систем.

Пов'язаною з вищезазначеною технологією є ідея ***молекулярних комп'ютерів*** (див. Рис. 22), принцип дії яких заснований на використанні низки властивостей молекул в процесі фотосинтезу. До прикладу, базою для цих пристроїв може стати нещодавня розробка вченими з Індійського інституту аналогової обчислювальної платформи. Прикметно, що остання здатна зберігати й обробляти дані в молекулярній плівці, яка підтримує 16500 станів провідності, відкриваючи нові можливості для енергоефективних обчислень.

Технологія молекулярних комп’ютерів базуватиметься на станах молекули під час фотосинтезу: вчені присвоять певні логічні значення кожному такому стану, тобто 0 або 1. Так, використовуючи певні молекули, вчені визначили, що їх фотоцикл складається всього з двох станів, «перемикати» які можна змінюючи кислотно-лужний баланс середовища. Останнє дуже легко зробити за допомогою електричного сигналу: адже сучасні технології вже дозволяють створювати цілі ланцюжки молекул, організовані подібним чином.

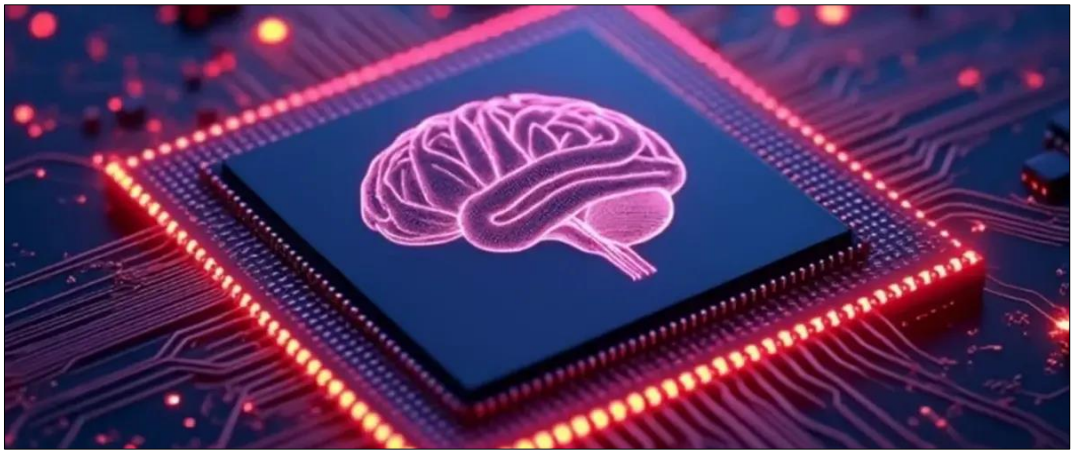


Рис. 22. Молекулярний комп’ютер.

Системи розподіленого реєстру **(*Distributed Ledger Technology***; далі – СРР) – технологія розподіленого реєстру. Для унаочнення такої системи уявімо Google-документ, у якому одночасно сидить певна кількість працівників якоїсь установи (відділу, департаменту тощо). Коли один співробітник вносить певні зміни, то це оновлення файлу миттєво відображається на екранах пристроїв усіх користувачів, які знаходяться онлайн у ньому. Це гарантує, що кожен працівник має доступ до найновішої та найповнішої інформації щодо роботи команди, що сприяє кращій взаємодії та координованості подальших дій і планів.

Технологія розподіленого реєстру втілює цей підхід одночасного ведення записів у великій мережі. Сутнісно вона є цифровим реєстром, який не зберігається в одній центральній точці, а відтворюється в кількох вузлах або місцях, які беруть участь у мережі. Відтак, немає єдиного зберігача, натомість кожен учасник має доступ до всіх проведених змін і може відстежити їх.

Специфічною для цієї технології є її *децентралізована природа*, завдяки якій відбувається розподіл відповідальності між користувачами певної мережі.

Таким чином, відсутня рольова диференціація ролей (юзер, адмін, редактор та інші), а значить – ніхто не має ексклюзивного доступу. У такій технології кожен вузол у мережі зберігає копію реєстру, що усуває потребу в центральній авторитетній версії.

СРР, а за розширенням блокчейн, який є різновидом технології розподіленого реєстру, працює в межах цієї децентралізованої архітектури. Перевагою такої технології є те, що перевірка та запис дій (тут – здебільшого, транзакцій) виконується незалежно кожним вузлом. Ця архітектура змінює підхід до збору та розповсюдження даних: так, щоразу, коли ініціюється певна дія, вона транслюється по всій мережі.

Після цього вузли, дотримуючись алгоритму консенсусу, працюють над перевіркою та синхронізацією нової дії зі своїми індивідуальними реєстрами. Криптографічні підписи слугують еквівалентом перевірених облікових даних у записах, гарантуючи безпеку та незмінність останніх. Технологія розподіленого реєстру разом зі своєю найвідомішою формою, – блокчейном, – пропонує надійну, незмінну та прозору структуру для реєстрації даних.

СРР є потужним каталізатором інновацій, трансформуючи галузі шляхом децентралізації контролю, посилення безпеки та підвищення прозорості. Цей прогрес виходить за межі традиційного ведення записів, тобто реєстрів, представляючи зміну парадигми у виконанні таких узвичаєних дій як: транзакції, автентифікація особи та управління цифровою інформацією в усьому світі.

Компоненти робототехніки й сенсорика. Історія розвитку ЦТ тісно переплітається з еволюцією робототехніки та сенсорних систем. Ці напрямки відіграють ключову роль у створенні машин, здатних виконувати завдання, що раніше були доступні лише людині. Розуміння компонентів робототехніки та сенсорики дозволяє оцінити, як вони вплинули на розвиток технологій і суспільства загалом. Робототехніка базується на інтеграції декількох ключових елементів, що забезпечують її функціональність. Серед них:

– *механічна структура:* роботи мають каркас, виготовлений із матеріалів, які забезпечують легкість, міцність і гнучкість (це основа, що дозволяє виконувати фізичні дії);

– *актори (приводи):* пристрої, які перетворюють енергію на рух, дозволяючи роботам взаємодіяти з навколишнім середовищем (до прикладу, електродвигуни, гідравлічні й пневматичні приводи);

– *контролери* є основною частиною, яка забезпечує обчислення й управління роботами, саме завдяки їм машини здатні виконувати складні операції за заданими алгоритмами;

– *ПЗ* синхронізує всі компоненти, оптимізує їх роботу та адаптує до змінюваних умов.

Сенсорика є невід'ємною частиною сучасної робототехніки, що пов'язано з її ключовою роллю в сприйнятті роботами (мовиться про їх «зір», «слух» тощо). Відповідно, сенсори доцільно розділяти за їх функціональним призначенням:

а) *візуальні* (камери й системи комп'ютерного зору) забезпечують розпізнавання об'єктів, орієнтацію в просторі та відстеження рухів;

б) *тактильні сенсори* дозволяють оцінювати силу натискання та текстуру поверхні та інше;

в) *аудіосенсори* (мікрофони й системи обробки звуку) формують сприйняття й розпізнавання мовлення або звукових сигналів;

г) *сенсори положення й руху* (гіроскопи, акселерометри та інші пристрої) локалізують позицію машини в просторі тощо.

Отже, компоненти робототехніки й сенсорика є фундаментальними елементами в історії розвитку ЦТ. Вони не лише забезпечують новий рівень функціональності машин, а й створюють основу для подальших проривів у науці й техніці. Розвиток цих напрямків формує майбутнє, де робототехніка стане невід'ємною частиною повсякденного життя людства, що безпосередньо пов'язане з *технологією віртуальної й доповненої реальності*.

Так, історія розвитку ЦТ відображає безперервний прогрес у створенні інтерактивних середовищ, які забезпечують нові способи взаємодії з інформацією. ***Технології віртуальної*** (далі – ВР) ***та доповненої реальності*** (далі – ДР) стали невід'ємною частиною цієї еволюції, змінюючи наше уявлення про можливості цифрового світу.

Перші концепції ***ВР*** виникли ще в середині XX століття:

– *стереоскопи* (див. Рис. 23)*:* пристрої для перегляду зображень із тривимірним ефектом, створені ще в XIX столітті, заклали фундамент для майбутніх VR-технологій;



Рис. 23. Лінзовий стереоскоп.

– *Sensorama (1962 р.)* (див. Рис. 24)*:* Мортон Гейліг створив пристрій, який дозволяв користувачам зануритися у віртуальне середовище за допомогою зору, звуку, запаху та тактильних відчуттів;



Рис. 24. Sensorama.

– *The Sword of Damocles' (1968 р.)* (див. Рис. 25): Іван Сазерленд розробив перший пристрій для занурення у віртуальну реальність, який був підключений до комп'ютера й дозволяв переглядати 3D-зображення.



Рис. 25. The Sword of Damocles.'

Натомість розвиток ***ДР*** поєднав у своїх межах цифрові елементи з реальним світом: зокрема, її генеза пов'язана з репрезентацією реальності. Так, у 1990 році Том Коделл уперше використав цей термін, описуючи технології, які забезпечують накладання інформації на реальні об'єкти. Більше того, друге дихання ця технологія отримала з розвитком смартфонів (згадаймо, застосунок Pokémon GO та інші).

Ключові технології й пристрої:

а) шоломи та окуляри ВР: сучасні пристрої (до прикладу, Oculus Rift, HTC Vive, PlayStation VR тощо), забезпечують повне занурення у віртуальні світи;

б) окуляри ДР: до прикладу, Microsoft HoloLens і Magic Leap дозволяють накладати цифрові зображення на реальне середовище.

в) сенсори та контролери: пристрої відстеження рухів і жестів, такі як Leap Motion, забезпечують інтуїтивну взаємодію з віртуальними об'єктами.

Технології бездротового зв'язку. Історія ЦТ нерозривно пов'язана з еволюцією технологій бездротового зв'язку. Від перших радіопередач до сучасних 5G-мереж, бездротові комунікації відіграють ключову роль у трансформації способів взаємодії людей і машин, забезпечуючи швидкий обмін інформацією на будь-якій відстані.

Перші спроби передачі інформації без використання дротів були виявлені до кінця XIX століття. Знаковими віхами стали:

– *відкриття радіохвиль:* у 1888 році Генріх Герц експериментально підтвердив існування електромагнітних хвиль, передбачених теоретично Джеймсом Клерком Максвеллом;

– *перша радіопередача:* у 1895 році Гульєльмо Марконі здійснив передачу радіосигналу на відстань кількох кілометрів, поклавши початок епосі радіозв'язку;

– *телеграф без дротів:* до початку XX століття радіотехнології почали активно застосовуватися для морських комунікацій, що значно підвищило безпеку судноплавства;

– *розвиток радіо і телебачення:* 1920-ті роки стали періодом розквіту радіо як засобу масової інформації (згодом, у 1930-х роках, з'явилося телебачення, яке також використовувало бездротові технології для трансляції аудіо- та відеосигналів).

Вищезазначені досягнення не лише змінили способи передачі інформації, а й заклали фундамент для майбутніх цифрових систем зв'язку. Іншою значущою подією стала ***поява мобільного зв'язку*** – це був справжній прорив у бездротових технологіях, основні етапи якого узвичаєно поділяють на такі:

Перше покоління (1G): у 1980-х роках аналоговий мобільний зв'язок забезпечував голосові виклики, хоча й із низькою якістю звуку та обмеженим покриттям.

Друге покоління (2G): у 1990-х роках цифровий мобільний зв'язок дозволив передавати текстові повідомлення (SMS) і покращив якість голосових дзвінків.

Третє покоління (3G): на початку 2000-х років з'явилися можливості для передачі даних, що проклало шлях мобільному Інтернету.

Четверте покоління (4G): з 2010-х років технологія LTE забезпечила високу швидкість передачі даних, що зробило можливими потокове відео, онлайн-ігри та інші ресурсоємні застосунки.

П'яте покоління (5G): сучасні мережі 5G забезпечують не лише високу швидкість, а й мінімальні затримки, що є критично важливим для згаданого Інтернету речей, автономного транспорту та інших передових технологій.

Wi-Fi та Bluetooth*: революція локального зв'язку.* На додаток до мобільного зв'язку, технології Wi-Fi та Bluetooth забезпечили локальну бездротову комунікацію. Зокрема, вперше з'явившись у 1997 році, Wi-Fi став стандартом для підключення до Інтернету в домашніх і офісних мережах, забезпечуючи зручність і високу швидкість. Натомість Bluetooth був запущений у 1998 році, зробивши можливою передачу даних між пристроями на невеликих відстанях і ставши основою для бездротових гарнітур, аксесуарів тощо.

Супутниковий зв'язок. Супутникові системи зв'язку, розроблені в середині XX століття, забезпечили глобальне покриття для передачі сигналів. Вони стали важливими для телекомунікацій, навігації (GPS) та моніторингу Землі.

Таким чином, бездротові технології радикально змінили підхід до передачі даних, забезпечивши:

а) *глобальне охоплення:* Інтернет і мобільні комунікації стали доступними в найвіддаленіших куточках світу;

б) *мобільність:* завдяки смартфонам і ноутбукам люди отримали можливість працювати й спілкуватися будь-де;

в) *нові послуги:* потокове відео, хмарні обчислення та онлайн-сервіси стали реальністю завдяки бездротовому зв'язку.

Саме тому технології бездротового зв'язку стали невід'ємною частиною історії ЦТ, відкривши нові можливості для комунікації, бізнесу та повсякденного життя. Завдяки постійному вдосконаленню, бездротовий зв'язок продовжує формувати майбутнє, роблячи світ більш взаємопов'язаним і доступним.

1.3. Вплив цифрових технологій на суспільство

Природно, що швидкий розвиток ЦТ спричинив значущі зміни в суспільстві, які торкнулися всіх сфер життя. Так, у соціальному вимірі їх розвиток спродукував:

а) *культурний прорив:* мовиться про вільний доступ до культурних надбань світу, який набув універсальної і безперешкодної природи;

б) *підвищення доступності освіти:* так, дистанційна освіта забезпечила неперервність навчання навіть у часи пандемії COVID-19;

в) *посилення поляризаційних тенденцій у суспільстві та інтенсифікацію соціальної нерівності:* вони базовані на нерівному доступі до технологій в окремих країнах (до прикладу, Індії), що є виявом дискримінації певних соціальних груп чи мешканців тих чи тих регіонів країни.

Відзначимо, що, на сьогодні, немає такої сфери життєдіяльності, якої б не торкнулась цифровізація (диджиталізація): так, завдяки інтенсивній імплементації ЦТ ми спілкуємось на великих відстанях у режимі реального часу, отримуємо доступ до величезних обсягів інформації будь-якої природи (від рецептів випічки до вирощування цуценят).

Окрім того, використання ЦТ дозволяє автоматизувати побут через використання персональних помічників, роботу в Customer Relationship Management (CRM-)системах, які дозволяють покращити командну роботу й мають широкий функціонал. Також завдяки технологічному розвитку з’являються нові засоби комунікації, які, своєю чергою, дозволяють створювати нові продукти та послуги. Так, генерація тексту ШІ дозволяє створювати цілі книжки, які недобросовісні користувачі можуть монетизувати тощо.

Запитання для самоконтролю:

1. Що таке ЦТ? Чим вони відрізняються від аналогових?

2. Назвіть і поясніть чотири основні риси ЦТ.

3. У чому полягає принцип функціонування ручних обчислювальних пристроїв? Наведіть приклади.

4. Які ключові винаходи були здійснені на етапі механічних обчислювальних пристроїв і хто їх автори?

5. Чим відрізняються ЕОМ на електронних лампах від транзисторних комп’ютерів?

6. Які переваги забезпечили широке поширення ЕОМ на інтегральних схемах у третьому поколінні?

7. Яку роль у становленні персональних комп’ютерів відіграли компанії Intel, IBM та Microsoft?

8. Якими були основні етапи розвитку Інтернету? Як він вплинув на суспільство?

9. У чому полягає сутність роботи нейрокомп’ютерів та їх відмінність від традиційних ЕОМ?

10. Які перспективи відкривають квантові та оптоелектронні технології у розвитку цифрових технологій майбутнього?

Тема 2.

ТЕКСТОВІ ДАНІ ТА РОБОТА З НИМИ В КОНТЕКСТІ ЦИФРОВИХ ТЕХНОЛОГІЙ

2.1. Цифрові технології для роботи з корпусами текстів

Корпус текстів – структурована електронна збірка текстів, яка створюється для аналізу природної мови (розглянемо його детальніше у наступній лекції). Текстові корпуси можуть містити писемні або усні (транскрибовані) тексти , що відібрані за певними критеріями (жанром, стилем, часовим періодом, соціокультурним контекстом тощо), їх використовують в корпусній лінгвістиці для практичного дослідження мови.

Основними характеристиками є:

а) репрезентативність – корпуси повинні охоплювати широкий спектр мовних варіацій, щоб забезпечити об’єктивність аналізу;

б) анотованість – багато корпусів супроводжує лінгвістична розмітка (морфологічна, синтаксична, семантична), що полегшує їх автоматичний аналіз;

в) динамічність – деякі корпуси постійно оновлюються, особливо якщо вони відображають сучасні мовні зміни (до прикладу, корпуси соціальних мереж або новинні корпуси);

г) доступність – відкриті корпуси використовують для досліджень у лінгвістиці, ШІ, машинному перекладі та інших сферах.

Типи корпусів:

– загальні корпуси: містять тексти різних жанрів (до прикладу: British National Corpus, Національний корпус української мови тощо);

– спеціалізовані корпуси: зосереджені на певній сфері (до прикладу: медичний, юридичний та інший корпус);

– паралельні корпуси: містять тексти та їхні переклади, що використовують в машинному перекладі (до прикладу: OPUS, Europarl і тому подібні);

– частотні корпуси: створені для вивчення частотності вживання слів та фраз;

– синхронні та діахронні корпуси: перші відображають сучасний стан мови, а другі – її еволюцію (до прикладу: Corpus of Historical American English і так далі).

Роль корпусів у лінгвістиці:

а) аналіз граматичних та лексичних закономірностей – вивчення синтаксичних структур, семантичних зв'язків і варіативності слів;

б) формування лексикографічних ресурсів – створення словників, тезаурусів та граматик;

в) дослідження мовних змін – аналіз еволюції мовних конструкцій та неологізмів;

г) автоматична обробка природної мови (Natural Language Processing; далі – ОПМ) – навчання моделей для машинного перекладу, розпізнавання тексту, класифікації документів тощо;

ґ) соціолінгвістичні дослідження – аналіз варіативності мови залежно від регіону, соціального статусу, вікової групи.

Отже, корпуси текстів є важливим ресурсом для сучасної лінгвістики та автоматичної обробки мовних даних. Вони забезпечують емпіричну базу для досліджень, дозволяють виявляти закономірності у мовленні, а також сприяють розвитку технологій ШІ.

Головними цифровими інструментами для роботи з корпусами є:

– конкордансери та пошукові системи (до прикладу, AntConc та інші);

– аналіз частотності слів і колокацій (до прикладу, WordSmith Tools і так далі);

– автоматичний морфологічний аналіз (до прикладу, TreeTagger і тому подібні);

– аналіз синтаксису та семантики (до прикладу, Stanford NLP тощо);

ґ) великі корпуси та хмарні технології (до прикладу, Google Books Ngram та інші).

2.2. Лінгвістична експертиза тексту

Лінгвістична експертиза тексту – це комплексний аналіз текстів із застосуванням наукових методів лінгвістики для встановлення їхніх змістових, структурних, стилістичних та прагматичних характеристик. Вона використовується у перекладознавстві, правозастосуванні, медіааналізі, рекламі та наукових дослідженнях. Застосування такої експертизи пов’язане з виявленням авторства, перевіркою відповідності нормам мови, аналізу стилю, локалізацією прихованих смислів чи маніпулятивних прийомів (до прикладу, мови ворожнечі тощо).

Сучасні програмні продукти значно підвищують ефективність і точність вищезазначеної експертизи: зокрема, *автоматизовані системи аналізу тексту* здатні:

– швидко обробляти великі масиви даних;

– здійснювати частиномовну розмітку;

– визначати частотність та сполучуваність лексем;

– виявляти стилістичні та граматичні особливості;

– інтегрувати статистичний і семантичний аналіз.

До прикладу, вебзастосунок TextAttributor був створений у межах проєкту «Визначення авторства анонімних українськомовних текстів з використанням методів штучного інтелекту в мережі Інтернет» і представлений у 2024 році (див. Рис. 26).

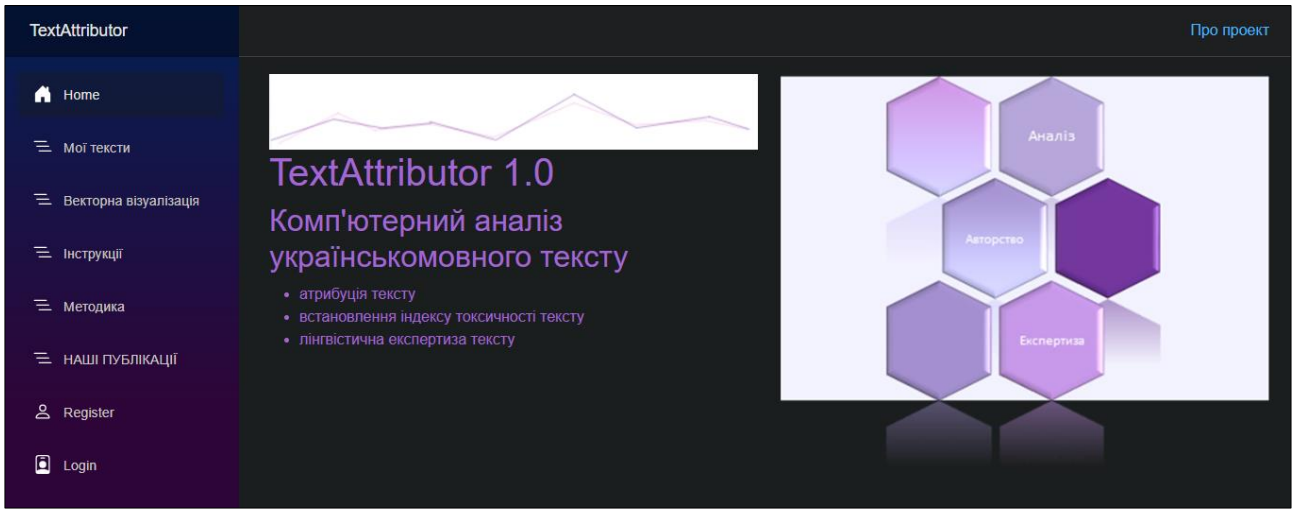


Рис. 26. TextAttributor.

Система є інтерактивним аналітичним інструментом для лінгвостатистичного аналізу, призначеним для атрибуції (визначення достовірності) українськомовних медійних текстів у завданнях стилеметрії (локалізації стилю тексту) та встановлення авторства. Вона також використовується для формування висновків згаданої лінгвістичної експертизи на основі статистичних параметрів атрибуції та оцінки негативної тональності (токсичності) тексту.

У системі автоматичного лінгвостатистичного аналізу TextAttributor застосовано методи компонентного та дистрибутивного аналізу, аналізу безпосередніх складників, квантування, аналізу настроїв, комп'ютерного моделювання та глибокого МН, про які ми поговоримо на наступних лекціях.

Таким чином, вебзастосунок TextAttributor демонструє результати роботи системи у двох напрямах:

а) визначення авторства тексту;

б) виявлення токсичності українськомовного медійного контенту.

Обидва завдання реалізовано з використанням двох методичних підходів комп'ютерної лінгвістики:

– аналіз за словником і правилами;

– методи глибокого МН.

Робота системи у режимі стилеметричного порівняння обмежена медійними текстами (публіцистичним та іншим стилем). Водночас лінгвостатистичний аналіз без визначення індивідуальних авторських ознак можна здійснювати для українськомовних текстів будь-якого стилю.

2.3. Поняття плагіату та його виявлення

Плагіат (буквально з грецької означає «крадіжка») – дія, мета якої – присвоєння чужої інтелектуальності власності (у нашому випадку – тексту) з метою отримання певної вигоди. Простими словами, це використання «чужого» під виглядом «свого», без зазначенням авторства (див. Рис. 27). Назвемо такі *види плагіату*:

а) дослівний – вставлення чужого тексту;

б) мозаїчний – компіляція текстів кількох авторів;

в) ідейний – крадіжка ідеї та її інтерпретування;

г) перекладений – переклад чужого тексту задля підвищення унікальності кінцевого варіанту;

ґ) автоплагіат – запозичення у самих себе, без вказівки на те, що цей текст вже був опублікований.

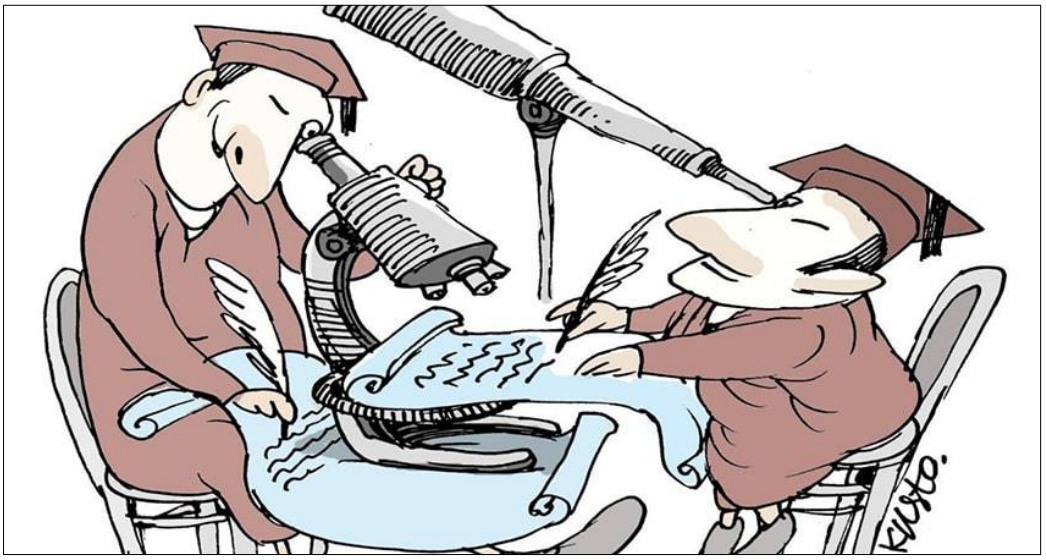


Рис. 27. Плагіат.

У контексті лінгвістики та ЦТ плагіат нас може цікавити безпосередньо щодо роботи програм з його виявлення (AntiPlagiarism, Cognitive Text Analyzer, StrikePlagiarism, Plag, Unicheck та інших) та, власне, лінгвістичного аналізу таких текстів. Окрім того, показовим для нас є і те, і те: з одного боку, цікавим є вивчення тексту та можливостей його привласнення й виявлення цього факту вручну, з іншого – машинне (алгоритмічне) встановлення цього факту.

Головними методами виявлення плагіату є:

а) аналіз лексики – безпосередній пошук точних збігів за рахунок використання пошуковиків (до прикладу, Google);

б) вивчення семантики – виявлення несанкціонованих запозичень, навіть за умови, якщо їх перефразували, переклали тощо;

в) корпусний аналіз – порівняння тексту з великими базами даних та пошук збігів у них;

г) стилістичний – може інтегрувати всі попередні, оскільки у нього входить лексика, довжина речень, обсяг абзаців, логіка думок тощо.

Безпосередньо сама перевірка тексту відповідним ПЗ відбувається так: документ репрезентують у текстовому редакторі й паралельно відбувається сам процес аналізу. Така перевірка виконується по *шинглах* (шингл – розбивка цілого тексту на фрагменти), які є структурно-логічним складником аналізованого тексту й здебільшого складається з кількох слів.

У попередній лекції зазначено, що аналіз в Інтернеті за таких умов відбувається на базі пошуковиків (кількість та деталізацію, зазвичай, користувач налаштовує самостійно (якщо мовиться про десктопну версію програми)). Результатом такої перевірки є список ресурсів із візуалізацією ступеня збігу з кожним із них, який може бути маркований певним кольором.

Базові характеристики можливостей антиплагіатних систем включають:

– можливість мережевого пошуку;

– пошук у локальних базах даних;

– безкоштовний (обмежений) та платний (повний) доступ;

– підтримка слов'янських мов;

– підтримка різних форматів документів;

– функція генерації звіту.

Розглянемо роботу такого ПЗ на прикладі Cognitive Text Analyzer, який вивчає синтаксис, семантику, структуру (розбір і аналіз тексту) тощо. Це дозволяє програмі виявити близько 100 питомих властивостей аналізованого тексту. Швидкість її роботи на «середньому» комп'ютері – близько 500 сторінок у хвилину.

Головною особливістю цього ПЗ є потреба в «ідеальному» з'єднанні, а також менша продуктивність роботи зі слов'янськими мовами, тобто вона орієнтована на англійськомовні тексти. Прикметно, що на відміну від інших систем програма локалізує структурні зв'язки, підлеглість конструкцій, відмінювання та інше. Вищезазначене не змінюється у випадку перестановки

слів і речень, абзаців та пропозицій усередині аналізованого тексту, що дозволяє порівнювати його з іншими, визначаючи кількісний відсоток їх подібності.

2.4. Фактчекінг: методологія та інструменти

Гібридна російська-українська війна яскраво виявила безпековий потенціал медіаграмотності щодо мізінформаційних (поширення неправдивих даних ненавмисно), дезінформаційних (навмисне поширення таких даних із чіткою метою й розумінням наслідків), пропагандистських (поширення певних поглядів у суспільстві) матеріалів. Дезінформація й пропаганда можуть базуватися на певних наративах (до прикладу, у випадку з російською федерацією це ідея «руського міра»), які формують «потрібний» світогляд у населення. Саме тому для нас зараз таким важливим є потенціал цифрових наук у боротьбі з вищезазначеними виявами.

Фактчекінг – це перевірка певних фактів, які викликають сумнів (див. Рис. 28). На сьогодні, він може стосуватися будь-якого контенту: так, ми бачили низку картинок, зроблених ШІ. Від картинки з відважним льотчиком або братами-близнюками-захисниками до бабусі, яка продає сир… об’єднує всі ці підробки певне спонукання до дії. Тут стануть у пригоді ЦТ та лінгвістичні методи, з використанням яких процес перевірки набуває нових форм, дозволяючи нам убезпечити себе й рідних.

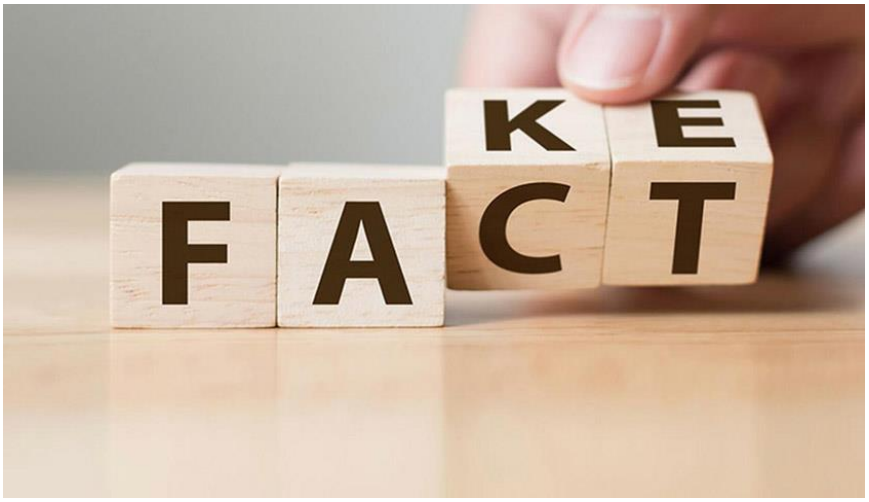


Рис. 28. Фактчекінг.

Зокрема, фактчекінг базований на таких *видах аналізу текстових даних*:

1. Семантичний аналіз: полягає в локалізації значення слів, словосполучень та речень у певному контексті. Так, фраза «Я дурію з цією прикормки» більшою мірою пов'язана з контекстом вживання, ніж безпосередньо зі значенням компонентів, які до неї входять.

2. Прагматичний аналіз: розгляд певних намірів автора щодо впливу на свою аудиторію. До прикладу, в Інтернеті або соціальних мережах (зокрема, у Тік-Ток) часто можна бачити публікації або коментарі, зроблені з метою створення певного хайпу. Згадаємо славнозвісні відео Пана Романа та багатьох інших, які створені спеціально для того, щоб викликати певні реакції аудиторії й набрати перегляди.

3. Дискурсивний аналіз репрезентативний щодо стилістики та риторичних прийомів, які можуть містити елементи мови ненависті, дискримінації, маніпулятивних стратегій та іншого. Зокрема, це може бути аналіз послідовності певних речень, пропозицій тощо, здійснюваний у межах психолінгвістики, соціолінгвістики, мовної прагматики і тому подібного.

До прикладу, в політичний інтернет-дискурс входять дописи відповідних акторів у соціальних мережах, їхні промови, інтерв'ю та інше. В плані дискурсу важливе розуміння контексту, проте в ширшому, ніж у першому випадку варіанті. Так, у травні 2022 року королева Великої Британії вбралася в одяг кольору українського прапору, чим продемонструвала свою позицію, яку безпосередньо не озвучувала.

4. *Корпусний аналіз* показовий у вимірі частотності, а саме – кількості появи у текстових даних певного терміна, словосполучення та іншого. Так, можна взяти дисертації, написані в Україні у період 2023–2025 років і дослідити написання слів «росія», «російська федерація», «москва» та інші. Навіть не заглядаючи у ці дослідження можна зробити впевнене припущення, що в більшості, якщо не в усіх, ці слова будуть написані з малої літери (як вияв особистого ставлення авторів та неповаги до ворога тощо).

Закономірно, що ЦТ надають зручні інструменти для фактчекінгової діяльності через автоматизацію рутинних процесів. Серед найпоширеніших назвемо такі:

1. Алгоритми ОПМ, у межах яких застосовують визначення фактологічної достовірності певних текстових даних.

Моделі на основі МН, здатні класифікувати текст або визначати достовірність фактів. Найпоширенішими методами тут є:

1.1. Наївний байєсівський класифікатор. Цей класичний метод часто використовують для попереднього аналізу текстів на достовірність, коли йдеться про їх класифікацію за категоріями «достовірний» / «недостовірний». Проте його застосування обмежене, оскільки він не враховує складні контексти у таких даних.

1.2. ШНМ та глибоке навчання. Сучасні методи, зокрема рекурентні нейронні мережі (Recurrent Neural Network; далі – РНМ), трансформери (до прикладу, BERT, ChatGPT та інші), дають змогу точніше визначати достовірність інформації завдяки своїй здатності до розпізнавання контексту та синтаксичних зв’язків у текстах. Ці моделі можуть бути навчені на спеціалізованих корпусах текстів для подальшого застосування до нових документів.

2. Алгоритми на основі фактчекінгу.

Для перевірки фактів використовуються алгоритми, які порівнюють висловлені твердження в текстах із наявними базами знань або авторитетними джерелами. Ось кілька підходів:

2.1. Методи порівняння з базами знань. Для таких алгоритмів властивим є використання вже структурованих баз знань: Wikidata, DBpedia, Freebase тощо. Алгоритм здійснює пошук відповідних тверджень у цих базах і порівнює їх із контекстом заяви в тексті, що, своєю чергою, дозволяє визначити фактологічну коректність останньої.

2.2. Метод порівняння з кількома джерелами. Використовується кілька джерел інформації для перевірки правильності фактів, коли дані з різних джерел порівнюють: таким чином алгоритм ухвалює рішення на основі множини

джерел. Цей підхід допомагає оцінити фактологічну достовірність, коли одне джерело може бути ненадійним або спотвореним.

3. Алгоритми на основі аналізу текстової репутації (Reputation-Based Models).

Іншим підходом є використання алгоритмів, які оцінюють репутацію джерела. До прикладу, аналіз вебсайтів, новинних агенцій або наукових публікацій на надійність. Важливим аспектом є визначення «довіреності» джерел за допомогою метрик (до прикладу, PageRank), кількості покликань або взаємодії з авторитетними партнерами. Такий підхід дозволяє ефективніше відрізняти надійні джерела від ненадійних.

4. Алгоритми виявлення маніпуляцій і фальсифікацій.

Цей напрямок спеціалізується на виявленні спроб маніпуляції фактами через зміщення контексту або викривлення подій. Алгоритми намагаються визначити, чи є певні висловлювання в текстах маніпулятивними, шляхом аналізу їх лінгвістичних ознак (до прикладу: зміни емоційного тону, підбору слів). Для оцінки результатів вищезазначених алгоритмів застосовують стандартні метрики:

а) *точність (accuracy)* – відношення правильно класифікованих елементів до загальної кількості;

б) *повнота (recall)* – здатність алгоритму виявити всі істинно достовірні або недостовірні факти;

в) *F1-міра* – комбінована метрика, яка враховує і точність, і повноту;

г) *матриця плутанини* – для розрізнення правильних і помилкових категорій.

5. Пошукові системи та агрегатори новин: Google Fact Check Explorer, Snopes, PolitiFact.

5.1. Google Fact Check Explorer – це інструмент, створений для полегшення пошуку фактчекінгових публікацій, що стосуються різноманітних новин, заяв і подій (див. Рис. 29). Він дозволяє користувачам шукати перевірені дані, пов’язані з конкретними запитами або заявами, та оцінювати правдивість інформації через покликання на джерела фактчекінгових організацій.

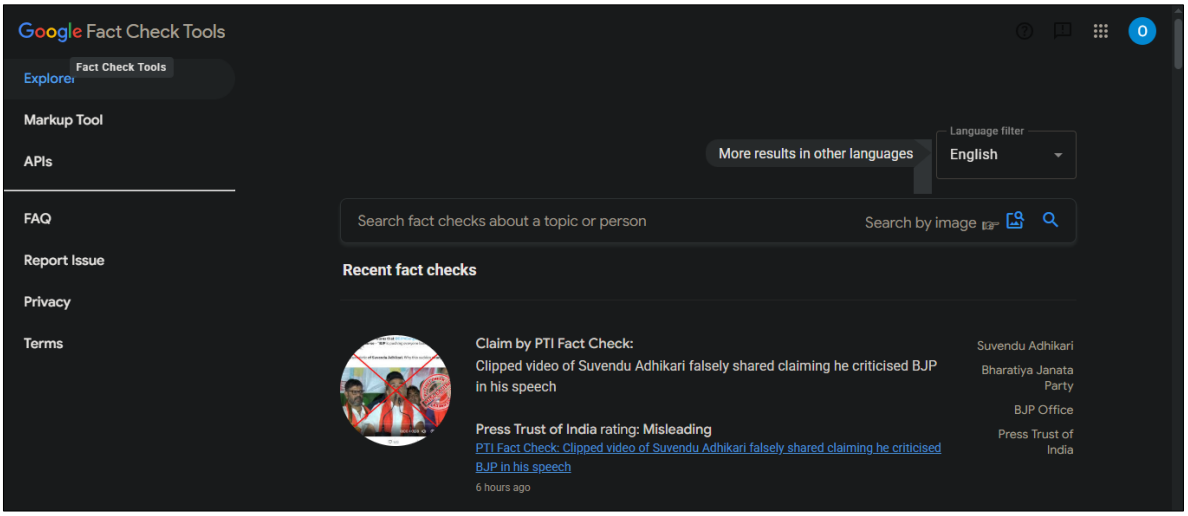


Рис. 29. Google Fact Check Explorer.

Основні характеристики:

– інтерфейс пошуку: користувачі можуть ввести ключові слова або запити, і система надасть інформацію, де була проведена фактчекінгова перевірка;

– масштабність та доступність: оскільки Google Fact Check Explorer має доступ до великої кількості фактчекінгових публікацій з усього світу, він дозволяє знаходити перевірену інформацію на основі найавторитетніших джерел;

– *прозорість джерел:* кожен результат супроводжує покликання на конкретну фактчекінгову публікацію з визначенням її джерела (до прикладу, Reuters, AFP, PolitiFact тощо).

Переваги:

а) миттєвий доступ до широкого спектра перевірених даних;

б) прозорість джерел та точність перевірок завдяки співпраці з відомими фактчекінговими організаціями;

в) легкість у використанні, що дозволяє користувачам швидко знаходити перевірену інформацію.

Обмеження:

– покриття лише тих заяв або фактів, які були зафіксовані фактчекінговими платформами;

– залежність від того, наскільки активно фактчекери працюють із певними типами новин.

5.2. Snopes – один із найстаріших і найвідоміших сайтів для фактчекінгу, що спеціалізується на перевірці міфів, чуток, фейків і маніпуляцій в Інтернеті (див. Рис. 30). Сайт був заснований у 1994 році Бобом й Барбарою Стаперс і став основним ресурсом для перевірки правдивості новин та заяв.

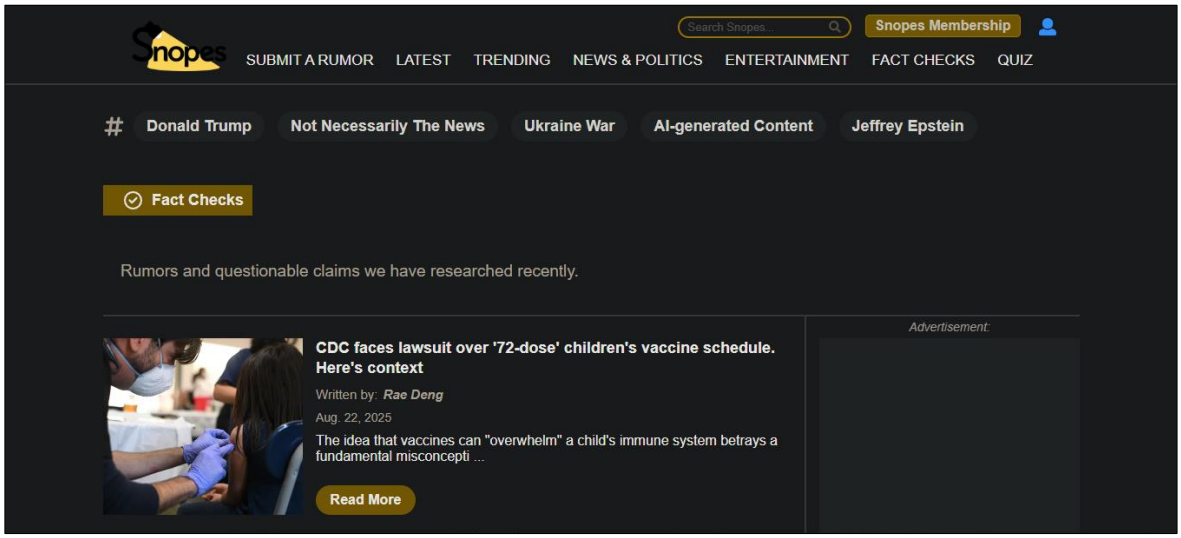


Рис. 30. Snopes.

Основні характеристики:

а) різноманітність перевірок: Snopes проводить фактчекінг у широкому спектрі тем – від політики та здоров'я до забобонів, міфів і чуток;

б) детальний аналіз: кожен фактчекінг на Snopes супроводжує детальне пояснення, чому конкретна інформація є або неправдивою, або частково правдивою;

в) категоризація: публікації групують за типами, такими як «Міфи», «Фальшиві новини», «Шахрайства» та інші, що допомагає швидко знайти потрібну перевірку.

Переваги:

– наявність великої кількості перевірок різних типів інформації;

– простий і зрозумілий інтерфейс для користувачів;

– докладний опис кожної перевірки, що дозволяє зрозуміти причини результату.

Обмеження:

Іноді Snopes зосереджується на менш значущих питаннях або на чутках, що зменшує його ефективність у серйозніших випадках фактчекінгу, таких як політичні маніпуляції чи фейкові новини. Може бути обмеженим у порівнянні з професійнішими фактчекінговими платформами для академічного чи політичного аналізу.

5.3. PolitiFact – онлайн-ресурс, орієнтований на перевірку заяв, зроблених політиками та діячами відповідної сфери (див. Рис. 31). Він був заснований у 2007 році в США й має за мету оцінити правдивість публічних висловлювань, спираючись на факти.



Рис. 31. PolitiFact.

Основні характеристики:

а) політична орієнтація: основна мета PolitiFact – перевірка фактів, пов'язаних із політикою, вебсайт займається аналізом публічних заяв, політичних обіцянок і промов, оцінюючи їх достовірність;

б) рейтинг правдивості: PolitiFact має свою систему оцінки правдивості заяв, яка включає категорії від «Чиста правда» до «Потребує подальшої перевірки», що дозволяє швидко оцінити точність або надійність заяви;

в) поглиблені дослідження: сайт також має детальніші статті та аналізи на основі фактів, що дозволяють глибше зрозуміти контекст заяв.

Переваги:

– оцінка заяв за допомогою зрозумілої шкали правдивості;

– сфокусованість на важливих політичних подіях і публічних висловлюваннях, що дозволяє використовувати PolitiFact як надійне джерело в політичному контексті;

– доступність для ширшої аудиторії, оскільки ресурси надаються безкоштовно.

Обмеження:

а) політична орієнтація може обмежувати використання PolitiFact у контексті інших сфер (науки чи культури);

б) часом може бути обмежений обсяг перевірених заяв, особливо в порівнянні з великими агрегаторами фактчекінгу такими, як вищезазначений сервіс від Google.

Отже, кожен із цих інструментів має свої переваги та слабкі сторони в контексті фактчекінгу. *Google Fact Check Explorer* надає швидкий доступ до великої кількості перевірених фактів від різних джерел, що робить його зручним для загальних перевірок. *Snopes* надає детальний контекст для розповсюдженіших міфів і чуток, тоді як *PolitiFact* є незамінним інструментом для перевірки політичних заяв і обіцянок. Разом ці ресурси можуть забезпечити різносторонній підхід до виявлення дезінформації та фальшивих новин, знижуючи ймовірність маніпуляцій і допомагаючи користувачам орієнтуватися в потоці інформації.

6. Інструменти розпізнавання маніпуляцій: OsoMeNet, Botometer для аналізу дезінформаційних кампаній у соціальних мережах.

6.1. OsoMeNet – інструмент для візуалізації й вивчення поширення фейкових новин, дезінформаційних кампаній та маніпуляцій у соціальних мережах (зокрема, в BlueSky). Він дозволяє аналізувати, як швидко й широко розповсюджується певна інформація в соціальних мережах, і чи є в цьому процесі маніпуляції або фальсифікація (див. Рис. 32).

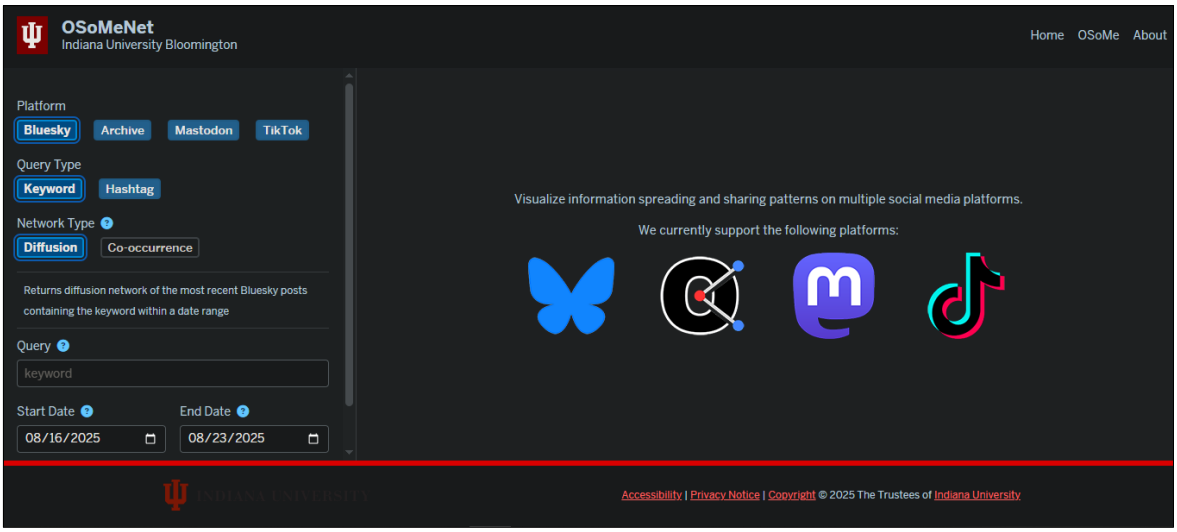


Рис. 32. OsoMeNet.

Основні характеристики:

– моніторинг поширення дезінформації: OsoMeNet спеціалізується на аналізі того, як фейкові новини чи маніпуляції поширюються через соціальні мережі; інструмент дозволяє відстежувати шлях новини, визначати, скільки людей та акаунтів були залучені в її поширення, й чи було це здійснено через автоматизовані канали;

– візуалізація поширення: інструмент використовує візуальні діаграми, щоб показати, як інформація (позитивна чи негативна) поширюється в мережі, що допомагає виявити аномалії або неприродні патерни, які можуть вказувати на маніпулятивні дії;

– інтерфейс користувача: зручний інтерфейс дозволяє дослідникам та журналістам швидко аналізувати певну новину або тему, відстежуючи динаміку її поширення.

Переваги:

а) візуалізація: здатність наочно показати, як інформація поширюється серед користувачів, що дає уявлення про етапи та канали розповсюдження дезінформації;

б) моніторинг за ключовими словами: можливість відслідковувати поширення певних заяв або тем у соціальних мережах;

в) прозорість: відкритий доступ до результатів аналізу дозволяє дослідникам незалежно вивчати дані про поширення інформації.

Обмеження:

– обмеженість на X: наразі інструмент в основному орієнтований на окреслену соціальну мережу, що може звужувати його застосування для аналізу інших платформ, таких як Facebook або Instagram;

– залежність від даних: точність результатів може бути лімітована наявними даними про конкретну новину чи кампанію.

6.2. *Botometer* – інструмент, що дозволяє виявляти ботів і автоматизовані акаунти в соціальних мережах (зокрема, у згаданій X). Botometer аналізує акаунти та оцінює ймовірність того, що вони є ботами, використовуючи більш, ніж 100 різних ознак поведінки акаунтів (див. Рис. 33).



Рис. 33. Botometer.

Основні характеристики:

а) оцінка поведінки акаунтів: Botometer аналізує різні характеристики акаунтів (частотність публікацій, час між ними, схожість контенту та взаємодія з іншими акаунтами), на базі чого надає оцінку ймовірності того, що акаунт є ботом;

б) *аналіз соціальних мереж:* інструмент може аналізувати великі набори даних у реальному часі, що дозволяє визначати бот-активність у певних кампаніях або під час обговорення важливих тем;

в) визначення впливу ботів: інструмент дозволяє не лише ідентифікувати боти, а й визначити, наскільки великий вплив вони (автоматизовані акаунти) мають на поширення дезінформації в соціальних мережах.

Переваги:

– детальний аналіз акаунтів: Botometer може дати точну оцінку ймовірності бот-активності в кожному акаунті, що дозволяє фахівцям зрозуміти, як саме боти впливають на розповсюдження інформації;

– масштабованість: цей інструмент здатний обробляти великі обсяги даних, що робить його ефективним для аналізу масштабних кампаній;

– інтеграція з іншими інструментами: Botometer може бути використаний у комбінації з іншими інструментами для комплекснішого аналізу дезінформаційних кампаній.

Обмеження:

а) оцінка на основі статистичних даних: інструмент не може 100% точно ідентифікувати кожен акаунт як бот або людину, оскільки його оцінки базуються на алгоритмах, які можуть мати помилки, або бути недостатньо точними в деяких випадках;

б) обмеженість у розпізнаванні складних маніпуляцій: хоча Botometer виявляє ботів, він не завжди може визначити складнішу маніпуляцію, коли вони діють у комбінації з реальними акаунтами.

Таким чином, *OsoMeNet* і *Botometer* є важливими інструментами для фактчекінгу, особливо в аналізі дезінформаційних кампаній у соціальних мережах. *OsoMeNet* дозволяє відстежувати поширення фейкових новин і маніпуляцій, виявляючи аномалії у патернах поширення інформації, тоді як *Botometer* спеціалізується на виявленні автоматизованих акаунтів, що активно маніпулюють контентом у соціальних мережах.

7. Бази даних перевірених фактів: використання верифікованих джерел для перевірки тверджень.

Найвідоміші фактчек-медіа в Україні: інтернет-проєкт “Stop Fake” (див. Рис. 34), аналітична платформа “Vox Check” та фактчек-масмедіа «Слово і Діло» та інші. Сайт для перевіряння фактів stopfake.org (проєкт “Stop

Fake") запустили 2 березня 2014 року, його засновниками стали випускники Могилянської школи журналістики та програми для журналістів і редакторів "Digital Future of Journalism".

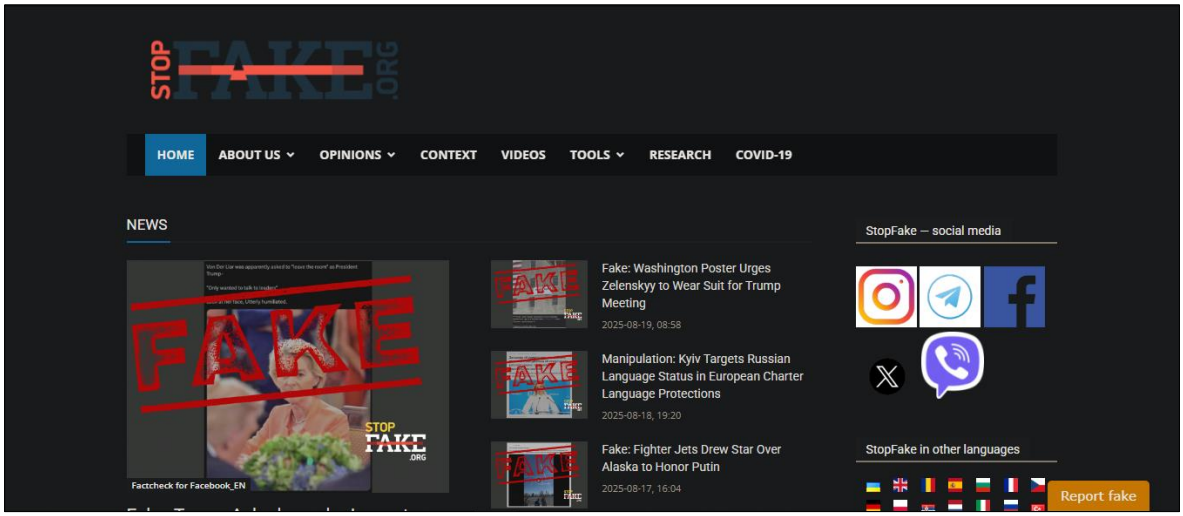


Рис. 34. Інтернет-проєкт "Stop Fake"

Основною метою проєкту активісти називають перевіряння та спростування перекрученої інформації й пропаганди про події в Україні, що поширюють через масмедіа. Зі слів учасників, діяльність їхньої групи спрямована на збереження України як самостійної та незалежної держави. Окрім впливу пропаганди на Україну, проєкт також досліджує її методи роботи й в інших країнах: від Сирії й Туреччини до країн ЄС і колишнього Союзу Радянських Соціалістичних Республік (далі – СРСР).

"Vox Check" заснований і працює на базі незалежної аналітичної платформи "Vox Ukraine" і систематично перевіряє заяви топових політиків. Аналітична платформа «Слово і Діло» – єдине в Україні масмедіа, яке виконує системний моніторинг обіцянок публічних людей з метою підвищення рівня підзвітності та прозорості влади.

На порталі щоденно публікують найважливіші новини України і світу, коментарі, авторські статті, інфографіку та аналітику на суспільно-політичні теми. За роки роботи колектив «Слова і Діла» зібрав понад 40 тис. обіцянок. На порталі можна перевірити рівень відповідальності будь-якого політика / чиновника, дізнатися, які обіцянки він виконав, а які – ні, і порівняти між собою.

2.5. Машинний текст та його аналіз: автоматизовані та ручні підходи виявлення

В умовах стрімкого розвитку технологій ШІ та автоматизованого генерування тексту, важливим завданням стає виявлення машинних текстів, тобто створених за допомогою алгоритмів (мовних моделей, ШНМ та інших технологій). Це питання постає особливо гостро у сфері боротьби з дезінформацією, виявленням фальсифікацій та автоматичного аналізу контенту. Виявлення машинного тексту є складним завданням, яке вимагає поєднання як автоматизованих, так і лінгвістичних підходів для забезпечення високої точності та ефективності.

1. Машинний текст: характеристика та контекст

Машинний текст визначається як текст, який був згенерований або частково створений автоматизованими системами, що використовують алгоритми ШІ для синтезу мовлення. Зазвичай такі тексти не містять людського авторського внеску, натомість вони згенеровані моделями, здатними відтворювати природну мову. Моделі, що створюють машинний текст, можуть варіюватися від простих шаблонних систем до складних глибоких нейронних мереж (далі – ГНМ), таких як ChatGPT та інші.

Основними ознаками машинного тексту є:

а) автоматична генерація: текст створюється без активної участі людини;

б) відсутність когерентності або емоційного забарвлення: тексти часто не містять емоційного контексту чи виписаних логічних переходів;

в) використання шаблонів і повторень: тексти можуть містити шаблонні структури та/або повторювані фрази.

2. Автоматизовані методи виявлення машинного тексту.

Автоматизовані методи виявлення машинного тексту використовують численні алгоритми та техніки МН для аналізу великих обсягів текстових даних і визначення їх походження. Основні підходи включають:

2.1. Алгоритми на основі стилістичного аналізу.

Стилістичний аналіз є одним із ключових інструментів виявлення машинного тексту. Алгоритми, що реалізують цей метод, порівнюють різні стилістичні характеристики текстів, такі як:

а) частотний аналіз лексичних одиниць: згенеровані тексти часто аномально використовують певні слова чи фрази, що відрізняються від мовної поведінки людей;

б) структура речень та абзаців: у машинному тексті зазвичай спостерігається однотипна структура речень, без складних синтаксичних побудов.

Алгоритми МН, які застосовують для стилістичного аналізу, включають підтримку векторних машин, дерева рішень, а також методи глибокого навчання (далі – ГН), такі як вищезазначені ШНМ.

2.2. Методи на основі ШНМ.

Моделі, засновані на ГНМ, такі як довготривала короткочасна пам’ять (Long Short-Term Memory; далі – ДКП) та трансформери (до прикладу, згадуваний ChatGPT), можуть автоматично навчатися на великих корпусах текстів для визначення відмінностей між машинними й людськими текстами. Ці моделі зазвичай аналізують:

– контекстні аномалії: машинний текст часто не здатен ефективно обробляти контекст, що може призвести до вищезазначених недоречних чи непослідовних зв’язків між фразами;

– пасивні звороти: машинний текст часто містить велику кількість пасивних конструкцій або схожих граматичних форм, що не властиво людським текстам (виключення – науковий стиль).

3. Лінгвістичні методи виявлення машинного тексту.

Лінгвістичні методи виявлення машинного тексту ґрунтуються на традиційному аналізі мови, а також на використанні мовних теорій і структур для розпізнавання аномалій, що характерні для текстів, створених алгоритмами.

3.1. Семантичний аналіз.

Лінгвістичні методи на основі семантики займаються виявленням аномалій у змісті тексту. Вищезазначено, що згенеровані тексти часто мають обмежене або неповне розуміння контексту, що призводить до семантичних помилок або нелогічних переходів між ідеями. Семантичний аналіз дозволяє виявляти такі помилки через:

а) логічні суперечності: машинний текст може містити висловлювання, що суперечать реальним фактам або порушують логіку;

б) невідповідність контексту: випадкові або неприродні зміни теми, що виникають через відсутність «здорового» контексту, можуть бути сигналом про автоматичне генерування.

3.2. Синтаксичний та морфологічний аналіз.

Машинні тексти можуть бути неповністю або неправильно скомпільовані з точки зору синтаксису або морфології. Лінгвісти можуть виявити такі аномалії за допомогою традиційних методів морфологічного та синтаксичного аналізу, визначаючи незвичні граматичні структури або нетипові словоформи.

3.3. Визначення емоційного забарвлення.

Лінгвісти також можуть застосовувати аналіз емоційного забарвлення для виявлення машинного тексту, оскільки багато машинних моделей мають труднощі з адекватною передачею емоційного тону. Емоційно нейтральні або неприродні висловлювання можуть бути ознакою такого тексту.

4. Перспективи розвитку методів виявлення машинного тексту.

З огляду на те, що технології генерації тексту постійно удосконалюють, розробка нових методів виявлення машинного тексту стає необхідною для забезпечення ефективної боротьби з дезінформацією та збереженням якості інформаційного простору. Основні напрямки, які можуть змінити ситуацію в найближчому майбутньому, включають:

4.1 Розвиток мультимодальних моделей.

Одним із нових напрямків є використання мультимодальних моделей, які поєднують текстову й нетекстову (аудіо, графіку або відео) інформацію. Це дозволить створювати комплексніші моделі для виявлення машинного тексту,

що можуть аналізувати контекст на кількох рівнях. До прикладу, перевіряти відповідність між текстом і зображенням або відео, що додається. Це може бути корисно для виявлення фальсифікацій у медіа та соціальних мережах.

4.2. Використання глибоких моделей для крос-доменного виявлення.

Глибокі моделі (зокрема, трансформери) можуть бути адаптовані у різних сферах: науці, масмедіа, економічній царині тощо. Застосування такої адаптації дозволяє ефективніше визначати згенеровані тексти, які можуть виглядати природно в контексті певної галузі, проте містити специфічні лінгвістичні ознаки або помилки, притаманні автоматичному генеруванню.

4.3. Використання пояснювальних моделей (Explainable AI).

З розвитком технологій ШІ виникає необхідність у створенні пояснювальних моделей, які дозволяють зрозуміти, чому алгоритм ухвалив те чи те рішення. Включення таких функцій у процес виявлення машинного тексту допоможе не лише виявляти фальсифікації, а й надавати пояснення щодо того, які саме характеристики тексту стали підставою для визначення його як згенерованого. Це створить більше довіри до таких систем, адже користувачі зможуть краще розуміти логіку алгоритмів.

4.4. Поліпшення інтеграції лінгвістичних знань у моделях МН.

Хоча ГН дозволяє створювати дуже потужні моделі для аналізу тексту, інтеграція традиційних лінгвістичних знань може значно покращити точність виявлення згенерованого тексту. Використання мовних теорій, синтаксичних правил та семантичних мереж у комбінації з алгоритмами МН дозволяє створювати точніші моделі, які здатні виявляти навіть найменші аномалії, характерні для таких текстів.

4.5. Використання міжмережевих баз даних для виявлення фальшивих джерел.

У майбутньому може з’явитися можливість використовувати бази даних різних типів (включаючи ті, що містять історичну інформацію, соціальні мережі, медіа-ресурси, наукові публікації) для ефективнішої перевірки фактів та виявлення згенерованого тексту і його потенційного джерела.

Запитання для самоконтролю:

1. Які лінгвістичні методи використовуються для виявлення плагіату в текстах?

2. Що таке шингл у контексті антиплагіатної перевірки та яку роль він відіграє в аналізі тексту?

3. Які функції виконує Cognitive Text Analyzer під час перевірки тексту і які є його обмеження?

4. Які рівні лінгвістичного аналізу (семантичний, прагматичний, дискурсивний) використовуються у процесі фактчекінгу?

5. Чим відрізняються Google Fact Check Explorer, Snopes та PolitiFact за функціональністю та сферою застосування?

6. Які алгоритми МН застосовуються для автоматизованого фактчекінгу? Які їхні переваги та недоліки?

7. Які лінгвістичні ознаки можуть свідчити про те, що текст створений автоматично, а не людиною?

8. Як мультимодальні моделі допомагають у виявленні фальсифікацій і машинного тексту в сучасному цифровому середовищі?

9. Які основні завдання виконує вебзастосунок TextAttributor у межах лінгвістичної експертизи тексту?

10. Які методи аналізу застосовуються в системі TextAttributor для визначення авторства та виявлення токсичності тексту?

Тема 3.

ТЕКСТОВІ КОРПУСИ В ЦИФРОВІЙ ЛІНГВІСТИЦІ: ВИДИ, МОЖЛИВОСТІ, ЗАСТОСУВАННЯ

3.1. Поняття текстового корпуса, його основні характеристики

Упродовж XX–XXI століть лінгвістика перейшла від споглядально-дескриптивного до обчислювального, цифрово орієнтованого підходу. Так званий «цифровий поворот» дав змогу формалізувати аналіз мови, зробити його масштабованим і повторюваним. Цифрова гуманітаристика (зокрема, лінгвістика) сьогодні охоплює широкий спектр дослідницьких практик, що спираються на комп'ютерну обробку мовного матеріалу. Основою такої трансформації стала згадана у минулій лекції *корпусна лінгвістика*, яка працює з великими масивами текстових даних.

У центрі уваги цієї дисципліни – *текстовий корпус* як репрезентативне джерело мовних одиниць у їхньому реальному вжитку. Корпус став центральний інструментом, що забезпечує доступ до структурованих, репрезентативних і масштабних мовних даних. Це відкрило нові горизонти у когнітивних, перекладацьких, соціолінгвістичних та компаративних студіях.

Корпус текстів – це структурована електронна колекція текстів, зібраних за певним принципом і призначених для лінгвістичного аналізу природної мови. На відміну від довільного набору текстів корпуси формуються так, щоб бути репрезентативними, рівноважними й аналітично придатними (згадайте початок попередньої лекції).

Окрім того, корпуси можуть містити писемні або усні (транскрибовані) тексти , що відібрані за певними критеріями (жанром, стилем, часовим періодом, соціокультурним контекстом тощо), їх використовують у корпусній лінгвістиці для практичного дослідження мови.

Основними характеристиками є:

а) електронний формат – корпус має бути представленим у машиночитаному форматі, доступному через спеціальні інтерфейси, для

можливості пошуку, аналізу та обробки даних за допомогою комп'ютерних програм;

б) репрезентативність – корпуси повинні охоплювати широкий спектр мовних варіацій і ситуацій, щоб забезпечити об'єктивність аналізу;

в) анотованість – тексти в корпусі повині мати лінгвістичну розмітку (морфологічну, синтаксичну, семантичну), що полегшує їх організацію та автоматичний аналіз;

г) метарозмітка – кожен текст, включений у корпусний масив, має супроводжуватися метаданими (інформацією про авторство, жанр, дату створення тощо);

ґ) прагматичність – корпус укладається з певною метою;

д) збалансованість – корпус має містити пропорційну кількість текстів різних жанрів, стилів та тематик задля адекватного відображення мови загалом й уникнення упередженості;

е) динамічність – деякі корпуси постійно оновлюються, особливо якщо вони відображають сучасні мовні зміни (до прикладу, корпуси соціальних мереж або новинні корпуси);

є) доступність – відкриті корпуси використовують для досліджень у лінгвістиці, ШІ, машинному перекладі та інших сферах.

3.2. Типологія корпусів

Корпуси класифікують за різними критеріями:

1. *За кількістю мов:*

а) монолінгвальні – корпуси однієї мови (до прикладу, British National Corpus (BNC), Corpus of Contemporary American English (COCA), Digitales Wörterbuch der deutschen Sprache (DWDS), Корпус української мови (КУМ), Генеральний регіонально анотований корпус української мови (ГРАК) та інші);

б) двомовні – корпуси, що містять тексти двома мовами (до прикладу, Open Parallel Corpus (OPUS), EuroParl, ParaRook тощо);

в) багатомовні – корпуси з текстами кількома мовами (до прикладу, Leipzig Corpora Collection, UN Corpus, Tatoeba, MULTEXT-East та інші).

2. *За жанром:*

– загальні: містять тексти різних жанрів (до прикладу, BNC, COCA, КУМ, ГРАК і так далі);

– спеціалізовані: включають тексти певної сфери: юридичні, наукові, художні тощо (до прикладу, EconBiz, SciELF тощо).

3. *За часом:*

а) синхронні – відображають мову певного періоду або часовий «зріз» (до прикладу, COCA, NOW Corpus та інші);

б) діахронні – охоплюють тексти різних епох, відображають мовні зміни упродовж історичного періоду (до прикладу, Helsinki Corpus of English Texts (HCET), Corpus of Historical American English (COHA), Deutsches Textarchiv (DTA) і тому подібні).

4. *За ступенем обробки й рівнем розмітки:*

– сирі корпуси: містять текст без розмітки (до прикладу, частина OPUS та тому подібні);

– анотовані корпуси, до яких додано морфологічну (частини мови, леми, граматичні категорії – MULTEXT-East), синтаксичну (дерева залежностей або складових – Penn Treebank), семантичну (теми, концепти, семантичні ролі – FrameNet), прагматичну (комунікативна функція, наміри мовця, інтенції – RST Discourse Treebank (RST-DT) розмітку.

5. *За типом текстів:*

а) корпуси писемних текстів – містять тексти художніх творів, публіцистики, наукових праць тощо (до прикладу, BNC, Brown Corpus та інші);

б) корпуси усних текстів – містять транскрибовані інтерв'ю, телевізійні виступи тощо (до прикладу, Spoken BNC, Santa Barbara Corpus of Spoken American English (SBCSAE) та тому подібні);

в) корпуси змішаних текстів – містять і писемні, і усні тексти (до прикладу, International Corpus of English (ICE) і так далі).

6. *За обсягом:*

– малі корпуси: до 1 мільйона слів; часто створюються для конкретних проєктів (до прикладу, Brown Corpus ≈ 1 млн. слів);

– середні корпуси: 1–100 млн. слів (до прикладу, BNC ≈ 100 млн. слів);

– великі корпуси: 100+ млн. слів (до прикладу, *COCA* ≈ 1 млрд слів).

Наведені класифікації корпусів текстів не вичерпують усіх можливих підходів. Так, залежно від дослідницьких завдань, технологічних можливостей або галузевої спеціалізації, корпуси також можуть класифікуватися за іншими критеріями. До прикладу:

а) за рівнем доступності (відкриті / обмеженого доступу / комерційні);

б) за типом носія (онлайн / офлайн / гібридні);

в) за способом представлення даних (у форматі *xml*, *json*, *tei*, *conll* тощо);

г) за інтеграцією з іншими мовними ресурсами (словниками, тезаурусами, базами даних).

Таким чином, класифікація корпусів є динамічною й може розширюватися відповідно до потреб сучасної цифрової гуманітаристики (зокрема, лінгвістики), нових технологій та міждисциплінарних практик.

3.3. Структура корпуса: одиниці організації, розмітка, формати

У корпусі лінгвістичні дані організовують та маркують на різних рівнях. Основними одиницями організації є такі:

1. *Текстові рівні* (ієрархічні одиниці):

– документ: найбільша одиниця, що відповідає цілісному тексту (стаття, книга, файл);

– абзац: логічний структурний елемент документа;

– речення: повна синтаксична одиниця;

– фрагмент / діалог: може використовуватися для маркування частин тексту (до прикладу, у художній літературі чи розмовних корпусах).

2. *Токени* – окремі мовні елементи, на які поділяється текст: ним може бути слово, скорочення, число, розділовий знак, символ. До прикладу, у реченні «Я перемагаю.» токенами є «Я», «перемагаю», «.».

3. *Леми* – словникова, або початкова, форма слова, яка представляє всі його граматичні варіанти. До прикладу, для слів «перемагаю», «перемагаєш», «перемагали» лемою є «перемагати».

4. *Метаінформація (метадані)* – інформація про сам текст, а не про його мовний зміст, яка допомагає впорядковувати та фільтрувати корпус. Включає такі характеристики:

а) автор, дата, джерело: хто створив текст, коли й звідки він узятий;

б) жанр, тип тексту: науковий, художній, розмовний тощо;

в) інша атрибутика: стать, вік автора, регіон, мовна варіація тощо.

Корпусна розмітка – один із найважливіших аспектів корпуса, для якої найчастіше використовують автоматичні інструменти. При цьому результат цього процесу завжди потребує людського контролю, а серед її типів виокремлюють такі:

– морфологічна: інформація про частину мови, граматичні категорії (рід, число, відмінок, час, вид тощо), лематизація (зведення до базової форми);

– синтаксична: структура речення, синтаксичні залежності, дерева залежностей;

– семантична: значення слів, синонімія, полісемія, тематичні ролі;

– прагматична: роль у дискурсі, наміри мовця (аргументація, вплив, переконання тощо).

Зазвичай процедура корпусного аналізу включає в себе три *основні кроки*:

а) ідентифікація мовних даних за допомогою категоріального аналізу;

б) визначення співвідношення мовних даних за допомогою статистичних методів;

в) інтелектуальна інтерпретація результатів.

Первинне автоматичне розмічування тексту для подальшої обробки в межах корпуса передбачає виконання таких основних операцій: сегментація тексту (визначення абзаців та речень), токенізація, морфологічний аналіз: лематизація та визначення граматичних категорій. При цьому *головними цифровими інструментами для роботи з корпусами* є:

1. *Платформи та пошукові системи:*

Sketch Engine / NoSketch Engine – потужні онлайн-сервіси, що поєднують пошукові системи, конкордансери й інструменти для аналізу частотності та колокацій (див. Рис. 35).

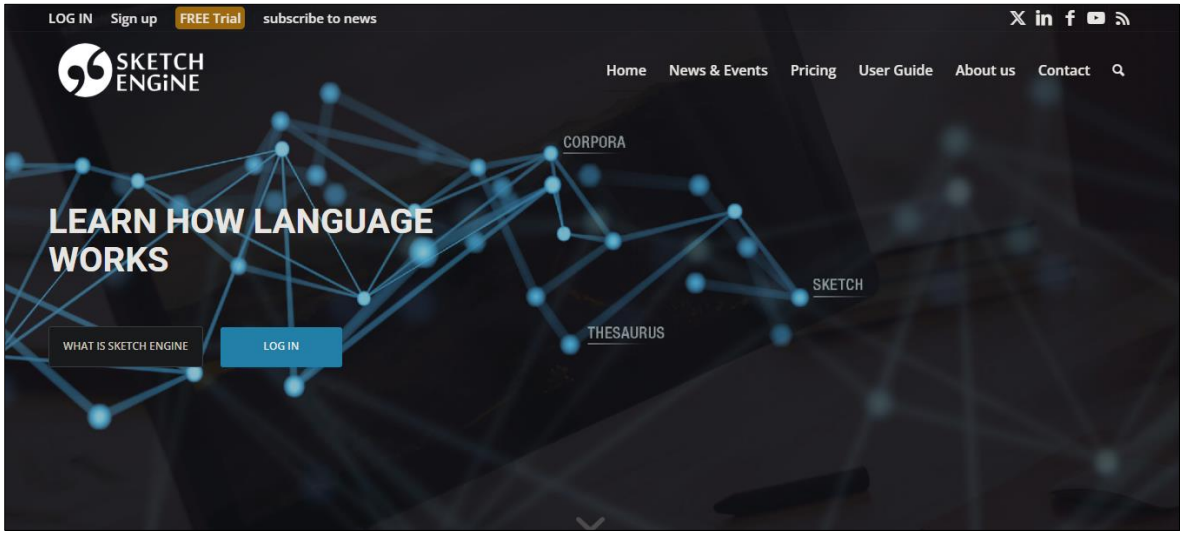


Рис. 35. Sketch Engine.

AntConc – популярний безкоштовний десктопний конкордансер та інструмент для корпусного аналізу (див. Рис. 36).

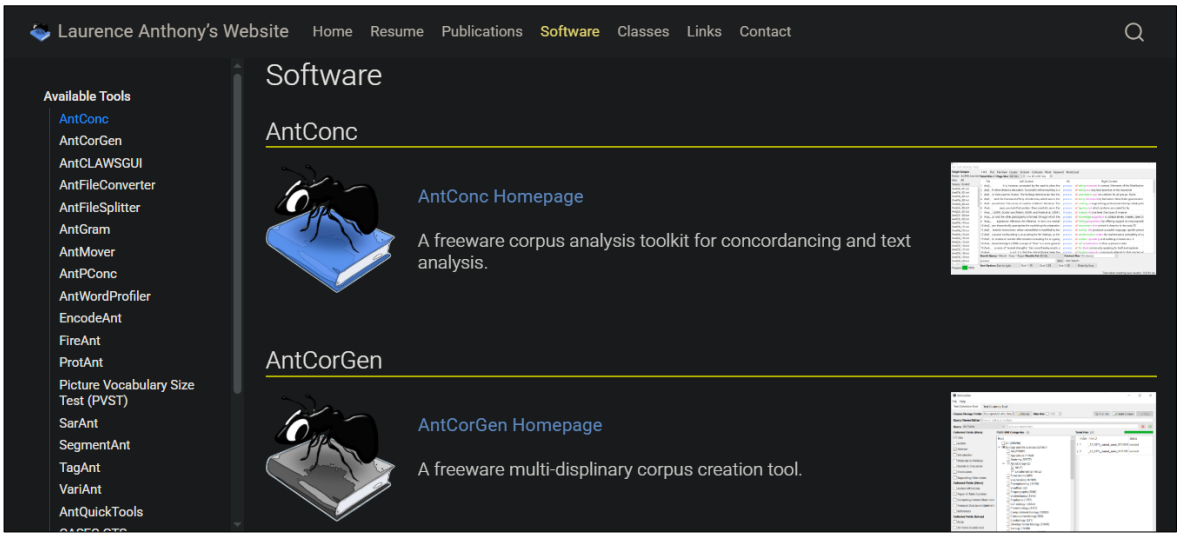


Рис. 36. AntConc.

2. *Інструменти для вирівнювання текстів:*

Nova Text Aligner – ПЗ для створення паралельних корпусів, яке автоматично вирівнює речення в текстах, написаних різними мовами, що дозволяє проводити порівняльний (контрастивний) аналіз (див. Рис. 37).



Рис. 37. Nova Text Aligner.

InterText – гнучкий менеджер і редактор для вирівнювання паралельних текстів на рівні речень; доступний як у вигляді десктопної програми, так і як багатокористувацького серверного додатка (див. Рис. 38).

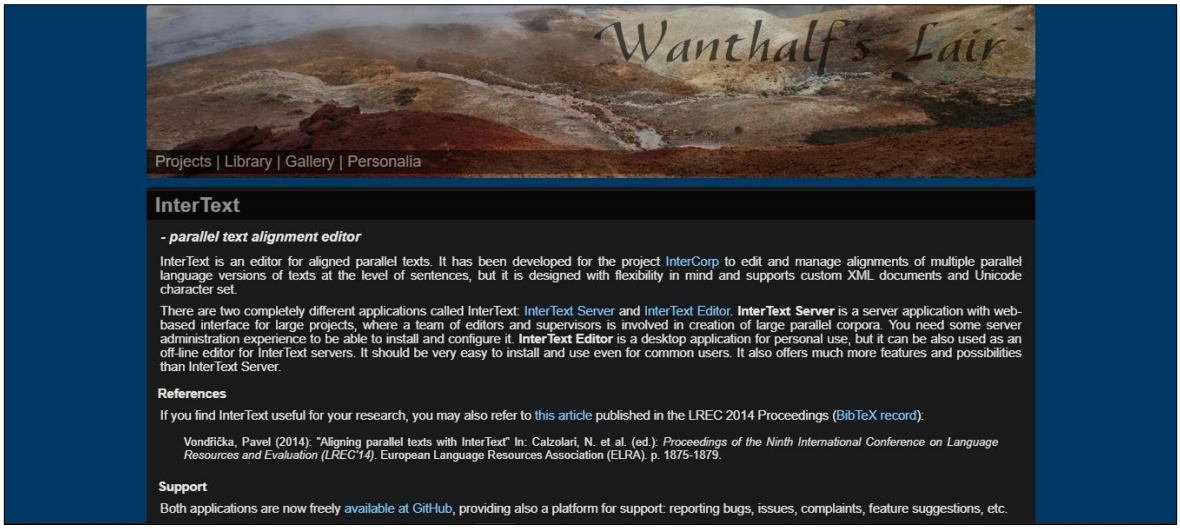


Рис. 38. InterText.

3. *Інструменти для аналізу частотності слів, колокацій та візуалізації*:

WordSmith Tools – ПЗ для створення списків частотності, ключових слів і виявлення колокацій (див. Рис. 39).

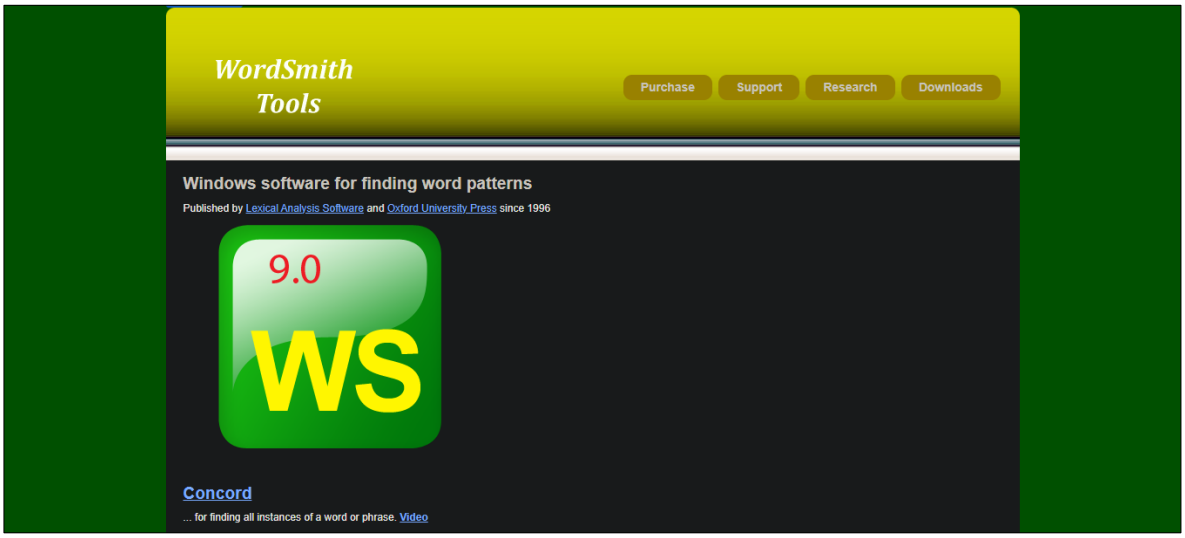


Рис. 39. WordSmith Tools.

Voyant Tools – онлайн-платформа для візуалізації та аналізу даних корпуса (див. Рис. 40).

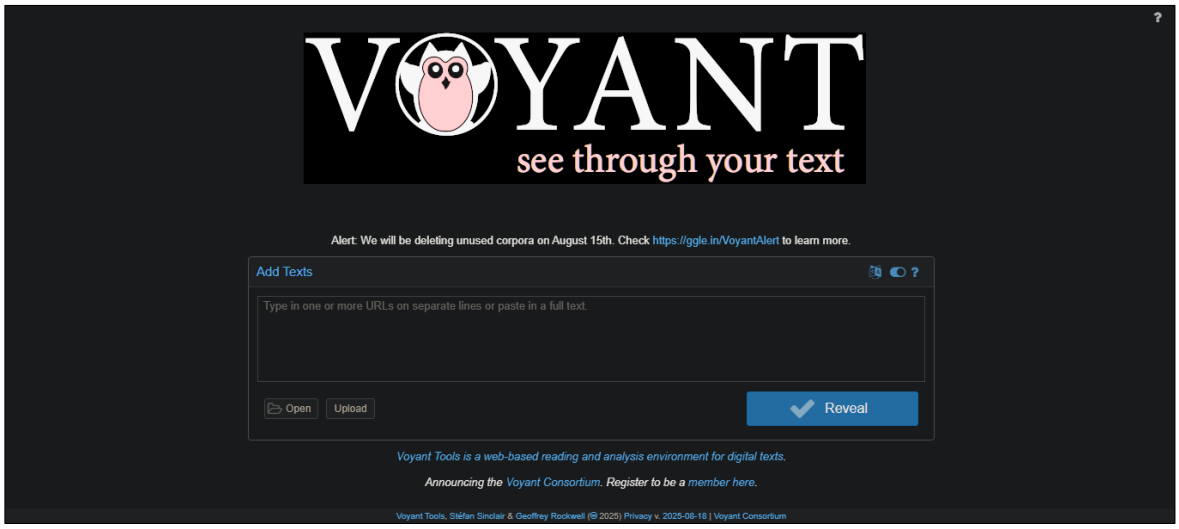


Рис. 40. Voyant Tools.

Google Books Ngram Viewer – інструмент для аналізу частотності слів у великому корпусі книг за певний часовий проміжок (див. Рис. 41).

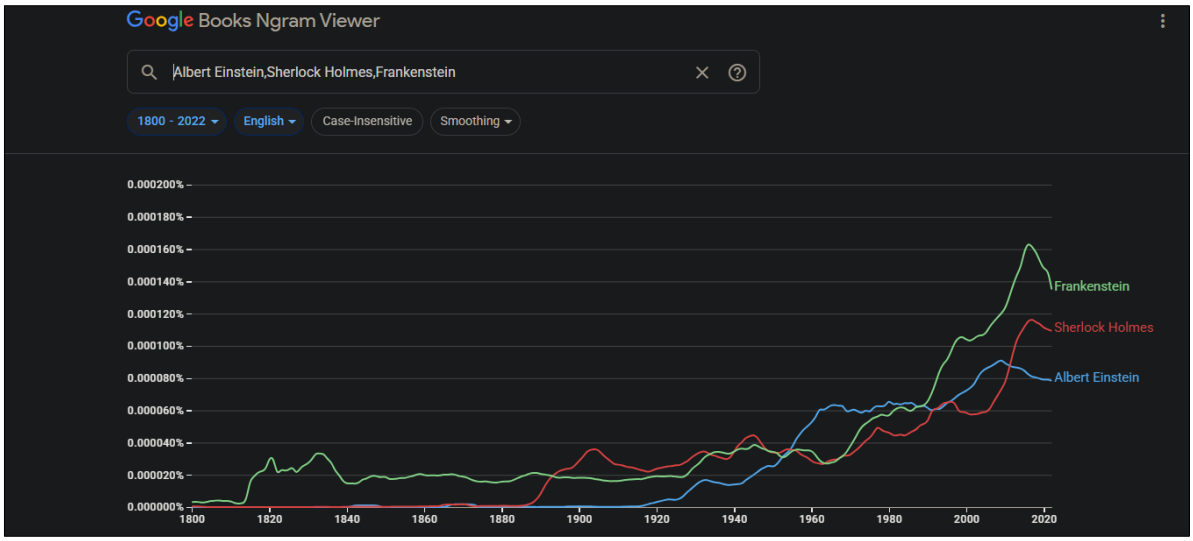


Рис. 41. Google Books Ngram Viewer.

4. *Інструменти для автоматичного лінгвістичного аналізу (фреймворки ОПМ)*:

spaCy, NLTK – потужні бібліотеки Python для ОПМ, що дозволяють виконувати морфологічний, синтаксичний та семантичний аналіз (див. Рис. 42).

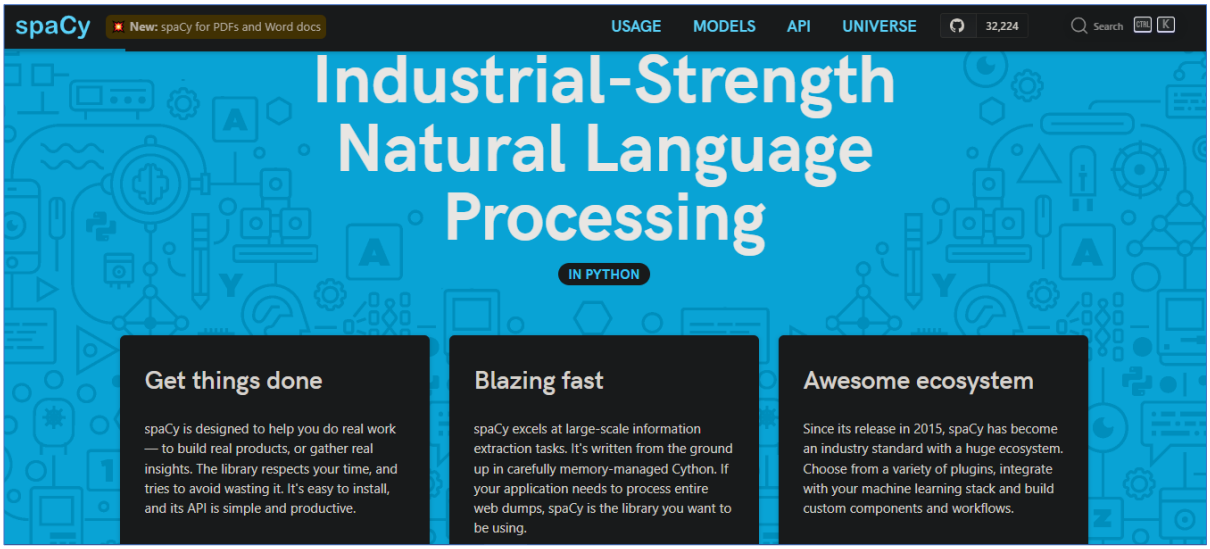


Рис. 42. spaCy.

TreeTagger – програма для розмітки текстів за частинами мови (POS-tagging) та лематизації (див. Рис. 43).



TreeTagger - a part-of-speech tagger for many languages

The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The TreeTagger has been successfully used to tag German, English, French, Italian, Danish, Swedish, Norwegian, Dutch, Spanish, Bulgarian, Russian, Portuguese, Belarusian, Ukrainian, Galician, Greek, Chinese, Swahili, Slovak, Slovenian, Latin, Estonian, Polish, Persian, Romanian, Czech, Albanian, Coptic and old French texts and is adaptable to other languages if a lexicon and a manually tagged training corpus are available.

Sample output:

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

The TreeTagger can also be used as a chunker for English, German, French, and Spanish.

The tagger is described in the following two papers:

- Helmut Schmid (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.
- Helmut Schmid (1994): Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK.

Рис. 43. TreeTagger.

Stanford NLP – набір моделей для синтаксичного та семантичного аналізу текстів.

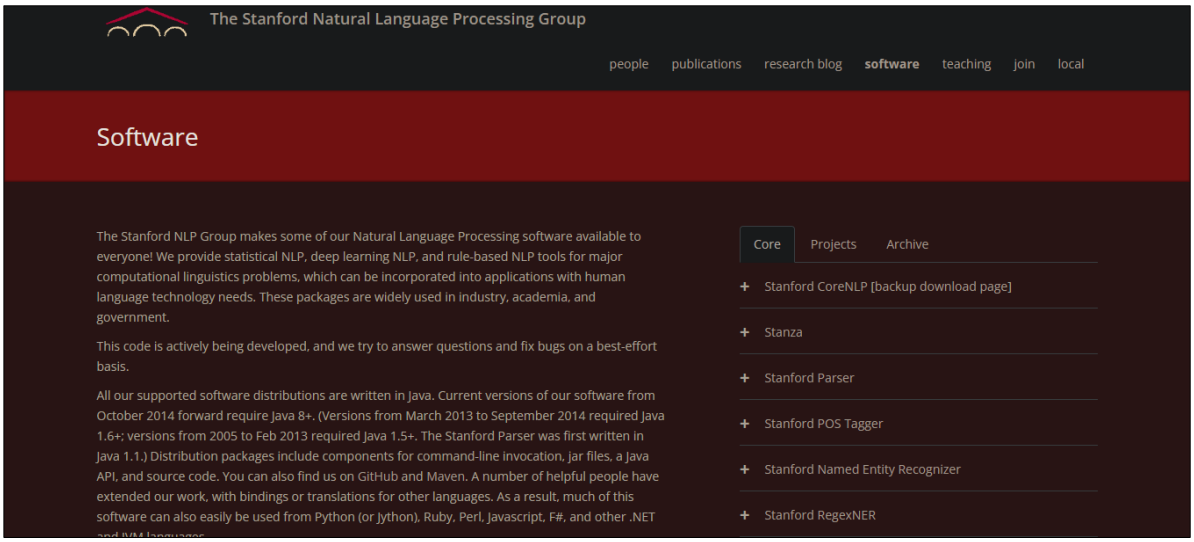


Рис. 44. Stanford NLP.

UDPipe – програма для обробки тексту на основі Universal Dependencies (див. Рис. 45).

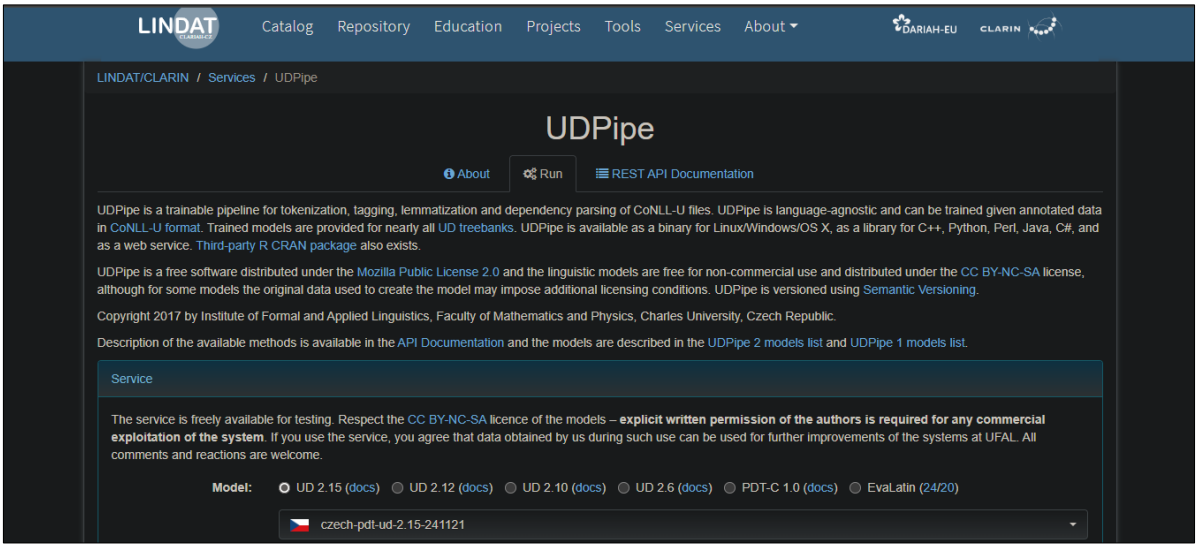


Рис. 45. UDPipe.

Найуживанішими форматами представлення корпусів є:

а) *tei/xml* – формат для анотації літературних текстів;

б) *conll-u* – для корпусів із синтаксичною розміткою;

в) *json, csv, tsv* – для обміну та аналізу.

Пошукові запити здійснюються за допомогою спеціальних мов, до прикладу *Corpus Query Language (CQL)* – дозволяє формулювати запити за частинами мови, лемами, позиціями; *Regular expression tester (RegEx)* – регулярні вирази для пошуку патернів.

3.4. Застосування корпусів у філології та перекладознавстві

Для сучасного філолога вміння працювати з корпусами є професійною необхідністю. Так, корпуси допомагають:

а) аргументовано обґрунтовувати лінгвістичні твердження й висновки;

б) проводити емпіричні дослідження;

в) позиціонувати мову як динамічну систему в реальному вжитку;

г) готувати академічні й популярні лінгвістичні тексти.

Застосування корпусів є надзвичайно широким і охоплює як фундаментальні лінгвістичні дослідження, так і практичні галузі (зокрема,

перекладознавство та лексикографію). Мовиться про те, що корпуси у дослідженнях:

– *з лінгвістики* дозволяють: проводити частотний аналіз; виконувати аналіз сполучуваності й варіативності; порівнювати діалекти, стилістичні варіанти; вивчати синтаксис, морфологію; аналізувати мовні зміни – еволюцію значень і конструкцій;

– *з перекладознавства* дають можливість: проводити порівняльний аналіз оригіналу й перекладу; ідентифікувати сталі відповідники; вивчати перекладацькі трансформації та виявляти їх закономірності; створювати навчальні корпуси для перекладачів;

– *з лексикографії* репрезентативні щодо автоматичного вилучення прикладів; визначення актуальної частотності; вибудовування словникових статей на підґрунті реального уживання; формування лексикографічних ресурсів (словників, тезаурусів та граматик).

3.5. Обмеження та виклики при роботі з текстовими корпусами

Хоча корпусна лінгвістика є потужним і незамінним інструментом, важливо розуміти її ключові обмеження, щоб уникнути помилок в інтерпретації даних. Зокрема, виділимо такі проблеми:

1. *Репрезентативність*. Якість корпусного аналізу безпосередньо залежить від того, наскільки представленим є корпус. Так, якщо добір текстів необ'єктивний (до прикладу, включає лише академічні статті), то результати дослідження не можна екстраполювати на всю мову. Це пов'язано з тим, що вони не відображатимуть розмовну мову, художню літературу тощо, тобто не будуть об'єктивними. Корпус завжди є лише вибіркою, і його репрезентативність потрібно чітко обґрунтовувати.

2. *Точність автоматичної анотації*. Автоматичні інструменти для розмітки текстів (морфологічний, синтаксичний аналіз, POS-тегування) не є ідеальними. Відповідно, вони можуть помилятися, особливо зі складними або неоднозначними мовними конструкціями. Саме тому для критично важливих

досліджень результати автоматичної розмітки часто потребують ручної перевірки та корегування.

3. *Нерівномірність ресурсів*. Не всі мови мають достатньо великі, збалансовані та якісно розмічені корпуси. Для деяких мов або вузьких жанрів такі ресурси можуть бути відсутніми взагалі, що унеможливлює або значно ускладнює проведення досліджень.

4. *Відсутність контексту*. Корпуси надають дані про те, як слова використовуються, однак не завжди можуть надати весь спектр їхнього значення. Своєю чергою, це пов'язано з тим, що такий контекст залежить від ситуації спілкування, позамовних характеристик мовців, культурних алюзій, реалій та лакун тощо. Наслідком цього є те, що дослідник має самостійно доповнювати корпусні дані глибшим якісним аналізом (зокрема, враховувати культурні практики тощо).

5. *Вимоги до технічних навичок*. Для ефективного аналізу великих корпусів часто потрібні не лише лінгвістичні знання, а й певні технічні навички: вміння працювати зі спеціалізованим ПЗ, знання мов програмування (до прикладу, Python), щоб автоматизувати процеси збору та аналізу даних.

Отже, корпусна лінгвістика відкриває нові шляхи вивчення мови: надійні, масштабні, об'єктивні. Зокрема, вона дає змогу філологу не лише описувати мовні явища, а й доводити їх на підставі реальних даних. При цьому текстові корпуси є основою цифрової гуманітаристики (у тому числі й лінгвістики).

Саме тому природно, що їх застосування охоплює як теоретичні, так і прикладні напрями. Відповідно, для сучасного філолога володіння корпусними інструментами й навичками роботи з корпусами – це не просто технічна навичка. Мовиться про нову форму наукового мислення, базованого на доказах і даних, а також невід'ємний складник цифрової грамотності лінгвіста XXI століття.

Запитання для самоконтролю:

1. Що таке текстовий корпус і які його основні характеристики? Зверніть увагу на репрезентативність, анотацію, метадані, динамічність тощо.

2. Які основні типи корпусів виокремлюють за кількістю мов? Наведіть приклади.

3. Чим відрізняються синхронні й діахронні корпуси? Яке їхнє лінгвістичне значення?

4. Що таке корпусна розмітка і які її основні види? Наведіть приклади.

5. Які інструменти й програмні засоби використовують для роботи з корпусами? Назвіть по два з різних категорій.

6. У чому полягають переваги використання корпусів у перекладознавстві?

7. Які структури організації тексту та одиниці аналізу використовують в корпусах?

8. Чим корпуси усних текстів відрізняються від писемних? Наведіть приклади.

9. Які проблеми можуть виникати при автоматичній розмітці корпусів? Як їх можна розв’язувати?

10. Чому робота з корпусами є невід’ємним складником цифрової грамотності сучасного філолога?

Тема 4.

ШТУЧНИЙ ІНТЕЛЕКТ ТА ШТУЧНІ НЕЙРОННІ МЕРЕЖІ

4.1. Основні поняття, принципи роботи та класифікація

В останні роки разом із *нечіткими дискретними системами* (описують динаміку ситуації ухвалення рішень (машиною або людиною) за допомогою формалізації певних дій (алгоритмів)) великий інтерес викликає проблематика ШІ, ШНМ і генетичних алгоритмів (далі – ГА), тобто таких, де використовують механізми, що нагадують біологічні процеси або структури. Прикметно, що вищезазначені напрями належать до наукової сфери, їх здебільшого окреслюють за допомогою терміна «обчислювальний інтелект» (“computational intelligence”).

Отже, *ШІ* – галузь комп’ютерних наук, що займається створенням систем, здатних виконувати завдання, які традиційно потребують людського інтелекту. Серед таких завдань: розпізнавання мовлення та зображень, ухвалення рішень, прогнозування та автоматизоване навчання.

Основні характеристики ШІ:

а) адаптивність – здатність навчатися на основі попереднього досвіду та корегувати поведінку;

б) автономність – виконання завдань без постійного втручання людини;

в) обчислювальна потужність – можливість опрацювання великих обсягів даних із високою швидкістю;

г) генеративні можливості – здатність створювати нові зображення, тексти, музику тощо (до прикладу, ChatGPT, Midjourney, Sora тощо).

Класифікація ШІ:

– слабкий (вузький): системи, що спеціалізуються на конкретних завданнях (до прикладу: розпізнавання мовлення в Siri; машинний переклад у Google Translate і тому подібне);

– загальний: гіпотетичний рівень, на якому система зможе розуміти, навчатися й виконувати будь-які завдання на рівні людського мозку;

– суперінтелект: гіпотетична стадія розвитку, коли ШІ перевершить людський у всіх аспектах.

ШНМ – математичні моделі, натхненні біологічною будовою мозку, які використовують для розпізнавання складних закономірностей у даних. Вони складаються зі штучних нейронів (далі – ШН), які взаємодіють між собою через вагові зв'язки (див. Рис. 46).

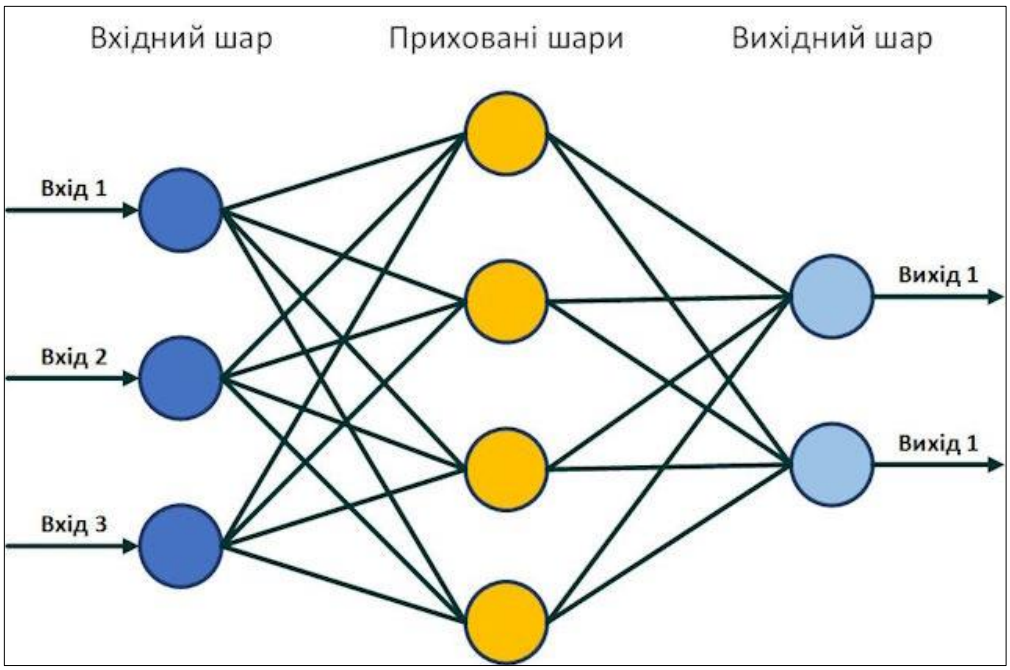


Рис. 46. Структура щтучних нейронних мереж.

Компоненти ШНМ:

а) згадані штучні нейрони – базові елементи, що приймають, обробляють та передають інформацію;

б) вхідний шар – набір таких нейронів, що приймає вхідні дані;

в) приховані шари – їхні шари, які виконують обчислення та розпізнають закономірності;

г) вихідний шар – нейрони, що формують кінцевий результат роботи мережі;

ґ) функції активації – механізм, який визначає, чи буде нейрон активований для передачі сигналу далі (до прикладу, ReLU, Sigmoid, Softmax і тому подібні).

Безпосередньо *типи ШНМ* корелюють із їх архітектурою (топологією), яку буде розглянуто пізніше, проте назвемо головні різновиди:

– перцептрон (класичний) є найпростішою нейромережею, здатною виконувати базову класифікацію даних, а також перцептрон (багатошаровий, MLP), що складається з кількох прихованих шарів, які дозволяють виконувати складніші обчислення;

– згорткові нейронні мережі (Convolutional Neural Network; далі – ЗНМ), ефективні для аналізу зображень і відео;

– рекурентні нейронні мережі (РНМ, ДКП, ЗРБ), використовувані для аналізу послідовностей даних, таких як текст або аудіо;

– генеративні змагальні мережі (ГЗМ, ДМ), застосовувані для створення нових зображень, музики, відео.

Природно, що ШІ та ШНМ взаємопов’язані: так, перший є широкою галуззю, в якій актуалізовано низку методів, підходів (зокрема, МН, яке буде розглянуто далі, ШНМ). Натомість другі є одним із найпотужніших інструментів ШІ, який дозволяє розпізнавати складні закономірності, прогнозувати події та автоматично навчатися без жорстких обмежень (програмування правил цього процесу). Завдяки розвитку ШНМ сучасний ШІ досяг значного прогресу в автоматизації аналітики, творчих процесів і ухваленні рішень, відкриваючи нові можливості для науки, економіки та суспільства.

Оскільки ШІ інтегрується в різні галузі (генеративні моделі (до прикладу, ГЗМ, ДМ та інші), медична діагностика, фінансовий аналіз, автоматизація виробництва тощо), то ми зупинимось на ШНМ докладніше. Так, завдання ШНМ, ГА та нечітких систем можна розглядати як у зв’язку, так і відокремлено, проте саме у першому випадку це є найпродуктивнішим.

Зокрема, вищезазначені ГА можна застосовувати для підбору ваг (про них будемо говорити окремо) і архітектури ШНМ, а також для формування бази правил та функцій приналежності нечіткої системи, в тому числі й функціонування ГА (про це йшлося на початку цієї лекції).

Логічно, що філософію ШНМ можна закласти у фундамент таких систем, оскільки саме вона виформовує у них здатність до навчання й розвитку. Окрім того, методи теорії нечітких множин дозволяють підбирати як вищезазначені параметри ГА, так і коефіцієнти, що визначають швидкість навчання таких

мереж. Натомість адаптивні системи на базі ШНМ дозволяють з успіхом вирішувати проблеми розпізнавання образів, виконання прогнозів, оптимізації, асоціативної пам’яті й керування.

У попередніх лекціях було згадано, що ШНМ базовані на механізмі людського мислення, суть якого – у збереженні даних у вигляді образів. Особливістю такого утворення (мереж) є можливість роботи, а також створення паралельних мереж, їх навчання та вирішення інтелектуальних завдань без використання традиційного програмування. Саме тому в лексиконі розробників та користувачів ШНМ присутні слова «поводити себе», «реагувати», «самоорганізовувати», «навчати», «узагальнювати» та «забувати». Прикметно, що розробниками таких мереж можуть бути вчені з даних, політологи, лінгвісти та інші.

4.2. Архітектура та специфіка роботи штучних нейронних мереж

ШНМ конструюють із базового блоку – ШН, іншою їх особливістю є використання величезної кількості зв’язків, які слугують містками між окремими нейронами. Пам’ятаємо, що групування нейронів у мозку людини забезпечує обробку інформації динамічним, інтерактивним та самоорганізованим шляхом. Вищезазначені біологічні нейронні мережі складаються з мікроскопічних компонентів, які існують у тривимірному просторі й здатні до різноманітних з’єднань.

Подібно до них, актуалізація ШН у ШНМ пов’язана з низкою фізичних обмежень: так, об’єднуючись у мережі, такі нейрони утворюють систему обробки даних, яка забезпечує ефективну адаптацію моделі до постійних змін із боку зовнішнього середовища. Зокрема, в процесі функціонування мережі відбувається перетворення вхідного вектора сигналів у вихідний, натомість конкретний вид такого процесу (перетворення) визначає архітектура або топологія нейромережі (згадаємо вищеназвані типи: РНМ, ЗНМ, ГНМ та інші).

Безпосередньо архітектуру й питомі особливості кожної розглянемо нижче: саме у них криється самобутність нейронних елементів, засоби керування та синхронізації інформаційних потоків між нейронами. Цікаво, що ШН є

базовим модулем ШНМ: так, його роль вкрай важлива, оскільки під час роботи він одночасно отримує багато вхідних сигналів.

Кожен такий вхід має свою власну *синаптичну вагу*, яка надає йому (входу) *вплив*, необхідний для загальної обробки поставленого розробником завдання. Таким чином, ці ваги є *мірою сили вхідних зв'язків*, які моделюють різноманітні синаптичні сили нейронів. Ваги суттєвого входу підсилюються і, навпаки, вага несуттєвого входу примусово зменшується, що визначає інтенсивність вхідного сигналу.

Окреслені ваги можуть змінюватись відповідно до навчальних прикладів, архітектури ШНМ, мети її створення, специфіки даних та іншого. Саме тому важливим фактором ефективності мережі є встановлення оптимальної кількості нейронів та типів зв'язків між ними. Так, для опису ШНМ використовують усталену термінологію, куди входять такі поняття як:

а) структура ШНМ – спосіб зв'язків нейронів у ній;

б) архітектура, або топологія ШНМ – параметризація нейромережі та типів нейронів;

в) парадигма ШНМ – спосіб навчання та використання, іноді містить поняття архітектури.

Цікаво, що на базі однієї архітектури може бути реалізовано різні парадигми нейромережі й навпаки. Серед відомих архітектурних рішень виокремлюють групу *слабозв'язаних нейронних мереж*, у випадку, коли кожний нейрон мережі пов'язаний лише із сусідніми. Натомість в *повнозв'язаних нейронних мережах* входи кожного нейрона зв'язані з виходами всіх решти нейронів.

Найпоширенішим варіантом архітектури є багатошарові нейронні мережі, з якими ви всі знайомі безпосередньо: тут нейрони об'єднують у шари з єдиним вектором вхідних сигналів. Зовнішній вхідний вектор подається на вхідний шар ШНМ (рецептори), а виходами такої мережі є вихідні сигнали останнього шару (ефектори).

Окрім вхідного та вихідного шарів, нейромережа має один або кілька прихованих шарів нейронів, які не мають контактів із зовнішнім середовищем.

Важливо розуміти, що зв’язки між нейронами називають *проєктивними* (якщо це різні шари) та *латеральними, або бічними* (якщо шар лише один). Незважаючи на те, що потенційно другий варіант можливий (можуть існувати ШНМ, які мають лише один шар і навіть один елемент), проте здебільшого їх три: вхідний, прихований та вихідний.

Пам’ятаємо з попередніх лекцій, що основна робота відбувається саме у прихованому шарі або шарах, оскільки вхідний шар лише отримує певні дані для роботи, а вихідний – пересилає її результати. При цьому між вхідним і вихідним шарами можуть бути розміщено багато прихованих різної архітектури: саме природа та послідовність їх розташування впливає на результат роботи ШНМ. Окрім того, входи та виходи кожного зі ШН у таких шарах сполучені одне з одним у різних напрямках, що є вкрай важливим. Так, зв’язки, скеровані від вхідних шарів до вихідних, називаються *аферентними*, натомість зв’язки в зворотному напрямку називаються *еферентними*.

В більшості мереж кожен нейрон прихованого шару отримує сигнали від усіх нейронів попереднього шару чи від нейронів вхідного. Після виконання операцій над сигналами, нейрон передає свій вихід до всіх нейронів наступних шарів, забезпечуючи передачу вперед (feedforward) на вихід.

Натомість за зворотного зв’язку, вихід нейронів шару скеровують до нейронів попереднього шару. Природно, що напрям зв’язків нейронів має значний вплив на роботу конкретної ШНМ, саме тому більшість мереж дозволяють розробнику додавати, вилучати та керувати з’єднаннями як завгодно.

Так, корегуючи параметри, можна налаштувати зв’язки як на посилення, так і на послаблення величини сигналів. За архітектурою зв’язків, більшість відомих ШНМ розділяють на дві великі групи: *мережі прямог*о (з односкерованими послідовними зв’язками) та *зворотного* (з рекурентними зв’язками) *поширення*.

1. **До мереж прямого поширення належать:**

Одношаровий перцептрон, здатний розпізнавати найпростіші образи: у ньому окремий нейрон обчислює певну суму сигналів вхідних елементів,

віднімає значення зсуву й пропускає результат через жорстку порогову функцію, вихід якої дорівнює +1 чи –1 (відповідно, кожному з цих значень присвоюють клас (А або Б)).

Така ШНМ складається з одного нейрона, формує дві вирішальні області, розділені *гіперплощиною*: у МН вона є межею певного рішення, оскільки сепарує вхідний простір на дві або більше областей, кожна з яких відповідає окремому класу, або вихідній мітці (вищезазначені класи А та Б; див. Рис. 47).

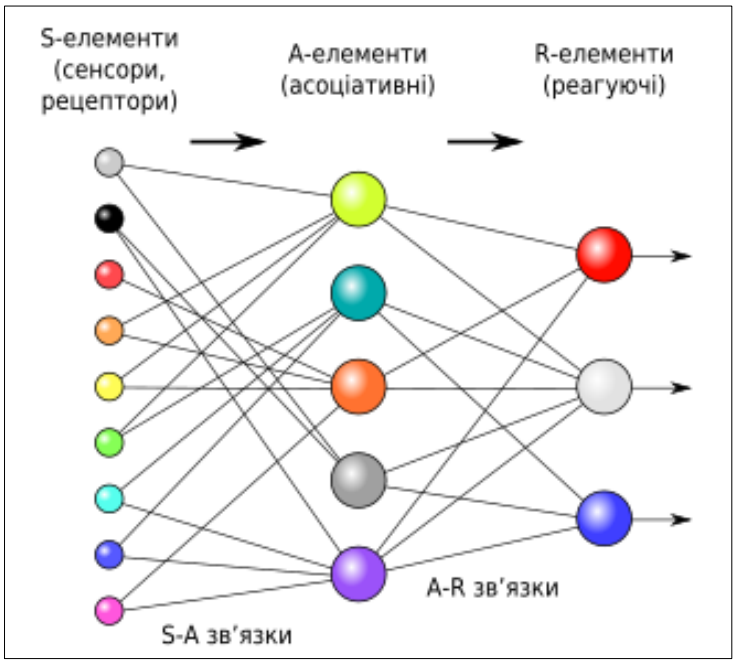


Рис. 47. Логічна схема перцептрона з трьома виходами.

Багатошаровий перцептрон (далі – БП) – по суті, класична структура ШНМ, оскільки саме ця фундаментальна архітектура заклала базу для складніших моделей ГН й продовжує використовуватися в різних програмах та застосунках із ШІ (див. Рис. 48). Важливо, що БП та ШНМ є важливими поняттями в галузі ШІ, проте це не однорідні й не синонімічні поняття.

Так, БП – особливий тип ШНМ: до прикладу, це певна серія Tesla Ілона Маска в ширшій категорії елекромобілів. Натомість, всі БП є ШНМ, однак не всі ШНМ є БП. Це пов'язано з тим, що самобутність структури БП у тому, що це пряма архітектура з взаємопов'язаними шарами нейронів, тоді як ШНМ може мати різні архітектури, включаючи пряму, рекурентну та згорткову).

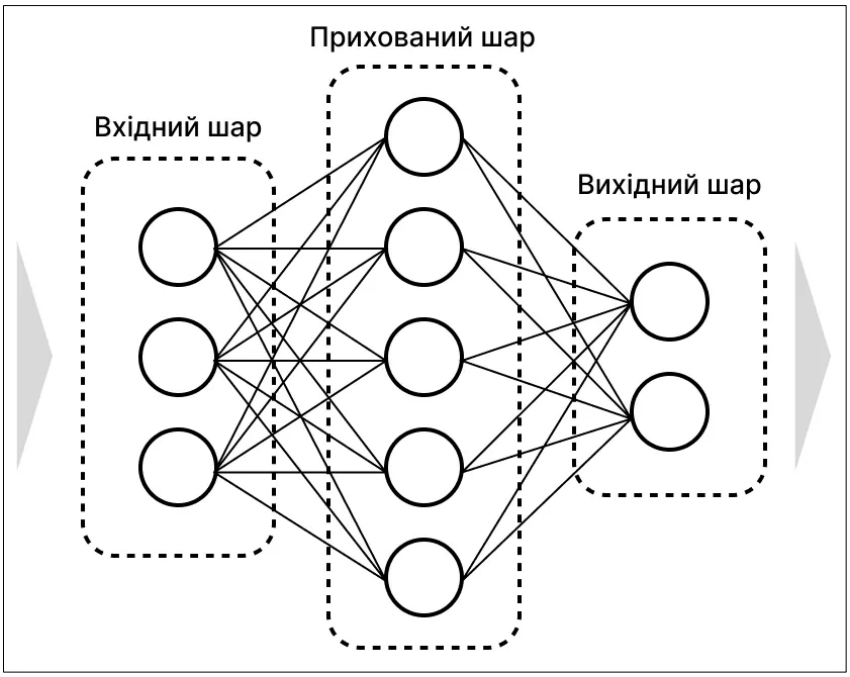


Рис. 48. Схематична модель багатошарового перцептрона.

Окрім того, БП повинен мати щонайменше три шари: вхідний, прихований і вихідний, тоді як ШНМ може мати будь-яку кількість шарів, включаючи одношарові мережі. Показово, що ШН БП в кожному шарі повністю з'єднані з нейронами в наступному, в той час як поєднання в ШНМ можуть бути повністю з'єднані, частково з'єднані або мати цикли. Внаслідок цього, БП добре працює в:

– *розпізнаванні зображень* (ідентифікація облич, об'єктів і сцен на зображеннях);

– *ОПМ* (машинний переклад, аналіз настроїв і класифікація тексту);

– *обробці сигналів* (розпізнавання мовлення, аналіз зображень та фінансове прогнозування);

– *системах управління* (робототехніка й сенсорика, автономні транспортні засоби, ігровий ШІ).

Натомість ШНМ – в різних програмах та застосунках, залежно від конкретної архітектури, включаючи розпізнавання мовлення, робототехніку та ігровий ШІ.

Отже, *сильні сторони БП:*

а) простота й зрозумілість: їх структуру та функції відносно легко осягнути порівняно зі складнішими моделями ГН;

б) гнучкість: вони доволі легко адаптуються до різних завдань, змінюючи кількість шарів і нейронів, вибираючи різні функції активації та використовуючи різні алгоритми МН;

в) фундаменталізм: такий тип перцептронів слугує будівельними блоками для багатьох передових архітектур ГН.

Натомість *слабкі сторони БП:*

– обмежена здатність обробляти складні дані: може мати проблеми з даними високої розмірності, такими як природна мова, або великі зображення;

– обчислювально дорогий: його навчання здебільшого доволі повільне і вимагає значних ресурсів, особливо для великих наборів даних;

в) природа «чорної скриньки»: хоча до певної міри можна інтерпретувати, проте повне розуміння внутрішньої роботи навченого БП є складним завданням.

Мережа на основі радіальних базисних функцій (Radial Basis Function Network) – ШНМ, де прихований шар використовує радіальні базисні функції як активаційні. Вона складається з ШН, що реагують на вхідні вектори довільної розмірності (див. Рис. 49). Для ефективного навчання такої мережі необхідно попередньо визначити центри радіальних функцій (до прикладу, за допомогою методів кластеризації).

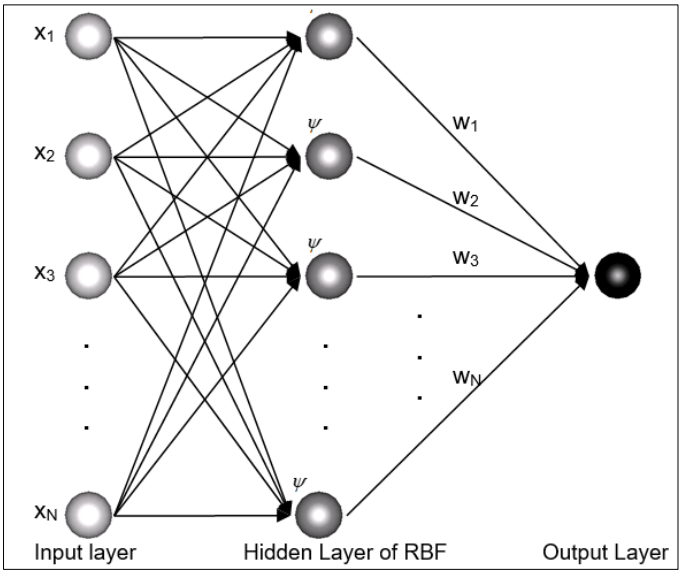


Рис. 49. Схема мережі на основі радіальних базисних функцій.

Здебільшого така ШНМ містить два шари, де перший відповідає за об’єднання вхідних ознак із функцією радіальних базисів у внутрішньому шарі.

Натомість другий шар є вихідним щодо вищезазначених функцій. Цікаво, що в процесі її роботи враховують відстані від точки до центру: цю відстань визначають від центру до точки в площині, аналогічно до визначення радіуса окружності. Модель залежить від радіуса окружності та максимального охоплення при класифікації точок у різні категорії.

2. **До мереж зі зворотними зв'язками належать:**

Самоорганізаційна Карта Кохонена – один із різновидів нейромережевих алгоритмів, який вирішує завдання кластеризації й проєктування багатовимірного простору в простір із нижчою розмірністю (див. Рис. 50). Його часто застосовують для розв'язання найрізноманітніших завдань: від відновлення пропусків у даних до аналізу та пошуку закономірностей.

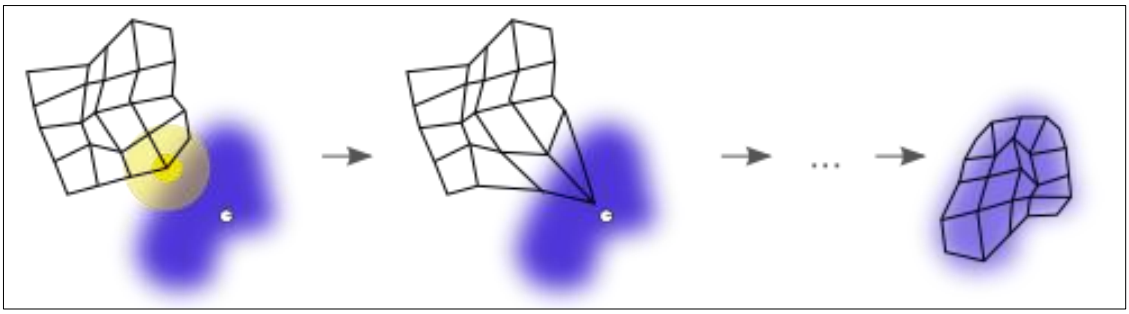


Рис. 50. Ілюстрація МН самоорганізаційної карти.

Такі карти є одним із різновидів нейромережевих алгоритмів, водночас основною відмінністю цієї технології від ШНМ, що навчаються за алгоритмом зворотного розповсюдження, є те, що під час навчання використовують метод навчання без учителя, результат якого залежить тільки від структури вхідних даних (про це частково йшлося на попередніх лекціях, а також поговоримо окремо далі на цій).

ШНМ цього типу часто застосовують для розв'язання найрізноманітніших завдань, від відновлення пропусків у даних до їх аналізу і пошуку закономірностей у них (до прикладу, у випадку аналізу контенту в Х щодо виявлення мови ненависті).

Нейронні мережі Кохонена – клас ШНМ, основним елементом яких є шар Кохонена, який складається з адаптивних лінійних суматорів («лінійних формальних нейронів»). Здебільшого вихідні сигнали шару Кохонена

оброблюють за правилом «Переможець отримує все»: найбільший сигнал перетворюється у 1, інші – обертаються у 0.

Мережа Хопфілда є асоціативною ШНМ (див. Рис. 51): РНМ, запропонована Джоном Хопфілдом у 1982 році, яку використовують переважно для *асоціативної пам'яті* та *оптимізаційних задач*.

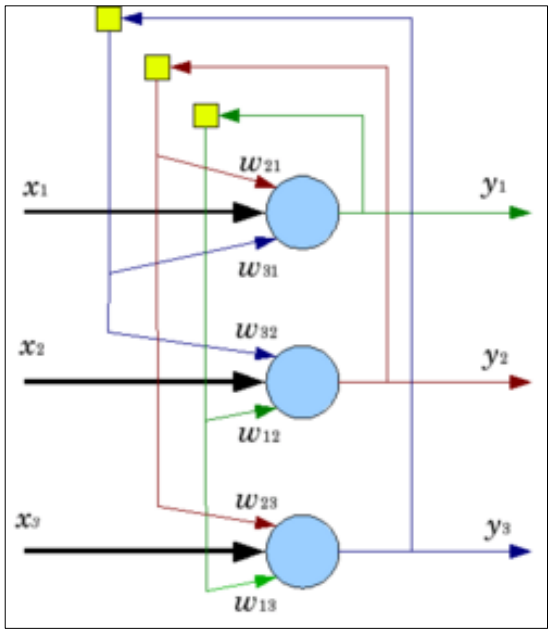


Рис. 51. Схема мережі Хопфілда з трьома нейронами.

Основними характеристиками цієї ШНМ є:

а) вищезазначена *асоціативна пам'ять* (мережа зберігає патерни (образи) у своїх вагах і може відновлювати їх навіть із частково пошкоджених або зашумлених вхідних даних);

б) *рекурентна структура* (на відміну від БП, вона повністю взаємозв'язана: кожен нейрон пов'язаний з усіма іншими (крім самого себе));

в) *біполярні нейрони* (використана біполярна активаційна функція);

г) *симетричні ваги* (ваги зв'язків між ШН є симетричними, що забезпечує наявність енергетичної функції, яка дозволяє описувати динаміку мережі як процес мінімізації енергії).

Таким чином, ця мережа особливо продуктивна у випадку відновлення інформації та оптимізації, проте вона має обмеження щодо ємності та стійкості до помилок.

Моделі ART (Adaptive Resonance Theory) – клас ШНМ, запропонований у 1987 році Стівеном Гроссбергом і Ґейлєм Карпентером. ART-моделі вирішують проблему стабільності-пластичності, тобто те, як ШНМ може навчатися нової інформації, не забуваючи попередньо засвоєної (див. Рис. 52).

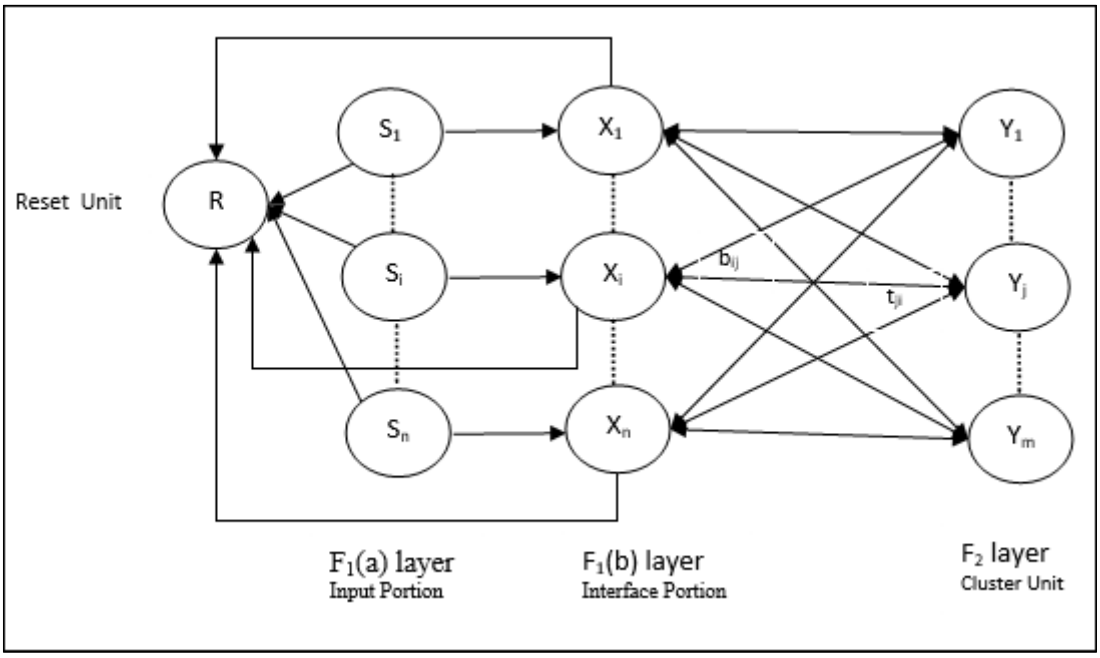


Рис. 52. Схема моделі ART.

Головними особливостями такої моделі є:

– *онлайн-навчання* (ART-мережі навчаються безперервно, не вимагаючи повного оновлення ваг, як у багатьох інших ШНМ);

– *категоризація вхідних даних* (мережа автоматично класифікує нові вхідні зразки, створюючи нові класи, або відносячи їх до наявних);

– *регульована пластичність* (використання параметра ρ (поріг резонансу) дозволяє контролювати чутливість мережі: високе значення цього параметра – детальніший аналіз, натомість низьке – узагальнені дані);

– *резонансний механізм* (перед тим, як мережа навчиться на новому прикладі, вона перевіряє його сумісність з уже сформованими категоріями: якщо новий вхідний зразок не відповідає жодному з наявних, то створюється нова категорія).

Таким чином, ART-моделі унікальні тим, що поєднують *адаптивне навчання* та *категоризацію* без втрати попередньої інформації. Це робить їх

корисними у завданнях, де важливі *стабільність навчання* та *гнучкість адаптації до нових даних*.

Отже, ШНМ прямого поширення відносять до *статичних*, оскільки тут на входи нейронів надходять вхідні сигнали, які не залежать від попереднього стану мережі. Натомість РНМ вважаються *динамічними*, бо за рахунок зворотних зв'язків (петель) входи ШН модифікують у часі, що призводить до зміни станів мережі.

4.3. Навчання штучних нейронних мереж (машинне навчання)

Вищезазначене *МН* – це підгалузь ШІ, базована на використанні алгоритмів, які дозволяють комп'ютерам навчатися на основі даних без явного програмування. При цьому оригінальність ШНМ, як відповідника мозку людини або тварини, полягає у здатності навчатись на прикладах, що складають навчальну множину (наборах даних).

Такий процес (навчання ШНМ) розглядається як налаштування архітектури (топології) та вагових коефіцієнтів синаптичних зв'язків відповідно до даних навчальної множини для ефективного вирішення поставленого завдання. Існують такі *варіанти навчання ШНМ:*

1. Навчання з вчителем (контрольоване навчання).

Більшість реалізацій ШНМ використовують контрольоване МН, де вихід, що змінюється, постійно порівнюють із бажаним виходом (результатом аналізу машиною певних даних із конкретною метою). При цьому вагові коефіцієнти зв'язків на початку встановлюють випадково (ініціалізація мережі). Однак під час наступних взаємодій (ітерацій) їх корегують задля досягнення близької відповідності між бажаним та поточним виходами. Такі методи навчання націлені на мінімізацію наявних похибок усіх елементів обробки, що відбувається завдяки неперервній зміні синаптичних ваг до досягнення прийнятної точності моделі.

Важливо, що перед використанням, ШНМ із контрольованим навчанням повинна бути *навченою*. Фаза навчання, як і *препроцесинг* (підготовка) набору даних для дослідження, займає певний час. Так, МН вважають завершеним лише

у випадку досягнення ШНМ визначеного розробником, тестувальником або користувачем рівня ефективності й бажаної статистичної точності. Прикметно, що після навчання вагові коефіцієнти зв'язків фіксують для подальшого застосування.

Деякі типи ШНМ дозволяють під час використання продовжувати МН, і це допомагає їм адаптуватись до змінних умов. Навчальні множини повинні бути достатньо великими, щоб містити всю необхідну інформацію для виявлення важливих особливостей і зв'язків. Прикметно, що навчальні приклади мають містити широке різноманіття даних, якщо цього потребує дослідження або, навпаки, бути локально зосереджені на певних типах інформації. Згадаємо славнозвісні фотосесії ШІ, які набули популярності у соціальних мережах (зокрема, Тік-Ток), для яких потрібно завантажити не менше 5 фото майбутньої «моделі».

Якщо ШНМ навчається лише з одного прикладу, то вагові коефіцієнти, що старанно встановлені для цього прикладу, радикально змінюються у навчанні для наступного. Попередні приклади при навчанні наступних просто забуваються, окрім певних видів ШНМ, які їх враховують у наступних аналізах. Однак навіть у цьому випадку, чим старішим є певний приклад від поточного, тим більшою є ймовірність того, що він не вплине на нього. Таким чином, система має навчатись всього разом, знаходячи найкращі вагові коефіцієнти для низки прикладів.

До прикладу, в навчанні системи розпізнавання піксельних образів для десяти цифр, які представлені двадцятьма прикладами кожної цифри, всі приклади цифри «сім» не доцільно представляти послідовно. Краще надати мережі спочатку один тип представлення всіх цифр, потім інший тип і так далі. Головним складником успішної роботи мережі є представлення й кодування вхідних і вихідних даних. ШНМ працюють лише з числовими вхідними даними: отже, необроблені дані, що надходять із зовнішнього середовища, повинні перетворюватись.

Важливою є *нормалізація даних* (про неї поговоримо нижче), тобто приведення всіх значень даних до єдиного діапазону. Нормалізацію виконують

шляхом ділення кожного складника вхідного вектора на його довжину, що призводить до набуття ним певного значення (тут – 1). Попередня обробка зовнішніх даних, отриманих за допомогою сенсорів, у машинний формат є спільною й легко доступною для стандартних комп'ютерів.

Якщо після контрольованого навчання ШНМ ефективно опрацьовує дані навчальної множини, важливим стає її продуктивність роботи з тією їх частиною, яка не використовувалась для навчання. У випадку отримання незадовільних результатів для тестової множини, навчання продовжують. Тестування використовують для забезпечення запам'ятовування не лише даних заданої навчальної множини, а й створення загальних образів, що можуть міститись у них.

2. Навчання без вчителя (неконтрольоване навчання).

Неконтрольоване навчання є перспективним, оскільки воно знаменує можливість автономії цього процесу. На сьогодні, таке навчання використовують у ШНМ, що є самоорганізованими картами (self organizing maps; згадаємо мережу Кохонена). Вищезазначені мережі не використовують зовнішніх впливів для корегування своїх ваг і внутрішньо контролюють власну ефективність, шукаючи регулярність або тенденції у вхідних сигналах, та здійснюють адаптацію відповідно до навчальної функції.

Навіть без повідомлення правильності чи неправильності дій, мережа повинна мати інформацію щодо власної організації, яка закладена в її архітектуру та навчальні правила. Таким чином, алгоритм неконтрольованого навчання скеровано на знаходження подібності між групами нейронів, які працюють разом. Так, якщо зовнішній сигнал активує будь-який вузол в групі ШН, то дія (вплив) усієї групи загалом збільшується.

Відповідно, якщо зовнішній сигнал у групі зменшується, це приводить до гальмуючого ефекту на всю групу. Основу для навчання формує конкуренція між нейронами. Так, навчання конкуруючих нейронів підсилює відгуки певних груп на деякі сигнали. Це пов'язує групи між собою та відгуком, також при конкуренції змінюються ваги лише нейрона-переможця.

3. ***Навчання з підкріпленням*** – алгоритм навчається шляхом спроб і помилок, отримуючи винагороду за правильні рішення (до прикладу, роботи, що вчаться ходити).

4. ***ГН*** – це підхід у МН, який використовує багатошарові нейронні мережі, він дозволяє знаходити складні закономірності в даних, що зробило можливими такі досягнення, як:

а) розпізнавання зображень (нейромережі в камерах смартфонів);

б) генерація тексту (ChatGPT, Gemini тощо);

в) аналіз ДНК у біотехнологіях.

Оцінка МН залежить від кількох керованих факторів, а саме:

– *ємності* (показує, скільки зразків може запам'ятати мережа, і які межі ухвалення рішень можуть бути на ній сформовані);

– *складності зразків* (визначає кількість навчальних прикладів, необхідних для досягнення здатності мережі до узагальнення);

– *обчислювальної складності* (безпосередньо пов'язана з потужністю комп'ютера).

Основні етапи розв'язання задач за допомогою ШНМ:

1. Збір даних для навчання.

Вибір даних для навчання ШНМ і їхня обробка є найскладнішим етапом розв'язання задачі. Так, набір даних для навчання повинен задовольняти декільком *критеріям*, а саме – бути:

а) репрезентативним: дані мають ілюструвати поточний стан справ (розробленість проблеми, яку досліджує вчений);

б) несуперечливим (однозначним): суперечливі дані в навчальній вибірці призведуть до поганої якості навчання ШНМ;

в) обсяговим: здебільшого, число записів у вибірці повинне перевершувати кількість зв'язків між нейронами в мережі, інакше – вона просто «запам'ятає» усю навчальну вибірку й не зможе виконати узагальнення.

2. Підготовка й нормалізація даних.

Вихідні дані набудуть вигляду, у якому їх можна подати на входи мережі: так, кожен запис у наборі даних називається *навчальною парою* або *навчальним*

вектором. Навчальний вектор містить по одному значенню на кожен вхід мережі і, залежно від типу навчання (із вчителем, або без), по одному значенню для кожного виходу мережі. Так, навчання ШНМ на «сирому» наборі, зазвичай, не дає якісних результатів, оскільки тоді не відбувається поліпшення її «сприйняття» таких даних.

Серед таких *засобів:*

2.1. Вищезазначене *нормування*: його виконують, коли на різні входи подають дані довільної розмірності. До прикладу, на перший вхід мережі дають набір лексем, які позначають характеристики (холодний, важкий, вологий тощо), а на другий – модальність (ненормальний, пришелепкуватий, дурнуватий і тому подібні).

За відсутності нормування значення на другому вході будуть завжди робити істотно більший вплив на вихід ШНМ, ніж значення на першому вході. Це пов'язано з тим, що такий набір негармонійний, натомість у процесі нормування параметри всіх вхідних і вихідних даних поєднують.

2.2. *Квантування* виконують над безперервними величинами, для яких визначають кінцевий набір дискретних значень. До прикладу, його можна використати для аналізу частот звукових сигналів при розпізнаванні мовлення.

2.3. *Фільтрація*, яку виконують для «зашумлених» даних.

Окрім того, велику роль відіграє саме представлення як вхідних, так і вихідних даних. Припустимо, що ШНМ навчається розпізнаванню букв на зображеннях і має один числовий вихід – номер букви в алфавіті. У цьому випадку вона одержить неправильне уявлення про те, що букви з номерами 1 і 2 подібніші, за букви з номерами 1 і 3, що загалом неправильно. Для того, щоб уникнути такої ситуації використовують архітектуру ШНМ з великим числом виходів, коли кожен вихід має свій зміст. Чим більше виходів у такій моделі, тим більша відстань між класами й тим складніше їх поплутати.

3. Вибір архітектури ШНМ.

Вибирати тип мережі необхідно виходячи з постановки задачі й наявних даних для навчання. Так, для навчання з учителем потрібна наявність для кожного елемента вибірки «експертної» оцінки, тобто або мати промарковані

дані, або зробити це самостійно. При цьому одержання такої оцінки для великого масиву даних просто неможливе.

У цих випадках природним вибором є мережа, що навчається без учителя (до прикладу, вищезазначені самоорганізовані карти Кохонена, або мережа Хопфілда). При розв'язанні інших задач, таких як прогнозування часових рядів, експертна оцінка вже утримується у вихідних даних і може бути виділена у процесі їхньої обробки. У цьому випадку можна використовувати згаданий БП.

4. Експериментальний підбір параметрів навчання ШНМ.

Після вибору загальної структури потрібно експериментально підібрати параметри ШНМ. Так, для мереж, подібних до перцептрона, це буде число шарів, блоків у прихованих шарах, наявність або відсутність обхідних з'єднань, передатні функції нейронів. При виборі кількості шарів і нейронів у них варто виходити з того, що здатності мережі до узагальнення тим вище, чим більше сумарне число зв'язків між нейронами.

З іншого боку, число зв'язків обмежене кількістю записів у навчальних даних. Після вибору конкретної архітектури, необхідно обрати параметри навчання ШНМ: цей етап особливо важливий для мереж, які навчаються з учителем. Від правильного вибору параметрів залежать те, наскільки швидко відповіді мережі будуть сходитися до правильних відповідей.

До прикладу, вибір низької швидкості навчання збільшить час сходження, однак іноді дозволяє уникнути паралічу ШНМ (тупика) або не провокуватиме перенавчання, за якого потрібно починати все з початку. Натомість збільшення тривалості такого навчання може прискорити (навіть уповільнити) досягнення бажаного результату її роботи. Виходячи з такого суперечливого впливу параметрів, можна зробити висновок, що їх значення потрібно вибирати експериментально, керуючись при цьому критерієм завершення навчання (до прикладу, мінімізацією похибки або обмеженням його часу).

5. Власне навчання ШНМ.

У процесі навчання мережа у визначеному порядку переглядає навчальну вибірку: порядок перегляду може бути послідовним, випадковим тощо. Мережі, які навчаються без учителя, переглядають вибірку тільки один раз, натомість у

випадку навчання з учителем для ШНМ передбачено багаторазовий перегляд вибірки. Цікаво, що один такий повний прохід по вибірці називається *епохою навчання*: зазвичай набір вихідних даних поділяють на дві частини – власне навчальну вибірку й тестові дані; принцип поділу може бути довільним.

Навчальні дані подають мережі для навчання, а перевірочні – використовують для розрахунку її похибки (перевірочні дані ніколи для навчання мережі не застосовують). Таким чином, якщо на перевірочних даних похибка зменшується, то ШНМ дійсно виконує узагальнення. Якщо похибка на навчальних даних продовжує зменшуватися, а похибка на тестових даних збільшується, виходить, модель перестала виконувати узагальнення й просто «запам'ятовує» навчальні дані.

Це явище називається *перенавчанням мережі або оверфіттінгом*: у таких випадках навчання, зазвичай, припиняють. Окрім того, у процесі МН можуть проявитися інші проблеми, такі як *параліч або влучення мережі в локальний мінімум поверхні похибок*. Так, неможливо заздалегідь вирахувати прояв тієї або тієї проблеми, так само як і дати однозначні рекомендації щодо її розв'язання.

6. Перевірка адекватності навчання.

Навіть у випадку успішного, на перший погляд, навчання мережа не завжди опановує саме те, чого від неї хотів розробник. Відомий випадок, коли ШНМ навчалася розпізнаванню зображень певних предметів за фотографіями, однак пізніше з'ясувалося, що всі предмети були сфотографовані на тому самому тлі. У результаті мережа навчилася розпізнавати цей тип ландшафту, замість того, щоб набути навичок ідентифікувати предмети на фото. Таким чином, мережа розуміє не те, що від неї було потрібно, а те, що найпростіше узагальнити.

7. Корегування параметрів, остаточне навчання.

Після первинного навчання нейромережі (набору початкових ваг та основного тренування) настає етап *корегування параметрів* (*fine-tuning*) та *остаточного навчання*, які є вирішальними для досягнення максимальної продуктивності моделі. До цього етапу входять:

а) власне, *корегування параметрів* (використання попередньо навченої моделі (замість тренування «з нуля» часто застосовують трансферне навчання, де вже навчена модель адаптується до нових задач), *оптимізація гіперпараметрів* (мовиться про швидкість навчання й опрацювання набору даних), *додаткове донавчання* (у випадку, якщо результат все ще не задовольняє, то використовують додаткові цикли навчання) та *тонке налаштування окремих шарів* (це відбувається шляхом замороження частини шарів, у той час як інші навчають заново);

б) *остаточне навчання* (валідація моделі (перевірка на валідаційному наборі даних, що не використовувалися під час навчання, використання перехресної перевірки (cross-validation) для уникнення вищезазначеного перенавчання));

в) *регуляризація* (*overfitting control*) (для уникнення перенавчання також використовують: вимикання випадкових нейронів у прихованих шарах; нормалізацію активацій всередині моделі; припинення навчання, якщо точність більше не покращується);

г) *фінальне тестування* (оцінка моделі на тестовому наборі, який не брав участі у навчанні);

ґ) *завершальний етап* (інтеграція моделі у виробниче середовище (deployment) і моніторинг продуктивності та можливе повторне донавчання на нових даних).

Отже, корегування параметрів і остаточне навчання – критично важливі для стабільності та продуктивності ШНМ. Так, саме вони дозволяють досягти оптимальної узагальнюючої здатності, уникнути перенавчання та адаптувати модель до нових задач.

8. Вербалізація мережі з метою подальшого використання.

Вербалізація нейромережі – це процес пояснення роботи моделі у зрозумілій для людини формі. Цей етап важливий для переведення абстрактних параметрів моделі (ваг, активацій, правил) у структуровані знання, що можуть бути використані для:

– інтерпретації її рішень;

– впровадження моделі у реальні застосування;

– перевірки правильності роботи та довіри до її рішень.

Основні аспекти вербалізації.

8.1 Формалізація правил ухвалення рішень: перетворення параметрів ШНМ у логічні правила або експліцитні закономірності. До прикладу, якщо мережа розпізнає рукописні цифри, вона може бути представлена як набір ознак (кут нахилу ліній, форма кривих тощо).

8.2. Візуалізація внутрішніх процесів: використання теплових карт для розуміння, які пікселі важливі при розпізнаванні образів; аналіз активацій у прихованих шарах для пояснення, як модель розпізнає структури вхідних даних.

8.3. Спрощення складних представлень: переклад ШНМ у логічні або символічні правила. До прикладу, ГНМ може бути спрощена до дерева ухвалення рішень, або набору лінійних функцій.

Вербалізація для подальшого використання пов’язана з інтеграцією у:

а) *бізнес-процеси:* модель пояснюють у вигляді простих правил ухвалення рішень, а спрощені представлення дозволяють бізнес-аналітикам перевірити модель;

б) *наукові дослідження:* результати пояснюють у формі закономірностей, які можуть бути перевірені іншими методами: так, ШНМ може бути використана для генерації гіпотез у науці; окрім того, цей процес може бути використаний для *контролю прозорості моделі*: пояснення важливих для довіри (зокрема, у лінгвістиці, медицині та фінансах) тез; виявлення можливих упереджень та помилкових закономірностей.

Таким чином, вербалізація ШНМ перетворює її рішення на зрозумілі логічні або візуальні пояснення, що робить останню придатною для використання в реальних задачах, забезпечує прозорість та довіру, а також сприяє покращенню продуктивності та адаптації до нових умов.

Запитання для самоконтролю:

1. Що таке ШІ і які його основні характеристики?

2. Які існують рівні розвитку ШІ?

3. У чому полягає зв’язок між ШІ та ШНМ?

4. Які основні компоненти має ШНМ?

5. Які основні архітектури ШНМ існують? В чому їх особливості?

6. Що таке МН? Які його типи використовують у ШНМ?

7. Чим БП відрізняється від інших ШНМ?

8. Що таке мережі прямого й зворотного поширення сигналу? Наведіть приклади.

9. Як відбувається нормалізація та підготовка даних для навчання ШНМ?

10. Які фактори впливають на ефективність МН ШНМ?

Тема 5.

АНАЛІЗ МЕТОДІВ ОБРОБКИ ПРИРОДНОЇ МОВИ

5.1. Загальне поняття та сфери застосування обробки природної мови

В останні десятиріччя спостерігається значний інтерес до вирішення завдань обробки тексту саме у практичному ключі. Так, із розвитком голосових помічників та чат-ботів все більше зростає необхідність у розумінні та інтерпретації природної мови. Однозначність сприйняття тексту суттєво впливає на процес обробки закладених даних та отримання результатів.

ОПМ – підрозділ ЦТ, ШІ та лінгвістики, метою якого є вивчення проблем комп'ютерного аналізу та синтезу природного мовлення (див. Рис. 53). Повне розуміння та відтворення смислів останнього – надзвичайно складне завдання, оскільки людська мова має цілу низку особливостей.



Рис. 53. Обробка природної мови.

Людська мова – спеціально сконструйована система передачі смислу сказаного або написаного, якій притаманні дискретні, категоріальні або символьні властивості. Така система володіє особливим кодуванням та усвідомленою передачею даних, що вирізняється стійкістю та надійністю. Категоріальні символи мови кодують сигнали для спілкування по декількох каналах (звук, жести, письмо, зображення тощо).

На сьогодні, використання ОПМ зводять до вирішення задач пошуку (письмовий або усний), підбору контекстної онлайн-реклами, автоматичного або напівавтоматичного перекладу, аналізу настроїв, розпізнавання мовлення, чат-ботів та голосових помічників. З розвитком голосових інтерфейсів і діалогових систем, ОПМ стала однією з найважливіших технологій ШІ.

Технології ОПМ давно вирішують вузькі завдання: пошукові системи обробляють запити з урахуванням семантики мови, сучасні перекладачі пристойно відображають загальний зміст фрази іншими мовами, голосові асистенти роблять все менше помилок, а спеціальні архітектури ШНМ дозволяють витягувати актуальну інформацію з довільних текстів.

Саме тому основним завданням в ОПМ у наші дні постає створення універсальних мовних моделей і архітектур, які будуть оперувати з текстом та голосом за допомогою однієї системи End-To-End. Системи, які будуть розуміти інформацію й зможуть взаємодіяти з користувачем подібно до людини з певним багажем знань.

Основні напрямки досліджень ОПМ:

1. Мовні технології:

– розпізнавання мовлення (Speech-to-Text);

– синтез мовлення (Text-to-Speech);

– голосова біометрія (Voice Biometry).

2. Розуміння природної мови (Natural Language Understanding):

а) голосовий інтерфейс (Voice User Interface);

б) діалогові системи, голосові помічники, чат-боти, системи «Питання-Відповідь»;

в) машинний переклад (Machine Translation).

Давайте детальніше розглянемо *мовні технології*.

5.2. Розпізнавання мовлення

Розпізнавання мовлення – процес перетворення мовного сигналу в цифрову інформацію (див. Рис. 54), що є одним із найскладніших технічних завдань, яке стоїть перед системами ШІ. Вищезазначені розпізнавання і синтез мовлення, голосова біометрія увійшли в повсякденне життя й поставлені на комерційний потік.

Так, стрибок у розвитку мовних технологій стався завдяки зменшенню вартості обчислювальних ресурсів та використанню ШНМ для вирішення багатьох завдань. У розпізнаванні мовлення використовують останні досягнення в царині МН, що сприяють високій точності розпізнавання.

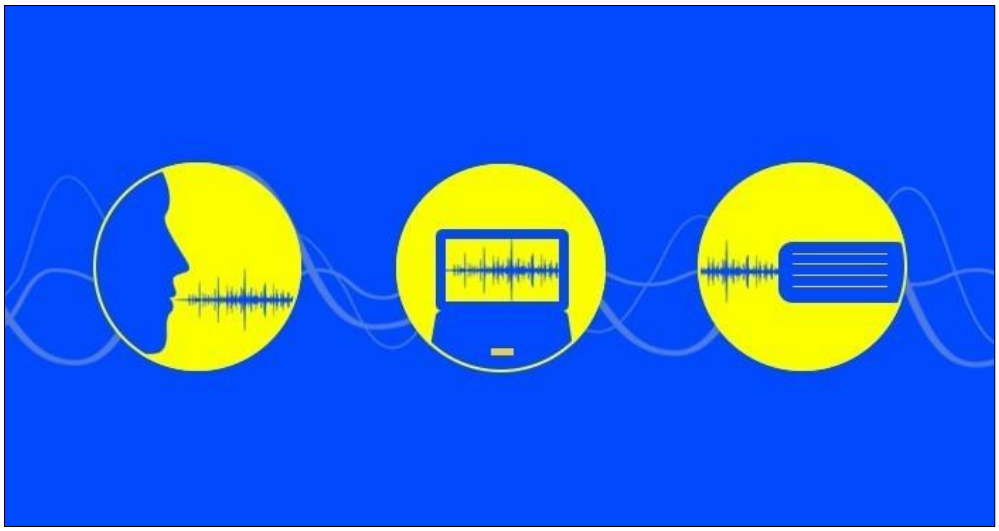


Рис. 54. Розпізнавання мовлення.

Основні сфери застосування машинного розпізнавання мовлення:

– системи голосового обслуговування та інтерактивні автовідповідачі: найбільшого поширення набули в контакт-центрах, сервісах самообслуговування, онлайн-банкінгу (привітання, голосове меню, озвучення стану рахунків тощо);

– ідентифікація особи (використовують великі банки для підтвердження особи користувача за голосовим відбитком, для голосового підпису, а також у системах безпеки);

– аналітика дзвінків і переговорів (призначена для оцінки відгуків клієнтів, підвищення якості роботи операторів, виявлення трендів при зверненнях до служби підтримки, відділів із продажу);

– голосове управління (застосовують у багатьох сферах людської діяльності: в Інтернеті для пошуку інформації, в побуті для керування домашніми приладами, в системах «розумний будинок», навігаторах, управлінні автотранспортом, у мобільних застосунках для виконання певних завдань).

Методи розпізнавання мовлення

На сьогодні сформувалося багато методів розпізнавання мовлення, які можна розділити на *такі типи*:

а) розпізнавання голосових міток та окремих команд, побудоване на роздільному проголошенні й подальшому ідентифікуванні слова або словосполучення на базі невеликого тематичного словника; його використовують у відносно простих системах, призначених для виконання заздалегідь записаних голосових команд;

б) пошук ключових слів: розпізнавання окремих ділянок мовлення, які містять задані слова чи словосполучення; у цьому процесі широко використовують контекст, підлаштування за рахунок додаткових даних (до прикладу, контекст фрази або використання тематичного словника для певної царини), застосовують у пошукових системах, а також у системах моніторингу мовлення;

в) розпізнавання за граматикою: ідентифікування фраз, що відповідають певним заданим правилам (граматиці); широко застосовують у системах голосового самообслуговування (користувача можуть попросити вимовити дату, будь-які номери, прізвища, адреси, підтвердити будь-яку дію словами «так» або «ні»);

г) розпізнавання лексичних елементів мови: найскладніша технологія, яка переводить мову в текст, не обмежуючись при цьому заданою граматикою; в

процесі її реалізації з мовленнєвого потоку виокремлюють звуки й з них складаються слова; ситуацію ускладнює те, що та сама фраза, вимовлена різними людьми в довільних умовах, буде відрізнятись за звучанням й давати несхожі одне на одного сигнали; саме цей підхід і використовують у сучасних системах розпізнавання мовлення, що дозволяє забезпечити високу якість та достовірність цього процесу.

Основні терміни ОПМ, пов'язані з природою мови: так, загальновідомо, що речення складаються зі слів, у той час як слова – зі складів і морфем, а вони, своєю чергою, – з фонем і алофонів.

Морфема – найменша частина слів, що має певне змістовне значення. У такому розумінні морфеми є підґрунтям для словотвору (корінь, префікс, суфікс, постфікс, основа, закінчення, інтерфікс).

Фонема – найменша звукова одиниця мови, фонемою може бути як окремий звук, так і поєднання декількох звуків, що на слух сприймається як один звук. У слов'янських мовах є близько 50 фонем, сенони сталі послідовності фонем (початок, середина або кінець).

Алофон – конкретний мовний звук, зумовлений певним фонетичним оточенням, усі реально вимовлені звуки – алофони, об'єднані у відносно невелике число фонем.

5.3. Попередня обробка звукових сигналів

Попередня обробка мовного сигналу призначена для видалення шумів і сторонніх сигналів, частотний спектр яких знаходиться поза спектром людської мови. Відфільтрований звуковий сигнал піддається аналого-цифровому перетворенню і оцифруванню. Зазвичай, системи розпізнавання мовлення не вимагають для своєї роботи спеціальних пристроїв.

Попередню обробку звукового сигналу можна зробити за допомогою стандартного звукового адаптера, встановленого в пристрої чи комп'ютері. Додаткова цифрова обробка звукового сигналу (до прикладу, частотна фільтрація) може виконуватися центральним процесором комп'ютера.

Важливим етапом попередньої обробки вхідного сигналу є *нормалізація рівня сигналу*. Це дозволяє зменшити похибки розпізнавання, пов’язані з тим, що диктор може вимовляти слова з різним рівнем гучності. Так, якщо вхідний звуковий сигнал має занадто низький рівень гучності, то після нормалізації може з’явитися шум.

Для успішної роботи системи розпізнавання мовлення необхідно відрегулювати оптимальним чином чутливість мікрофона. Втім, надмірна чутливість може привести до спотворень сигналу й до збільшення похибки розпізнавання мовлення.

Зокрема, вимовлений голосовий запит пристрій (комп’ютер чи смартфон) записує й надсилає до сервера. Там запис розділяється на багато маленьких фрагментів – *фреймів* (структур, які описують деякий складний об’єкт або абстрактний образ, або модель для представлення певної концепції).

На кожну секунду мови доводиться сто фреймів, вони тривають по 25 мілісекунд і слідують один за одним, перекриваючи попередній фрейм, як черепиця, щоб інформація на стиках не втрачалася. Спектр сигналу масштабують і, після низки перетворень, для кожного фрейму обчислюють коефіцієнти.

Кожен фрейм піддається певним перетворенням, у результаті яких виходить близько 40 коефіцієнтів, що описують його частотні характеристики. Правильно інтерпретувати оброблені звуки допомагають акустична та мовна ШНМ.

5.3.1. Акустична нейромережева модель

Фрейми надходять на вхід акустичної моделі, яка обчислює розподіл імовірностей, тобто сполучення фонем закладене в кожній частині цього звуку. Цього ШНМ вчиться на великому корпусі начитаних дикторами текстів та їх транскрипцій (сотні годин розміченої мови). Збільшення бази сприяє зростанню якості МН ШНМ, проте кращою є правильна підготовка даних і акуратний запис транскрипцій, що дозволяє з прийнятною якістю навчатися на відносно невеликій базі.

За матеріал для прикладів беруть різноманітні аудіозаписи: офіційні промови на публічних заходах, діалоги з фільмів, аудіокниги, любительські відеоролики в Інтернеті та багато інших різних джерел. Спеціально підготовлені фахівці слухають записи, пишуть, що було сказано в цій промові, і позначають, де були шуми та фонові розмови.

Найскладнішим завданням у розпізнаванні мовлення є *відокремлення основного спікера від фонових*. Це завдання називається «вечірка з коктейлем»: коли багато народу, всі говорять, чутно уривки фраз, гримить музика на фоні, і треба здогадатися, хто є основним спікером, чий голос потрібно розпізнати.

Для розпізнавання особливостей мовлення і, зокрема, акцентів створюють бази текстів від людей, що мають певні особливості вимови. Великої допомоги надає автоматичний збір аудіозаписів від різних сервісів: навігаторів, голосових пошукових запитів, ігрових та розважальних застосунків.

В результаті накопичують максимально можливі варіанти звучання довільної природної мови. Ці дані використовують в навчанні акустичної нейромережевої моделі, що дозволяє їй у підсумку правильно розуміти будь-які акценти, логопедичні недоліки мови та особливості темпу вимови.

Алгоритм такої моделі зіставляє промову з відповідним текстом. Після ознайомлення з певною кількістю матеріалів, і накопичивши достатню статистику збігів, вона визначає, якому звуку з такими характеристиками відповідає яка фонема, іншому звуку – інша тощо.

Рівні розпізнавання мовлення акустичної нейромережевої моделі. Розпізнавання мовлення є багаторівневим процесом; на кожному з трьох рівнів сигнал кодують представниками попередніх рівнів, а саме:

– звукові елементи мовлення (фонеми та алофони);

– склади й морфеми;

– слова, речення та повідомлення.

Під час переходу з рівня на рівень передають і деякі додаткові ознаки, тимчасові залежності й відносини між сигналами. Збираючи сигнали з попередніх рівнів, вищі рівні містять більший обсяг інформації, й можуть здійснювати управління процесами на нижчих рівнях.

Звуки мови є різноманітними, оскільки звучання фонеми залежить від того, в якій частині слова вона знаходиться (на початку, усередині, наприкінці), а також того, які фонеми її оточують. До прикладу, [а] між двома голосними у поєднанні «на аудіозаписі» відрізняється від [а] між приголосними в слові «бак». Тому, для якісного розпізнавання фонема є занадто грубою одиницею.

Відтак, для точнішого моделювання вимови певної фонеми:

1. Кожну фонему поділяють на три частини: умовні початок, середину й кінець (вищезазначені сенони).

2. Використовують фонетичний алфавіт, який враховує позицію й контекст фонем. Такий алфавіт складається з тисяч елементарних одиниць різноманітних поєднань і комбінацій. Власне, з цим набором і працює технологія розпізнавання мовлення.

Прикметно, що для розпізнавання фонем та алофонів застосовують нейромережеві моделі та їх ансамблі. До останніх застосовують вищезазначене навчання без вчителя, при якому відбувається статистична обробка всіх сигналів, що надходять на вхід такої мережі. Ансамблі формують відповідно до найчастотніших сигналів, а запам’ятовування рідкісних сигналів відбувається пізніше й вимагає контролю вищого рівня.

В алгоритмах розпізнавання також закладається *голосова біометрія* – класифікатор, який навчають на багатьох різних розмічених аудіозаписах із певною ймовірністю ідентифікувати особу диктора (стать, вік тощо). Звичайно, іноді бувають помилки – голоси деяких жінок схожі на дитячі, а у літньої людини – може бути молодий, бадьорий голос.

5.3.2. Мовна нейромережева модель

Якщо людина не може зрозуміти, або не розчула всі звуки, тоді вона добудовує слово, виходячи з контексту. Для цього вона спирається на власний мовний досвід: якщо співрозмовник скаржиться на закладений ніс, можна зрозуміти, що у нього «нежить», а не «дежить» абощо. Аналогічно працює система розпізнавання, тільки замість мовного досвіду вона використовує мовну нейромережеву модель.

Остання описує відповідну тематичну область та містить загальні відомості. До прикладу, для навігатора – тексти на географічну тематику (адреси, назви організацій і так далі), для пошуку в Інтернеті – тематику пошукових запитів (все, що люди запитують у пошуковому рядку) тощо. Для того, щоб знати все різноманіття слів та словоформ, наголоси, стійкі та поширені поєднання слів, контекст вживання, мовна модель навчається на текстових корпусах (терабайтах даних).

Під час навчання мовна нейромережева модель звертає увагу не стільки на відповідність звуків і фонем, скільки на *їхні послідовності* – тобто на *слова й фрази, які з них формуються*. Така модель працює з ланцюжком імовірних фонем. Так, акустична нейромережева модель, після обробки частотних ознак фрейму, видає не одну конкретну фонему, а кілька – і в кожної з них свій коефіцієнт імовірності. До прикладу, в разі «нежиті» акустична модель надає дві приблизно однакових імовірних фонеми на початку слова – [н] і [д].

Натомість мовна модель призначена вирішити цю суперечність: так, під час навчання слово «дежить» зустрічалося їй нечасто, а от слово «нежить» – багато разів. Відповідно, модель зробить висновок, що на початку слова, найімовірніше, знаходиться фонема [н]. Приблизно так само, виходячи з контексту, мовна нейромережева модель визначає послідовності слів. До прикладу, вибираючи між «ніч яка місячна» і «нічья комісячна», вона надасть перевагу першому словосполученню, оскільки воно знайоме їй з навчального корпуса (набору даних) текстів.

5.4. Синтез мовлення

Синтез мовлення (перетворення тексту в голос) на сьогодні застосовують у найрізноманітніших сферах: голосових асистентах, IVR-системах (Interactive Voice Response), «розумних будинках» та інших (див. Рис. 55). Написаний текст повинен вимовлятися природно, тобто правдоподібно до людської вимови. В напрямку синтезу голосу пройдено довгий шлях і застосовано багато методів, першим із яких був примітивний підхід склеювання окремих звуків у слова.

Метод «Unit-Selection»

Якщо набір текстів, які треба озвучити, є відносно невеликим і там зустрічаються повторні висловлювання (до прикладу, в оголошеннях про відправлення й прибуття поїздів на вокзалі), достатньо запросити диктора, записати в студії потрібні слова й фрази, а потім «збирати» з них потрібне повідомлення.

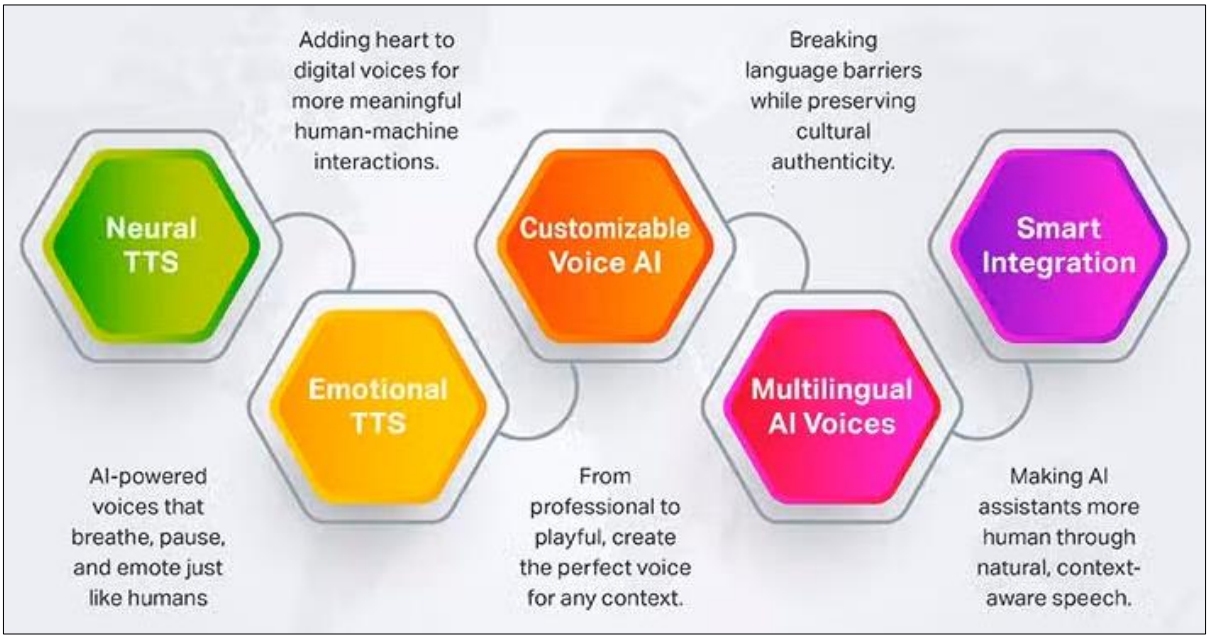


Рис. 55. Синтез мовлення.

Природно, що диктор у студії наговорює багато текстів, усі промовлені слова розділяють на елементарні складники (до прикладу, на фонеми). Після цього складають велику базу даних: як звучить кожна фонема, залежно від того слова, в якому вона знаходиться. Таким чином, коли виникає завдання озвучити якийсь текст, із цієї бази збираються й склеюються звуки.

На коротких текстах це працює пристойно: стандартні фрази: «Вас вітає компанія *така-то*» або «Сьогодні погода *така-то*». Проте це не зовсім синтез мовлення, а просто відтворені частинки реальної мови. До всього, такий підхід не працює з довільними текстами: до прикладу, якщо таким синтезом читати щось довге (рекламне оголошення на два речення).

Тоді у синтезі починає спотворюватися інтонація або голос, оскільки немає алгоритму, який відповідає за правильне об’єднання інтонації на стиках

змістовних фрагментів фраз чи речень. Так, синтез довгого аудіотексту (до прикладу, озвучення книги) швидко починає дратувати, оскільки інтонація є монотонною й людина втомлюється її слухати. На сьогодні, для синтезу мовлення використовують МН та ШНМ, що надають достатню якість і наближений до людського синтезований голос.

Статистичний (параметричний) метод

Статистичний (параметричний) синтез мовлення є аналогічним до розпізнавання, тільки відбувається в зворотному напрямку. Є акустична модель, яка навчається на даних – записах мовлення з текстовими транскрипціями. Для збору даних диктор читає текст, проте в значно меншому обсязі.

До прикладу, записують 600 речень, промовлених непрофесійним диктором за один день у студії, здійснюють синтез і отримання голосу. Статистичний метод дозволяє швидко розширювати систему, натомість у попередньому варіанті для додавання нового голосу потрібно створити окрему мовну базу. Для цього диктор мав записати багато годин мовлення, при цьому бездоганно витримуючи однакову інтонацію і гучність.

Ці недоліки усунені у статистичному методі, оскільки для створення нового голосу достатньо записати хоча б дві години мовлення – приблизно тисяча вісімсот спеціальних, фонетично-збалансованих речень. Такий метод має безсумнівні переваги:

а) швидкість відтворення різних голосів;

б) завжди чітка й однакова інтонація на будь-якій довжині тексту.

Голос у такому синтезі виходить із роботизованим забарвленням, хоча з розбірливістю та інтонаціями все добре. Саме тому актуальним завданням є знайти алгоритми, які допоможуть зробити синтезований голос максимально натуральним.

5.4.1. Етапи створення системи синтезу

Для створення системи синтезу мовлення, потрібно виконати багато завдань, що потребує команди фахівців із різних напрямків. З кожного етапу реалізації існує ціла маса алгоритмів і підходів. Зокрема, вищезазначена у попередніх лекціях *нормалізація* тут означає використання спеціального

алгоритму, який готує текст, щоб роботу було зручно його читати: розгорнути всі скорочення, перевести числа й дати в текст.

Так, «50-ті роки XX століття» мають перетворитися в «п'ятдесяті роки двадцятого століття», а «м. Київ, просп. Т. Шевченка» в «місто Київ, проспект Тараса Шевченка». Це повинно відбуватися природно: наче людиною, яку попросили прочитати написане. Далі текст поділяють на фрази, тобто на словосполучення з безперервною інтонацією – для цього комп'ютер орієнтується на знаки пунктуації та стійкі конструкції. Для всіх слів складається фонетична транскрипція.

Підготовка словника наголосів. Розставляння наголосів може проводитися за правилами сучасної літературної мови: так, в англійській наголос часто ставиться на перший склад, а в іспанській – на передостанній. При цьому обов'язково потрібно враховувати цілу низку виключень, які не відповідають загальним правилам. Так, для української мови в загальному сенсі правил щодо наголосів узагалі не існує, тому наявність словника з розставленими наголосами є обов'язковою.

Таким чином, щоб зрозуміти, як читати слово й де поставити в ньому наголос, синтезатор спочатку звертається до класичних, складених вручну словників, які вбудовані в систему. Якщо потрібне слово в словнику відсутнє, синтезатор будує транскрипцію самостійно – спираючись на правила, які запозичено з академічних довідників.

Нарешті, якщо звичайних правил виявляється недостатньо – а таке трапляється, адже будь-яка жива мова постійно змінюється, – він використовує статистичні правила. Якщо слово зустрічалося в корпусі навчальних текстів, система запам'ятовує, на якому складі в ньому зазвичай наголошували диктори.

Зняття омографів – слів, які збігаються в написанні, проте різняться у вимові: так, носій мови легко розставить наголоси: «дверний замок» і «замок на горі». А ось «ключ від замка» є значно складнішим завданням, адже повністю «зняти» омографи без урахування контексту неможливо.

Виокремлення синтагм і розстановка пауз. Синтагма – слово або група слів, що представляють цілісний відрізок мовлення, який несе завершену за

змістом думку. Так, коли людина говорить, вона зазвичай вставляє паузи між фразами, саме тому в синтезі мовлення необхідно чітко розділяти текст на такі синтагми.

Автопунктуація. Диктуючи що-небудь, в отриманому тексті варто, щоб автоматично розставлялися знаки пунктуації. Роль розділових знаків у мовленні грають інтонаційні паузи, саме тому створюються алгоритми для їх розпізнавання. Паузи між словами отримують одну з міток: пробіл, крапка, кома, знак питання, знак оклику, двокрапка. Для того, щоб передбачити, яка мітка відповідає конкретній паузі, використовується контекст та враховуються три попередніх і два наступні слова.

Ці нескладні правила дозволяють із досить високою точністю розставляти знаки. Кожному знаку пунктуації призначається фонема, і з точки зору системи це є нові «слова», що складаються з «пунктуаційних» фонем – там, де виникали паузи або певним чином змінювалася інтонація.

Визначення типу інтонації. Вираз завершеності, питання й вигуки – найпростіші інтонації, однак висловлення іронії, сумніву, наснаги тощо є складнішим завданням. Дослідження голосових характеристик надають можливість використовувати ці особливості для синтезу голосу.

Синтезатор через характеристики звукових хвиль спроможний розпізнати емоційне забарвлення: сум, розпач, радість, подив тощо. Для цього тексти позначають мітками з відповідною емоцією. Аби модель голосу змогла побудувати алгоритми, які відповідають різним емоціям, треба правильним чином її навчити. Так, під час запису диктор вимовляє свої репліки по черзі нейтральним голосом, радісним та роздратованим. В процесі МН система виділяє і описує параметри й характеристики голосу, що відповідають кожному з цих станів.

Отримання транскрипції. Для того, щоб прочитати підготовлений текст, використовується вищезазначена особливого роду акустична нейромережева модель, яка відрізняється від акустичної моделі, застосовуваної при розпізнаванні мовлення.

Оскільки тут важливішим є промовляння, а не написання, то замість букв (графем), логічно використовувати звуки (фонеми), зокрема:

– у випадку з розпізнаванням мовною нейромережевою моделлю потрібно встановити відповідність між звуками з певними характеристиками і фонемами;

– у випадку із синтезом мовлення акустична модель, повинна, навпаки, за описами фонем скласти описи звуків.

Перетворення графемного запису у фонемний – окреме завдання, що складається з багатьох правил і винятків. Так, оригінальний текст передають у модуль G2P (Grapheme-To-Phoneme), де відбувається перетворення останнього в послідовність фонем. Далі, коли транскрипція готова, модель розраховує, як довго буде звучати кожна фонема, тобто скільки в ній фреймів.

Потім кожен фрейм описується за багатьма *параметрами:*

а) частиною якої фонеми він є, яке місце в ній займає;

б) в який склад входить ця фонема;

в) якщо це голосна, то чи на ній є наголос;

г) місце, яке вона займає в складі, склад у слові, слово у фразі, фраза в реченні;

ґ) розділові знаки, наявні до й після цієї фрази;

д) знак, який стоїть наприкінці речення, і яка його головна інтонація.

Таким чином, для синтезу кожних 25 мілісекунд мовлення використовують багато даних. Так, інформація про найближче оточення забезпечує плавний перехід від фрейму до фрейму й від складу до стилю, а дані про фразу й речення потрібні для створення правильної інтонації синтезованого мовлення.

ШНМ намагається відтворювати унікальні характеристики голосу людини для впізнаваності синтезованого результату (голосу). При цьому акустична нейромережева модель видає числа, що задають параметри звукової хвилі, а спеціальний модуль «вокодер» (далі згадаємо його) отримує на вхід послідовність чисел і перетворює їх в звук, у мову. Після цього модель, отримавши дані про те, як розмовляє певна людина, спроможна читати вголос будь-який текст доволі подібно до неї.

Природно, що впізнати голоси, першою чергою, можна за тембром, який залежить від особливостей будови органів мовного апарата кожної людини (до прикладу, голос Володимира Зеленського доволі впізнаваний). Тембр голосу можна змоделювати, тобто описати його характеристики: для цього достатньо начитати в студії невеликий корпус текстів.

Після цього дані про тембр можна використовувати при синтезі мовлення, навіть мовою, що невідома диктору. Для фізичного промовляння тексту використовують вищезазначений генератор звукових хвиль – *вокодер*. У нього завантажується інформація про частотні характеристики фрази, що отримана від акустичної нейромережевої моделі, а також дані про тембр, який надає голосу впізнаване забарвлення.

Обчислення параметрів інтонації. Тембр такого голосу дещо «комп'ютерний», зате у нього природні й плавні інтонації. При цьому гладкість мовлення не залежить від обсягу й довжини тексту, що промовляється, а голос легко налаштовувати. В параметрах запиту можна задати ключ, і модуль синтезу буде продукувати голос із відповідним емоційним забарвленням.

Потрібно вирішити, як буде змінюватися висота основного тону й швидкість промовляння залежно від розставлених пауз, підібраної послідовності фонем і типу вираженої інтонації. Окрім основного тону й швидкості є інші параметри, з якими можна довго експериментувати. Чим більше даних, на яких навчається модель, тим краще її вимова й інтонування.

Модифікація і звукові ефекти. Для надання більшої природності застосовують постобробку, спеціальні фільтри, що наближають синтезоване мовлення до людського та виправляють виявлені дефекти.

5.4.2. Застосування штучних нейронних мереж

На сьогодні, в сфері синтезу мовлення, як і в багатьох інших царинах, широко застосовують МН. Так, цілу низку компонентів всієї системи можна замінити на ШНМ, що дозволить не просто наблизитися за якістю до наявних алгоритмів, а навіть значно їх перевершити. Популярними архітектурами для синтезу є вищезазначені РНМ. На них вирішують усі завдання, пов'язані з послідовностями: голосовими, текстовими або музичними.

Сучасні голосові помічники навчають говорити не буквами, а фразами. Відразу змусити нейромережу цілком видавати їх (фрази) не вийде, адже тоді їй треба буде запам’ятати всі фрази в мові, розмір яких є велетенським. Тут на допомогу приходить те, що текст, мова або музика – послідовності. Кожне слово чи звук – самостійна одиниця, однак вона залежить від попередніх і наступних.

Так, досить легко навчити ШНМ вимовляти окремі слова та букви. Для цього беруть низку аудіофайлів, розмічених на слова, і навчають її за вхідним словом видавати послідовність сигналів, схожих на його вимову. При цьому порівнюють з оригіналом від диктора, намагаючись максимально наблизитися до ідеалу: для цього підійде навіть вищезазначений перцептрон.

Втім, перцептрон не запам’ятовує, що він генерував раніше (для нього кожен запуск буде, як вперше), тому з’явилася ідея додати до кожного ШН пам’ять. Так були винайдені РНМ, в яких кожен нейрон запам’ятовував усі свої попередні відповіді й при наступному запуску використовував їх як додатковий вхід. Тобто, нейрон міг «сказати» самому собі в майбутньому, як повинен звучати наступний звук.

Першою чергою, це спродукувало іншу проблему: внаслідок того, що кожен нейрон запам’ятовував усі минулі результати, в мережі утворювалася надвелика кількість входів. Своєю чергою, це призводило до неможливості навчання такої кількості зв’язків: адже, якщо ШНМ не вміє забувати, то її й не можна нічого навчити.

Спочатку проблему вирішили просто – відімкнули для кожного нейрона пам’ять. Проте згодом придумали як для неї використовувати спеціальні комірки, що схожі на пам’ять комп’ютера або регістри процесора. Кожна комірка дозволяла записати в себе числа, прочитати чи скинути: їх назвали комірками згаданої ДКП.

Таким чином, коли ШН потрібно було поставити собі «нагадування» на майбутнє, то він просто заносив це в комірку. Натомість, коли, навпаки, вся історія ставала непотрібною (до прикладу, речення закінчилося), то такі комірки просто «скидалися», залишаючи тільки «довготермінові» зв’язки, як у класичному перцептроні.

Іншими словами, мережа навчалася не тільки встановлювати поточні зв’язки, а й ставити нагадування. Озвучені тексти для навчання почали брати звідки завгодно. Так, на прикладі багатьох відеозаписів із виступами Барака Обами непогано навчили ШНМ розмовляти його голосом, що засвідчило потенціал сучасних синтезаторів до імітації голосів. Природно, що ви неодноразово чули озвучення класичних українських поетичних творів ШНМ (до прикладу, канал в Telegram «Нейробандура»).

5.5. Голосовий інтерфейс

Голосові інтерфейси (Voice-User Interface) – програмний продукт, який за допомогою голосової або мовної платформи дозволяє взаємодіяти користувачу й комп’ютеру, запускаючи автоматизовані процеси. Завдання таких інтерфейсів – розпізнати голос людини й виконати певне завдання. Так, показовою у цьому контексті є концепція «розумного будинку», у якому можна роздавати голосові команди (максималізували цю ідею в серіалі від Netflix – «Касандра»).

Природно, що голосові інтерфейси зручні, якщо у вас немає змоги в той чи інший момент вводити текст, проте часто-густо останні можуть написати щось доволі дивне. Разом із тим, такі інтерфейси продуктивні, скажімо, під час водіння автомобіля, коли користувач може проговорити свій запит, продиктувати потрібну адресу, перевірити завантаженість маршруту в додатку навігатора.

Складники голосового інтерфейсу:

– голосове введення: запити вимовляють голосом, а не вводять за допомогою клавіатури чи графічних елементів екранного інтерфейсу;

– природне мовлення: користувачі не повинні обмежуватися використанням певного, оптимізованого для комп’ютера словника або синтаксису й можуть структурувати речення будь-якими способами, подібно до розмови з людиною;

– голосове виведення: результат виконання вимовляють голосом, а не виводять текстом на екрані;

– інтелектуальна інтерпретація: для справжнього розуміння запитів користувача голосовий інтерфейс повинен використовувати додаткову інформацію, таку як контекст використання та попередні дії користувача;

– сприяння: голосовий інтерфейс вчиняє дії, необхідні для виконання завдання користувача, які користувач не запитував.

Голосовий інтерфейс із вищезазначеними п’ятьма (голосове введення, природне мовлення, голосове виведення, інтелектуальна інтерпретація, сприяння) властивостями надає взаємодію з *двома суттєвими перевагами*:

а) формулювання мети природною мовою, завдяки якій зникає необхідність у вивченні інтерфейсу й натисканні кнопок;

б) передбачення цілей користувача, наслідком чого стає пропозиція додаткових дій на основі контекстної інформації, або попередніх кроків.

5.5.1. Голосове управління

Голосове управління – вербальний процес спілкування з машиною. В хорошому голосовому інтерфейсі таке спілкування має бути таким же природним, як і з людиною. Саме тому в проєктуванні таких систем значно більше психології та розуміння особливостей людського мислення.

Так, розробники при проєктуванні мають дотримуватися таких принципів, аби уникнути відторгнення таких технологій у широкому спектрі: від несприйняття логіки до ефекту «зловісної» долини. Максималізували ідею такого спілкування у фантастичній драмі «Вона», де Скарлет Йохансон озвучила операційну систему, з якою в героя Хоакіна Фенікса зав’язується роман.

Зокрема, мовиться про такі *принципи*:

1. *Довіра* – не технічне питання, проте без неї користувач просто не буде послуговуватись голосовим інтерфейсом для виконання більш-менш значущих завдань (до прикладу, для оплати товару, бронювання квитків тощо). Так, якщо людина не розуміє ступінь можливої помилки системи, то вона її просто не використовує, не кажучи вже про старше покоління, у представників якого ЦТ викликають побоювання.

2. *Невидимий інтерфейс.* Невидимість – принципова відмінність голосового інтерфейсу: так, користувач не бачить його елементи, в якій частині

останнього, або на якому етапі він знаходиться. Зокрема, у кожної людини свої уявлення про можливості системи та її логіку роботи, а тому реакція системи на дії людини допомагає їй корегувати свої запити потрібним чином.

Відтак, коли питання та відповіді на них обмежені (до прикладу, так або ні), то користувач може зробити висновок, що система доволі примітивна. Отже, всі подальші команди й відповіді він буде формулювати відповідним чином. Якщо ж система ставить питання й розуміє відповіді користувача, то він (користувач) буде формулювати їх як завгодно.

3. *Людяність* – для того, щоб взаємодія з голосовим інтерфейсом була природною, треба зробити систему, яка, отримуючи зворотний зв'язок, сама буде визначати, що зроблено правильно, і корегувати характеристики, які є важливими для природної взаємодії.

4. *Індивідуальність* – сучасні реалізації голосового інтерфейсу дозволяють імітувати його характер, серед його проявів дружність, почуття гумору, інтелектуальність та інше. Це досить різнопланові характеристики, і підхід різних компаній до їх реалізації відрізняється.

Так, Siri – проєкт компанії Apple, ідеологія якої базується на тезі про те, що все повинно просто працювати: і все дійсно працює, якщо користувач вгадує з граматикою й словником. Якщо не вгадує, то система, без жодних пояснень про те, що пішло не так і як скорегувати поведінку, просто перестає працювати. При цьому великий акцент зроблено на індивідуальності: якість голосу, жарти, смішні коментарі при виконанні звичайних завдань вражають.

Користувач розслабляється й намагається взаємодіяти із Siri як з особистістю, однак, якщо коли система починає реагувати не так, як він очікує, сприйняття різко змінюється. Користувач думає, що його дії не схвалюються, або над ним просто сміються, і це значно гірше, ніж якби він спочатку сприймав її як машину. Саме тому в Google роблять акцент на презентацію, першою чергою, високотехнологічного програмного продукту, якому від початку навіть не надали імені, а вже потім він отримав назву Gemini.

Розуміння природного мовлення

На етапі розуміння природного мовлення пристрій має справу вже не із самою мовою, а з текстом, у який вона була перетворена. Системи розпізнавання мовлення можна використовувати не тільки в голосових інтерфейсах, а й для створення чат-ботів, здатних спілкуватися або відповідати на листи.

Звичайно, самі по собі слова для комп'ютера нічого не означають. Подібно до людини, яка бачить перед собою текст, написаний невідомою мовою, не може уявити, про що йдеться, якщо тільки не побачить знайомі слова. Приблизно таке ж завдання виконує комп'ютер – за знайомими словами й формулюваннями зрозуміти вимовлене.

Для того, щоб зрозуміти, чого хоче користувач, і активувати потрібну функцію, система має впізнати в мовленнєвому потоці ключові слова (до прикладу, «виклич таксі»). Цих слів її навчив розробник: вони занесені в програму серед інших формулювань виклику таксі: «замов таксі», «виклич машину», [...] «хочу поїхати додому» тощо. Природно, всіх варіантів розробник урахувати не може. Саме тому, зібравши тестову версію застосунку, пропонується покористуватися ним багатьом людям, аби з'ясувати, як ще вони можуть замовляти таксі.

Всі нові формулювання записуються в конфігураційний файл програми, а потім розмічаються: в них визначають слова-маркери й питомі синтаксичні схеми. Аналізуючи розмітку, програма сама вчиться розуміти навіть такі формулювання, яких немає в списку, однак вони схожі на відомі їй конструкції. До прикладу, якщо в списку є варіанти «виклич машину» і «замов таксі», то й формулювання «замов машину» програма теж зрозуміє.

5.5.2. Модуль голосової активації

Другий ключовий складник голосового інтерфейсу – це *система активації голосом*, яка запускає потрібну дію у відповідь на ключову фразу. Розробники Google для розпізнавання коронної фрази «Окей, Google» використовують ГНМ, яка надає високу якість. Проте система активації при цьому обмежена єдиною командою, й для навчання потрібна величезна кількість даних. Модель для

розпізнавання всім знайомої фрази навчалася на прикладі більш ніж 40 тисяч голосів користувачів, які зверталися до своїх смартфонів із Google Now.

В інших голосових інтерфейсах застосовують різні підходи. До прикладу, користувач самостійно обирає ключову фразу з наявної бібліотеки, або вигадує свою. Розпізнавання команди повинно відбуватися на самому пристрої, без звернення до сервера. Для системи активації потрібен зовсім невеликий словник, адже їй достатньо розуміти кілька ключових фраз, а всі інші можна просто ігнорувати.

Тому мовна модель активації є компактною: якщо повноцінна модель розпізнавання може мати десятки мільйонів станів, що описують як чітке мовлення, так і тишу, сторонній шум, фонові розмови та займати гігабайти, то для голосової активації вона обмежена сотнями станів і вміщується в кілька десятків кілобайт.

Модель для розпізнавання ключової фрази має враховувати особливості вимовляння навіть і при невеликому словнику. Команда повинна бути відносно довгою (бажано – більше одного слова) і нечасто зустрічатися в повсякденній мові, щоб виключити помилкові спрацьовування.

«Будь ласка» погано підійде для голосової активації, а «слухай мою команду» – цілком. У систему вбудовано детектор голосової активності, який стежить за появою ключової фрази у вхідному звуковому потоці людського голосу. Решта звуків ігнорується, тому в фоновому режимі енергоспоживання модуля активації обмежене тільки мікрофоном.

5.5.3. Виокремлення змістовних об’єктів

Слова, які вимовляє людина, важливо не лише перевести в літери, а й наповнити змістом. Ця технологія починає працювати вже після того, як вимовлені слова розпізнано. Повний набір мовних технологій містить модуль виокремлення змістовних об’єктів у природному мовленні, і на виході отримується не просто розпізнаний, а вже розмічений текст.

Як змістовні об’єкти виокремлюються дати, час, адреси, імена, прізвища, географічні назви. У гібридній системі поєднують контекстно-вільні граматики, словники ключових слів і статистичні дані пошуку й різних сервісів, а також

алгоритми МН. До прикладу, у фразі «поїхали на головний залізничний вокзал» слово «вокзал» допомагає системі визначити контекст, після чого в базі даних знаходиться відповідний об’єкт.

На вхід системи подається список різних типів об’єктів і приклади фраз з живої мови, що їх описують. Далі із цих прикладів формуються *патерни*, в яких враховуються початкова форма, корінь, морфологічні варіанти слів. Наступним кроком надаються приклади вживання обраних об’єктів у різних поєднаннях, які допоможуть системі розуміти контекст. На основі цих прикладів будується модель, де виокремлені в репліці користувача об’єкти стають станами, що відстежуються.

Запитання для самоконтролю:

1. Що таке ОПМ? Які основні завдання вона вирішує в сучасних ЦТ?
2. Які є основні напрями досліджень у сфері ОПМ? Наведіть приклади для кожного з них.
3. Які методи використовують для розпізнавання мовлення? У чому їх відмінність і які сфери застосування кожного?
4. Що таке акустична ШНМ? Як вона використовується в процесі розпізнавання мовлення?
5. У чому полягає роль мовної ШНМ у розпізнаванні значення фрази або слова в конкретному контексті?
6. Що таке фрейми у процесі попередньої обробки мовного сигналу? Яку інформацію вони несуть?
7. Які етапи потрібно пройти для створення системи синтезу мовлення на базі ШНМ?
8. У чому відмінність між методами синтезу мовлення “unit-selection” та статистичним (параметричним) методом?
9. Яким чином система обробки мови працює з омографами, наголосами та пунктуацією під час синтезу мовлення?
10. Що таке вокодер? Яку роль він виконує у фінальному етапі синтезу мовлення?

Тема 6.

ПАРАМЕТРИЗАЦІЯ ОБРОБКИ ПРИРОДНОЇ МОВИ

6.1. Підходи до обробки природної мови: спрощений та розширений

У попередній лекції було з’ясовано, що ОПМ дає змогу машинним технологіям сприймати та генерувати мови, якими ми спілкуємося. В своїй основі ОПМ, як і будь-яке програмування, базована на чітких алгоритмах, визначених структурах даних і твердій об’єктивності, яка дає змогу ефективно обробляти будь-яку інформацію. Окреслені алгоритми ОПМ керуються чіткими законами та взаємовідносинами (ієрархія, наступність, логіка), при цьому обробку вищезазначених даних базують на законах та методах математики, статистики, логіки.

Водночас на попередній лекції було зроблено висновок про те, що процес сприйняття природної мови людською свідомістю не підлягає чіткій алгоритмізації. Це пов’язано з тим, що людське мислення оперує як аналітичними здібностями, так і абстрактними зв’язками. Прикладом першого може бути формування речення з чітким порядком / відмінюванням слів з допомогою підстановки необхідних суфіксів та закінчень до кореня.

На сьогодні, найпродуктивнішими парадигмами є *спрощена* (робить мову простішою) та *розширена* (розглядає мову як суцільний об’єкт). Тут важливо, що здебільшого ОПМ втілювана за рахунок мов програмування, найпростішою для використання серед яких є Python. У межах першого підходу доцільно згадати, що лінгвістична екосистема цієї мови програмування блискуче показує себе у низці завдань, що можна описати в певних межах, правилах.

Так, аналіз різноманітних бібліотек ОПМ (Pandas, Spacy, NLTK, Scikit-learn тощо) дозволяє стверджувати, що кожна з них може виконати різноманітні завдання, які допомагають *спростити мову до обмеженої системи*, а саме:

– поділ на токени (токенізація (tokenisation));

– класифікація частин мови (part of speech tagging);

– лематизація (lemmisation): побудова синтаксичного дерева;

– розпізнавання іменованих сутностей;

– екстракція тексту за патернами (зокрема, Spacy Matcher і регулярні вирази);

– порівняння схожості між словами (similarity comparison).

Однак, мова, першою чергою, є інструментом спілкування між людьми, саме тому доцільно згадати *питомі види значень*:

а) пряме значення (концептуальне);

б) припущене (те, що не говориться, проте мається на увазі);

в) конотативне (те, що мовець асоціює зі словом);

г) колокативне (слово, яке часто вживають з іншим словом, утворюючи сталий вираз);

ґ) соціальне (вибір слів від неформальних до формальних).

У межах другого (розширеного) підходу продуктивним є згаданий метод МН, суть якого – в розробці методології обробки мови у ширшому ключі. Саме у межах останнього ШІ здатний впоратись із завданнями, що потребують аналізу та синтезу самих значень слів. Зокрема, мовиться про:

– моделювання теми (topic modelling);

– узагальнення тексту (text summirisation);

– машинний переклад (machine translation);

– фільтрацію тексту (text filtration);

– аналіз настрою (sentimental analysis).

Так, машинні алгоритми можуть вирішувати вищезазначені завдання з достатньо високою точністю, однак МН, попри низку переваг потребує, як зазначено в попередніх лекціях, величезної кількості відібраних даних. Останні потрібні навіть для тренування ШНМ, що, своєю чергою, означає потребу у використанні достатньо великих ресурсів, аби продукувати необхідні розробнику результати роботи.

Внаслідок цього таке навчання не здатне до спонтанної ОПМ на прикладі малих даних, так само неможливим у цьому випадку є і відслідковування причинно-наслідкових, логічних та інших зв’язків у його межах. Окреслені обмеження позбавляють машину можливості виконувати цілу низку завдань,

пов’язаних з інформативним аналізом і вивченням необроблених (згадаємо нормалізацію і так далі) певним чином даних (до прикладу, препроцесинг тощо).

Разом із тим, ОПМ є галуззю, головною метою якої є допомога комп’ютерам зрозуміти людську, або «природну» мову. На сьогодні, інтерес до цієї галузі лише зростає, оскільки її розвиток корелює з роботою ШНМ та іншими напрямами, які впроваджують у всі сфери діяльності.

Хоча природні мови не здаються чимось дивовижним, комп’ютерам дуже важко правильно їх інтерпретувати й використовувати. Жорсткий, обмежений правилами формат електронних таблиць і баз даних є ідеальним для машин, а випадкова природа людських мов, що залежить від контексту й не завжди пов’язана певними правилами, спричиняє нерозуміння смислів ШІ та ШНМ.

Природно, що ОПМ перебуває лише на початку свого поступу й розвитку, вже зараз демонструючи свій унікальний потенціал впровадження у життя людей. Так, експерти вважають, що наступний прорив у розвитку ОПМ буде колосальним, зумовлюючи перехід від структурованих (бази даних) до неструктурованих (текст) даних і значно поліпшить здатність машин розуміти людей у звичайній розмові. Технології розпізнавання та синтезу мовлення були згадані у попередній лекції, тому ви маєте уявлення про швидкість їх розвитку та вдосконалення у контексті обробки текстових даних.

6.2. Глибоке навчання в обробці природної мови

Вагома частина цих технологій базована на ГН, про яке також йшлося у попередніх лекціях. Отже, воно – галузь МН на базі низки алгоритмів, призначених для роботи з абстракціями. Така робота можлива у цих технологіях через застосування глибокого графа, поєднаного з декількома обробними шарами, які зазнають низки лінійних / нелінійних перетворень.

Переваги використання технологій ГН:

а) робота з великими обсягами тренувальних даних (до прикладу, вищезазначеними в лекціях текстовими корпусами та іншим);

б) наявні обчислювальні потужності (мовиться про багатоядерні центральний (CPU) і графічний (GPU) процесори);

в) створені нові моделі й алгоритми з розширеними можливостями й поліпшеною продуктивністю, з гнучким навчанням на проміжних уявленнях;

г) поява навчальних методів із використанням контексту, а також нової методології регуляризації та оптимізації.

Успішне використання більшості методів МН було досягнуто завдяки наявності репрезентативних даних, вхідних ознак, та оптимізації ваг, для підвищення точності фінального передбачення. У ГН алгоритм намагається автоматично «витягти» кращі ознаки (подання) з «сирих» вхідних даних. Створені вручну ознаки можуть бути занадто спеціалізованими, неповними та такими, що потребують часу для створення та затвердження.

На противагу цьому, виявлені ГН ознаки легко пристосовують, оскільки воно пропонує гнучкий, універсальний та такий, що навчається, фреймворк для подання інформації за допомогою візульного та лінгвістичного представлень.

Фреймворк – це інфраструктура програмних рішень, що полегшує розробку складних систем; спрощено її можна вважати своєрідною комплексною бібліотекою, що має низку обмежень правилами створення структури проєкту та написання коду.

6.2.1. Векторне представлення слів

У традиційній ОПМ слова розглядають як дискретні символи, що в подальшому представлені у вигляді одноразових векторів. Недолік такого способу представлення слів полягає у відсутності визначення схожості для таких векторів. Вирішити цю проблему можна за допомогою кодування схожості в самих векторах.

Отже, *векторне представлення слів* – метод репрезентації рядків у вигляді векторів зі значеннями: це означає, що для кожного слова створюють плаский (щільний) вектор так, щоб слова у схожих контекстах мали подібні вектори (див. Рис.56).

Такий спосіб представлення враховує відправну точку для більшості завдань ОПМ та робить ГН ефективним на малих наборах даних. Головними техніками векторних представлень є, створені Google, Word2vec і GloVe, які є доволі популярними й часто використовувані у завданнях ОПМ.

Як це працює простою мовою:

1. *Слова як точки в просторі*. Уявіть, що всі слова розміщені в багатовимірному просторі, де близькі за значенням слова знаходяться поруч (до прикладу, «король» і «королева» будуть близько, а «король» і «автомобіль» – далеко).

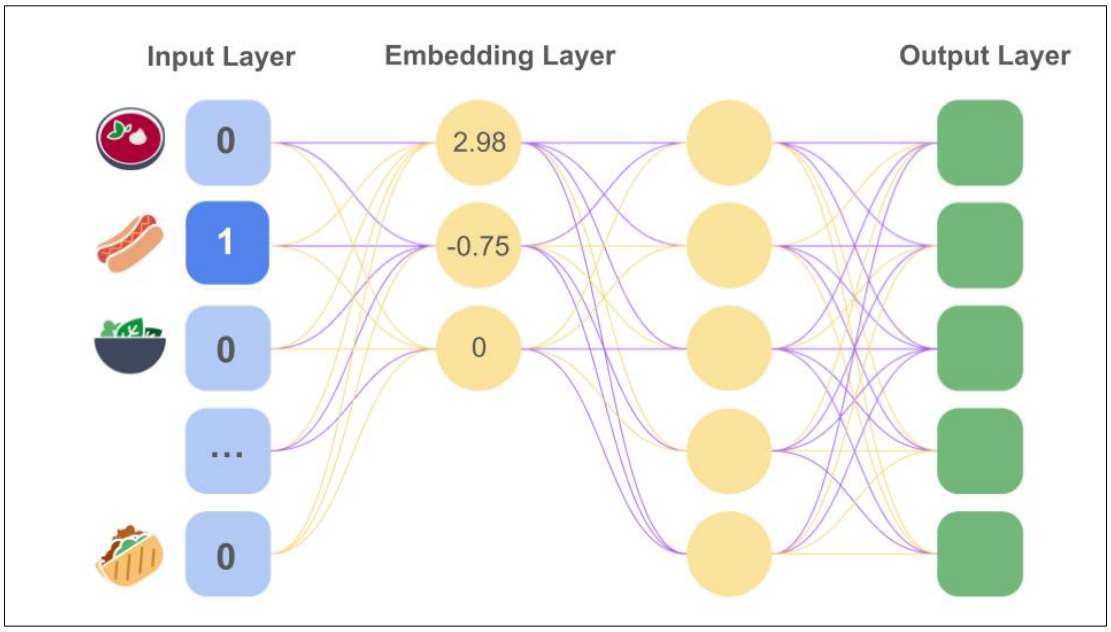


Рис. 56. Унітарне кодування слова “hot dog” подається
як вхідні дані для глибинної нейронної мережі.

2. *Навчання через контекст*. Алгоритм аналізує великі тексти й вчиться, які слова частіше трапляються поряд. До прикладу, якщо «кіт» часто з’являється поруч зі словами «мурчить», «м’який» або «рибка», то їх вектори будуть схожими.

3. *Можливість робити аналогії.* Завдяки цьому підходу можна навіть виконувати прості «арифметичні» операції зі словами: так, «король» – «чоловік» + «жінка» = «королева». Тобто алгоритм розуміє, що зв’язок між «король» і «королева» подібний до зв’язку між «чоловік» і «жінка».

Основна ідея – навчити комп’ютер сприймати мову не просто як набір букв, а як структуру, де слова мають значення й зв’язки між собою. Це стало основою для багатьох сучасних моделей ШІ, які працюють з текстами.

Word2Vec

Word2vec приймає великий корпус тексту, в якому кожне слово у фіксованому словнику подано у вигляді вектора (див. Рис. 57). Алгоритм проходить по кожній позиції *t* у тексті, яка має центральне слово *c* і контекстне слово *o*. Далі використовують схожість векторів слів для *c* та *o*, щоб розрахувати ймовірність *o* при заданому *c* (або навпаки), і триває регулювання векторів слів для максимізації цієї ймовірності.

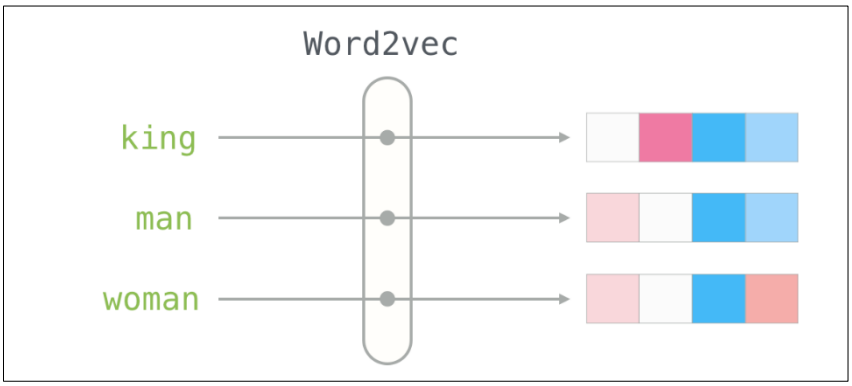


Рис. 57. Word2vec.

Для досягнення кращого результату роботи Word2vec, із набору даних (dataset) видаляють непотрібні слова (або слова з великою вживаністю), в англійській мові – це артиклі (*a*, *the*). Своєю чергою, це допомагає збільшити точність моделі та скоротити час на її тренування. Окрім того, використовують негативну вибірку (negative sampling) для кожного входу, оновлюючи ваги для всіх правильних міток, проте тільки на невеликій кількості некоректних міток.

Так, при використанні моделі SKIPGRAM розглядають контекстне вікно, що містить *k* послідовних слів. Після цього пропускають одне слово й відбувається процес навчання ШНМ, що містить усі слова, крім пропущеного. Алгоритм намагається передбачити його: так, якщо два слова періодично поділяють схожий контекст у корпусі, то ці слова будуть мати близькі вектори.

При використанні моделі Continuous Bag of Words береться багато пропозицій у корпусі. Кожен раз, коли алгоритм зустрічає певне слово, то береться у розрахунок і сусіднє з ним. Далі на вхід нейромережі подають контекстні слова й передбачають слово в центрі цього контексту.

У випадку великої кількості тематичних слів і центрального слова, отримуємо один екземпляр набору даних для ШНМ. Остання тренується та подає на виході закодований прихований шар, що є вкладенням (embedding) для певного слова. Така ж ситуація відбувається, якщо нейромережу тренують на великій кількості пропозицій та слів.

У схожому контексті для них формують подібні вектори. Прикметно, що недоліком цих методів є приналежність до класу window-based моделей, для яких притаманна низька ефективність використання статистики збігів у корпусі, що призводить до неоптимальних результатів.

GloVe

GloVe має на меті вирішити вищезазначену проблему захопленням значення одного вбудованого слова (word embedding) зі структурою всього доступного для огляду корпуса. Для цього ШНМ шукає глобальні збіги числа слів і використовує статистичну методологію, мінімізуючи середньоквадратичне відхилення, й видає простір вектора слова з розумною субструктурою. Окреслена схема достатньою мірою дозволяє ототожнювати схожість слова з векторним представленням.

До прикладу: для того, щоб кількісно відобразити нюанси, необхідні для розрізнення чоловіка і жінки, модель повинна пов’язувати з парою слів більше, ніж одне число. Природним і простим рішенням для розширеного набору розрізнювальних чисел тут стане різниця між двома векторами слів. GloVe розроблений таким чином, щоб такі векторні різниці якомога точніше відображали значення, визначене поєднанням двох слів (див. Рис. 58).

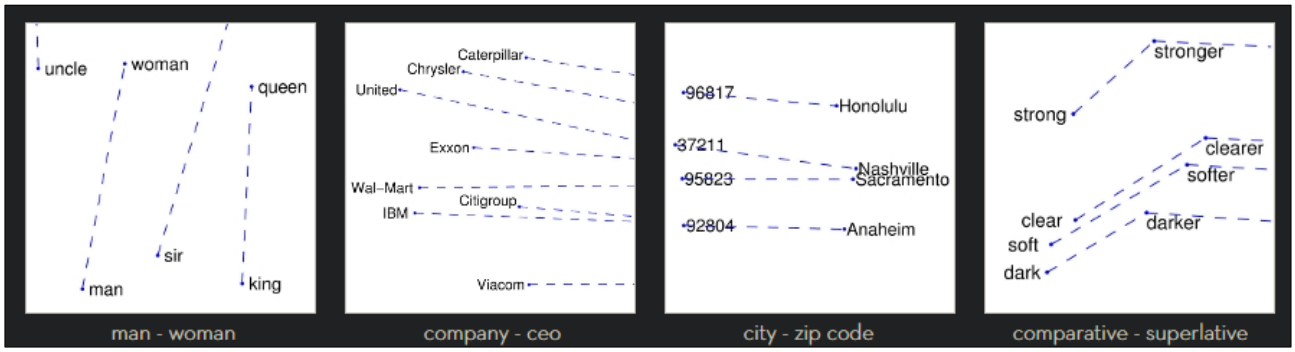


Рис. 58. GloVe.

6.2.2. Нейромережевий машинний переклад

Машинний переклад (*Machine translation*) – перетворення тексту мовою оригіналу (природною мовою) в еквівалентний за змістом текст іншою такою мовою (мовою перекладу). Машинний переклад виконують програми або машини без участі людини; він базований на статистиці використання сусідніх слів.

Зокрема, у традиційних системах машинного перекладу доводиться використовувати згаданий паралельний корпус – набір текстів, кожен із яких перекладено на одну або декілька інших мов. До прикладу, маючи вихідні природні мови: *f* (французьку) і цільову – *e* (англійську), потрібно побудувати статистичну модель, що включає вірогіднісне формулювання для правила згаданого у інших лекціях Томаса Байєса.

Таким чином, мовиться про модель перекладу $p(f \mid e)$, навчену на паралельному корпусі, і модель мови $p(e)$, навчену тільки на корпусі з англійською мовою. Вищезазначений підхід пропускає сотні важливих деталей та вимагає великої кількості спроєктованих вручну ознак, а також складається з різних і незалежних завдань МН.

Нейромережевий машинний переклад (*Neural Machine Translation*) – підхід до моделювання перекладу за допомогою вищезазначеної в інших лекціях РНМ (див. Рис. 59). Остання є нейромережею із залежністю від попередніх станів, що має зв’язки між ітераціями (тренувальними циклами МН). ШН отримують інформацію з попередніх шарів та із самих себе на попередньому етапі.

Це означає, що порядок, у якому подають на вхід дані й тренують ШНМ, важливий: результат подання «Міккі» – «Маус» не збігається з результатом подання «Маус» – «Міккі». При цьому стандартна модель такого перекладу є наскрізною нейромережею, де вихідну пропозицію кодують РНМ (encoder), а цільове слово передбачають за допомогою іншої РНМ (decoder).

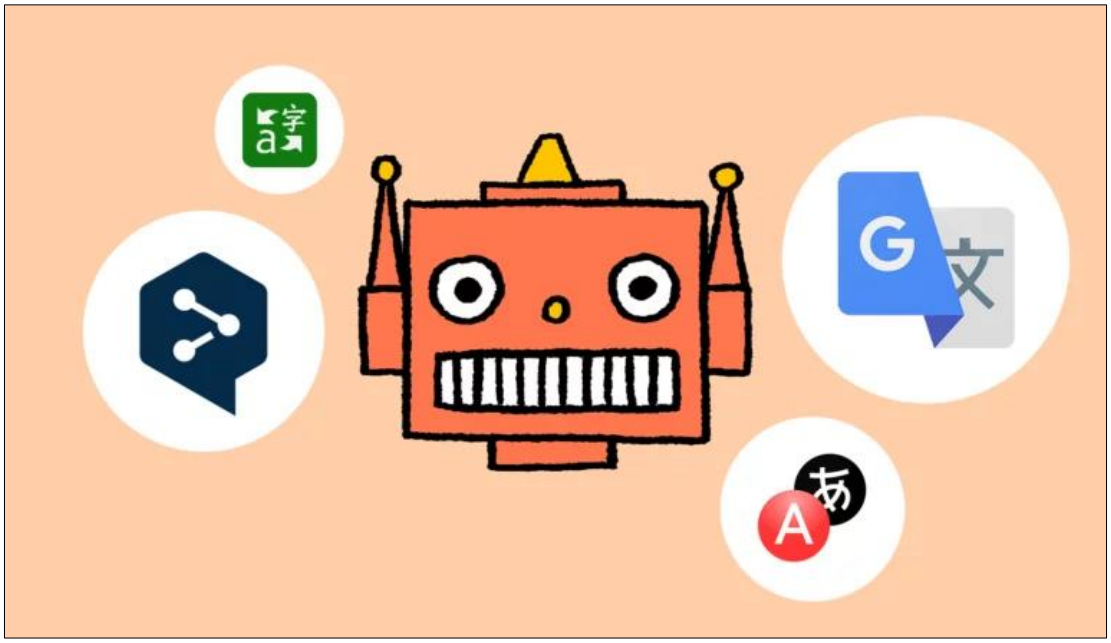


Рис. 59. Машинний переклад (DeepL, Google Translate тощо).

Кодувальник зчитує вихідну пропозицію зі швидкістю один символ в одиницю часу, після чого об’єднує вихідну пропозицію в останньому прихованому шарі. Декодер використовує зворотне поширення помилки для вивчення цього об’єднання й повертає перекладений варіант.

Важливо, що головною проблемою таких ШНМ є зникнення градієнта, коли інформація втрачається з часом. На перший погляд ця проблема не є критичною, оскільки це ваги, а не стани нейронів. Однак, із часом, ваги стають місцями для зберігання інформації з минулого, а відтак, якщо вага прийме значення 0 або 100000, то попередній стан втратить свою інформативність. Наслідком цього стане те, що РНМ буде зазнавати труднощів у запам’ятовуванні слів, що стоять далі в послідовності, а передбачення робимуться на базі крайніх слів.

Натомість неоднарозово згадувані у попередніх лекціях ***мережі ДКП*** борються з проблемою градієнта зникнення шляхом введення гейтів (gates) та згаданих комірок пам’яті. Таким чином, тут кожен нейрон є коміркою пам’яті з трьома гейтами: на вхід, вихід та забування (forget). Своєю чергою, ці стани виконують функцію охоронців інформації, дозволяючи, або забороняючи потік даних:

– *вхідний гейт* визначає кількість інформації, що буде зберігатися в цьому осередку з попереднього шару;

– *вихідний* локалізує, яка частина наступного шару дізнається про стан поточної комірки;

– *гейт забування* контролює міру збереження значення в пам’яті.

До прикладу, якщо під час аналізу книги починається новий розділ, то ШНМ необхідно забути деякі слова з попереднього. Прикметно, що мережі ДКП здатні:

а) навчатися на складних послідовностях (останні поширені й використовують в машинному перекладі);

б) писати тексти в конкретному стилі та можуть використовуватись для складення простої музики.

Окрім цього, це стандартна модель для більшості завдань маркування (labeling) послідовності, яка складається з великої кількості даних (текстових корпусів).

Закриті рекурентні блоки (*Gated Recurrent Units*, далі – ЗРБ) відрізняються від мереж ДКП, хоча теж є розширеннями для МН. У ЗРБ міститься на один гейт менше та інший принцип роботи: відсутні вхідний та вихідний гейти, а також гейт забування. Замість них у ЗРБ є *гейт поновлення (update gate)*, який визначає, скільки інформації необхідно зберегти з останнього стану й скільки даних пропускати з попередніх шарів.

Цікаво, що функція *скидання гейта (reset gate)* схожа на вищезазначене забування у мережі ДКП, проте відрізняється розташуванням: так, ЗРБ завжди передає свій повний стан, не маючи вихідного затвору. За функціональністю цей затвор схожий на відповідний у ДКП, однак, у ЗРБ він працює швидше й легший в управлінні.

На практиці вони прагнуть нейтралізувати одне одного, оскільки потрібна велика ШНМ для відновлення виразності (expressiveness), яка зводить нанівець прирости результативності. Зокрема, у випадках, де не потрібно екстравиразності, ЗРБ показують кращий результат, ніж мережі ДКП.

Окрім вищезазначених головних архітектур (РНМ, мереж ДКП та ЗРБ) за останні кілька років з’явилося багато покращень в нейромережевому машинному перекладі:

1. *Послідовне навчання за допомогою нейронних мереж (Sequence-to-Sequence Learning with Neural Networks)* довело ефективність мереж ДКП для нейронного машинного перекладу. Цей метод використовує багатошарову мережу ДКП, щоб відобразити вхідну послідовність у вигляді вектора з фіксованою розмірністю. При цьому відбувається застосування іншої такої мережі для декодування цільової послідовності з вектора.

2. *Нейронний машинний переклад шляхом спільного навчання вирівнювання та перекладу (Neural Machine Translation by Jointly Learning to Align and Translate)* представив механізм уваги (attention mechanism) в ОПМ. Так, використання вектора фіксованої довжини є локальним засобом щодо поліпшення результативності вищезазначеного процесу (ОПМ). При цьому було дозволено моделі автоматично шукати частини вихідної пропозиції, що релевантні до передбачення цільового слова (без необхідності явного формування цих частин).

3. Google створила власну систему ОПМ, яка називається *Google’s Neural Machine Translation*, що вирішує завдання точності й простоти застосування. Модель складається з глибокої мережі ДКП, у якої 8 кодуючих та 8 декодуючих шарів, і використовує як залишкові зв’язку, так і attention-зв’язку від декодермережі до кодермережі (див. Рис. 60).

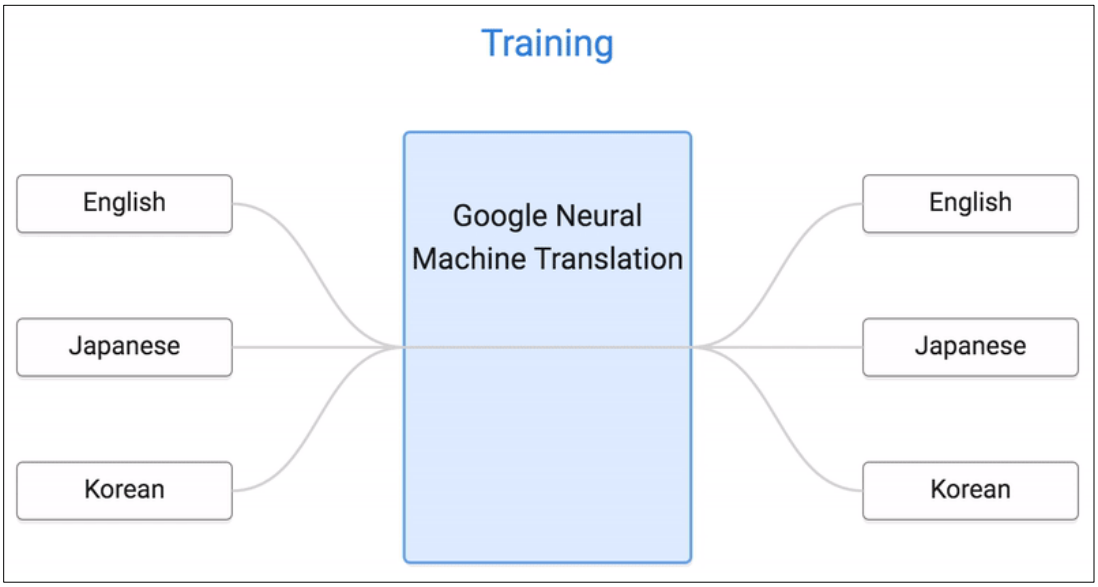


Рис. 60. Google’s Neural Machine Translation.

4. Замість використання РНМ, Facebook AI Researchers застосовують вищезазначену в другій та інших лекціях ЗНМ для задач sequence-to-sequence навчання в ОПМ.

6.2.3. Голосові помічники

Через обмеженість можливостей ШІ в ОПМ, створення повноцінного розмовного асистента залишається відкритим питанням. Так, завдяки спільному дослідженню вчених з Монреалю, Технічного Інституту Технологій Джорджії, Microsoft та Facebook було створено ШНМ, здатну генерувати чутливі до контексту відповіді у розмові. Тренування цієї нейромережевої моделі проводилося на великому наборі неструктурованих діалогів у соціальній мережі X.

Цікаво, що архітектура РНМ була використана у цьому випадку для створення відповідей на окремі питання, що виникають під час інтегрування контекстної інформації у класичну статистичну модель. Створена ШНМ показує істотне покращення результатів над контент-чутливою й контент-нечутливою базовою лінією машинного перекладу та пошуку інформації.

Нейронна машина для відповідей (Neural Responding Machine; далі – НМВ*)* – розроблений у Гонконзі генератор відповідей для коротких розмов, створений із використанням спільного кодер-декодер фреймворка. Так, спочатку формалізують створення відповіді, як процес розшифрування на основі

прихованого представлення вхідного тексту, в той час як кодування та декодування здійснюють за допомогою РНМ.

Навчання НМВ відбувалося на великих обсягах даних з однозначними діалогами, зібраними з мікроблогів. Емпірично встановлено, що НМВ генерує граматично коректні та контекстуально доречні відповіді у 75% наданих для обробки розмовах, що перевищує показники інших сучасних моделей схожої конфігурації.

Нейронна розмовна модель Google (Google's Neural Conversational Model) – модель, що пропонує підхід до моделювання діалогів на основі вищезазначеного sequence-to-sequence фреймворку. ШНМ здатна вести розмову, висуваючи припущення стосовно наступної репліки, які базовані на попередніх висловлюваннях.

Прикметно, що МН такої моделі є наскрізним, що зменшує потребу в створенні великої кількості штучних правил і обмежень. Натомість прості діалоги можуть конструюватися на основі діалогового тренувального набору даних, вузькоспеціалізованих та зашумлених наборів даних із субтитрами до фільмів.

6.2.4. Системи типу «питання – відповідь»

Системи типу «питання-відповідь» видобувають інформацію безпосередньо з джерела (документа, розмови, онлайн пошуку, набору картинок тощо). Такі системи надають короткі й лаконічні відповіді замість розгорнутого тексту, оскільки більшість завдань їхньої роботи класифікують за типом «питання-відповідь». У таких завданнях користувач надсилає запит (до прикладу, через інтегрованого чат-бота) й отримує відповідь від системи.

Для вирішення таких завдань існує спеціальна оптимізована архітектура ГН – *Мережа динамічної пам'яті* (*Dynamic Memory Network*; далі – МДП). Така ШНМ навчається на тренувальному наборі вхідних даних та питань і формує епізодичні «спогади» про них, які потім використовують для генерації доречних відповідей.

Архітектура МДП складається з таких компонентів:

1. Модуль семантичної пам’яті, аналогічний базі знань. У нього входять попередньо підготовлені GloVe вектори, які використовуються для створення послідовностей згаданих векторних представлень слів із вхідних висловлювань (це вхідні дані нейромережевої моделі).

2. Вхідний модуль переробляє пов’язані з питанням вхідні вектори у їх наборах (фактах). Цей модуль реалізований за допомогою *керованого рекурентного блоку* (Gated Recurrent Unit, далі – КРБ), що дозволяє визначити релевантність розглянутого висловлювання.

3. Модуль питань обробляє питання слово за словом, і генерує вектор із використанням КРБ (такі самі, як на вхідному модулі й вагах).

4. Модуль епізодичної пам’яті зберігає визначені на вході вектори фактів і питань, закодованих як вкладення.

5. Модуль відповідей генерує доречну відповідь з епізодичної пам’яті, що містить необхідну для цього інформацію на останньому етапі. Цей модуль використовує інший КРБ, натренований із класифікацією крос-ентропійної помилки коректної послідовності, що конвертується назад у природну мову.

Ентропія – це усереднена швидкість, із якою продукує інформацію певне джерело даних.

Важливо, що МДП мають вищу ефективність у вирішенні вищезазначених завдань порівняно з іншими архітектурами для семантичного аналізу та частиномовної розмітки (part-of-speech tagging) про яку поговоримо нижче.

6.2.5. Стислий переказ тексту

Стислий переказ тексту (Text Summarization) – інструмент інтерпретації текстової інформації, що дозволяє створювати лаконічні резюме великих текстових фрагментів (див. Рис. 61). Скорочення тексту формується у декілька *етапів*:

– підрахунок частотності слова в текстовому документі;

– визначення 100 найчастотніших слів;

– сортування визначених слів;

– оцінювання кожного речення з такими словами з наданням їм більшого вагового коефіцієнту;

– сортування перших *x* речень з урахуванням їхнього положення в оригінальному тексті.

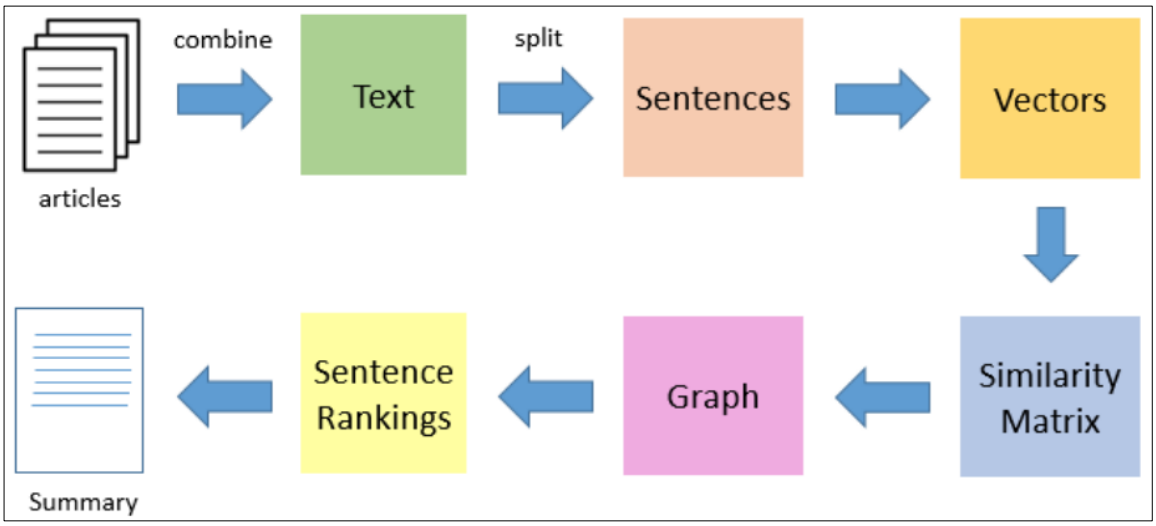


Рис. 61. Стислий переказ тексту.

Визначають *два основних підходи до скорочення тексту:*

1. *Видобувний підхід* добуває слова та фрази з оригінального тексту для створення стислого перекладу. Показовими в плані першого підходу є LexRank і TextRank, які використовують варіації алгоритму сортування сторінок Google PageRank.

LexRank – алгоритм навчання без учителя на основі графів, який застосовує модифікований косинус зворотної частоти використання слова як міри схожості двох речень. Такий алгоритм запобігає тавтологіям шляхом інтелектуальної післяобробки, покликаної перевіряти головні речення переказу на низький коефіцієнт подібності.

TextRank схожий на LexRank, проте містить низку переваг, а саме:

а) використання лематизації (групування відмінюваних форм слова так, щоб їх можна було проаналізувати як єдиний елемент, ідентифікований за лемою (доведеним твердженням, корисним не саме по собі, а для доказу інших тверджень) слова або словниковою формою) замість стемінгу (процес скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс, його результати іноді дуже схожі на визначення кореня

слова, проте алгоритми базують на інших принципах, саме тому слово після обробки алгоритмом стемінгу (стематизації) може відрізнятися від морфологічного кореня слова);

б) застосування частиномовної розмітки та розпізнавання імені об’єкту;

в) виділення важливих ключових фраз разом зі стислим переказом тексту.

2. *Абстрактний підхід* вивчає внутрішнє мовне представлення тексту, щоб створити, подібну до людської, інтерпретацію шляхом перефразування. Моделі для абстрактного стислого переказу тексту використовують ГН, що призвело до прориву в цій сфері.

Отже, моделі рекурсивного ГН можуть вирішувати безліч мовних завдань, що включають передбачення на рівні слова та речення як безперервної, так і дискретної природи. Одним із багатьох можливих рішень для досягнення цієї мети є застосування рекурсивних або рекурентних методів для обчислення елементів даних рівня абзацу або документа.

Другий виклик для таких моделей – *логічні міркування першого порядку*: логіка першого порядку (числення предикатів) є формальною системою в математичній логіці, в якій допускають висловлення відносно змінних, фіксованих функцій, і власне предикатів. Така логіка є розширенням логіки висловлювань, а також частковим випадком логіки вищого порядку.

Вищезазначені логічні міркування першого порядку можуть знадобитися для отримання правильної інформації з баз знань за допомогою питань природної мови. Таким чином, розглянуті моделі можуть бути розширені, щоб урешті-решт спільно моделювати мову, образи та бази знань в одній цілісній семантичній структурі.

6.3. Використання методів обробки природної мови для вилучення визначень слів із контексту

У сучасну епоху величезного цифрового контенту точне розуміння значення слів у їхньому природному контексті має вирішальне значення для завдань мовної обробки. Актуальною також є проблема вилучення точних визначень слів за допомогою передових методів ОПМ. Ця проблема може бути

розкрита через можливості вищезазначених популярних бібліотек ОПМ, таких як SpaCy та NLTK, а також потужної моделі BERT та бібліотеки bert-extractive-summarizer.

SpaCy надає лінгвістичні анотації та функціональні можливості, включаючи вищезазначену токенізацію та синтаксичний аналіз залежностей, тоді як bert-extractive-summarizer використовує BERT, найсучаснішу мовну модель-трансформер (до прикладу, знайомий усім нам ChatGPT), для виявлення ключових речень.

Ці технології дозволяють ідентифікувати контекстно-релевантні речення, допомагаючи «витягувати» визначення слів. Для цього завдання можна було б розглянути альтернативні технології, такі як WordNet. NLTK пропонує інструменти для різних завдань ОПМ, тоді як WordNet надає визначення й семантичні зв’язки між словами. Попередньо навчені мовні моделі, такі як ChatGPT і RoBERTa, також можуть бути використані для вилучення визначень на основі контексту.

Так, для проведення лінгвістичного дослідження потрібно ретельно відібрати колекцію з певної кількості (до прикладу, 12) текстових уривків, які охоплюють різноманітні сфери та мовні стилі. Ці уривки оброблюємо за допомогою бібліотеки SpaCy для токенізації речень, що забезпечить детальне представлення тексту. Згодом використовуємо бібліотеку bert-extractive-summarizer для виявлення речень, найсхожіших на визначення цільового слова.

Далі, використовуючи модель BERT, згенеруємо контекстні пропозиції, які вдосконалять процес вилучення дефініції слова. Результати такої роботи демонструють, що запропонований підхід дає точні та релевантні визначення слів, враховуючи семантичні й контекстуальні аспекти даних.

Поєднання SpaCy, bert-extractive-summarizer і BERT значно підвищує точність та повноту вилучення дефініцій слів порівняно з традиційними методами. Продовжуючи дослідження в цій галузі, можна вдосконалювати й розширювати можливості систем ОПМ, що зрештою дозволить точніше й нюансованіше розуміти природну мову.

Прикметно, що хоча сучасні технології ОПМ досягли значного прогресу, вони не можуть працювати оптимально для всіх мов, включаючи українську. Це обмеження підкреслює необхідність розробки та впровадження національних технологій, спеціально адаптованих для української мови. Така потреба пов’язана з тим, що наша мова має унікальні лінгвістичні характеристики та нюанси, які вимагають спеціальної уваги та лінгвістичних ресурсів для ефективного опрацювання та аналізу тексту. До прикладу, згадаємо нещодавню ініціативу Міністерства цифрової трансформації України щодо створення першої українськомовної моделі національного масштабу.

Отже, ми розглянули сучасні підходи до визначення слів у їх природному контексті, ознайомились із застосуванням технологій SpaCy, bert-extractive-summarizer та BERT, що показало себе як ефективний підхід для виявлення семантичних нюансів та контекстних варіацій значень слів. Таким чином, важливим є контекстно-орієнтований аналіз для точної ідентифікації дефініцій слів, що, своєю чергою, прокладає шлях до покращення розуміння мови та інформаційно-пошукових систем.

6.4. Труднощі в роботі систем обробки природної мови та основні методи їх вирішення

Вищезазначене на цій та попередній лекціях засвідчило, що проблему ОПМ не можна назвати простою. Так, труднощі виникають через низку об’єктивних причин: зокрема, існування сотень природних мов, у кожній із яких діють свої синтаксичні правила. У межах однієї мови побутують слова, які можуть мати різний зміст залежно від контексту вживання тощо. Більше того, навіть на рівні окремих символів зустрічаються певні труднощі.

Відтак, в процесі ОПМ завжди необхідно враховувати кодування, яке використовують у конкретному документі. Текст може зберігатися в різних кодуваннях: ASCII, UTF-8, UTF-16 або Latin-1. Особливі види обробки можуть знадобитися для знаків пунктуації та для чисел. Іноді доводиться окремо обробляти використання знаків, які відображають емоції (комбінації символів або спеціальні символи), гіперпосилань, розділових знаків, що

повторюються (... або ---), розширень файлів та імен користувачів, що містять крапки.

Важливим у цьому контексті є також, власне, поділ тексту на фрагменти або елементи, під якими здебільшого розуміють вищезазначене представлення останнього у вигляді послідовності слів. У цьому випадку слова позначають знайомими нам поняттями – «лексичний елемент», «лексема», згадувана «лема», або просто «токен» (“token”), відповідно позиціонуючи поділ тексту як процес «токенізації».

Такий процес узвичаєно не викликає особливих труднощів у мовах, що використовують пробільні символи для розділення слів. Однак в мовах, на кшталт китайської, він проходить набагато важче, оскільки ієрогліфи можуть позначати як склади, так і цілі слова. Та й в англійській мові з процесом токенізації можуть виникнути певні труднощі. Це пов’язано з тим, що існує велика кількість альтернативних варіантів, коли одне слово може писатися разом, окремо або через дефіс.

Для нас, як лінгвістів, знайомою є концепція, у межах якої текстові дані досліджують таким чином: найменшою одиницею приймають слова, які, своєю чергою, об’єднуються в словосполучення й речення. При цьому визначення меж речень теж може бути пов’язано з певними труднощами, хоча на перший погляд здається, що достатньо лише знайти крапку, яка позначає кінець речення.

Проте крапки можуть зустрічатися й в середині речень (до прикладу, після скорочених слів, або у випадку певного «обриву» смислу тощо). Відповідно, це складає доволі серйозні проблеми автоматизації: так, усі ми знаємо про існування мов програмування, у яких саме знаковий складник виступає головним. Так само нам відомо, що в таких мовах надзвичайно важливими є послідовність певних знаків, які покликані актуалізувати ті чи ті функції. Зокрема, саме за допомогою певного синтаксису в HTML можна впливати на тип заголовку, візуалізацію даних та інше, що, природно, має значущість і у випадку ОПМ.

Не менш важливим є й граматичний розбір, із яким ми також знайомі, оскільки не раз робили його вручну: тут так само виникають доволі серйозні

проблеми з точністю. По-перше, багато в такому розборі залежить від якості морфологічної (частиномовної) розмітки (part-of-speech tagging): остання має бути достатньо високою (97–98%). При цьому розповсюдженим у довгих реченнях є поява неправильно ідентифікованої частини мови, внаслідок чого суттєво зростає кількість подальших помилок такого розбору.

По-друге, вищезазначений розбір у автоматичному вимірі ніколи не сягне необхідних (97–98%), оскільки «стелею» його можливостей є 90–93%, що, своєю чергою, гарантує наявність помилкового розбору довгого речення. До прикладу, окреслена точність автоматичного розбору (90–93%) у реченні довжиною 10 слів без жодної похибки складе лише 35%, тобто не буде прийнятною.

Разом із тим, сучасний стан досліджень дає надію на покращення якості розбору, проте часто правильний синтаксичний розбір включає також розуміння семантики речення, що узвичаєно викликає низку труднощів у будь-якій мові (зокрема, англійській). Так, у реченні “He saw a man with glasses” може бути два різних варіанти синтаксичного розбору залежно від того, чи вважаємо ми, що людину побачили за допомогою окулярів, або побачили людину з окулярами. Звичайно, якщо потрібно отримати максимально точний синтаксичний розбір, то має сенс лишати кілька найімовірніших варіантів, а потім визначати правильний за сукупністю різних факторів, у тому числі семантичних.

Так, для лінгвістів звичайною є практика, у межах якої доводиться визначати певні зв’язки між словами. Зокрема, встановлення кореферентності (coreference resolution) визначає зв’язки між конкретними словами. Останні позначають один і той самий об’єкт, тобто мають однаковий референт в одному або в декількох реченнях. До прикладу, в реченнях “The city is large but beautiful. It fills the entire valley” слово “it” кореферує, тобто є референційно тотожним слову “city”.

Цікаво, що явища кореферентності зумовлені фундаментальними закономірностями організації тексту. Оскільки текст має лінійну будову, а ситуація, яку він описує, зазвичай, нелінійна, то наслідком цього є те, що в ньому майже неминуче повинні міститися повторні згадки елементів описуваної ситуації.

При кожній новій згадці того ж об’єкта проводиться його нова номінація, яку базують на тому, що вже було сказано про цей об’єкт, і на тих знаннях, які в тексті не вербалізовані (екстралінгвальні знання мовця про контекст певної царини). Хоча проблема кореферентності в лінгвістиці доволі докладно досліджена, втілення цих теоретичних знань на практиці є складним, оскільки бачення тих чи тих зв’язків, смислів та іншого – вельми суб’єктивне.

Зокрема, слово може мати кілька значень, тому для його визначення в цьому випадку потрібно вирішити лексичну багатозначність (word sense disambiguation), що, своєю чергою, може становити певні труднощі. До прикладу, в реченні “John went back home” слово “home” може означати “housing that someone is living in” або “the country or state or city where someone lives”. Саме тому однією з найбільших проблем у процесі обробки природномовних текстів є неоднозначність (багатозначність) його одиниць. Остання виявляється на всіх текстових рівнях та виражається в явищах полісемії, омонімії та синонімії.

Так, говорячи про неоднозначність, виокремимо:

а) *лексичну* (існування більше, ніж одного значення слова: до прикладу, “bank”);

б) *синтаксичну, або структурну* (коли одне речення має декілька ймовірних варіантів граматичної структури і, відповідно, інше значення: до прикладу, проблема приєднання (attachment ambiguity), коли прийменникова група (prepositional phrase) може входити як до дієслівної (verb phrase), так і до іменникової групи (noun phrase) в межах одного речення, що спричиняє зміну значення: “The police shot the rioters with guns”);

в) *семантичну* (коли те саме речення можна по-різному розуміти в різних контекстах, хоча лексична чи структурна багатозначність відсутня: “All linguists prefer a theory”);

г) *прагматична* (коли одне речення має декілька інтерпретацій в тому контексті, в якому воно існує: “Every student thinks he is a genius”).

Зокрема, сучасним системам вирішення лексичної багатозначності властива точність у діапазоні 60–70%, яка, здебільшого, досягається за рахунок окремих методів. Своєю чергою, вирішення проблеми зняття вищезазначеної

неоднозначності вимагає інтеграції декількох джерел інформації та методів. Незважаючи на всі окреслені труднощі, технологія ОПМ більшою мірою здатна доволі успішно впоратися зі своїми завданнями, а тому набуває все більшої популярності у різних сферах: від побуту до виробництва.

Приблизно в половині випадків має місце будь-яка форма омонімії, тож набір морфологічних ознак виявляється недостатнім для її вирішення. Зменшити неоднозначність можна за допомогою синтаксичного й семантичного аналізу з використанням статистичних методів, які дають змогу відкинути вкрай малоймовірні варіанти.

Природна мова, хоч і є символічною, проте обробляти її виключно на базі символічних, базованих на логіці, правил та об'єктивних моделей не зовсім доцільно. Це пов'язано з тим, що природна мова є вкрай неоднозначною та мінливою, тому для її обробки необхідно застосовувати й статистичні алгоритми. Відповідно, домінантними підходами до сучасної ОПМ є базовані на статистичному МН (statistical machine learning) підходи.

Історія розвитку таких підходів є показовою: так, на початку 1990-х років стали розвиватися методи МН й одночасно було проведено низку досліджень з статистичної лінгвістики. В МН чудово показали себе алгоритми класифікації для різних завдань, пов'язаних з обробкою текстів: визначення спаму, сортування документів за тематикою, виокремлення іменованих сутностей (named entities).

Також продуктивними вони виявилися у комп'ютерній лінгвістиці: зокрема, мовиться про приховані ланцюги А. Маркова (система, що моделюється, розглядається як марковський процес із неспостережуваними (прихованими) станами) та моделі максимальної ентропії (нагадаємо, що *ентропія* – усереднена швидкість, із якою продукує інформацію певне джерело даних, а в нашому випадку мовиться про швидкість обробки текстових даних ШНМ) тощо, які дозволили зробити визначення частин мови високоточним.

З'явилися *парсери* – частини програми, що перетворює вхідні дані (зазвичай, текстові) у певний структурований формат, потрібний для завдань їх подальшого аналізу й використання. Таким чином, технічно парсер виконує

синтаксичний аналіз даних (до прикладу, тексту), а для завдань з ОПМ продуктивними є парсери на базі стохастичних контекстно-вільних граматик, які створюють проєкти зі статистичного машинного перекладу.

Окрім того, були закладені основи вищезазначеного ГН, яке набуває все більшого розитку, що зумовлено прогресом у галузі високопродуктивних систем і появою великих обсягів даних, які використовують для такого навчання. Щодо практичного використання, то Instant Data Scraper дозволяє швидко видобувати необхідні дані з вебсайтів. Такі дані у подальшому можна обробити у текстовому Notepad++ та подібних і використовувати.

У 2010 році була запропонована модель лексикалізованої ймовірнісної (стохастичної) граматики, яка дозволила підвищити точність граматичного розбору до вищезазначених бажаних 93%, що, звичайно, далеко від ідеалу. Пам'ятаємо значущість точності такого розбору, яку визначають відсотком правильно визначених граматичних зв'язків, а також вірогідністю того, що довге речення буде розібрано правильно, що, як зазначалося вище, має доволі низьку вірогідність.

Водночас, завдяки новим алгоритмам і підходам, включаючи згадане ГН, збільшилася швидкість такого автоматизованого граматичного розбору. Також усі провідні алгоритми й моделі стали доступні широкому колу дослідників: напевно, найвідомішою роботою в царині ГН для ОПМ став вищезазначений алгоритм чеха Томаса Міколова (векторне представлення слів).

Таким чином, завдяки новим методам ГН, сьогодні можна отримати якісні семантичні представлення для слів, фраз і речень, причому навіть без навчальної вибірки (набору даних). Все менше зусиль потрібно для створення власних семантичних словників і баз знань, тому розробляти системи автоматичної обробки текстів стало простіше.

Однак ми все ще дуже далекі від адекватного вирішення завдання розуміння взаємопов'язаних подій, представлених у вигляді послідовності речень або образів, а також діалогів. Усі відомі сьогодні методи успішно працюють або при вирішенні завдань «поверхневого» розуміння мови, або при істотному обмеженні предметної царини.

Методи ГН є точнішими, ніж поверхневі методи, які не намагаються осмислити текст, однак на практиці бази знань, необхідні для їх роботи, існують для вкрай обмежених предметних напрямів. Саме тому на сьогодні досить часто використовують і поверхневі методи: останні беруть до уваги найближчі слова, видобуваючи потрібну інформацію й досліджуючи сполучуваність слів.

Правила можуть бути автоматично отримані за допомогою комп'ютера, при використанні навчальної текстової бази даних слів, доданих з їх значеннями. Цей метод теоретично не такий дієвий, як глибокі методи, хоча на практиці він дає кращі результати, внаслідок обмеженості знань комп'ютера про світ.

Відтак, процес розуміння та генерації природної мови з використанням ЦТ є надзвичайно складним: так, найвідомішими методами роботи з текстовими даними є методи з використанням алгоритмів МН з учителем, коли система намагається видобути мовні структури та правила з анотованих даних текстового корпуса.

До прикладу, завдання класифікації документів за категоріями (спорт, політика, економіка, розваги тощо) є досить простим, адже слова, які використовуються в документах таких тематик, є підказкою. Читач може з легкістю віднести текст до однієї із тем, базуючись і на власному досвіді, проте малоймовірно, що він зможе назвати конкретні правила, якими він послуговувався.

Натомість автоматизація цього процесу шляхом виформування правил такої категоризації даних є складною та марудною справою. Так, використовуючи алгоритми МН з учителем можна дозволити машині визначити мовні структури, які дадуть змогу категоризувати документи. Методи МН є відмінним рішенням для галузей, де важко визначити конкретні правила, а створити анотований корпус досить просто.

Отже, ОПМ активно досліджують у наш час; існує багато методів, за допомогою яких теоретично можна подолати труднощі в роботі її систем. Однак на практиці застосування таких методів є доволі далеким від ідеалу та потребує подальшого покращення й вдосконалення.

Запитання для самоконтролю:

1. У чому полягає різниця між спрощеним та розширеним підходами до ОПМ? Які приклади завдань реалізуються в кожному з них?

2. Які функції виконують основні бібліотеки Python для ОПМ (таких як NLTK, Spacy, Pandas, Scikit-learn)? Наведіть приклади завдань, які з ними можна реалізувати.

3. Назвіть і коротко охарактеризуйте 7 типів значення слова в лінгвістичному аналізі. Чому ці значення важливі для ОПМ?

4. У чому полягають основні переваги та обмеження ГН в ОПМ?

5. Поясніть принцип роботи векторного представлення слів. Як працюють моделі Word2Vec та GloVe? У чому полягає їх відмінність?

6. Як працює нейромережевий машинний переклад на основі РНМ? Які проблеми виникають при цьому? Як їх вирішують моделі ДКП і ЗРБ?

7. У чому полягає суть attention-механізму? Як він покращує якість нейромережевого перекладу порівняно з традиційною моделлю Seq2Seq?

8. Які архітектурні елементи має МДП? Як вони взаємодіють для вирішення завдань типу «питання – відповідь»?

9. Чим нейронні розмовні моделі (NRM, Google Neural Conversational Model) відрізняються від традиційних діалогових систем? Які особливості навчання таких моделей?

10. Що таке Text Summarization? Які етапи включає побудова стислого переказу? Які методи й метрики можуть бути використані для оцінювання його якості?

ЛАБОРАТОРНІ РОБОТИ

ЛАБОРАТОРНА РОБОТА № 1/2

Тема: Розробка ефективних промптів для ШІ: створення, оптимізація, аналіз.

Мета: навчитись створювати точні й зрозумілі промпти для взаємодії з ШІ; удосконалити навички формулювання завдань для отримання бажаного результату; оцінити вплив формулювання промпта на якість відповіді ШІ.

Завдання:

1. Ознайомитись із принципами роботи моделей ШІ щодо сприйняття текстових запитів.

2. Навчитись формулювати промпти для різних типів завдань (аналіз, генерація, переклад тощо).

3. Опанувати оптимізацію запитів для покращення якості результатів.

4. Порівняти відповіді на різні промпти й оцінити їх ефективність.

Матеріали:

а) наукові тексти: Google Академія (https://tinyurl.com/5n6axph2)

б) художні тексти: Adobe.com (https://tinyurl.com/zcbwc8rj)

Gutenberg.org (https://tinyurl.com/3zzv8e23)

SDFO.org (https://tinyurl.com/3wzxc4ee)

Derevo-kazok.org (https://tinyurl.com/4eecbp4s)

в) публіцистичні тексти: BBC.com (https://tinyurl.com/3ya5nccj)

Kunsht.com.ua (https://tinyurl.com/2ve979dn)

Інструменти: AI Presentation Maker (https://tinyurl.com/4yrwdb43)

ChatGPT (https://tinyurl.com/ycypp7h3)

DALL E mini (https://tinyurl.com/356pfs3k)

Gemini (https://tinyurl.com/yf9xapfj)

Результати роботи:

1. Сформульовані промпти для різних типів завдань.

2. Порівняльний аналіз результатів базових (первинних) і оптимізованих промптів.

3. Звіт із висновками щодо впливу правильного формулювання запиту на якість відповіді ШІ.

ПЛАН.

1. Вступна частина.

Промпт (з англ. prompt) – текстовий запит, який ми надаємо ШІ (ChatGPT, Gemini тощо) для виконання певного завдання. Він може містити запитання, інструкції, опис задачі або навіть приклади даних. Правильно сформульований промпт визначає, наскільки точно і якісно ШІ виконає поставлене завдання. Роль промпта у взаємодії з ШІ порівнюють з роллю інструкції для людини: чим зрозуміліша й конкретніша інструкція, тим якіснішим та швидшим буде виконання завдання.

1.1. Питомі характеристики ефективного промпта:

1.1.1. *Конкретика формулювань*.

До прикладу: Розкажи про текст, напиши текст тощо – поганий промпт.

Проаналізуй текст, визнач ключові теми, вкажи приклади для кожної підтеми у ньому тощо – гарний промт.

1.1.2. *Актуалізація контексту*. ШІ набагато краще опановує завдання, якщо йому надати умови його функціонування, параметризацію тощо.

До прикладу: Це уривок із наукової статті. Проаналізуй стиль тексту та визнач, чи відповідає він формату академічного письма.

1.1.3. *Архітектура промпта, його логічність*. Доцільно структурувати багатокрокове завдання, використовуючи невдалі попередні взаємодії як базу для уточнень.

До прикладу: 1. Створи текст за тими ж умовами, ускладнивши завдання у ньому.

2. Визнач підтеми проблеми в своєму тексті, вкажи виклики й перспективи у вирішенні кожної.

3. Зверни увагу на стиль написаного тобою тексту, зроби його науково-популярним.

1.1.4. *Мета та очікуваний результат*. Промпт має чітко вказувати на ваші очікування щодо результату роботи ШІ.

До прикладу: Зроби короткий підсумок тексту (до 100 слів) і вкажи його 3 ключові ідеї.

1.2. Типи завдання для роботи з ШІ:

1.2.1. *Аналіз тексту:*

а) частотний аналіз слів і фраз;

б) визначення ключових тем і стилістичних особливостей;

в) емоційний або тональний аналіз тексту.

1.2.2. *Генерація текстів:*

а) написання резюме, мотиваційних листів, анотацій та тому подібного;

б) створення сценаріїв, есе, романів і так далі за певними визначеними користувачем параметрами.

1.2.3. *Мовний аналіз:*

а) переклад, рерайт тощо тексту;

б) локалізація низки граматичних, лексичних, стилістичних та інших помилок.

1.2.4. *Робота зі структурованими даними* – представлення згенерованих даних у різних формах (таблиця, текст та інше), їх унаочнення, удосконалення та тому подібне.

1.3. Поширені помилки у створенні промптів:

1.3.1. *Загальність формулювання*.

До прикладу: Розкажи про кінологію.

Такий промпт не дає чітких вказівок, що призводить до поверхневих або неповних відповідей.

1.3.2. *Двозначність*.

До прикладу: Переклади текст. Проаналізуй текст.

Це формулювання потенційно не має ніяких обмежень, тому продукуватиме неоднозначні результати.

1.3.3. *Ускладненість*.

Якщо промпт надто довгий або має кілька завдань без логічної структури, ШІ може заплутатися.

1.4. Покроковий алгоритм створення промпта:

1.4.1. *Визначте мету запиту:* спрогнозуйте для себе кінцевий результат, його форму та обсяг.

1.4.2. *Сформулюйте контекст:* чи потрібні додаткові дані для ШІ, які б пояснили суть промпта?

1.4.3. *Структуруйте запит:* розпишіть етапність, обсяговість кожного кроку та кількість прикладів, підтем та іншого.

1.4.4. *Актуалізуйте етап тестування:* врахуйте попередні невдалі спроби тематичної сесії, аби вдосконалити завдання й зробити оптимальнішим результат.

1.5. Як покращити результат за допомогою оптимізації промптів:

1.5.1. Надайте приклад для ШІ.

1.5.2. Формулюйте запити й уточнюйте їх «по-живому».

До прикладу: Перефразуй текст, зберігаючи нейтральний стиль, проте зроби його коротшим.

1.5.3. Перевіряйте результат і вдосконалюйте промпт, актуалізуючи хибні елементи.

До прикладу: Проаналізуй текст повторно, виключаючи службові слова.

1.5.4. Локалізуйте питому стилістику ШІ: лексику, актуалізовані смисли.

1.6. Значення навичок формулювання промптів для лінгвістів:

Для сучасних фахівців-лінгвістів уміння формулювати запити до ШІ є ключовою навичкою, оскільки це дозволяє:

а) автоматизувати аналіз текстів;

б) працювати з багатомовними даними;

в) ефективно використовувати цифрові інструменти в дослідженнях, перекладах, викладанні та іншому.

Отже, якісно сформульований промпт є основною продуктивної роботи зі ШІ, а відтак – лише через розуміння принципів і практичне вдосконалення запитів можна досягти точних і надійних результатів.

2. Основна частина.

Коментар: кожна робота виконується на прикладах українськомовних та англійськомовних текстів трьох стилів – наукового, публіцистичного та художнього (див. шапку лабораторної роботи). У результаті має бути 6 уривків: 3 – англійськомовних, 3 – українськомовних (один уривок – півсторінки у Microsoft Word (Times New Roman, 14 шрифт, 1.5 інтервал)).

Завдання 1–2 виконуються з тими ж уривками (спочатку формулюєте промпт (завдання 1), згодом – удосконалюєте його, ускладнюєте (завдання 2)), завдання 3 – виконуєте з паралельною генерацією двох ШІ: спочатку – ChatGPT, а потім – Gemini. Робите аналіз 6 уривків за визначеними параметрами, потім – даєте пропмт по кожному з них для ШІ, робите порівняння в табличці. Після неї викладаєте розлогу аналітику (висновок щодо спільного, відмінного, тенденцій тощо). Така аналітика має ті самі мінімальні умови, що і уривок (див. кінець попереднього абзацу).

Зверніть увагу, що в процесі пошуку наукової статті потрібно використати статтю цього року (ліворуч поставити галочку навпроти поточного року) та

правильно формулювати запит. До прикладу: «Особливості машинного навчання в контексті цифрових гуманітарних наук» і так далі.

2.1. ChatGPT та Gemini.

Завдання 1. Формулювання базового промпта.

2.1.1. Пишете запит для ШІ, який дозволяє провести частотний аналіз тексту.

Приклад промпта: Проаналізуй цей текст і вкажи 10 найуживаніших слів.

2.1.2. Тестуєте промпт, отримуєте результати та аналізуєте їх.

2.1.3. Вивчаєте як можна зробити запит точнішим (до прикладу, конкретизувати, виключити службові слова абощо).

Завдання 2. Оптимізація промптів для складних завдань.

2.2.1. Формуєте складніше завдання.

До прикладу: Проаналізуй текст і визнач 5 ключових тем, вкажи приклади речень, які ілюструють кожну тему.

2.2.3. Етапи:

а) написати базовий промпт;

б) покращити промпт, щоб уточнити контекст і очікуваний результат;

в) порівняти відповіді ШІ для різних версій запиту.

Завдання 3. Генерація текстів за допомогою ШІ.

2.3.1. Написати промпт для генерації тексту: до прикладу, статті про новітні технології у лінгвістиці.

Базовий промпт: Напиши коротку статтю про використання цифрових технологій у лінгвістиці.

Удосконалений промпт: Напиши коротку статтю (300–400 слів) для наукового журналу про використання цифрових технологій у лінгвістиці. Зосередься на текстовому аналізі, машинному перекладі та корпусній лінгвістиці. Використовуй формальний академічний стиль.

2.3.2. Проаналізувати різницю в отриманих результатах і зробити висновки.

Завдання 4. Помилки у промптах.

2.4.1. Одна частина студентів формує набір погано сформульованих запитів: двозначних, неконкретних, загальних тощо або надто загальні, а інша – тестує й удосконалює його, локалізуючи помилки, переписує та обґрунтовує зміни.

Зразок:

Невдалий промпт: «Скажи щось про цей текст».

Виправлений промпт: «Зроби короткий аналіз цього тексту, визначаючи стиль, жанр і основну ідею».

2.2. Інші інструменти.

Коментар: кожне завдання робите англійською та українською мовами. Використайте *AI Presentation Maker* (потрібно зареєструватись) та *DALL E mini* (те саме). Мають бути згенеровані 2 презентації та 2 зображення (промпти англійською та українською).

2.2.1. Сформулюйте промт для генерації зображень, презентацій.

2.2.2. Проаналізуйте результати й визначте основні переваги та недоліки.

3. Заключна частина.

3.1. Аналіз отриманих результатів:

3.1.1. Чи вдалося отримати очікувані результати?

3.1.2. Як зміна формулювання вплинула на відповіді ШІ?

3.2. Дискусія:

3.2.1. Які характеристики роблять промпт ефективним?

3.2.2. Які обмеження ШІ можуть впливати на результат навіть при правильному промпті?

3.3. Огляд практичного застосування навичок створення промптів у науковій діяльності (письмово, наприкінці лабораторної роботи).

ЛАБОРАТОРНА РОБОТА № 3/4.

Тема: Корпуси текстів, використання цифрових технологій для їх аналізу (на прикладі AntConc).

Мета: ознайомитись на практиці з принципами відбору матеріалів для практичного дослідження мови за допомогою корпусів текстів; охарактеризувати сучасні цифрові технології, які застосовують у таких дослідженнях на прикладі вищезазначеної програми; навчитись головним принципам роботи інструментів для аналізу таких даних.

Завдання:

1. Ознайомитись з особливостями відбору даних для корпусних досліджень.

2. Навчитись формувати корпуси текстів із чітким баченням їх призначення.

3. Протестувати AntConc для роботи з окресленими даними, а також визначити алгоритм її використання.

4. Опанувати широкий спектр актуалізації корпусів текстів у контексті сучасних лінгвістичних досліджень.

Матеріали:

а) наукові тексти: Google Академія (https://tinyurl.com/5n6axph2)
ResearchGate (https://tinyurl.com/yue4jm6a)

б) художні тексти: Adobe.com (https://tinyurl.com/zcbwc8rj)
Gutenberg.org (https://tinyurl.com/3zzv8e23)
SDFO.org (https://tinyurl.com/3wzxc4ee)
Derevo-kazok.org (https://tinyurl.com/4eecbp4s)

в) публіцистичні тексти: BBC.com (https://tinyurl.com/3ya5nccj)
Kunsht.com.ua (https://tinyurl.com/2ve979dn)

Інструмент: AntConc (https://tinyurl.com/44nawsnp)

Результати роботи:

1. Формування та аналіз корпусів текстів: можливість створення власних корпусів текстів з обґрунтуванням вибору, а також визначення особливостей відбору тих чи тих даних залежно від дослідницьких завдань.

2. Виявлення лексичних та синтаксичних особливостей текстів: аналіз частотності слів, який дозволить локалізувати ключову лексику (поняття та їх варіації), та колокаційний аналіз, репрезентативний щодо самобутності структури текстів різних жанрів та мов.

3. Опанування самобутності алгоритму автоматизованої обробки таких даних: розуміння специфіки пошуку в корпусах репрезентативних даних, особливостей і складностей візуалізації тощо.

ПЛАН.

1. Вступна частина.

1.1.1. Визначення корпусної лінгвістики та її ролі в сучасних мовознавчих дослідженнях.

Корпусна лінгвістика – це галузь мовознавства, що займається збиранням, опрацюванням та аналізом великих обсягів текстів (корпусів) для вивчення мовних явищ і закономірностей. Вона відіграє важливу роль у сучасних мовознавчих дослідженнях, дозволяючи:

а) виявляти частотність вживання слів і фраз;

б) аналізувати стилістичні та граматичні особливості текстів;

в) використовувати цифрові технології для глибокого мовного аналізу;

г) розвивати автоматизовані системи обробки природної мови.

1.1.2. Принципи відбору текстових матеріалів для корпусних досліджень.

Для того, щоб корпус текстів був корисним для аналізу, він повинен відповідати певним критеріям, тобто бути:

а) репрезентативним – тексти мають відображати особливості певного стилю, жанру або епохи;

б) збалансованим – корпус має містити різні типи текстів (наукові, публіцистичні, художні тощо);

в) доступним – тексти повинні бути у відкритому доступі або з можливістю отримання дозволу на їх використання;

г) форматованим – бажано використовувати стандартизовані формати (*txt*, *xml* тощо) для подальшої обробки цифровими інструментами.

2. Основна частина.

Завдання 1. Формування корпуса текстів:

2.1.1. Вибір текстів.

Для аналізу потрібно вибрати три типи текстів:

а) науковий текст (академічна або спеціалізована література);

б) публіцистичний текст (газети, журнали, блоги, аналітика);

в) художній текст (оповідання, роман, новела).

Вибираєте по одному тексту українською та англійською мовами, тобто у вас знову має вийти 6 текстів на «виході».

До прикладу: зберігаєте сторінку кожного стилю, яку вважаєте найпоказовішою в плані вашого дослідження (лексика, морфологія, синтаксис та інше).

2.1.2. Обґрунтування вибору текстів.

Має відповідати меті дослідження. До прикладу:

– науковий текст допомагає зрозуміти колокації в академічному дискурсі;

– публіцистика відображає актуальні тенденції мови, публічну комунікацію;

– художня література демонструє стилістичні особливості сполучуваності слів.

Також доцільно акцентувати на тому, чому вибрано саме ці джерела.

До прикладу:

1. Вони мають високу якість та авторитетність.

2. Репрезентують сучасне мовлення (для публіцистики).

3. Художні тексти є класичними й містять багатий мовний матеріал.

2.1.3. Підготовка корпуса.

Формати корпуса.

Вищезазначено, що текстовий корпус текстів має бути у форматі, сумісному з аналітичними інструментами: ***txt*** (простий текст без форматування); ***xml*** (структурований формат для збереження метаінформації про тексти).

Доцільно використати звичайний «Блокнот» (також можна послуговуватись подібними програмами: Notepad++ та іншими): відкриваєте програму, копіюєте у неї потрібні фрагменти і зберігаєте файл (файл буде мати потрібне вам у виконанні роботи розширення – ***txt***).

Ручний вибір текстів:

1. Завантажити тексти у відкритому доступі (див. попередні матеріали).

2. Скопіювати тексти вручну.

2.1.4. Фінальний крок.

Перевірити, чи корпус правильно відображається в інструментах аналізу (на прикладі AntConc).

Завдання 2. Аналіз частотності лексики.

2.2.1. Спочатку вручну визначити, а потім – за допомогою AntConc 10 найчастотніших слів у кожному корпусі.

2.2.2. Виключити стоп-слова (лише для української мови) та повторити ручний і машинний аналізи.

2.2.3. Представити результати у структурованому вигляді (у таблиці довільної форми, де головною є аргументованість і стислість представлених даних).

2.2.4. Зробити висновок щодо продуктивності вищезазначеного інструменту для роботи лінгвістів, аргументувати думку.

Завдання 3. Алгоритм роботи з AntConc.

AntConc – безкоштовний інструмент для корпусного аналізу, який дозволяє досліджувати частотність слів, знаходити колокації, будувати конкорданс (алфавітний перелік усіх слів будь-якого тексту з наведенням контекстів їх використання) і виконувати інші статистичні операції.

2.3.1. Підготовка даних.

Вищезазначені у 2.3 дії.

Важливо:

а) кодування файлів має бути звичайне (UTF-8), щоб уникнути проблем із відображенням символів;

б) необхідно перевірити наявність будь-яких зайвих символів форматування (до прикладу: спецсимволів з PDF або HTML).

Якщо етап 2.3 виконано правильно, то з кодуванням проблем не буде, а наявність спеціальних символів доцільно перевірити, перечитавши й видаливши їх з текстів для аналізу. Знову-таки, краще обрати якісне джерело, щоб з нього «витягти» тексти, які не потрібно багато «чистити».

2.3.2. Завантаження корпуса, пошук по ньому.

1. На Laurence Anthony's Website (https://tinyurl.com/44nawsnp) завантажте Windows (Portable) (4.3.1) в "Official releases".

2. Після завантаження архіву, розпакуйте його та запустіть однойменний файл у ній (за потреби від імені адміністратора).

3. У програмі ліворуч натисніть меню "File", далі оберіть – "Open as Quick Corpus".

4. Після цього виберіть усі файли, які потрібно аналізувати (ті, які були підготовлені вами й збережені у форматі ***txt*** під час виконання попередніх кроків). Ці тексти з'являться у вікні знизу (ліворуч).

5. Оберіть праворуч вкладку "File View" і зачекайте завантаження вмісту певного файлу.

6. Після того, як вміст був завантажений, спробуйте пошукати певне ключове слово в цьому файлі, вписавши його у поле пошуку.

7. Спочатку впишіть певне слово, яке точно є у файлі, а потім спробуйте вписати його форму із * (це дозволить відобразити різні форми слова: вони будуть позначені кольором у вмісті файлу). Поекспериментуйте з частинами мови, службовими словами, щоб зробити певні висновки та розкажіть про них наприкінці виконання лабораторної роботи й аргументуйте.

2.3.3. Аналіз слова у контексті.

1. Перейдіть на вкладку “KWIC”, яка знаходиться праворуч.

2. Так само працюйте з пошуковим рядком, експериментуючи з * та різними формами, частинами мови тощо. Мета – узагальнити можливості автоматизованої роботи з контекстом. При цьому зверніть увагу на опції знизу “Sort 1” та “Sort 2”, за допомогою яких можна визначити позицію слова (третє зліва й так далі), а також сортування за “Freq”, тобто частотністю.

3. Ставте галочку на “Adv search” (на одному рівні з пошуковим рядком, праворуч від кнопки “Start”).

4. У вікні, що відкрилось, ставте галочку на “Search a Query List” і ліворуч від кнопки “Add” вписуйте те, що ви хочете пошукати. Після того, як вписали потрібне слово (краще з *), натисніть кнопку “Add”: якщо все зроблено правильно, то воно з’явиться у полі нижче.

5. Натискайте “Context Search” і налаштовуйте кількість слів ліворуч та праворуч, а також їх поєднання (“OR” або “AND”). Далі – у рядку, що знаходиться нижче, задайте потрібне поєднання шуканого слова з іншим (не забувайте про *, бо так відшукаєте й форми слова), що буде репрезентативним у контексті теми поточної роботи.

6. Натискайте “Add”, а внизу вікна “Apply”, а у вкладці “KWIC” –“Start”, щоб параметри пошуку видозмінили результати аналізу програми. Зробіть висновки, за потреби – поекспериментуйте декілька разів, щоб досягти бажаного результату.

2.3.4. Вивчення слова у парі.

1. Оберіть вкладку “Plot”, використайте пошуковий рядок, сформулювавши у ньому запит із *, натисніть “Start”.

2. Мають з’явитись смуги різної широти, які показують місце й тривалість появи шуканого слова. На кожну смугу можна клікнути, щоб подитивись місце, де слова знаходяться у вашому корпусі.

3. Дисперсія (отримані дані відносно середнього значення у вибірці) в результаті вказує на розподіл слова по аналізованому тексту. Зокрема, висвітлено рівномірність його появи (чим ближче до 1, тим такий розподіл гармонійніший, чим ближче до 0 – уривчастий). На вашу думку: чи враховано тут значення певного слова? Якщо так, то яким чином? Відповідь аргументуйте. Чи потрібно звіряти вручну? Поясніть свою думку.

2.3.5. Зв’язок між парами слів.

1. Вкладка “Cluster” дозволяє вивчити зв’язок між словами. Можна шукати за розташуванням ключового слова (ліворуч, праворуч), оскільки це впливає на його значення. Зверніть увагу, що можна налаштувати частотність появи: до прикладу, від двох разів і більше.

2. Вкладка “N-Gram”: схожа на попередню, проте відрізняється від неї тим, що такі слова можна запускати без пошуку ключових слів. Поекспериментуйте з обома вкладками, щоб мати змогу зробити певні висновки щодо їх функціоналу.

2.3.6. Стоп-слова (лише для української мови!!!).

1. Для того, щоб результати роботи програми були точнішими доцільно додати стоп-слова.

2. Заходьте у “Settings” – “Global Settings” – “Tool Filters” – ставте галочку на “Cluster/N-Gram…”.

3. Для цього візьміть список, створений Сергієм Купрієнком на його сторінці у GitHub (https://tinyurl.com/5azx32t9): завантажте звідти файл “stopwords_ua.txt”, додайте його за допомогою кнопки “Add File” у вищезазначеному вікні й натисніть на “Start”.

4. Зверніть увагу на параметр “Case”, використання якого дозволяє шукати запити за чіткою відповідністю великої/малої літери. До прикладу, в шуканому словосполученні локалізуватись будуть лише ті, де перше слово буде писатись із великої букви.

2.3.7. Аналіз частотності слів.

1. Для аналізу частотності використовуйте, розташовану праворуч вкладку “Word”, а також перевірте, щоб була галочка знизу на “Words”.

2. Натисніть кнопку “Start”, якою розпочинаєте процес аналізу всіх відібраних вами текстів. Результатом цього аналізу буде список слів з частотністю їхнього вживання (колонка “Freq” у цьому списку, при цьому знизу має бути обрано “Sort by Frequency” – стоїть за замовчуванням, проте перевірте).

2.3.8. Функціонал “Keyword”.

1. Першу з вищезазначених вкладок можна використовувати за умови наявності згаданого у лекціях контрольного корпуса або референсу. Для того, щоб отримати його, обміняйтесь текстами між собою: таким чином, контрольним стане корпус, створений колегою.

2. Візьміть цей файл і додайте його: “File” – “Open Corpus Manager” – “Reference Corpus” (праворуч).

3. Наступний крок – ліворуч виберіть “Raw Files” і натисніть “Add Files” або “Add Directory”, а потім клікніть на кнопку “Create” (зліва).

4. Тепер мав змінитися інтерфейс (ліворуч знизу з’явилося додаткове віконце). Після цього запуск “Keyword” буде показувати слова, характерні для створених вами корпусів, проте не питомі для контрольного.

5. Аби зробити зворотну дію (переставити корпуси місцями), перейдіть у “File” – “Open Corpus Manager” – “Reference Corpus” – “Swap (with target).”

2.3.9. Функціонал “Wordcloud”.

Для того, щоб створити «Хмару слів» необхідно лематизувати слова – це процес вимагає вже глибших знань, тому не розглядайте її.

3. Заключна частина.

3.1. Аналіз отриманих результатів:

3.1.1. Чи вдалося досягти поставлених цілей?

3.1.2. Які труднощі виникли при роботі з програмою?

3.1.3. Чи перспективною є автоматизація обробки тексту?

3.2. Дискусія:

3.2.1. Які переваги та недоліки цифрових методів аналізу корпусів?

3.2.2. Як можна покращити алгоритм аналізу текстів?

ЛАБОРАТОРНА РОБОТА № 5/6

Тема: Створення бітексту та паралельного корпуса з фрагмента художнього тексту.

Мета: ознайомитися з принципами укладання бітексту та створення паралельного корпуса; вивчити алгоритм роботи з цифровими інструментами для вирівнювання текстів, їх лематизації, тегування та подальшої роботи з корпусом.

Завдання:

1. Підготувати тексти оригіналу та перекладу для подальшої обробки.

2. Виконати автоматичне вирівнювання текстів за допомогою програми NOVA Text Aligner.

3. Опанувати процес лематизації та тегування текстів на платформі UDPipe.

4. Створити паралельний корпус текстів на платформі Sketch Engine.

5. Протестувати функціонал пошуку в укладеному паралельному корпусі.

Матеріали:

а) текст роману «Harry Potter and the Deathly Hallows» (англійською мовою: https://tinyurl.com/ms5w7sew) або рівнозначний йому на ваш вибір;

б) текст роману «Гаррі Поттер і Смертельні реліквії» (українською мовою: https://tinyurl.com/37st7vbu) або рівнозначний йому на ваш вибір.

Інструменти:

NOVA Text Aligner (програма для вирівнювання текстів – 30-денна безкоштовна пробна версія: https://tinyurl.com/4b8487rb);

UDPipe (платформа для лематизації та тегування: https://tinyurl.com/3m78yhu5);

Sketch Engine (платформа для укладання корпусів – 30-денна безкоштовна пробна версія: https://tinyurl.com/4b3zs3bk).

Результати роботи:

1. Створення бітексту: отримання вирівняного паралельного тексту, що поєднує оригінал і переклад.

2. Формування корпуса: на основі вирівняних текстів укладання двомовного паралельного корпуса, готового до аналізу.

3. Аналіз даних: опанування алгоритму автоматизованої обробки текстів, що включає підготовку, вирівнювання, лематизацію та створення корпуса.

4. Пошук одиниць (слів і конструкцій) у створеному корпусі й виявлення перекладацьких відповідників: опанування корпусного пошукового інструментарію.

ПЛАН.

1. Вступна частина.

1.1. Визначення бітексту та паралельного корпуса.

Бітекст – це текстова структура, що складається з двох версій одного й того ж тексту, оригінальної та перекладеної, вирівняних на рівні абзаців або речень.

Паралельний корпус – це колекція текстів, що складається з документів, які є перекладами одне одного, що дозволяє досліджувати перекладацькі відповідники та мовні закономірності.

1.2. Принципи відбору текстових матеріалів.

Для роботи обрано художні тексти, що мають високу якість як тексту оригіналу, так і тексту перекладу та репрезентують сучасну мову (англійську та українську). Тексти роману та його перекладу оцифровані та підготовлені до автоматичної обробки.

2. Основна частина.

Завдання 1. Підготовка текстів.

2.1. Збір метаданих.

Заповніть картки метаданих для оригіналу (англійською мовою) і перекладу (українською мовою):

Картка метаданих оригіналу художнього тексту

Критерій	**Метадані**
Назва мовою оригіналу	
Мова оригіналу	
Автор (повне ім’я мовою оригіналу)	
Дата створення тексту	
Дата публікації тексту	
Місце публікації (країна, місто)	
Видавництво	

Картка метаданих перекладу художнього тексту

Критерій	**Метадані**
Назва перекладу	
Мова перекладу	
Автор (повне ім’я мовою перекладу)	
Перекладач	
Дата публікації перекладу	
Місце публікації (країна, місто)	
Видавництво	

2.2. Очищення й технічна підготовка.

2.2.1. З текстів видаліть нетекстові елементи, зміст, бібліографію, внутрішньо- і післятекстові виноски; інформацію, що міститься на титульному аркуші, контртитулі, додатки (якщо є) і таке інше.

2.2.2. Виконайте корегування некоректних знаків (тире-дефіс, лапки тощо).

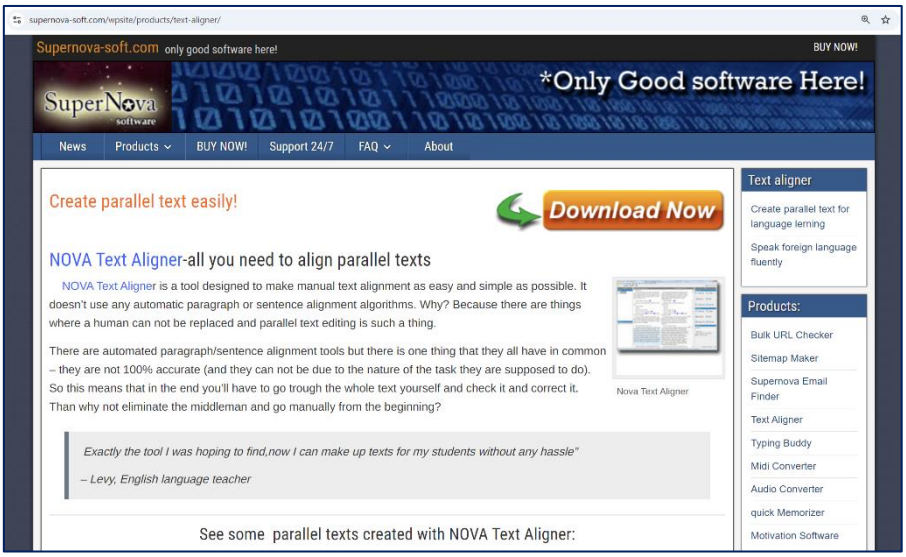
2.2.3. Проведіть автоматичне корегування (автозаміною) пунктуації й форматування: до прикладу, заміна «трикрапка + пробіл» на «трикрапка + знак абзацу» для покращення автоматичного вирівнювання.

2.3. Збереження.

Підготовлені тексти збережіть в окремих файлах у форматі ***txt*** та ***pdf*** для подальшої роботи з програмами. Обсяг текстів – будь-який фрагмент тексту оригіналу обсягом 5 000 слів + переклад цього фрагмента українською мовою.

Завдання 2. Вирівнювання та лематизація.

2.4. Вирівнювання за допомогою NOVA Text Aligner.



2.4.1. Імпортуйте тексти в програму одним із способів:

а) імпорт файлів у форматі ***txt***;

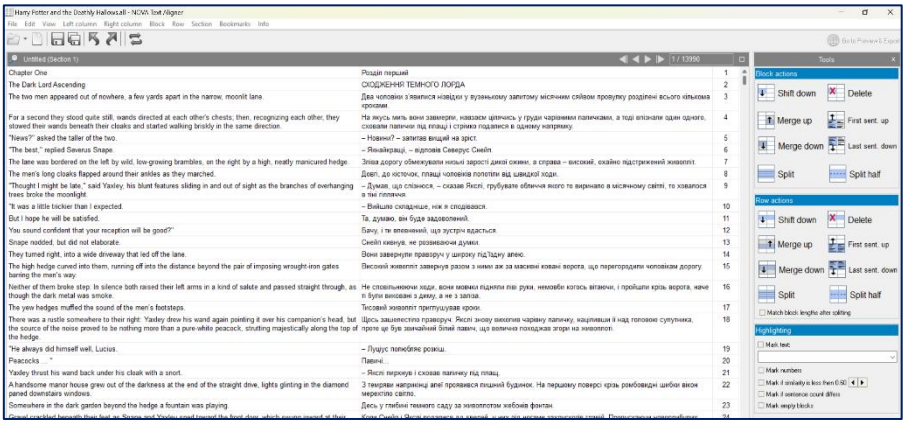
б) імпорт файлів у форматі ***pdf***;

в) копіювання й вставлення текстів.

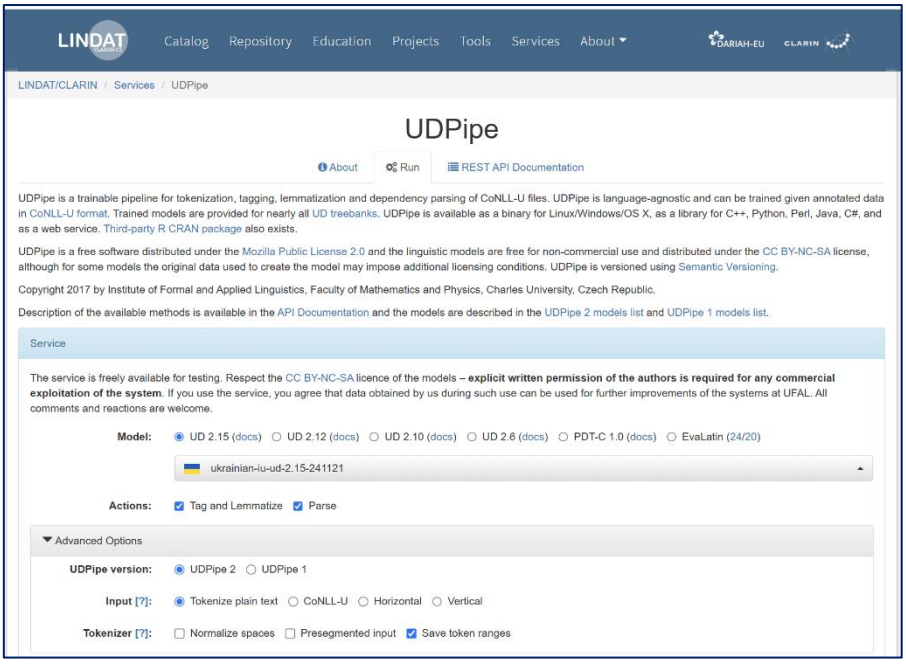
2.4.2. Проведіть автоматичне вирівнювання на рівні речень.

2.4.3. Виконайте ручне корегування результатів, де це необхідно.

2.4.4. Після завершення вирівнювання експортуйте паралельний текст у форматах ***xls*** та ***rtf***.



2.5. Лематизація та тегування за допомогою UDPipe.

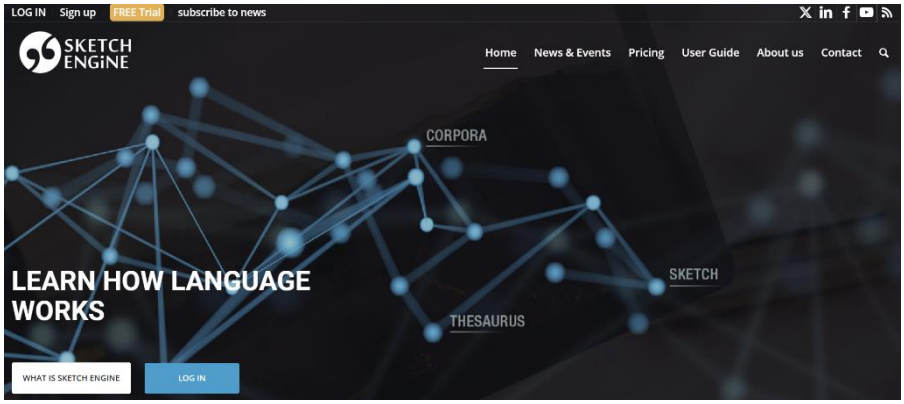


2.5.1. Обидва тексти по черзі завантажте на платформу.

2.5.2. Виконайте токенізацію, лематизацію та тегування.

2.5.3. Результати збережіть у форматі ***conllu***.

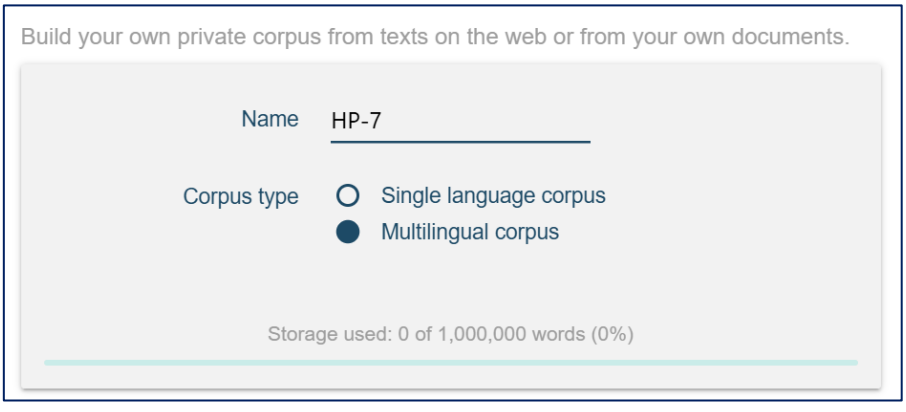
Завдання 3. Створення паралельного корпуса в Sketch Engine.



2.6. Налаштування корпуса.

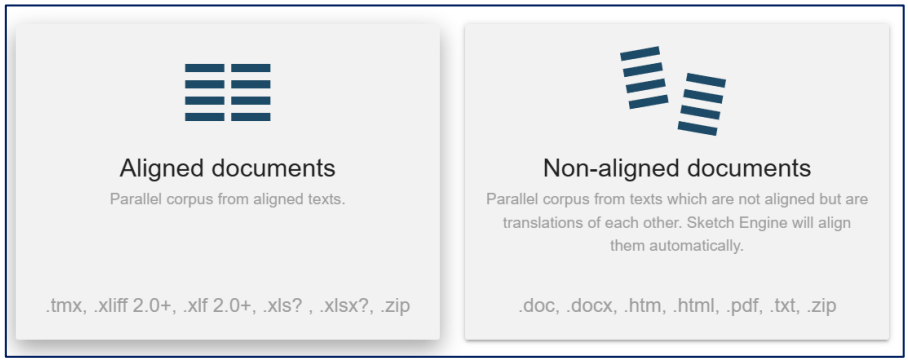
2.6.1. Задайте назву корпуса.

2.6.2. Оберіть тип «Багатомовний».

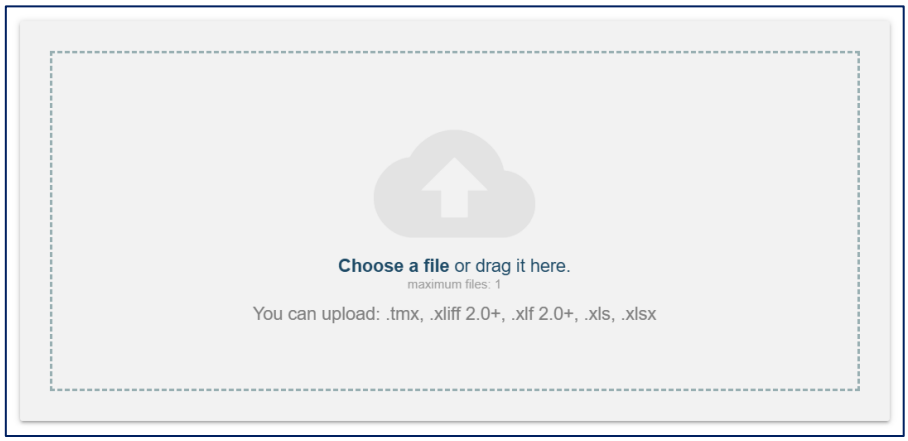


2.7. Завантаження даних.

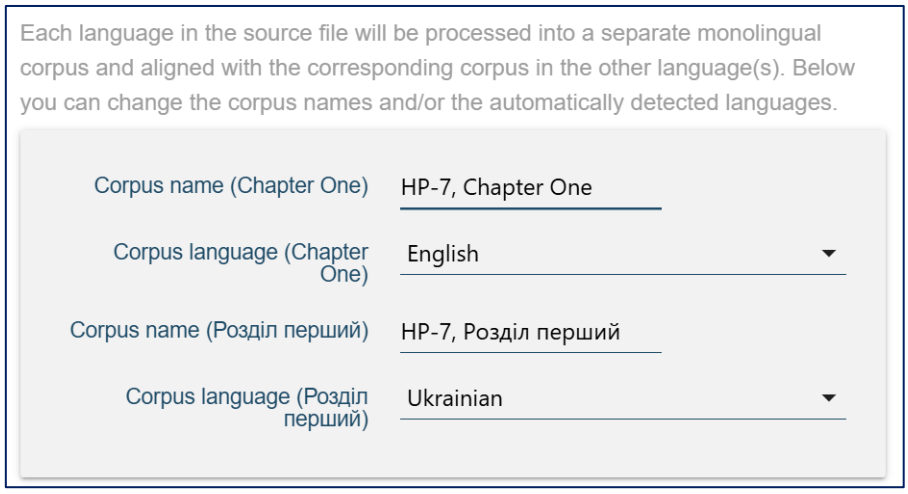
2.7.1. Оберіть опцію «вирівняні тексти».



2.7.2. Завантажте файл ***xls*** із вирівняним текстом.



2.7.3. Призначте мову для кожного тексту: англійську та українську.

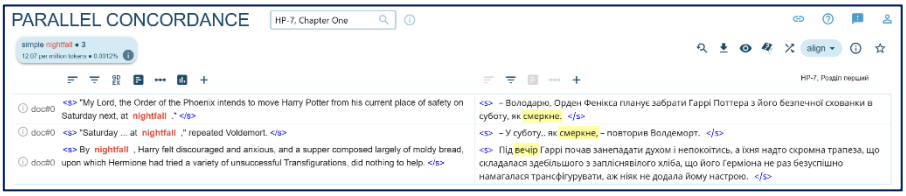


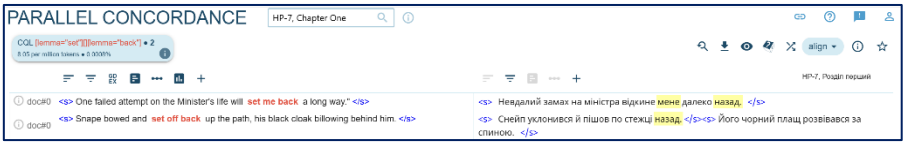
2.8. Робота з корпусом.

2.8.1. Скористайтеся опцією «Відкрити паралельний конкорданс».



2.8.2. У створеному корпусі протестуйте функціонал запитів: (1) простий пошук іменників та (2) пошук дієслівної конструкції з трьох слів.





3. Заключна частина.

3.1. Аналіз отриманих результатів:

3.1.1. Чи були виконані завдання: створено бітекст, проведено лематизацію та укладено паралельний корпус?

3.1.2. Які труднощі виникли на етапі ручного корегування автоматичного вирівнювання?

3.1.3. Чи перспективним є процес автоматизації обробки тексту? Чи пришвидшує він роботу з великими обсягами даних?

3.2. Дискусія:

3.2.1. Які переваги та недоліки цифрових методів обробки великих масивів текстових даних?

3.2.2. Як можна вдосконалити алгоритми очищення текстів та автоматичного вирівнювання, щоб мінімізувати потребу в ручному втручанні?

ЛАБОРАТОРНА РОБОТА № 7/8

Тема: Робота антиплагіатних та інших систем.

Мета: ознайомитися з особливостями роботи антиплагіатних та інших систем (видами й алгоритмами роботи з текстовими даними); навчитися виявляти тексти за допомогою галузевих відповідних сервісів; порівняти їх ефективність та оцінити продуктивність використання для виявлення плагіату, спаму, «води», стилістичної й граматичної перевірки; сформулювати висновки щодо ефективності, точності та обмежень кожної з таких систем.

Завдання:

1. Ознайомитись із специфікою роботи антиплагіатних та інших систем у контексті прикладної лінгвістики.

2. Навчитись працювати з текстовими даними для такого аналізу (оригінальними та з елементами плагіату: перекладеним текстом, після рерайту, з великою кількістю «води» тощо).

3. Провести перевірку текстів за допомогою онлайн-сервісів.

4. Проаналізувати результати проведеної перевірки, зробити висновки та аргументувати точність та ефективність кожної системи.

Матеріали:

а) нормативно-правові тексти:

Верховна Рада України (https://tinyurl.com/45yssd4t)

Президент України (https://tinyurl.com/4aj4tm8p)

European Commission (https://tinyurl.com/2zzuvusn)

EUR-Lex (https://tinyurl.com/hedm9kd8)

б) наукові тексти: Google Академія (https://tinyurl.com/5n6axph2)

ResearchGate (https://tinyurl.com/yue4jm6a)

в) публіцистичні тексти: BBC.com (https://tinyurl.com/3ya5nccj)

Kunsht.com.ua (https://tinyurl.com/2ve979dn)

Інструменти:

програми: Copyscape (https://tinyurl.com/nhmxmb9u)
Copywritely (https://tinyurl.com/4rk98n39)
Grammarly (https://tinyurl.com/2hykncsw)
Plag (https://tinyurl.com/9s394ycb)
Plagiarism Checker (https://tinyurl.com/y58xf549)
PlagiarismChecker.co (https://tinyurl.com/jrhmjyuk)
Plagiarism Detector (https://tinyurl.com/43j99kcj)
PlagiarismSearch.com (https://tinyurl.com/58yfsmwc)
Quetext (https://tinyurl.com/45npywfh)
StrikePlagiarism (https://tinyurl.com/2ewvxd8b)

ШІ: ChatGPT (https://tinyurl.com/ycypp7h3)
Gemini (https://tinyurl.com/yf9xapfj)
Grok (https://tinyurl.com/437smc3c).

Результати роботи:

1. Формування, пошук та аналіз даних для їх перевірки на плагіат, помилки, «воду» тощо: зокрема, набуття навичок удосконалення тексту.

2. Розуміння принципів та алгоритмів обробки текстових даних та «ухвалення» рішень вищезазначеними системами тощо.

3. Осягнення специфіки автоматизованої обробки текстових даних щодо виявлення їх подібності (логіки викладу, порядку, структури тощо) та розуміння самобутності видів текстів в контексті цього процесу.

ПЛАН.

1. Вступна частина.

Здебільшого робота з текстовими даними представлена у практичній площині: це виявлення помилок, «води», а також машинної (нелюдської) генерації тексту. Зосередимося більшою мірою на *антиплагіатних системах*, які є цифровими інструментами. Призначення таких інструментів – локалізація

збігів, відстеження запозичень та, власне, аналіз унікальності тексту. Застосування таких систем є важливим для цілої низки галузей, серед яких виокремимо освіту та науку.

Головними *принципами роботи таких систем є*:

1. Пошук за збігами – виявлення фрагментів (речень, абзаців, сторінок тощо), які дослівно подібні іншому контенту з бази даних або відкритих джерел (репозитарів, порталів та іншого).

2. Семантичний аналіз – вивчення лексико-семантичної подібності наявних у досліджуваних текстових даних фраз для виявлення рерайту, перекладу й тому подібного.

3. Аналіз структури – відстеження синтаксичних змін, синонімії, пасивних конструкцій та іншого, використовуваних із метою маскування плагіату.

Типами окреслених систем є: а) вузькоспеціалізовані (до прикладу, StrikePlagiarism тощо); б) масові (до прикладу, Plag і такі інші); в) гібридні (до прикладу, Grammarly і так далі).

2. Основна частина.

Коментар: за можливості не просто аналізуйте матеріал, а й виявляйте потенційні помилки, «воду» та інше. Зазначайте правильність таких результатів у аналізованих вами інструментах.

Завдання 1. Ознайомитись із вищезазначеними інструментами.

2.1.1. Відвідайте вебсайти кожного такого інструмента та коротко охарактеризуйте, заповнивши Табл. 1. Зверніть увагу на значущість їх алгоритмів, роботи тощо в контексті прикладної лінгвістики та зробіть під таблицею висновок щодо їх роботи лише **на підставі опису розробників**, аргументуйте. ШІ не включайте у Табл. 1:

Табл. 1. Порівняння інструментів (на базі даних розробників).

№	Назва	Характеристика	Оцінка

Завдання 2. Підберіть тексти для аналізу (перелік – у шапці лабораторної).

Коментар: оскільки у програм є обмеження на кількість символів, то **перевіряйте тексти частинами**. Не забувайте також викладати у таблицях покликання на підібрані тексти.

2.2.1. Підберіть тексти (див. *Матеріали*): англійськомовні та українськомовні (обсяг – 1 сторінка: Times New Roman, розмір – 14, інтервал – 1.5, абзац – 1.25).

2.2.3. Проаналізуйте їх, використавши Copywritely, Grammarly, Plag, Plagiarism Checker, PlagiarismChecker.co, Plagiarism Detector, PlagiarismSearch.com, Quetext.

2.2.4. Запишіть результати у Табл. 2, а під нею зробіть висновок щодо ефективності роботи та іншого виключно **на підставі власних дослідів із ними**, аргументуйте.

Табл. 2. Порівняння інструментів (на базі власних спостережень).

№	Назва	Характеристика	Оцінка

Завдання 3. Створіть власні тексти й проаналізуйте їх за допомогою тих самих інструментів.

Коментар: оскільки у програм є обмеження на кількість символів, то **перевіряйте тексти частинами**.

2.3.1. Підготувати власні англійськомовні та українськомовні тексти для аналізу вищезазначеними програмами (вимоги такі самі, як і у попередньому завданні).

2.3.2. Нормативно-правові тексти робити за зразком будь-якого документа (заходите на портал у матеріалах і наслідуєте його стиль, зміст, видозмінюючи під свій задум) у двох варіантах (англійськомовний та українськомовний). Наукові та публіцистичні робите за таким же принципом, або лише спираючись

на власну фантазію щодо тематики (головне – дотримати стиль). Так само в двох варіантах кожен.

2.3.3. Проаналізувати зроблені тексти та заповнити Табл. 3, подібну до Табл. 1 та 2, що була у 2.1.1 та 2.2.4. Під таблицею зробити висновок щодо ефективності роботи та іншого виключно **на підставі власних дослідів із ними**, аргументувати.

Таблиця 3. Аналіз інструментів (на базі власних спостережень).

№	Назва	Характеристика	Оцінка

Завдання 4. Проаналізувати власні тексти, поділивши їх попередньо за мовами (1 файл – англійськомовні, 1 – українськомовні), *за допомогою ШІ* (див. перелік в *Інструменти*).

2.4.1. Сформулювати запит щодо їх оригінальності у трьох ШІ та записати результати у Табл. 4. Під таблицею зробити висновок щодо ефективності роботи та іншого виключно **на підставі власних дослідів із ними**, аргументувати.

Табл. 4. Порівняння ШІ.

№	Назва	Характеристика	Оцінка

3. Заключна частина.

Загальний аналіз отриманих результатів на базі ваших висновків під Табл. 1–4:

3.1. Чи вдалося досягти поставлених цілей?

3.2. Які труднощі виникли при роботі? Наскільки корисним був кожен інструмент?

3.3. Чи перспективним є використання ШІ як інструмента роботи з текстом? Аргументуйте свою відповідь у загальному висновку.

ЛАБОРАТОРНА РОБОТА № 9/10

Тема: Методи фактчекінгу та аналізу достовірності інформації у цифрових текстах.

Мета: ознайомитися зі специфікою фактчекінгової перевірки текстів; навчитися перевіряти факти та виявляти дезінформацію в них за допомогою цифрових інструментів і лінгвістичних методів; удосконалити розуміння принципів роботи фактчекінгу та критичного аналізу джерел.

Завдання:

1. Ознайомитись із теоретичними основами фактчекінгової перевірки текстових даних.

2. Засвоїти фактчекінгові методи й практики у широкому спектрі: від перевірки поодиноких фактів до виявлення маніпуляцій у певних кампаніях.

3. Здійснити перевірку текстів за допомогою цифрових інструментів (див. *Інструменти*).

4. Здійснити перевірку текстів вручну (на підставі власних знань про верифікацію даних).

4. Проаналізувати результати обох перевірок (ручної та машинної), зробити висновки та аргументувати значення й позитивний/негативний досвід роботи з такими інструментами.

Матеріали: медіатексти (публікації, новини, аналітика, інтерв'ю).

Інструменти: Google Fact Check Explorer: https://tinyurl.com/2ds3szb4

PolitiFact: https://tinyurl.com/mubw4ffn

Snopes: https://tinyurl.com/3rnyhk2u

StopFake: https://tinyurl.com/7v6z32y2

VoxCheck: https://tinyurl.com/37tdketd

За бажанням, для поглибленого аналізу:

Botometer: https://tinyurl.com/5e52r3jw

OsoMeNet: https://tinyurl.com/y27pxnes

Результати роботи:

1. Формування, пошук та аналіз даних для їх фактчекінгової перевірки: вручну та автоматично.

2. Розуміння принципів, алгоритмів та іншого такої обробки даних щодо достовірності наведених у них тверджень.

3. Аналітичний висновок (мінімум півсторінки у Microsoft Word, 14 шрифт, 1.5 інтервал, відступ – 1.25) щодо особливостей такої перевірки, її актуальності та складності ручного та автоматичного аналізу.

ПЛАН.

1. Вступна частина.

Фактчекінг – це процес перевірки достовірності фактів, викладених у тексті, медіа, відео чи зображенні. Його мета – відрізнити правдиву інформацію від маніпулятивної, фальшивої або викривленої.

Види дезінформації:

а) фейки (відверта вигадка);

б) маніпуляції (зміщення контексту);

в) дезінформація (навмисне викривлення);

г) мізінформація (ненавмисне викривлення).

Методи фактчекінгу:

1. Семантичний аналіз тексту.

2. Прагматичний аналіз контексту.

3. Перевірка тверджень через фактчекінгові платформи.

4. Порівняння тверджень з авторитетними джерелами.

5. Аналіз дискурсу для виявлення маніпуляцій.

2. Основна частина.

Коментар: зверніть увагу на оформлення кожного етапу лабораторної роботи й набір інструментів (див. *Інструменти*), який потрібно використати.

Завдання 1. Перевірка фактів у цифрових текстах.

2.1.1. Знайдіть 20 тверджень із текстів (українськомовних й англійськомовних) із відібраних власноруч ресурсів. Кожне твердження проаналізуйте у Табл. 1 за поданим зразком.

Таблиця 1. Порівняння тверджень з текстів.

№	Твердження	Ресурс	Результат перевірки	Коментар
			Автоматична перевірка:	
			Ручна перевірка:	

2.1.2. Під час перевірки використайте мінімум 2 різні інструменти, а також перевірте факт вручну (українськомовні й англійськомовні тексти можуть мати різні твердження). Під час ручної перевірки «пробивайте» дані двома мовами, щоб порівняти ефективність інструментів на кожній із них.

2.1.3. Намагайтесь урізноманітнити використані інструменти й ресурси, орієнтуючись так, що одне твердження перевірити за першими, а інше – за другими.

2.1.4. Знайдіть 10 прикладів мовних маніпуляцій у вибраних вами текстах та проаналізуйте їх у Табл. 2, у якій має бути описано: маніпуляцію (назвати) – її суть (означити) – альтернативу (оригінальний текст новини, яку було перекручено, з покликанням) – коментар (що саме було змінено і як це вплинуло на сприйняття тексту, чому це важливо і як вплине на людей без належного рівня медіаграмотності тощо). Під таблицею зробіть невеликий висновок, аргументуйте його.

Таблиця 2. Мовні маніпуляції.

№	Маніпуляція	Характеристика	Оригінал	Коментар

Завдання 2. Інверсивний аналіз.

2.2.1. Перейдіть на Телеграм-канал (https://tinyurl.com/2t592h7s) або сайт «Центру протидії дезінформації» (https://tinyurl.com/384ejewj) і знайдіть там останні 10 спростувань певних фейків або фактів та, покладаючись на ваше розуміння суті дезінформації, мізінформації та пропаганди, «реставруйте» оригінальні тексти (без їх реального пошуку).

2.2.2. Опишіть ті зміни, які, на вашу думку, було внесено в тексти або інший контент, та поясніть їх суть та значення в контексті вищезазначених явищ (дезінформації, мізінформації та пропаганди).

2.2.3. Відшукайте оригінальну новину або інший контент та порівняйте зі своїми міркуваннями, зробіть аналіз правильності ходу ваших думок щодо цього, аргументуйте.

Завдання 3. Віднаходження фейку.

2.3.1. Знайдіть три зображенням або відео в соціальних мережах, які є фейком. Дайте покликання на кожний і охарактеризуйте: як саме можна зрозуміти, що це підробка.

До прикладу: фото українського солдата, який просить поставити вподобайку та поширити, що насправді є контентом, створеним ШІ.

Завдання 4. Використання OsoMeNet та/або Botometer **(за бажанням).**

2.4.1. Відслідкуйте поширення фейку або новини у соцмережах.

2.4.2. Виявіть кількість джерел, особливості поширення (визначте його тип: природне чи кероване).

3. Заключна частина.

3.1. Аналіз результатів:

3.1.1. Які твердження виявились фейковими, а які – достовірними? Чому? Яка різниця між ручним та автоматичним фактчекінгом? Який із них швидший та зручніший? Аргументуйте.

3.1.2. Які методи виявились найефективнішими? Які методи ви б хотіли спробувати або розказати про них рідним/друзям/знайомим, аби убезпечити їх?

3.1.3. Які джерела заслуговують довіри? Чому саме вони? Чи вважаєте ви урядові канали інформації надійними й достовірними?

3.2. Дискурсія:

3.2.1. Чи допомагає фактчекінг знизити ризики маніпуляцій? Чи може мова ворожнечі бути частиною маніпуляцій? Наведіть приклади.

3.2.2. Які обмеження має навіть найкращий інструмент фактчекінгу? Чи має обмеження ручний аналіз? Який із них зручніший на вашу думку?

ЛАБОРАТОРНА РОБОТА № 11

Тема: Машинний текст та його виявлення: автоматизовані та лінгвістичні підходи.

Мета: ознайомитися з специфікою локалізації машинного (згенерованого, алгоритмізованого) тексту, методами його ручного й автоматичного та лінгвістичного аналізу; удосконалити розуміння принципів роботи безкоштовних інструментів виявлення вищезазначеного тексту та проведення базового лінгвістичного аналізу.

Завдання:

1. Ознайомитись з теоретичними основами перевірки текстових даних на машинну та ручну генерацію.

2. Засвоїти базові методику ручної та автоматизованої перевірки таких текстів.

3. Здійснити перевірку текстів вручну (на підставі власних знань) та використавши окреслені у роботі інструменти (див. *Інструменти*).

4. Проаналізувати результати вищезазначених перевірок та зробити висновки, оцінивши власний досвід взаємодії з ними.

Матеріали: медіатексти (публікації, новини, аналітика, інтерв'ю).

Інструменти: GPTZero: https://tinyurl.com/5hj87uvx
Sapling AI Detector: https://tinyurl.com/2s3u53v3
AI Detector Pro (Free Plan): https://tinyurl.com/3uada3en
Content at Scale AI Detector: https://tinyurl.com/2s4cevc4

Результати роботи:

1. Ознайомитися з питомими підходами до виявлення машинного (згенерованого тексту).

2. Виявити точність автоматизованих сервісів у контексті ручного лінгвістичного аналізу.

3. Локалізувати семантичну та стилістичну самобутність згенерованого тексту, яка лишається поза увагою автоматичного аналізу.

ПЛАН.

1. Вступна частина.

Машинний текст – це текст, згенерований ШІ (зокрема, великою мовною моделлю: ChatGPT, Gemini тощо). Виявлення такого тексту є актуальною проблемою в контексті академічної доброчесності, журналістики, медіаграмотності.

Головними підходами до виявлення такого тексту є:

а) *автоматизовані* – базовані на актуалізації спеціального ПЗ, сервісів і тому подібного для аналізу стилістичних, лексичних та структурних маркерів аналізованого тексту;

б) *лінгвістичні* мають за основу ручний аналіз мовної самобутності – шаблонності, узагальненості, відсутності конкретики, рівного тону викладу, семантичних та інших артефактів.

2. Основна частина.

Коментар: зверніть увагу на оформлення кожного етапу вашої лабораторної роботи і той набір інструментів (див. *Інструменти*), який потрібно використати.

Завдання 1. Підготовка англійськомовних текстів для аналізу.

2.1.1. Напишіть вручну 5 коротких есе на будь-яку тему (150–200 слів) у таких стилях: художньому, публіцистичному, епістолярному, науковому, розмовному.

2.1.2. Згенеруйте 5 текстів того ж стилю і тематики схожого обсягу за сформульованим промптом для ШІ.

Завдання 2. Аналіз за допомогою онлайн-детекторів.

2.2.1. Перейдіть на сайт GPTZero (https://tinyurl.com/5hj87uvx), вставте тексти та проаналізуйте результат.

2.2.2. Повторіть це з усіма вищезазначеними у «шапці» лабораторної роботи інструментами.

2.2.3. За результатами перевірки заповніть Табл 1 й зробіть під нею висновок та аргументуйте свою думку.

Таблиця 1. Аналіз інструментів.

Текст	**GPTZero**	**Sapling AI Detector**	**AI Detector Pro (Free Plan)**	**Content at Scale AI Detector**
людський				
машинний				

Завдання 3. Лінгвістичний аналіз.

2.3.1. Проаналізуйте тексти вручну за такими критеріями:

а) лексика;

б) стилістика;

в) конкретність;

г) клішованість, повторюваність думок;

ґ) зв’язність, наступність.

2.3.2. Зробіть висновок щодо відмінностей між текстами та визначте ключові моменти кожного з них.

3. Заключна частина.

3.1. Аналіз результатів:

3.1.1. У чому полягає різниця між текстом, створеним людиною, і текстом, згенерованим машиною?

3.1.2. Які недоліки мають автоматизовані детектори ШІ-текстів?

3.1.3. Назвіть лінгвістичні ознаки, за якими можна виявити машинне походження тексту.

3.1.4. Чому важливо використовувати комбіновані підходи до виявлення? Аргументуйте.

3.2. Дискусія:

3.2.1. Наскільки продуктивним є автоматичний аналіз текстів на машинну генерацію? Ручний? Що краще й чому?

3.2.2. Які обмеження ручого та автоматичного аналізу? Якими є переваги? Який з інструментів продуктивніший? Чому? У чому саме?

ПІСЛЯМОВА

Цифрові технології є значною частиною життя сучасної людини, саме тому їх використання у гуманітарному секторі є логічною вимогою часу. Відповідно, знання таких технологій стало невід'ємним складником професійної підготовки гуманітарія. Саме тому в пропонованому виданні автори намагались поєднати лекційний матеріал, лабораторні роботи, а також широкий спектр аналітики: перелік актуальних нормативно-правових документів, словник основних термінів та тому подібне.

Така широта представлених у підручнику даних дозволила сформувати цілісну модель опанування цифрових технологій у координатах прикладної лінгвістики, перекладознавства та цифрової гуманітаристики загалом. Структура видання не лише ознайомлює з теоретичними засадами, передумовами та широким спектром сучасних технологій, вжитку у професійній та побутовій діяльності, а й закріплює їх шляхом практичного застосування у реальних дослідницьких задачах. Представлені у підручнику теми – від аналізу текстових даних (мовних корпусів та іншого) до технологій обробки природної мови й фактчекінгу та ШІ – відображають ключові напрями цифрових компетентностей сучасного спеціаліста-гуманітарія європейського рівня.

Кожен розділ цього підручника має за мету формування здатності не лише працювати з цифровими ресурсами, а й критично оцінювати результати їх роботи, убезпечуючи себе від помилок та порушення академічної доброчесності й медіаграмотності. Самобутня природа навчального видання увиразнена його потенційною відкритістю й гнучкістю, які, своєю чергою, дозволяють розвивати книгу, доповнюючи її новими, актуальними темами й прикладами, репрезентативними щодо тих чи інших тенденцій, диджиталізаційних змін.

Сподіваємося, що запропоновані матеріали стануть не лише підґрунтям для професійного становлення, а й сприятимуть формуванню критичного мислення, медіаграмотності та цифрових компетентностей, без яких неможливе успішне функціонування в сучасному суспільстві знань.

ОРІЄНТОВНІ ПИТАННЯ НА ІСПИТ

Блок завдань закритого типу:

1. Якими значеннями оперують цифрові технології? Чому?

2. Які сигнали використовують аналогові технології для обробки даних?

3. Назвіть перший електронний пристрій.

4. Назвіть базові технології Всесвітньої павутини (World Wide Web).

5. Охарактеризуйте хмарні обчислення.

6. Назвіть пристрій, що став прототипом сучасних комп'ютерів.

7. Охарактеризуйте корпусну лінгвістику. Що таке корпус текстів. Які види корпусів існують?

8. Назвіть цифрові інструменти для роботи з текстовими даними (корпусами текстів).

9. Охарактеризуйте роль і значення транзисторів у історії цифрових технологій.

10. Яка компанія випустила перший комп'ютер?

11. Назвіть міжмережевий протокол передачі даних.

12. Що таке оперативна пам'ять? Що таке напівпровідникова пам'ять?

13. Який із пристроїв найбільше наблизив обчислювальні системи до принципів роботи людського мозку?

14. Який рік вважається роком запуску ARPANET?

15. Який вчений запропонував концепцію Всесвітньої павутини?

16. Яка технологія забезпечила можливість об'єднання мільйонів людей у віртуальному просторі?

17. Препроцесинг, стоп-слова тощо для чого використовують? У яких дослідженнях?

18. МН та його типи.

19. Що є оригінальним у штучній нейронній мережі, як відповіднику мозку людини або тварини?

20. Методи навчання штучної нейронної мережі.

21. Який метод навчання штучної нейронної мережі передбачає постійне порівняння вихідних даних з бажаним результатом?

22. Яка технологія дозволяє імітувати когнітивні функції людини, такі як самонавчання?

23. В чому різниця між штучним інтелектом та штучною нейронною мережею?

24. Штучний інтелект та корпусна лінгвістика: точки дотику.

25. У якому році був побудований перший нейрокомп'ютер?

26. Яка технологія використовує квантові властивості речовини для обробки інформації?

27. Яка технологія спрямована на створення систем, здатних розуміти, інтерпретувати та генерувати людську мову?

28. Яке завдання виконує акустична модель в системах розпізнавання мови?

29. Принципи роботи антиплагіатних систем. Основні антиплагіатні системи.

30. Що таке плагіат? Як його виявляють? Які тексти найскладніше аналізувати?

31. Яка програма поєднує перевірку граматики та аналіз унікальності тексту?

32. Що таке гібридна антиплагіатна система? Назвіть обмеження антиплагіатних систем.

33. Які типи штучних нейронних мереж ви знаєте? В чому особливість кожної з них?

34. Який підхід є панівним у сучасній обробці природної мови?

35. Яке завдання виконують парсери в обробці природної мови?

36. Що таке URL?

37. Який тип аналізу використовують лінгвісти для виявлення машинного тексту?

38. Яка з технологій відіграла ключову роль у глобальному поширенні цифрових сервісів?

39. Який протокол лежить в основі функціонування інтернету як глобальної мережі обміну даними?

40. Фактчекінг передбачає перевірку достовірності фактів у тексті чи медіа? Які інструменти фактчекінгу існують?

41. Що таке неправдиві дані? Які види таких даних існують?

42. Яка платформа спеціалізується на спростуванні дезінформації в Україні?

43. Що з переліченого є перевагою ручного фактчекінгу?

44. Який метод використовують лінгвісти для виявлення машинного тексту за емоційним забарвленням?

45. Яка технологія використовує МН для розпізнавання шаблонів у великих обсягах даних?

46. Що є метою формалізації правил ухвалення рішень у нейронних мережах?

47. Який тип пам’яті використовують в моделі семантичної пам’яті для створення векторних представлень слів?

48. Який модуль обробляє питання слово за словом і генерує вектор у моделі відповіді на запитання?

49. Який інструмент рекомендують для видобування даних з вебсайтів для подальшої обробки?

50. Який етап обробки природної мови передбачає перетворення вхідних даних у структурований формат?

51. Який протокол використовують для безпечної передачі даних через Інтернет?

52. Яка технологія забезпечила можливість миттєвого та глобального спілкування?

53. Що таке промпт? Які правила його створення?

54. Який тип завдання для штучного інтелекту передбачає створення тексту за заданими параметрами?

55. Який тип файлів найкраще підходить для збереження простого тексту без форматування?

56. Який інструмент часто використовується для візуалізації даних у корпусному аналізі?

57. Що є основною метою корпусного аналізу?

58. Які технології забезпечили розвиток комп'ютерів, ґрунтуючись на використанні дискретних сигналів?

59. Яка з проблем є найактуальнішими для сучасних лінгвістичних досліджень?

60. Який факторів забезпечив стрімкий розвиток саме платформ соціальних медіа?

Блок завдань відкритого типу:

1. Аргументуйте, чим цифрові технології відрізняються від аналогових.

2. Назвіть ключові характеристики цифрових технологій та наведіть приклади.

3. Опишіть роль інтернету в процесі диджиталізації суспільства.

4. Які виклики стоять перед сучасним розвитком інтернету?

5. Опишіть вплив цифрових технологій на сучасне суспільство (наведіть 5 прикладів).

6. Назвіть основні переваги використання штучного інтелекту в обробці даних (наведіть 5 аргументів).

7. Опишіть способи застосування корпусів текстів у лінгвістичних дослідженнях (наведіть 5 прикладів).

8. Які можливості надають цифрові технології для автоматизації рутинних процесів у лінгвістиці (наведіть 5 прикладів)?

9. Аргументуйте важливість розвитку навичок фактчекінгу в епоху цифрових технологій (наведіть 5 аргументів).

РЕКОМЕНДОВАНА ЛІТЕРАТУРА

1. Біскуб І. Quo vadis? Прикладна лінгвістика у другій декаді XXI століття. *Лінгвістичні студії*. 2024. Випуск 47. С. 137–148. DOI : https://doi.org/10.31558/1815-3070.2024.47.11 *Лінгвістичні студії* : вебсайт. URL: https://tinyurl.com/2v35puy4 (дата звернення: 01.08.25).

2. Бук С. Сучасні комп'ютерні інструменти опрацювання української мови в університетському курсі «Вступ до прикладної лінгвістики». Перша міжнародна наукова конференція «Лінгвосфера учора, сьогодні, завтра», присвячена вшануванню пам'яті професора Костянтина Тищенка (м. Київ, Україна, 10-12 жовтня 2024 року). Київ, 2024. С. 291–344. *Amazon CloudFront* : вебсайт. URL: https://tinyurl.com/yaeju6vr (дата звернення: 01.08.25).

3. Ваховська О. В. Основи комп'ютерної лінгвістики : навчально-методичний посібник. Київ : ВЦ КНЛУ, 2023. 112 с. *ResearchGate* : website. URL: https://tinyurl.com/4whyestp (дата звернення: 01.08.25).

4. Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту) : підручник. Київ : Видавничо-поліграфічний центр «Київський університет», 2008. 351 с.

5. Дарчук Н. П. Можливості семантичної розмітки корпусу української мови (КУМ). *Науковий часопис Національного педагогічного університету імені М. П. Драгоманова. Серія 9: Сучасні тенденції розвитку мов* : зб. наук. пр. Київ : Вид-во НПУ імені М. П. Драгоманова, 2017. Вип. 15. С. 18–28. *Інституційний репозитарій Українського державного університету імені Михайла Драгоманова* : вебсайт. URL: https://tinyurl.com/yndcedn8 (дата звернення: 01.08.25).

6. Демська О. М. Текстовий корпус : ідея іншої форми. Київ : ВПЦ НаУКМА, 2011. 342 с. *Academia* : вебсайт. URL: https://tinyurl.com/3ca4nexe (дата звернення: 01.08.25).

7. Дячук Н. Мовні технології штучного інтелекту в завданнях автоматичного анотування текстів. *Актуальні питання гуманітарних наук*. 2024. Вип. 82, том 1. С. 320–327. DOI : https://doi.org/10.24919/2308-4863/82-1-47

Електронна бібліотека Житомирського державного університету : вебсайт. URL: https://tinyurl.com/57dfm9nj (дата звернення: 01.08.25).

8. Жуковська В. В. Вступ до корпусної лінгвістики : навчальний посібник. Житомир : Вид-во ЖДУ ім. І. Франка, 2013. 142 с. *Електронна бібліотека Житомирського державного університету* : вебсайт. URL: https://tinyurl.com/3mp8pfys (дата звернення: 01.08.25).

9. Карпіловська Є. А. Вступ до прикладної лінгвістики : комп'ютерна лінгвістика : підручник. Донецьк : ТОВ «Юго-Восток, Лтд», 2006. 188 с. *Інститут мовознавства ім. О. О. Потебні Національної академії наук* : вебсайт. URL: https://tinyurl.com/3dj8hbyj (дата звернення: 01.08.25).

10. Новітні напрями сучасного мовознавства : навчально-методичний посібник для студентів 1 курсу другого (магістерського) рівня вищої освіти спеціальності 014 Середня освіта предметної спеціальності 014.01 Середня освіта (Українська мова і література) освітніх програм «Українська мова і література», «Українська мова і література. Англійська мова і література» / уклад. В. Ф. Баньої, В. А. Папіш. Ужгород, 2024. 108 с. *Електронний репозитарій ДВНЗ «УжНУ»* : вебсайт. URL: https://tinyurl.com/59apne7e (дата звернення: 01.08.25).

11. Олександрук І. В. Лінгвістичний корпус як інструмент семантичних досліджень. *Наукові записки. Серія : Філологічні науки*. 2024. № 210. С. 209–214. DOI : https://doi.org/10.32782/2522-4077-2024-210-31 *Центральноукраїнський державний університет імені Володимира Винниченка* : вебсайт. URL: https://tinyurl.com/2ekrrp5r (дата звернення: 01.08.25).

12. Панченко О. І., Попко Л. П., Ходоренко Г. В. Сучасні інформаційні технології у діяльності перекладача : навчальний посібник. Дніпропетровськ : Пороги, 2010. 168 с.

13. Паралельний корпус текстів ПарКУМ. Драгоманова / Н. П. Дарчук та ін. *Науковий часопис Національного педагогічного університету імені М. П. Драгоманова. Серія 9: Сучасні тенденції розвитку мов*. Київ : Вид-во НПУ імені М. П. Драгоманова, 2017. Вип. 15. С. 28–35. *Репозитарій Українського державного університету імені Михайла*

Драгоманова : вебсайт. URL: https://tinyurl.com/ye23f8pw (дата звернення: 01.08.25).

14. Партико З. В. Прикладна і комп'ютерна лінгвістика : Вступ до спеціальності : Навчальний посібник. Львів : Афіша, 2008. 224 с. *SCRIBD* : вебсайт. URL: https://tinyurl.com/5n87wczj (дата звернення: 01.08.25).

15. Перебийніс В. І., Сорокін В. М. Традиційна та комп'ютерна лексикографія : Навчальний посібник. Київ : ВЦ КНЛУ, 2009. 218 с.

16. Сизонов Д. Медійна е-лексикографія у фокусі сучасних глобалізаційних процесів. *Енциклопедичний вісник України*. 2024. Volume 16. P. 1–26. DOI : https://doi.org/10.37068/evu.16.6 *Енциклопедичний вісник України* : вебсайт. URL: https://tinyurl.com/m7jsuvsb (дата звернення: 01.08.25).

17. Сітко А. В., Ващенко Д. В. Локалізація вебконтенту як форма міжкультурної комунікації. *Наукові записки. Серія: Філологічні науки*. 2025. № 213. С. 365–370. DOI : https://doi.org/10.32782/2522-4077-2025-213-53 *Наукові записки. Серія: Філологічні науки* : вебсайт. URL: https://tinyurl.com/2n4bzzmr (дата звернення: 01.08.25).

18. Тарасенко Р. О., Амеліна С. М. Інформаційні технології у синхронному перекладі : навчальний посібник / Нац. ун-т біоресурсів і природокористування України. 2-ге вид., перероб. і допов. Київ : Ямчинський О. В. [вид.], 2023. 289 с.

19. Фокін С. Б. Комп'ютерні інноваційні технології в перекладацькій та перекладознавчій діяльності : навчальний посібник. Київ : КНУ імені Тараса Шевченка, 2019. 260 с. *Портал освітніх, словникових і термінологічних утиліт для перекладу (ПОСТУП)* : вебпортал. URL: https://tinyurl.com/yene2x45 (дата звернення: 01.08.25).

20. Цимбал С., Лужецька Т., Скрипник Н. Концептуальні та методологічні основи перекладознавства сучасності : проблематика та перспективний розвиток. Актуальні проблеми філології та перекладознавства. 2025. Випуск 33. С. 44–48. DOI : https://doi.org/10.31891/2415-7929-2025-33-7 *Актуальні проблеми філології та перекладознавства* : вебсайт. URL: https://tinyurl.com/4bu7usax (дата звернення: 01.08.25).

21. Широков В. А. Комп'ютерна лексикографія : [монографія]. Київ : Наукова думка, 2011. 351 с.

22. Яскулко Д. Д., Дубовик Н. М. Веб-застосунок для аналізу тональності тексту. *Закарпатські філологічні студії*. 2024. Вип. 38. С. 240–244. DOI : https://doi.org/10.32782/tps2663-4880/2024.38.44 Електронний репозитарій ДВНЗ «УжНУ» : вебсайт. URL: https://tinyurl.com/mt2ves54 (дата звернення: 01.08.25).

23. Dovhan O. Evaluation of the efficiency of artificial neural network models in the process of their integration into systems for monitoring political Internet discourse. *Теорія і практика викладання української мови як іноземної.* 2025. Вип. 19. С. 173–185. DOI : http://dx.doi.org/10.30970/ufl.2025.19.4815 *Теорія і практика викладання української мови як іноземної* : вебсайт. URL: https://tinyurl.com/4jedakjx (дата звернення: 01.08.25).

24. Dovhan O. V. Actualization of linguistic markers in political Internet discourse, and its peculiarities in the context of artificial neural network operation, and hate speech detection. *Вісник Луганського Національного університету імені Тараса Шевченка. Філологічні науки*. 2024. № 2(361). С. 41–50. DOI : https://doi.org/10.12958/2227-2844-2024-2(361)-41-50 *Вісник Луганського Національного університету імені Тараса Шевченка. Філологічні науки* : вебсайт. URL: https://tinyurl.com/bddk3t8z (дата звернення: 01.08.25).

25. Dovhan O. V. Architectural Originality of Artificial Neural Networks for Linguistic Markers Recognition. *Вісник науки та освіти.* 2024. Випуск № 7(25). С. 17–36. DOI : https://doi.org/10.52058/2786-6165-2024-7(25)-17-36 *Вісник науки та освіти* : вебсайт. URL: https://tinyurl.com/38vk6rn7 (дата звернення: 01.08.25).

26. Dovhan O. V. Features of data collection and processing for neural network modeling of language categories. *Trends in the development of philological education in the era of digitalization : European and national contexts : Scientific monograph*. Riga, Latvia : "Baltija Publishing", 2023. P. 2–17. DOI : https://doi.org/10.30525/978-9934-26-303-3-1 *Izdevniecība "Baltija Publishing"* : website. URL: https://tinyurl.com/3kzmhuzj (date of application: 01.08.25).

27. English-Ukrainian Parallel Corpus : Prerequisites for Building and Practical Use in Translation Studies / S. Matvieieva, N. Lemish, A. Zernetska et al. *Studies*

about Languages. 2022. № 40. P. 61–74. DOI : https://doi.org/10.5755/j01.sal.40.1.30735 *Studies about Languages* : website. URL: https://tinyurl.com/37e7228z (date of application: 01.08.25).

28. Hunston S. Corpora in Applied Linguistics. Cambridge : Cambridge University Press, 2022. 300 p. *SCRIBD* : website. URL: https://tinyurl.com/4cby84ja (date of application: 01.08.25).

29. Machine Translation and Society. The Case of Lithuania. R. Kasperė et al. Kaunas : KTU Leidykla "Technologija", 2024. 168 p. DOI : http://dx.doi.org/10.5755/e01.9786090218808 *Ebooks.edu* : website. URL: https://tinyurl.com/bdcpm6fy (date of application: 01.08.25).

30. Konstantinavičienė E., Kasperė R., Matvieieva S. Post-Editing Decisions in Machine Translation of Legal Contracts. *Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development*. 2023. № 26. P. 42–54. DOI : https://doi.org/10.31392/NPU-nc.series9.2023.26.05 *Journal of National Pedagogical Dragomanov University. Series 9. Current Trends in Language Development* : website. URL: https://tinyurl.com/27re5aay (date of application: 01.08.25).

31. Matvieieva S. Selection criteria and initial processing of empirical material for a parallel corpus of legal texts. *Forum Filologiczne Ateneum*. 2019. № 1(7). P. 167–181. DOI : http://dx.doi.org/10.36575/2353-2912/1(7)2019.167 *Forum Filologiczne Ateneum* : website. URL: https://tinyurl.com/mtae9p3d (date of application: 01.08.25).

32. Roszko R. On New Manually Aligned and Tagged Bilingual Parallel Corpora and Their Applications. *Acta Baltico-Slavica*. 2021. № 45. Article 2576. DOI : https://doi.org/10.11649/abs.2576 *Acta Baltico-Slavica* : website. URL: https://tinyurl.com/3nm3838j (date of application: 01.08.25).

СЛОВНИК ОСНОВНИХ ТЕРМІНІВ І ПОНЯТЬ[1]

Адаптер – апаратний або програмний пристрій, що забезпечує сумісність або з'єднання між різними системами чи компонентами.

Актор – елемент системи чи середовища, що виконує дію у процесі взаємодії з іншими компонентами.

Акустична нейромережева модель – модель, у якій представлено звукові характеристики мовлення для його автоматичного розпізнавання.

Алгоритм – послідовність формалізованих інструкцій або правил, що визначають порядок виконання операцій для розв'язання задачі.

Алофон – варіант вимови тієї самої фонеми, що не змінює значення слова.

Аналіз даних – систематичний процес обробки цифрової інформації з метою виявлення закономірностей, структур і залежностей.

Аналіз тексту – дослідження мовних одиниць та їхніх взаємозв'язків у текстових (цифрових) корпусах і базах даних.

Апаратні нейроморфні процесори – спеціалізовані апаратні системи, що імітують архітектуру та принципи роботи біологічних нейронних мереж, забезпечуючи обчислення з високою енергоефективністю.

Апроксимація – процес наближеного відтворення функції або залежності за допомогою математичної моделі чи штучної нейронної мережі, що дозволяє відображати складні нелінійні зв'язки у даних.

Архітектура інформаційних систем – структура програмних і технічних засобів, що забезпечує обробку, зберігання та передачу інформації.

Аферентний зв'язок – зв'язок в штучних нейронних мережах, що передає сигнали від рецепторів до центральних елементів обробки.

Багатошарова нейронна мережа – штучна нейронна мережа, яка складається з кількох послідовних шарів нейронів, що дозволяє моделювати складні залежності.

[1] Назви програмних інструментів (AntConc, Sketch Engine, Voyant Tools тощо) не включені до цього словника, оскільки вони подані й охарактеризовані в окремому розділі *«Цифрові інструменти та ресурси»*: тут вміщено лише поняття й категорії, що становлять теоретичний апарат цього курсу.

Багатошаровий перцептрон – різновид штучної нейронної мережі прямого поширення з кількома прихованими шарами, здатний апроксимувати довільні функції.

База даних – організоване сховище цифрової інформації, що забезпечує збереження, пошук і керування даними.

Біологічний комп'ютер – обчислювальна система, побудована на основі живих клітин, біомолекул або біологічних процесів, де носіями інформації виступають природні біоструктури.

Блокчейн – децентралізована технологія зберігання й передачі даних у вигляді послідовного ланцюга блоків, що гарантує їхню цілісність і захищеність від несанкціонованих змін.

Валідація – процес перевірки адекватності моделі або системи на незалежних даних.

Валідація штучної нейронної мережі – оцінювання продуктивності нейромережевої моделі на даних, що не використовувалися під час навчання, для запобігання перенавчанню.

Вебінструмент – програмний онлайн-засіб, який функціонує через інтернет-браузер (до прикладу, Google Chrome) і використовується для обробки даних (у тому числі, текстових).

Вебресурс – цифровий ресурс у мережі інтернет, що надає доступ до текстів, інструментів чи навчальних матеріалів.

Векторне представлення слів – метод відображення слів у багатовимірному просторі числових векторів, що зберігають семантичні відношення між ними; результат векторизації.

Великі дані – великі за обсягом та різноманітністю масиви цифрової інформації, що потребують спеціальних методів для обробки.

Вербалізація штучної нейронної мережі – представлення результатів роботи нейромережевої моделі у формі, придатній для інтерпретації людиною.

Видобування даних – методи автоматичного виявлення закономірностей у великих масивах даних.

Викопіювання (екстракція) даних – процес отримання інформації з цифрових текстів або вебсторінок для подальшого аналізу.

Віртуальна реальність – цифрове середовище, створене за допомогою комп'ютерних технологій, яке імітує реальний або вигаданий простір.

Вокодер – пристрій або алгоритм, що аналізує та синтезує мовлення, зберігаючи його основні характеристики.

Генеративна змагальна мережа – архітектура штучної нейронної мережі, що складається з двох частин – генератора та дискримінатора, які навчаються у змагальному режимі, забезпечуючи створення нових даних, подібних до навчальної множини.

Генетичний алгоритм – оптимізаційний метод, що імітує еволюційні механізми відбору, схрещування та мутацій.

Гіперпараметр – параметр моделі, що задається перед початком навчання й визначає архітектуру або процес оптимізації (до прикладу, швидкість навчання, кількість шарів).

Гіперплощина – геометричний об'єкт, що розділяє простір ознак у задачах класифікації.

Глибока нейронна мережа – штучна нейронна мережа з багатьма прихованими шарами, яка здатна моделювати складні нелінійні залежності та ефективно працювати з великими обсягами даних.

Глибоке навчання – підхід у машинному навчанні, що використовує штучні нейронні мережі (до прикладу, рекурентні).

Глибоке навчання – різновид машинного навчання, що базується на багатошарових нейронних мережах і забезпечує автоматичне виявлення ознак у даних.

Глибокий граф – структура даних, що описує складні багаторівневі зв'язки у системах штучного інтелекту.

Глобальна мережа (Інтернет) – світова система взаємоз'єднаних комп'ютерних мереж для обміну інформацією.

Голосова біометрія – технологія ідентифікації або автентифікації особи за індивідуальними характеристиками її голосу.

Голосове управління – спосіб керування пристроями чи програмами за допомогою голосових команд.

Голосовий інтерфейс – система взаємодії людини з комп'ютером або пристроєм через голосові команди.

Граф обчислень (**Граф обчислювальних операцій**) – модель представлення зв'язків між елементами даних у вигляді вершин і ребер.

Графема – мінімальна одиниця письма, що відповідає фонемі чи її комбінації.

Групування даних – класифікація цифрової інформації за визначеними параметрами для полегшення аналізу.

Дані (**Інформація**) – структуровані чи неструктуровані відомості, що підлягають цифровій обробці і мають смислове значення в контексті людської діяльності.

Дезінформація – навмисно поширена неправдива інформація з метою маніпуляції свідомістю.

Державний електронний ресурс – офіційна інформаційна система, створена органами влади для надання цифрових послуг (до прикладу, вебпортал «Дія»).

Дискретність – властивість інформаційних процесів або сигналів поділятися на окремі неподільні одиниці (кванти), що є основою цифрового обчислення.

Дискурс – форма існування мови, яка включає текст, комунікативну ситуацію та культурний контекст.

Дифузійна модель – клас генеративних моделей, що відтворюють дані шляхом поступового перетворення випадкового шуму у структуровану інформацію через процес дифузії та його обернення.

Додаткове навчання (Донавчання) – процес уточнення параметрів штучної нейронної мережі після початкового навчання на основі нових даних для їх врахування або адаптації до іншої вибірки.

Документ у цифровому форматі – текстовий, графічний чи мультимедійний матеріал, зафіксований у електронній формі.

Доповнена реальність – технологія, у якій інтегровано візуальні, аудіо- та сенсорні цифрові елементи у сприйняття фізичного світу в реальному часі для посилення або доповнення реального середовища.

Електронна комунікація – обмін інформацією за допомогою цифрових технологій та мереж (до прикладу, за посередництва месенджеру WhatsApp).

Електронна обчислювальна машина – універсальний програмований пристрій для автоматизованої обробки інформації, що здійснює операції за допомогою електронних схем.

Електронна обчислювальна машина на інтегральних схемах – пристрій, у якому обчислення реалізовано за допомогою інтегральних мікросхем.

Електронний документ – документ, створений і збережений у цифровій формі, що має юридичну силу відповідно до законодавства (до прикладу, ID-картка).

Електронний ламповий обчислювальний пристрій – перша електронна обчислювальна машина, побудована на базі вакуумних ламп, які забезпечували обробку сигналів до появи транзисторних технологій.

Електронний ресурс – цифровий матеріал (текст, зображення, аудіо, відео), доступний в Інтернеті чи локальних мережах.

Ентропія – міра невизначеності або непередбачуваності інформації в системі.

Епоха навчання – один повний прохід алгоритму навчання через усю навчальну вибірку.

Ефектор штучної нейронної мережі – вихідний елемент, що генерує результат на основі обчислених сигналів.

Еферентний зв'язок – зв'язок, що передає вихідний сигнал від нейронів до ефекторів.

Жанр (цифровий дискурс) – тип тексту або комунікативної практики в цифровому середовищі (до прикладу, блог, допис, коментар).

Загальний корпус – корпус текстів, які репрезентує різні сфери вживання мови без спеціалізованої спрямованості.

Закритий рекурентний блок – різновид рекурентної нейронної мережі зі спеціальними «комірками пам’яті» та механізмами керування потоками інформації, що дозволяє зберігати довгострокові залежності у даних.

Заплутаність – квантове явище кореляції між частинками, при якому стан однієї визначає стан іншої незалежно від відстані між ними; використовується у квантових обчисленнях.

Захист інформації – система правових, організаційних і технічних заходів, спрямованих на збереження конфіденційності та цілісності даних.

Збирання корпуса текстів – процес систематичного накопичення текстів у цифровій формі для лінгвістичних досліджень.

Згорткова нейронна мережа – штучна нейронна мережа, яка застосовує операцію згортки для автоматичного виділення ознак з даних (першою чергою, зображень), забезпечуючи їхню багаторівневу обробку.

Змістовий аналіз – метод дослідження текстів, спрямований на виявлення смислових структур та категорій.

Ініціалізація штучної нейронної мережі – встановлення початкових значень параметрів і вагових коефіцієнтів перед машинним навчанням.

Інтелект – здатність системи до адаптивного розв’язання задач, узагальнення досвіду та навчання.

Інтернет – глобальна система комп’ютерних мереж, що забезпечує передачу даних, комунікацію та доступ до інформаційних ресурсів на основі єдиних протоколів.

Інтонація – зміна висоти, сили й тривалості звуків у мовленні, що виражає смислові й емоційні відтінки висловлювання.

Інформаційна інфраструктура – сукупність апаратних, програмних і комунікаційних засобів, що забезпечують функціонування цифрових систем.

Інформаційна система – комплекс апаратних і програмних засобів, призначений для зберігання, обробки й поширення інформації.

Інформаційне суспільство – соціальний стан, у якому інформація та знання є ключовим ресурсом розвитку.

Ітерація – один крок у процесі оптимізації чи навчання моделі.

Карти Кохонена – різновид штучних нейронних мереж самоорганізації, що відображають багатовимірні дані на площину.

Квантова технологія – сукупність прикладних технологій, побудованих на використанні явищ квантової механіки (зокрема, у сфері обчислень, комунікацій та сенсорики).

Квантовий комп'ютер – обчислювальний пристрій, який виконує операції на основі принципів квантової механіки (суперпозиції, заплутаності), що дозволяє розв'язувати задачі недосяжні для класичних комп'ютерів.

Керований рекурентний блок – модифікація рекурентної нейронної мережі з менш складною архітектурою порівняно з довготривалою короткочасною пам'яттю, яка забезпечує ефективне навчання послідовних даних за допомогою механізмів керування потоками інформації.

Кібербезпека – стан захищеності інформаційних систем від кіберзагроз та несанкціонованих втручань.

Кіберпростір (**Медіапростір**) – віртуальне середовище функціонування інформаційних систем і комунікаційних мереж; сукупність каналів, платформ і засобів поширення інформації у цифровому середовищі.

Класифікація текстів – процес упорядкування текстових даних за визначеними ознаками для дослідницьких цілей.

Колокація – стійке поєднання слів, що вживаються разом частіше, ніж випадково.

Комп'ютерна лінгвістика – галузь лінгвістики, що вивчає мову з використанням методів і засобів інформатики, науки про дані та інших.

Комп'ютерна програма – набір інструкцій, що виконуються комп'ютером для реалізації певних функцій.

Конкорданс – перелік уживань слова або виразу в корпусі текстів із контекстом.

Корегування параметрів – процес уточнення значень ваг та інших характеристик моделі під час її навчання.

Критичне мислення – здатність оцінювати достовірність інформації та робити обґрунтовані висновки в умовах інформаційного потоку.

Латеальний (бічний) зв'язок – горизонтальний зв'язок між нейронами одного шару, що забезпечують узгоджену роботу мережі.

Лексема – основна словникова одиниця мови, яка репрезентує певне значення та формується з усіх словоформ.

Лексичний аналіз – вивчення структури словникового складу тексту з використанням цифрових інструментів.

Лема – початкова форма слова, до якої зводяться всі його словоформи.

Лематизація – приведення словоформи до початкової форми (леми) для забезпечення наукоємного й достовірного аналізу текстів.

Лінгвістична експертиза – аналіз текстів або мовних даних для встановлення фактів чи обґрунтування висновків у наукових і практичних цілях.

Лінгвістичний аналіз – комплекс методів вивчення мовних одиниць і структур на матеріалі цифрових корпусів.

Математична модель нейрона – формалізоване представлення роботи біологічного нейрона у вигляді рівнянь та алгоритмів, що відтворюють процеси активації та передачі сигналів.

Машинне навчання – методи штучного інтелекту, що дозволяють комп'ютерним системам удосконалювати свої алгоритми на базі досвіду.

Машинний (автоматичний) переклад – передача змісту тексту іншою мовою з використанням алгоритмів обробки природної мови (до прикладу, DeepL).

Медіаграмотність – здатність критично сприймати, аналізувати та використовувати інформацію з різних медіа.

Мережа динамічної пам'яті – нейронна архітектура, яка поєднує обчислювальні моделі з механізмами збереження та оновлення інформації у процесі навчання.

Мережа зворотного поширення – штучна нейронна мережа, яка навчається за алгоритмом поширення похибки назад.

Мережа на основі радіальних базисних функцій – нейромережева архітектура, що використовує радіальні функції як активаційні.

Мережа прямого поширення – штучна нейронна мережа, у якої сигнал рухається від входу до виходу без зворотних зв'язків.

Мережа Хопфілда – рекурентна нейронна мережа, здатна зберігати та відтворювати асоціативні образи.

Метаінформація (Метадані) – дані, що описують структуру, зміст і властивості цифрових ресурсів.

Механічний обчислювальний пристрій – прилад для виконання арифметичних операцій, де обчислення здійснюються за допомогою механічних елементів (важелів, зубчастих коліс тощо).

Мізінформація – ненавмисно поширена неправдива або неточна інформація.

Мікропроцесор – інтегральна схема, що містить центральний процесор комп'ютера на одному кристалі, забезпечуючи компактність і високу швидкодію.

Мова програмування – формальна система запису інструкцій, що задають алгоритми для комп'ютера (до прикладу, Python).

Мовна модель – алгоритм, здатний генерувати, аналізувати та прогнозувати текст на основі статистичних або нейромережевих методів.

Мовна нейромережева модель – модель на основі нейронних мереж, що прогнозує послідовності слів та структуру висловлювань у мові.

Мовний ресурс – корпус, словник, база даних тощо, використовувані у комп'ютерній, обчислювальній та іншій лінгвістиці.

Модель – абстрактне представлення системи, процесу чи явища у вигляді математичної, логічної чи комп'ютерної конструкції, що дозволяє аналізувати, прогнозувати та відтворювати їхні властивості.

Модель Adaptive Resonance Theory – модель штучних нейронних мереж, що забезпечує баланс між стабільністю та адаптивністю в машинному навчанні.

Моделювання мовних процесів – використання алгоритмів і програмних засобів для відтворення функціонування мови.

Молекулярний комп'ютер – експериментальний обчислювальний пристрій, у якому для зберігання та обробки даних використовуються молекули органічних сполук, переважно ДНК.

Морфема – найменша значуща одиниця мови, що виражає граматичне чи лексичне значення.

Мультимедійний ресурс – цифровий матеріал, що поєднує текст, звук, відео та графіку.

Навчальна вибірка (Навчальна множина, Набір даних) – частина даних, призначена для формування й налаштування параметрів штучної нейронної мережі.

Навчальна пара (вектор) – окремий приклад із навчальної множини, який складається з вхідного вектора та відповідного правильного виходу.

Навчання без учителя (Неконтрольоване навчання) – метод навчання, за якого модель самостійно виявляє структуру у даних без заздалегідь відомих правильних відповідей.

Навчання з підкріпленням – метод навчання, коли агент отримує винагороди або штрафи за свої дії й таким чином оптимізує стратегію поведінки.

Навчання з учителем (Контрольоване навчання) – метод навчання, при якому модель будується на основі прикладів з відомими правильними відповідями.

Наївний байєсівський класифікатор – імовірнісна модель класифікації, що ґрунтується на застосуванні теореми Томаса Байєса з припущенням незалежності ознак.

Нейрокомп'ютер – обчислювальна система, архітектура якої відтворює принципи організації та функціонування біологічних нейронних мереж, орієнтована на задачі розпізнавання та адаптивного навчання.

Нейрон – основна структурно-функціональна одиниця нервової системи, здатна приймати, обробляти та передавати сигнали; є прототипом для штучних нейронних моделей.

Нечітка дискретна система – система, що використовує апарат нечіткої логіки для роботи з даними, представленими у вигляді дискретних величин.

Нормалізація – процес упорядкування текстових даних (усунення варіантів написання, приведення до єдиної форми).

Нормалізація даних – перетворення даних до єдиного масштабу для забезпечення коректної роботи алгоритмів машинного навчання.

Обробка природної мови – галузь штучного інтелекту, що займається автоматичною обробкою, аналізом та генерацією людської мови у текстовій чи усній формі.

Обчислювальна лінгвістика – галузь лінгвістики, у межах якої застосовуються методи комп'ютерних наук і математичного моделювання для автоматичного аналізу, обробки та генерації природної мови.

Обчислювальний інтелект – галузь досліджень, що охоплює методи штучного інтелекту, які імітують природні механізми адаптації й навчання.

Одношаровий перцептрон – найпростіший тип штучної нейронної мережі, що складається з одного шару нейронів.

Омограф – слово, що має однакове написання, проте різну вимову й значення.

Онлайн-ресурс – цифровий матеріал чи інструмент, доступний через інтернет.

Оптимізація гіперпараметрів – процес пошуку найкращих значень останніх для підвищення ефективності нейромережевої моделі.

Оптоелектронний комп'ютер – обчислювальна система, у якій для обробки інформації використовується світло (фотони), що забезпечує потенційно вищу швидкодію порівняно з електронними пристроями.

Очищення даних – процес виявлення та усунення помилок, пропусків, дублювань і некоректних значень у наборі даних для забезпечення його точності та якості аналізу.

Паралельний корпус – корпус текстів, що містить матеріали двома чи більше мовами з вирівнюванням відповідних фрагментів.

Параліч (влучення мережі в локальний мінімум поверхні похибок) – ситуація, коли процес машинного навчання зупиняється в локальному мінімумі й не досягає глобального розв'язку.

Парсинг (граматичний розбір) – автоматизоване визначення морфологічних і синтаксичних характеристик (параметризації) тексту.

Патерн – типова структурна або функціональна модель, що повторюється у даних чи процесах.

Перенавчання (оверфіттінг) штучної нейронної мережі – ситуація, коли нейромережева модель добре працює на навчальних даних, проте втрачає здатність до узагальнення.

Персональний комп'ютер – універсальний обчислювальний пристрій, призначений для індивідуального використання, що забезпечує обробку даних, зберігання інформації та комунікацію.

Перцептрон – штучна нейронна мережа найпростішої архітектури, розроблена для моделювання процесів розпізнавання образів на базі принципів роботи біологічних нейронів.

Плагіат – привласнення чужих ідей, текстів або результатів досліджень без належного покликання на джерело.

Повнозв'язана нейронна мережа – мережа, у якій кожен нейрон одного шару з'єднаний із кожним нейроном наступного.

Поріг резонансу – критичне значення, при якому мережа резонансного типу фіксує новий зразок або відносить його до вже відомої категорії.

Пошукова система – програмний інструмент для знаходження інформації в інтернеті (до прикладу, Google).

Препроцесинг даних – етап підготовки даних для аналізу або моделювання, який включає очищення, нормалізацію, трансформацію та структурування інформації для підвищення якості та ефективності обробки.

Прикладна лінгвістика – галузь лінгвістики, у якій досліджують практичне застосування знань про мову для розв'язання соціальних, освітніх, комунікаційних та технологічних проблем.

Природна мова – система комунікації людей, яка відображає культурний і соціальний досвід (національну й мовну картини світу тощо).

Програмне забезпечення – сукупність програм і додатків, що забезпечують роботу комп'ютера та реалізацію завдань.

Проєктивний зв'язок – зв'язок між шарами мережі, що передає результати від одного шару до наступного.

Пропаганда – поширення ідей або поглядів із метою впливу на громадську думку.

Радіальні базисні функції – функції, значення яких залежить лише від відстані між точками; застосовуються як активаційні в штучних нейронних мережах.

Регуляризація – метод запобігання перенавчанню шляхом введення додаткових обмежень у процес машинного навчання.

Резонанс – стан відповідності між сигналом і внутрішніми параметрами системи, що забезпечує стійке розпізнавання.

Рекурентна нейронна мережа – штучна нейронна мережа, у якій реалізовано зворотні зв'язки, що дозволяє обробляти послідовні дані (текст, мовлення, часові ряди) із врахуванням попередніх станів.

Рецептори штучної нейронної мережі – вхідні елементи мережі, які сприймають інформацію ззовні.

Розпізнавання мовлення – процес перетворення усного мовлення людини у текст за допомогою алгоритмів та моделей обробки сигналів і мови.

Ручні обчислювальні пристрої – інструменти для виконання арифметичних операцій у ручному режимі (лінійки, абакуси, рахівниці), що передували автоматизованим системам обчислення.

Семантика – розділ мовознавства, що досліджує значення слів і висловлювань.

Сенон – мінімальна одиниця акустичної моделі у системах розпізнавання мовлення, що відповідає контекстно-залежним станам фонем.

Сервер – комп'ютер або програмне забезпечення, що надає послуги чи ресурси іншим комп'ютерам у мережі.

Синаптична вага – коефіцієнт, що визначає силу зв'язку між нейронами.

Синтагма – мінімальне смислове й інтонаційно виділене словосполучення у мовленні.

Синтаксис – система правил, що визначає поєднання слів у реченні.

Синтез мовлення – автоматизоване створення штучного усного мовлення на основі тексту чи інших мовних даних.

Синхронний та діахронний корпус – корпус текстів, що відображають мову на певному зрізі часу (синхронний) або у різні історичні періоди (діахронний).

Система розподіленого реєстру – технологія зберігання даних у вигляді розподілених баз, копії яких одночасно розміщуються на багатьох вузлах мережі, забезпечуючи прозорість і захист інформації.

Система розподіленого реєстру – цифрова система збереження та обміну даними, у якій інформація розподіляється між багатьма вузлами мережі та оновлюється за узгодженими правилами (блокчейн є її найвідомішим прикладом).

Слабозв'язана нейронна мережа – мережа з частковими зв'язками між нейронами, що знижує надмірність обчислень.

Соціальна мережа – цифрова платформи для обміну повідомленнями, медіаконтентом та підтримки комунікацій (до прикладу, Facebook).

Спеціалізований корпус – корпус текстів, зосереджених на певній сфері чи тематиці.

Статистичний аналіз тексту – методи кількісної обробки текстових даних для виявлення закономірностей.

Стемінг – процедура зведення слова до його основи шляхом відсікання закінчень і афіксів без урахування повних правил словозміни.

Стислий переказ тексту (Text Summarization) – автоматизований процес створення короткої версії тексту з відбором ключових смислових елементів.

Стоп-слова – слова у тексті (до прикладу, службові частини мови: сполучники, прийменники, займенники та інші), які несуть мінімум смислового навантаження і зазвичай видаляють під час обробки текстових даних.

Структурований текст – текст, поділений на розділи, абзаци й інші сегменти для зручності аналізу.

Суперпозиція – квантове явище існування фізичної системи одночасно в кількох можливих станах до моменту вимірювання, яке лежить в основі квантових обчислень.

Суспільна комунікація – обмін інформацією у межах соціуму (зокрема, в цифровому форматі).

Текстовий корпус – систематизоване зібрання текстів у цифровій формі, що є основою для лінгвістичного аналізу.

Термінологія – сукупність спеціальних слів і виразів, що описують певну галузь знань.

Тестові дані – частина даних, що використовується для підсумкової перевірки якості навченої моделі.

Технології цифрової гуманітаристики – цифрові методи й інструменти, що застосовують для вивчення гуманітарних наук.

Токен – мінімальна одиниця тексту, виділена під час токенізації (до прикладу, слово чи пунктуаційний знак).

Токенізація – процес поділу тексту на токени.

Транзисторна електронна обчислювальна машина – пристрій, у якому вакуумні лампи замінено транзисторами, що зменшило енергоспоживання і підвищило надійність обчислень.

Трансферне навчання – повторне використання вже навченої моделі для розв’язання нових задач.

Узгодження текстів – процес зіставлення текстів різними мовами для аналізу перекладу.

Фактчекінг – перевірка достовірності фактів, що містяться у повідомленнях.

Фінальне тестування – завершальний етап оцінки ефективності моделі на незалежних даних.

Фонема – мінімальна одиниця звукової системи мови, що розрізняє значення слів.

Формат даних – спосіб представлення інформації у цифровій формі.

Форматування тексту – процес упорядкування зовнішнього вигляду тексту (шрифт, відступи, структура).

Фрейм – структурована одиниця даних або знань, яка описує певну ситуацію чи об’єкт у системі штучного інтелекту.

Фреймворк – набір інструментів і бібліотек, що полегшує розробку й навчання нейромережевих моделей.

Функціональний стиль – різновид мови, що відповідає певній сфері спілкування (науковий, офіційний, художній тощо).

Хмарна технологія – цифрові сервіси зберігання та обробки даних через Інтернет.

Хмарне обчислення – модель надання обчислювальних ресурсів і сховищ даних через інтернет.

Цифрова (інформаційна) технологія – сукупність методів, засобів і процесів обробки інформації в електронному вигляді, що забезпечує її збереження, передачу, аналіз та використання у різних сферах діяльності.

Цифрова гуманітаристика – міждисциплінарна галузь, що поєднує гуманітарні науки з цифровими методами аналізу (до прикладу, науку про дані тощо).

Цифрова ідентичність – низка характеристик, що визначають користувача в онлайн-просторі.

Цифрова компетентність – сукупність знань, умінь та навичок, необхідних для ефективного використання цифрових технологій.

Цифрова трансформація (**Цифровізація**, **Диджиталізація**) – комплексна зміна соціальних і професійних практик під впливом цифрових технологій (до прикладу, концепція «Держава у смартфоні» від Міністерства цифрової трансформації України та інші).

Цифровий інструмент – програмний засіб, що використовується для аналізу, обробки та представлення даних.

Цифровий ресурс – тексти, бази даних, мультимедіа й інші матеріали у цифровому форматі.

Частотний корпус – корпус текстів, де підраховано частоти вживання мовних одиниць.

Чат-бот – програмний застосунок, що імітує діалог з користувачем у текстовій або голосовій формі.

Шингл – послідовність із кількох сусідніх слів або символів у тексті, що використовується для аналізу подібності (перевірці на плагіат).

Штучна нейронна мережа – обчислювальна модель, структура якої уподібнена людському мозку, що здатна до обробки (препроцесинг, класифікування та генерації) даних.

Штучний інтелект – галузь комп'ютерних наук, що моделює інтелектуальні функції людини.

Штучний нейрон – математична модель біологічного нейрона, що отримує вхідні сигнали, зважує їх, підсумовує та передає результат далі залежно від функції активації.

Якість даних – сукупність характеристик цифрової інформації, що визначають її достовірність і придатність до використання.

Application Programming Interface (API) – набір правил і методів взаємодії програмних компонентів.

POS-tagging (Part-of-Speech tagging) – автоматичне визначення частин мови у словах тексту.

Web scraping – метод автоматизованого збирання даних із вебсайтів.

ЦИФРОВІ ІНСТРУМЕНТИ ТА РЕСУРСИ

1. *AntConc / AntPConc*

Корпусний аналізатор, що локалізує частотність, колокації, ключові слова тощо текстових даних, здебільшого використовуваний у сучасних лінгвістичних дослідженнях.

2. *CATMA (Computer Assisted Text Markup and Analysis)*

Цифровий (веб) інструмент для розмітки та аналізу текстів, продуктивний для колективної роботи та аналітики.

3. *GATE (General Architecture for Text Engineering)*

Цифрова платформа для створення та тестування інструментарію обробки природної мови, зручна для великих проєктів у межах цифрової гуманітаристики.

4. *Google Books Ngram Viewer*

Вебінструмент для вивчення динаміки використання слів та фраз у великому текстовому корпусі оцифрованих книжок Google, який дозволяє відстежувати ті чи ті мовні тенденції.

5. *InterText*

Інструмент для аналізу інтертектуальних (міжтектових) зв'язків і порівняння таких даних (текстів) між собою, його застосовують для виявлення перекладацьких стратегій, відповідників і збігів у текстах перекладів.

6. *Istant Data Scraper*

Браузерне розширення для автоматичного збирання даних зі структурованих вебсторінок (таблиць, списків, каталогів та інших) без програмування. Особливо корисне для створення власних текстових корпусів у лінгвістичних та інших гуманітарних дослідженнях.

7. *Linguistic Inquiry and Word Count (LIWC)*

Інструментарій для психо- та соціолінгвістичного аналізу текстів, що виявляє емоційні і когнітивні маркери.

8. *NLTK*

Ефективна Python-бібліотека для обробки природної мови, у межах якої наявні інструменти токенізації, стемінгу, тегування, аналізу синтаксису та семантики текстових даних.

9. *NOVA Text Aligner*

Інструмент для автоматичного та напівавтоматичного вирівнювання текстів у паралельних корпусах, корисний для перекладознавчих та досліджень з прикладної лінгвістики.

10. *Sketch Engine / NoSketch Engine*

Цифрова платформа для корпусного аналізу з можливістю створення власних текстових корпусів, виявлення колокацій, граматичних структур і автоматичної побудови профілю слова.

11. *spaCy*

Ще одна Python-бібліотека, призначена для високопродуктивної обробки природної мови з підтримкою багатьох мов, яка забезпечує морфологічний і синтаксичний аналіз, локалізацію іменованих сутностей.

12. *Stanford NLP*

Набір інструментів від Стенфордського університету для глибинного синтаксичного та семантичного аналізу текстів, включно з побудовою дерев залежностей.

13. *TextAttributor*

Інструмент для визначення авторства текста, що працює на базі стилометричних показників, які застосовують в лінгвістичній експертизі та медіаграмотності.

14. *Transkribus*

Цифрова платформа для автоматичної обробки, розпізнавання, транскрибування й пошуку в рукописних і друкованих історичних текстах.

15. *TreeTagger*

Інструмент для морфологічного тегування та лематизації текстів, яка підтримує десятки мов і активно використовується у корпусній лінгвістиці.

16. *UDPipe*

Універсальний інструмент для лематизації, тегування, токенізації, синтаксичного парсингу текстів (відповідно до стандартів Universal Dependencies).

17. *Voyant Tools*

Цифрова платформа для візуальної й статичної обробки текстів, яка дозволяє швидко «витягути» параметризацію текстових даних (оглядові характеристики корпуса, графіки частот і хмари слів).

18. *WordSmith Tools*

Програмний пакет для корпусного аналізу, який використовують для роботи з конкордансами, частотними списками та колокаційними профілями.

19. *Wordnet*

Цифровий лексичний словник англійської мови, у якому матеріал погруповано за смисловими зв’язками, його використовують для аналізу семантики та обробки природної мови.

СПИСОК ДЖЕРЕЛ ТА ЛІТЕРАТУРИ

Джерела:

а) корпуси текстів:

1. Български национален корпус. *BNC* : вебсайт. URL: https://tinyurl.com/5yu32w4h (дата на обръщението: 01.08.25).

2. Генеральний регіонально анотований корпус української мови (ГРАК) / М. Шведова, Р. фон Вальденфельс, С. Яригін та ін. Київ ; Львів; Єна, 2017–2025. *ГРАК* : вебсайт. URL: https://tinyurl.com/59yr7h46 (дата звернення: 01.08.25).

3. Корпус української мови. *Mova.info* : вебсайт. URL: https://tinyurl.com/5c6776ym (дата звернення: 01.08.25).

4. *Корпус української мови* : загальномовний (або національний) неанотований та несистематизований корпус української мови. Містить 6,6 Гб україномовних текстів з електронної бібліотеки «Чтиво» : вебсайт. URL: https://tinyurl.com/2u7k4vud (дата звернення: 01.08.25).

5. *Проєкт "lang-uk"* : вебсайт. URL: https://tinyurl.com/njypnw92 (дата звернення: 01.08.25).

6. *Corpora Collection Leipzig* : website. URL: https://tinyurl.com/2f6yvzww (date of application: 01.08.25).

7. *Corpus de aprendices de español (CAES)* : sitio web. URL: https://tinyurl.com/ncaa42hb (fecha de solicitud: 01.08.25).

8. *Corpora and corpus tools* : website. URL: https://tinyurl.com/y3sppvtm (date of application: 01.08.25).

б) наочність:[2]

9. Векторне представлення : як його отримати. *Google.com* : website. URL: https://tinyurl.com/mw6s6u8u (дата звернення: 01.08.25).

10. Гнучкий диск. *Вікіпедія* : вебсайт. URL: https://tinyurl.com/bdzx7b64 (дата звернення: 01.08.25).

[2] Окрім цифрового інструментарію, який вміщено окремим тематичним блоком представленого списку джерел та літератури.

11. Історія розвитку розпізнавання мови. *Державний університет інформаційно-комунікаційних технологій* : вебсайт. URL: https://tinyurl.com/yc8netn7 (дата звернення: 01.08.25).

12. Комп'ютери помітно порозумнішають: скоро буде реалізована революційна пам'ять ULTRARAM. *Фокус* : вебсайт. URL: https://tinyurl.com/mvcze9na (дата звернення: 01.08.25).

13. Матвієнко С. Штучні нейронні мережі. *IT MASTER* : вебсайт. URL: https://tinyurl.com/5fn75v94 (дата звернення: 01.08.25).

14. Наукова робота без плагіату. Що рекомендує МОН? *Наука та метрика* : вебсайт. URL: https://tinyurl.com/3sabxfcv (дата звернення: 01.08.25).

15. Нейронна мережа Гопфілда. *Вікіпедія* : вебсайт. URL: https://tinyurl.com/4cttrj24 (дата звернення: 01.08.25).

16. Огляд процесорів Intel Core 13-го покоління. *Qbox computers* : website. URL: https://tinyurl.com/52ffz45v (дата звернення: 01.08.25).

17. Перцептрон. *Вікіпедія* : вебсайт. URL: https://tinyurl.com/475zb2pp (дата звернення: 01.08.25).

18. Рахункова машина Лейбница. *Galanix* : вебсайт. URL: https://tinyurl.com/m6udfpcd (дата звернення: 01.08.25).

19. Рефагі І. Британські шпигуни показали секретний комп'ютер Colossus, який обдурив Гітлера. Фокус : вебсайт. URL: https://tinyurl.com/ma273spa (дата звернення: 01.08.25).

20. Різницева машина Чарльза Беббіджа. *Galanix* : вебсайт. URL: https://tinyurl.com/yvft6tyx (дата звернення: 01.08.25).

21. Розроблено молекулярну плівку для нейроморфних комп'ютерів майбутнього. *Cikavosti.com* : пізнавальний інтернет-журнал : вебсайт. URL: https://tinyurl.com/yfhhwate (дата звернення: 01.08.25).

22. Стереоскоп. *Вікіпедія* : вебсайт. URL: https://tinyurl.com/585zs7zh (дата звернення: 01.08.25).

23. У 1984 випустили перший персональний комп'ютер Apple. *Хроніки Любарта* : вебсайт. URL: https://tinyurl.com/2699phvs (дата звернення: 01.08.25).

24. Фактчекінг. Листи до приятелів : журнал думки і чину : вебсайт. URL: https://tinyurl.com/yx9njpkb (дата звернення: 01.08.25).

25. Чернявська О. В. Абаки. *Черточка* : вебсайт. URL: https://tinyurl.com/mryavewz (дата звернення: 01.08.25).

26. Чернявська О. В. Антикітерський механізм. *Черточка* : вебсайт. URL: https://tinyurl.com/2n7swrk4 (дата звернення: 01.08.25).

27. Чернявська О. В. Квантові комп'ютери. *Черточка* : вебсайт. URL: https://tinyurl.com/44xj6mst (дата звернення: 01.08.25).

28. Чернявська О. В. Лічильна машина Леонардо да Вінчі. *Черточка* : вебсайт. URL: https://tinyurl.com/5bwfp92k (дата звернення: 01.08.25).

29. Чернявська О. В. Лічильний годинник Вільгельма Шиккарда. *Черточка* : вебсайт. URL: https://tinyurl.com/yu6cmj2r (дата звернення: 01.08.25).

30. Чернявська О. В. Машина Паскаля. *Черточка* : вебсайт. URL: https://tinyurl.com/596t7mrx (дата звернення: 01.08.25).

31. Шолудько Ю. Представлено перший у світі біологічний комп'ютер на основі клітин людського мозку. ТСН: вебсайт. URL: https://tinyurl.com/4an2dw7u (дата звернення: 01.08.25).

32. Що таке багатошаровий персептрон (Multilayer Perceptron, MLP) у машинному навчанні? *The Transmitted.com* : website. URL: https://tinyurl.com/ywkbd2xv (дата звернення: 01.08.25).

33. Що таке синтез мовлення? – Пояснення TTS. *Shaip* : вебсайт. URL: https://tinyurl.com/3nhe7y8t (дата звернення: 01.08.25).

34. 1968 Head-mounted VR display called 'The Sword of Damocles,' Ivan Sutherland. *Reddit* : website. URL: https://tinyurl.com/4vwycvsk (date of application: 01.08.25).

35. Adaptive Resonance Theory. *TutorialsPoint* : website. URL: https://tinyurl.com/mry4vf7k (date of application: 01.08.25).

36. Aik L. E., Wei Hong Tan W. H., Ahmad Kadri Junoh A. K. An Improved Radial Basis Function Networks Based on Quantum Evolutionary Algorithm for Training Nonlinear Datasets. IAES *International Journal of Artificial Intelligence (IJ-AI)*. 2019. Volume 8, № 2. P. 120–131. https://doi.org/10.11591/ijai.v8.i2.pp120-131

IAES *International Journal of Artificial Intelligence (IJ-AI)* : website. URL: https://tinyurl.com/ce5h656j (date of application: 01.08.25).

37. Alammar J. The Illustrated Word2vec. *Jalammar.github.io* : website. URL https://tinyurl.com/ym7vxvtm (date of application: 01.08.25).

38. Atanasoff-Berry Computer. *Computer History Museum* : website. URL: https://tinyurl.com/fsny6stk (date of application: 01.08.25).

39. ENIAC. *Computer History Museum* : website. URL: https://tinyurl.com/4juepe4x (date of application: 01.08.25).

40. Harford T. Origen de internet : la primera palabra con la que comenzó la fascinante historia de la red que cambió el mundo. *BBC News Mundo* : website. URL: https://tinyurl.com/4yfk249m (date of application: 01.08.25).

41. Mikroprocesor Intel 4004 – zapoczątkował rewolucję cyfrową. *eNews.pl* : website. URL: https://tinyurl.com/4dmmkvw2 (data zwrócenia się: 01.08.25).

42. PDP–1 Restoration Project. *Computer History Museum* : website. URL: https://tinyurl.com/y7bamh2j (date of application: 01.08.25).

43. Pennington J., Socher R., Manning C. D. GloVe : Global Vectors for Word Representation. *The Stanford NLP Group* : website. URL: https://tinyurl.com/6xjkp29a (date of application: 01.08.25).

44. Pereverzevs A. How to choose the best machine translation software for your company. *Lokalise* : website. URL: https://tinyurl.com/2enb76nz (date of application: 01.08.25).

45. Schuster M., Johnson M., Thorat N. Zero-Shot Translation with Google's Multilingual Neural Machine Translation System. *Google.com* : website. URL: https://tinyurl.com/3baf8kj3 (date of application: 01.08.25).

46. Text Summarization in NLP. *Top.Coder.com* : website. URL: https://tinyurl.com/nuask8zj (date of application: 01.08.25).

47. The Burroughs MOD 1 Guidance Computer Controls the Launch of the Atlas ICBM. *HistoryofInformation.com* : website. URL: https://tinyurl.com/msv9pdrf (date of application: 01.08.25).

48. The Sensorama : One of the First Functioning Efforts in Virtual Reality. *HistoryofInformation.com* : website. URL: https://tinyurl.com/mtbd7vuj (date of application: 01.08.25).

49. Top 10 Applications of Natural Language Processing. *SoluLab* : website. URL: https://tinyurl.com/5et2vz2a (дата звернення: 01.08.25).

50. ZX-Spectrum. *The EMG's ZX-SPECTRUM* : website. URL: https://tinyurl.com/yu66v7ks (date of application: 01.08.25).

в) нормативно-правові документи:

51. Деякі питання цифрової трансформації : розпорядження Кабінету Міністрів України від 17 лютого 2021 р. № 365-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/bdaf5j66 (дата звернення: 01.08.25).

52. Конституція України. *Президент України* : офіційне інтернет-представництво. URL: https://tinyurl.com/6m2sjcry (дата звернення: 01.08.25).

53. Положення про Міністерство цифрової трансформації України : постанова Кабінету Міністрів України від 18 вересня 2019 р. № 856. *Урядовий портал* : єдиний вебпортал органів виконавчої влади України. URL: https://tinyurl.com/4yskfxvy (дата звернення: 01.08.25).

54. Про вищу освіту : Закон України від 1 липня 2014 р. № 1556-VII. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/z6v4247e (дата звернення: 01.08.25).

55. Про електронні комунікації : Закон України від 16 грудня 2020 р. № 1089-IX. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/bddtabfw (дата звернення: 01.08.25).

56. Про забезпечення функціонування української мови як державної : Закон України від 25 квітня 2019 р. № 2704-VII. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/3c93xpd5 (дата звернення: 01.08.25).

57. Про затвердження плану заходів з реалізації Концепції розвитку штучного інтелекту в Україні на 2021–2024 роки : розпорядження Кабінету

Міністрів України від 12 травня 2021 р. № 438-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/26r9uzm3 (дата звернення: 01.08.25).

58. Про затвердження плану заходів з реалізації Концепції розвитку штучного інтелекту в Україні на 2025–2026 роки : розпорядження Кабінету Міністрів України від 9 травня 2025 р. № 457-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/fu6kheh7 (дата звернення: 01.08.25).

59. Про захист інформації в інформаційно-телекомунікаційних системах : Закон України від 5 липня 1994 р. № 80/94-ВР. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/bdc3sak5 (дата звернення: 01.08.25).

60. Про захист персональних даних : Закон України від 1 червня 2010 р. № 2297-VI. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/4ch6r6x4 (дата звернення: 01.08.25).

61. Про інформацію : Закон України від 2 жовтня 1992 р. № 2657-XII. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/3bv2d4h7 (дата звернення: 01.08.25).

62. Про медіа : Закон України від 13 грудня 2022 р. № 2849-IX. *Верховна Рада України :* офіційний вебпортал парламенту України. URL: https://tinyurl.com/mpeb7axk (дата звернення: 01.08.25).

63. Про основні засади забезпечення кібербезпеки України : Закон України від 5 жовтня 2017 р. № 2163-VIII. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/3pzh4k3e (дата звернення: 01.08.25).

64. Про рішення Ради національної безпеки і оборони України від 14 травня 2021 року «Про Стратегію кібербезпеки України» : указ Президента України від 6 серпня 2021 р. № 447/2021. *Рада національної безпеки і оборони України* : вебсайт. URL: https://tinyurl.com/mr3s67mn (дата звернення: 01.08.25).

65. Про схвалення Концепції розвитку цифрових компетентностей та затвердження плану заходів з її реалізації : розпорядження Кабінету Міністрів

України від 3 березня 2021 р. № 167-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/5n7m5btc (дата звернення: 01.08.25).

66. Про схвалення Концепції розвитку цифрової економіки та суспільства України на 2018–2020 роки та затвердження плану заходів щодо її реалізації : розпорядження Кабінету Міністрів України від 17 січня 2018 р. № 67-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/2b9zkpuz (дата звернення: 01.08.25).

67. Про схвалення Концепції розвитку штучного інтелекту в Україні : розпорядження Кабінету Міністрів України від 2 грудня 2020 р. № 1556-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/mryczrey (дата звернення: 01.08.25).

68. Про схвалення Стратегії відновлення, сталого розвитку та цифрової трансформації малого і середнього підприємництва на період до 2027 року та затвердження операційного плану заходів : розпорядження Кабінету Міністрів України від 30 серпня 2024 р. № 821-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/bdz4fkpw (дата звернення: 01.08.25).

69. Про схвалення Стратегії здійснення цифрового розвитку, цифрових трансформацій і цифровізації системи управління державними фінансами на період до 2030 року та затвердження плану заходів щодо її реалізації : розпорядження Кабінету Міністрів України від 17 листопада 2021 р. № 1467-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/4dxbucz2 (дата звернення: 01.08.25).

70. Про схвалення Стратегії цифрового розвитку інноваційної діяльності України на період до 2030 року та затвердження операційного плану заходів з її реалізації : розпорядження Кабінету Міністрів України від 31 грудня 2024 р. № 1351-р. *Верховна Рада України* : офіційний вебпортал парламенту України. URL: https://tinyurl.com/tds34xr5 (дата звернення: 01.08.25).

71. The Digital Europe Programme. *European Commission* : website. URL: https://tinyurl.com/mtk2ztfa (date of application: 01.08.25).

г) словникові проєкти:

72. *Горох* : онлайн бібліотека : вебсайт. URL: https://tinyurl.com/mse5p7en (дата звернення: 01.08.25).

73. *Cambridge Dictionary* : website. URL: https://tinyurl.com/skpfzkyz (date of application: 01.08.25).

ґ) цифровий інструментарій:

74. Архів української періодики онлайн. *Libraria.ua* : вебсайт. URL: https://tinyurl.com/tjfx2k63 (дата звернення: 01.08.25).

75. *Слово і діло* : вебсайт. URL: https://tinyurl.com/yv7dvc29 (дата звернення: 01.08.25).

76. AntConc. *Laurence Anthony's* : website. URL: https://tinyurl.com/2s3w83ym (date of application: 01.08.25).

77. *Botometer* : website. URL: https://tinyurl.com/5e52r3jw (date of application: 01.08.25).

78. *ChatGPT* : website. URL: https://tinyurl.com/ycypp7h3 (date of application: 01.08.25).

79. DALL E mini. *Hugging Face* : website. URL: https://tinyurl.com/356pfs3k (date of application: 01.08.25).

80. Gemini. *Google* : website. URL: https://tinyurl.com/yf9xapfj (date of application: 01.08.25).

81. Google Books Ngram Viewer. *Google Books* : website. URL: https://tinyurl.com/yen6zphj (date of application: 01.08.25).

82. Google Fact Check Explorer. *Google* : website. URL: https://tinyurl.com/2ds3szb4 (date of application: 01.08.25).

83. *Grok* : website. URL: https://tinyurl.com/437smc3c (date of application: 01.08.25).

84. InterText. *Wanthalf's Laiv* : website. URL: https://tinyurl.com/33zu47eu (date of application: 01.08.25).

85. Instant Data Scraper. *Chrome Web Store* : website. URL: https://tinyurl.com/44s4pc4a (date of application: 01.08.25).

86. *NLTK* : website. URL: https://tinyurl.com/3jjcmn63 (date of application: 01.08.25).

87. NOVA Text Aligner. *Supernova-Soft.com* : website. URL: https://tinyurl.com/4b8487rb (date of application: 01.08.25).

88. *OsoMeNet* : website. URL: https://tinyurl.com/y27pxnes(date of application: 01.08.25).

89. *PolitiFact* : website. URL: https://tinyurl.com/mubw4ffn (date of application: 01.08.25).

90. Sketch Engine / NoSketch Engine. *Sketch Engine* : website. URL: https://tinyurl.com/4b3zs3bk (date of application: 01.08.25).

91. *Snopes* : website. URL: https://tinyurl.com/3rnyhk2u (date of application: 01.08.25).

92. *spaCy* : website. URL: https://tinyurl.com/5bpvn5x7 (date of application: 01.08.25).

93. Stanford NLP. *The Stanford Natural Language Processing Group* : website. URL: https://tinyurl.com/4n4reh2m (date of application: 01.08.25).

94. *StopFake* : website. URL: https://tinyurl.com/7v6z32y2 (date of application: 01.08.25).

95. TextAttributor. *Ta.mova.info* : website. URL: https://tinyurl.com/mu9ctkwn (date of application: 01.08.25).

96. TreeTagger. *The Center for Information and Language Processing (CIS)* : website. URL: https://tinyurl.com/5f5p6mtn (date of application: 01.08.25).

97. UDPipe. *The Lindat/Clariah-cz Project* : website. URL: https://tinyurl.com/3m78yhu5 (date of application: 01.08.25).

98. *VoxCheck* : website. URL: https://tinyurl.com/37tdketd (date of application: 01.08.25).

99. *Voyant* : website. URL: https://tinyurl.com/2ssrx59r (date of application: 01.08.25).

100. *Wordnet* : A Lexical Database for English : website. URL: https://tinyurl.com/44f5mmcv (date of application: 01.08.25).

101. WordSmith Tools. *Lexically.net* : website. URL: https://tinyurl.com/2us56avf (date of application: 01.08.25).

Література:

102. Бокоч Т. Впровадження інформаційних технологій у навчання іноземних мов для фахівців-лінгвістів : сучасні підходи та виклики. *Актуальні питання гуманітарних наук*. 2024. Випуск 72, Том1. С. 191–195. DOI : https://doi.org/10.24919/2308-4863/72-1-26 *Актуальні питання гуманітарних наук* : вебсайт. URL: https://tinyurl.com/2n956hwa (дата звернення: 01.08.25).

103. Воротникова І., Дзябенко О., Морзе Н. Виклики впровадження персоналізованого навчання з використанням штучного інтелекту у вищій освіті. *Information Technologies and Learning Tools*. 2025. Volume 105, № 1. P. 144–157. DOI : https://doi.org/10.33407/itlt.v105i1.5893 *Information Technologies and Learning Tools* : website. URL: https://tinyurl.com/bddmebtw (date of application: 01.08.25).

104. Жуковська В. В. Застосування корпусних технологій у навчанні та вивченні іноземної мови. *Актуальні проблеми сучасної лінгвістики та методики викладання мови і літератури (за матеріалами онлайн-конференції, проведеної кафедрою міжкультурної комунікації та прикладної лінгвістики Навчально-наукового інституту іноземної філології)*. Житомир : Вид-во ЖДУ ім. Івана Франка, 2018. С. 39–59. *Електронна бібліотека Житомирського державного університету* : вебсайт. URL: https://tinyurl.com/959sf7yy (дата звернення: 01.08.25).

105. Захарченко О. Gpt Onlinecorrector як цифровий інструмент формування мовно-літературних компетентностей магістрів української філології. *Педагогічна наука і освіта XXI століття*. 2025. № 4. С. 374–382. *Педагогічна наука і освіта XXI століття* : вебсайт. URL: https://tinyurl.com/nhe55wv8 (дата звернення: 01.08.25).

106. Інтеграція штучного інтелекту в сферу освіти : проблеми, виклики, загрози, перспективи / Р. С. Гуревич та ін. *Сучасні інформаційні технології та інноваційні методики навчання в підготовці фахівців : методологія, теорія,*

досвід, проблеми. 2024. № 72. С. 170–186. DOI : https://doi.org/10.31652/2412-1142-2024-72-170-186 *Сучасні інформаційні технології та інноваційні методики навчання в підготовці фахівців : методологія, теорія, досвід, проблеми* : вебсайт. URL: https://tinyurl.com/54whbypf (дата звернення: 01.08.25).

107. Кремень В. Г., Спірін О. М. Роль Національної академії педагогічних наук України в цифровій трансформації освіти і науки. *Вісник Національної академії педагогічних наук України*. 2024. № 2, Том 6. С. 1–10. DOI : https://doi.org/10.37472/v.naes.2024.6228 *Вісник Національної академії педагогічних наук України* : вебсайт. URL: https://tinyurl.com/2rsppv5c (дата звернення: 01.08.25).

108. Скрипка Г. Штучний інтелект в освіті : удосконалення програм підвищення кваліфікації педагогів. *ITLT*. Vol. 101, № 3. P. 227–238. DOI: https://10.33407/itlt.v101i3.5639 *Електронна бібліотека НАПН України* : вебсайт. URL: https://tinyurl.com/ynedb5rp (дата звернення: 01.08.25).

109. Спірін О. М., Шишкіна М. П. Штучний інтелект в освіті та освітніх дослідженнях : виклики, ризики та перспективи інтеграції. Українське суспільство у перспективах розвитку : історичний, соціально-політичний, освітньо-педагогічний аспекти : Матеріали VI Всеукр. міжгалуз. науково-практ. онлайн-конф., м. Київ, 26–31 берез. 2025 р. Київ, 2025. С. 155–158. *Електронна бібліотека Житомирського державного університету* : вебсайт. URL: https://tinyurl.com/59akwvet (дата звернення: 01.08.25).

110. Швець Н. В., Швець О. В. Інтердисциплінарний підхід до підготовки перекладачів : синергія мовознавства, культурології та ІТ. *The 11th International scientific and practical conference "Science and technology : challenges, prospects and innovations" (June 19–21, 2025).* CPN Publishing Group, Osaka, Japan. 2025. P. 275–283. *Міжнародні наукові конференції* : вебсайт. URL: https://tinyurl.com/3x6nhe7u (дата звернення: 01.08.25).

111. Шишкіна М. П., Коваленко В. В. Про хід та результати досліджень, проведених в Інституті цифровізації освіти НАПН України, щодо використання штучного інтелекту в середній освіті : за матеріалами наукової доповіді на засіданні Президії Національної академії педагогічних наук України, 17 жовтня

2024 р. *Вісник Національної академії педагогічних наук України*. 2024. Том 6, № 2. С. 1–10. DOI: https://doi.org/10.37472/v.naes.2024.6217 *Вісник Національної академії педагогічних наук України* : вебсайт. URL: https://tinyurl.com/mzjyhrk5 (дата звернення: 01.08.25).

112. Aggarwal D. Exploring the Scope of Artificial Intelligence (AI) for Lifelong Education through Personalised & Adaptive Learning. *Journal of Artificial Intelligence, Machine Learning and Neural Network*. 2023. Volume 4, № 1. P. 21–26. DOI : https://doi.org/10.55529/jaimlnn.41.21.26 *Journal of Artificial Intelligence, Machine Learning and Neural Network* : website. URL: https://tinyurl.com/yt7y53az (date of application: 01.08.25).

113. Almasri F. Exploring the Impact of Artificial Intelligence in Teaching and Learning of Science : A Systematic Review of Empirical Research. *Research in Science Education*. 2024. Volume 54. P. 977–997. DOI : https://doi.org/10.1007/s11165-024-10176-3 *Springer Nature Link* : website. URL: https://tinyurl.com/3n9bt9pz (date of application: 01.08.25).

114. Artificial Intelligence in Higher Education : Enhancing Learning Systems and Transforming Educational Paradigms / M. Imran et al. *International Journal of Interactive Mobile Technologies (iJIM)*. 2024. Volume 18, № 18. P. 34–48. DOI : https://doi.org/10.3991/ijim.v18i18.49143 *International Journal of Interactive Mobile Technologies (iJIM)* : website. URL: https://tinyurl.com/35hxyhvr (date of application: 01.08.25).

115. Bayly-Castaneda K., Ramirez-Montoya M.-S., Morita-Alexander A. Crafting personalized learning paths with AI for lifelong learning : a systematic literature review'. *Frontiers in Education*. 2024. Volume 9. DOI : https://doi.org/10.3389/feduc.2024.1424386 *Frontiers* : website. URL: https://tinyurl.com/4vsp39ea (date of application: 01.08.25).

116. Gallego-Jiménez M. G., García-Barrera A. Analysis of the International Conception of Personalized Education. Difficulties for the Advancement in Scientific Educational Research. *Interchange*. 2023. Volume 54, № 3. P. 287–299. DOI : https://doi.org/10.1007/s10780-023-09495-7 Springer Nature Link : website. URL: https://tinyurl.com/y7frpycu (date of application: 01.08.25).

117. Kamalov F., Santandreu Calonge D., Gurrib I. New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability*. 2023. Volume 15, № 16. DOI : https://doi.org/10.3390/su151612451 *MDPI* : website. URL: https://tinyurl.com/8j87dx47 (date of application: 01.08.25).

118. Opportunities and Risks of Using AI-based Applications in Research : the Case of Ukrainian Universities / O. Borodiyenko, I. Drach, N. Bazeliuk et al. *Information Technologies and Learning Tools*. Volume 105, № 1. P. 125–143. DOI : https://doi.org/10.33407/itlt.v105i1.5794 *Information Technologies and Learning Tools* : website. URL: https://tinyurl.com/3wrf4c7w (date of application: 01.08.25).

119. Rakya Z. H. Exploring the Impact of Artificial Intelligence (AI) on Learner-Instructor Interaction in Online Learning (Literature Review). *International Journal of Emerging Multidisciplinaries : Computer Science & Artificial Intelligence*. 2023. Volume 2, № 1. DOI : https://doi.org/10.54938/ijemdcsai.2023.02.1.236 *International Journal of Emerging Multidisciplinaries : Computer Science & Artificial Intelligence* : website. URL: https://tinyurl.com/w47t66vx (date of application: 01.08.25).

120. Rani S., Kaur G., Dutta S. Educational AI Tools : A New Revolution in Outcome-Based Education. Explainable AI for Education : Recent Trends and Challenges. Cham : Springer Nature Switzerland, 2024. P. 43–60. DOI : https://doi.org/10.1007/978-3-031-72410-7_3 *ResearchGate* : website. URL: https://tinyurl.com/3dkb7fke (date of application: 01.08.25).

121. Revolutionising Distance Learning: A Comparative Study of Learning Progress with AI-Driven Tutoring / M. Möller et al. DOI : https://doi.org/10.48550/ARXIV.2403.14642 *arXiv.org* : website. URL: https://tinyurl.com/3f9v5fwb (date of application: 01.08.25).

122. Smart Education With Iot And Ai : Revolutionizing Learning In The Digital Age / S. Vijayalakshmi et al. *2nd International Conference on Disruptive Technologies (ICDT)*, Greater Noida, India, 2024. P. 1282–1286. DOI : https://doi.org/10.1109/ICDT61202.2024.10489741 *IEEE Xplore* : website. URL: https://tinyurl.com/5yv7c6v2 (date of application: 01.08.25).

ІНФОРМАЦІЯ ПРО АВТОРІВ



Довгань Олексій Валентинович

Google Scholar: https://tinyurl.com/26e4dedj

E-mail: o.v.dovhan@udu.edu.ua

кандидат філологічних наук, викладач кафедри прикладної лінгвістики, порівняльного мовознавства та перекладу факультету іноземної філології Українського державного університету імені Михайла Драгоманова

Гуманітарний науковець з міждисциплінарною підготовкою у сфері філології, культурології, літературознавства, філософії та соціогуманітарних наук. Має досвід академічного аналізу культурних ідентичностей, наративів пам'яті, постколоніального досвіду, а також соціальних трансформацій у сучасному українському контексті. Спеціалізується на роботі з текстами, дискурсивними практиками та культурними кодами, орієнтований на поєднання теоретичної глибини з етичним і суспільним виміром гуманітарного знання.



Матвєєва Світлана Анатоліївна

Google Scholar: https://tinyurl.com/495rvah8

E-mail: s.a.matvyeyeva@udu.edu.ua

доктор філологічних наук, професор, професор кафедри прикладної лінгвістики, порівняльного мовознавства та перекладу факультету іноземної філології Українського державного університету імені Михайла Драгоманова

Фахівець у галузі перекладознавства, зіставної та корпусної лінгвістики, когнітивного термінознавства. Має значний досвід дослідження мовних структур і термінологічних систем у контексті когнітивних підходів, поєднуючи теоретичні напрацювання з практичними аспектами перекладацької діяльності. Її наукові інтереси охоплюють аналіз мовних універсалій і специфічних рис національних мовних картин світу, розроблення корпусних ресурсів і методів автоматизованої обробки мовних даних, а також вивчення впливу когнітивних факторів на формування та функціонування термінології.

Навчальне видання

Довгань Олексій Валентинович
Матвєєва Світлана Анатоліївна

ЦИФРОВІ ТЕХНОЛОГІЇ В ЛІНГВІСТИЦІ: ВІД МОВНИХ КОРПУСІВ ДО ШТУЧНОГО ІНТЕЛЕКТУ

Підручник

Редактори: О. В. Довгань, С. А. Матвєєва

Електронне видання



Підписано до друку 19.09.2025 р. Формат 60x84.
Гарнітура Times.
Облік.видав.арк. 20,24
Зам. № 075
Віддруковано з оригіналів.

Видавництво Українського державного університету
імені Михайла Драгоманова.
01601, м. Київ-30, вул. Пирогова, 9
Свідоцтво про реєстрацію ДК 7896 від 25.07.2023.
(044) 239-30-26.



9 789669 313362