Міністерство освіти і науки України
Національний університет «Львівська політехніка»

В. В. Пасічник, М. В. Яромич

ГЕНЕРУВАННЯ, ОПРАЦЮВАННЯ ТА АНАЛІЗ ТЕКСТІВ:

інформаційні технології на основі великих мовних моделей

Навчальний посібник

Видавництво ПП «Магнолія 2006»
Львів 2026

УДК 004.8:81'322(075.8)

Затверджено Вченою радою
Національного університету «Львівська політехніка»

Рецензенти:

- **Н. Е. Кунанець** – д-р наук із соц. комунікацій, професор, професор кафедри інформаційних систем та мереж Національного університету «Львівська політехніка»;
- **В. І. Кут** – канд. техн. наук, доцент, завідувач кафедри інформатики та фізико-математичних дисциплін ДВНЗ «Ужгородський національний університет»;
- **О. П. Левченко** – д-р філол. наук, професор, завідувачка кафедри прикладної лінгвістики Національного університету «Львівська політехніка»;
- **Ю. В. Турбал** – д-р техн. наук, професор, завідувач кафедри комп'ютерних наук та прикладної математики Національного університету водного господарства та природокористування.

П19 **Генерування, опрацювання та аналіз текстів: інформаційні технології на основі великих мовних моделей** : навчальний посібник / В. В. Пасічник, М. В. Яромич. – Львів : Видавництво ПП «Магнолія 2006», 2026. – 383 с. – (Серія «Комп'ютинґ»).

ISBN 978-617-574-350-8

Навчальний посібник містить матеріал, необхідний для опанування теоретичних засад, функціональних можливостей та практичного застосування великих мовних моделей у задачах генерування, опрацювання й аналізу текстів. Розглянуто архітектурні принципи та обмеження мовних моделей, методи концептуального моделювання предметних галузей і побудови онтологій, зокрема темпоральних, технології автоматизованого реферування та керованого генерування наукових текстів, підходи до оцінювання якості згенерованого тексту, а також засади побудови оркестраторів агентів штучного інтелекту та мультиагентних середовищ наукового дослідження.

Розраховано на магістрів спеціальностей «Філологія» (B11) та «Інформаційні системи і технології» (F6), магістрів та докторів філософії спеціальності «Системний аналіз та науки про дані» (F4) галузей знань «Культура, мистецтво та гуманітарні науки» та «Інформаційні технології». Розглянуто базові засади та концепції великих мовних моделей, концептуальне моделювання предметних галузей, побудову онтологій та оцінювання якості згенерованих текстів, а також мультиагентні системи, інтелектуальну лінгвістичну лабораторію й методологічні питання відтворюваності, простежуваності та меж автоматизації наукового дослідження. Може бути використаний викладачами як дидактичний матеріал, а також для самостійного опрацювання та підвищення кваліфікації.

Публікується в авторській редакції.

ISBN 978-617-574-350-8 УДК 004.8:81'322(075.8)

ЗМІСТ

ПЕРЕДМОВА наукового редактора освітньо-наукової серії «КОМП'ЮТИНҐ»

Шановний Читачу!

Як науковий редактор освітньо-наукової серії книг КОМП'ЮТИНҐ і водночас один з авторів цього видання маю приємність звернутися до Тебе з нагоди виходу у світ навчального посібника «Генерування, опрацювання та аналіз текстів: інформаційні технології на основі великих мовних моделей».

Ця книга посідає в серії осібне місце, і мені хотілося б від початку пояснити, у чому саме полягає її відмінність. Упродовж двох десятиліть видання серії «Комп'ютинг» присвячувалися проблематиці систем баз даних та знань, інформаційному моделюванню, сховищам даних, інтелектуальним системам прийняття рішень – тобто предметним галузям, що мали вже усталену методологію, сформований понятійний апарат і багаторічну традицію викладання. Нинішнє видання вперше звертається до теми, якої ще кілька років тому просто не існувало як окремої галузі знань. Це накладає на авторів особливу відповідальність, адже писати доводиться не про усталене, а про те, що сьогодні формується динамічно в режимі реального часу.

У передмовах до попередніх видань освітньо-наукової серії я неодноразово звертався до системно-методологічного підходу, запропонованого академіком Віктором Михайловичем Глушковим у праці «Основи безпаперової інформатики» (1982), у якій було окреслено перші чотири цивілізаційні інформаційно-технологічні революції – появу мовлення, винайдення писемності, книгодрукування та комп'ютерів. У низці моїх публікацій цей ряд було продовжено ще двома революційними етапами, такими як винайдення Інтернету та широке упровадження методів і засобів штучного інтелекту. Отже, у цьому контексті на сучасному етапі людська цивілізація перебуває в умовах шостої інформаційно-технологічної революції, визначальною ознакою якої є масове практичне впровадження та використання технологій штучного інтелекту.

Проте в межах цієї шостої революції відбулася подія, яка, на моє переконання, заслуговує на окреме осмислення. Ідеться про появу великих мовних моделей. Уперше за всю історію інформаційних технологій інструмент опрацювання інформації став здатним оперувати не формалізованими даними, не структурованими записами й не заздалегідь визначеними схемами, а природномовним текстом у всій його багатозначності. Те, що впродовж півстоліття було предметом теоретичних сподівань і численних невдалих спроб формалізації, несподівано швидко перетворилося на повсякденний робочий інструмент, доступний кожному дослідникові.

Саме ця обставина й спонукала до підготовки цього навчального посібника. Великі мовні моделі увійшли в наукову практику значно швидше, ніж було вироблено методологію їхнього застосування, і між цими двома процесами утворився розрив, який відчуває кожен, хто працює зокрема з науковими текстами. Дослідник має потужний засіб, але не має відповіді на прості й водночас принципові питання: що саме можна доручити такому засобові, а що

ні; як відрізнити обґрунтований результат від правдоподібного; хто відповідає за зміст тексту, породженого машиною; і чи залишається науковою праця, у якій частину операцій виконано автоматично. Цей навчальний посібник є спробою дати відповіді принаймні на частину з них.

Особливо приємно мені відзначити, що це видання є спільною працею двох поколінь дослідників. Мій співавтор Максим Яромич – молодий талановитий дослідник, який здобув ступені бакалавра та магістра на факультеті прикладної математики та інформатики Львівського національного університету імені Івана Франка за спеціальністю «Комп'ютерні науки». Понад десять років працює в галузі інформаційних технологій як розробник програмного забезпечення, а нині він навчається в аспірантурі Національного університету «Львівська політехніка» на кафедрі прикладної лінгвістики Інституту комп'ютерних наук та інформаційних технологій за спеціальністю «Філологія», і провадить активну науково-дослідницьку діяльність за напрямом «Генерування наукових текстів великими мовними моделями на основі онтологічного підходу» – тобто саме з тієї проблематики, яка становить змістове ядро цього посібника. Прикметно, що його власний шлях від інженерної ІТ-освіти й багаторічної практики розробника до аспірантури з філології сам собою ілюструє ту міждисциплінарність, про яку йшлося вище. Щоб працювати з мовними моделями науково, недостатньо володіти лише технологією або лише мовознавчим апаратом.

Максим Яромич належить до того покоління, для якого великі мовні моделі є не новиною, що потребує осмислення, а природним середовищем професійної діяльності. У поєднанні двох перспектив – методологічної традиції, виробленої в наукових школах минулого століття, і безпосереднього досвіду роботи з новітніми технологіями – я вбачаю головну цінність нашої співпраці. Переконаний, що жодна з цих перспектив окремо не дала б повноцінного результату.

Питання, яким присвячено цей посібник, для мене особисто не є абстрактно-технологічними, а лежать виключно в науково-прикладній площині: що зміниться в цій праці тепер, коли частину операцій із текстом здатна виконувати машина, і що в ній не зміниться за жодних обставин.

Відповідь, до якої ми дійшли впродовж роботи над книгою, читач знайде в її останніх розділах, проте концептуально окреслю її вже тут. Великі мовні моделі істотно розширюють можливості опрацювання текстів, але не звільняють дослідника від обов'язку тлумачити й перевіряти. Машина здатна впорядкувати наявне; здобути ж нове знання й узяти за нього відповідальність може лише людина. Ця межа проходить не між простими й складними завданнями, а між процедурами перетворення даних та здобуттям нового знання, і усвідомлення цієї межі є, на мою думку, головною фаховою компетентністю дослідника нової доби.

Структуру посібника побудовано за принципом поступового ускладнення. Перші розділи знайомлять читача з базовими засадами та концепціями великих мовних моделей і з їхнім застосуванням до опрацювання й аналізу текстів. Далі розглядається концептуальне моделювання предметних галузей та побудова онтологій – тобто перехід від мовних одиниць до структурованого знання.

Наступні розділи присвячено генеруванню й аналізові наукових текстів, оцінюванню їхньої якості та практичному застосуванню в освітній галузі. Завершальні розділи описують побудову мультиагентних систем і концепцію інтелектуальної лінгвістичної лабораторії – середовища, у якому людина й програмні агенти працюють над спільними мовно-науковими завданнями як єдина гібридна інтелектуальна система.

Посібник адресовано магістрам спеціальностей «Філологія» (B11) та «Інформаційні системи і технології» (F6), а також магістрам і докторам філософії спеціальності «Системний аналіз та науки про дані» (F4) галузей знань «Культура, мистецтво та гуманітарні науки» та «Інформаційні технології» – і це поєднання не є випадковим. Тема, якій присвячено книгу, лежить точно на межі двох царин: без розуміння того, як улаштовано мову, неможливо коректно поставити завдання мовній моделі, а без розуміння того, як улаштована сама модель, неможливо правильно витлумачити її відповідь. Філолог, який працює з великими мовними моделями, і фахівець з інформаційних технологій, який працює з текстом, зустрічаються сьогодні на одному предметному полі, проте приходять на нього з різним рівнем знань і з різними прогалинами. Ми прагнули написати книгу, яка була б корисною обом, а тому пояснювали технологічні поняття без припущення про попередню математичну підготовку, а мовознавчі – без припущення про філологічну освіту.

Матеріал видання розраховано на магістрів та докторів філософії. Перші розділи подають базові засади та концепції, потрібні для свідомого користування мовними моделями в навчальній і фаховій роботі. Магістрам адресовано розділи, присвячені концептуальному моделюванню предметних галузей, побудові онтологій та оцінюванню якості згенерованих текстів, – тобто те, що потрібне для виконання кваліфікаційної роботи. Докторам філософії призначено насамперед завершальні розділи, в яких подано відомості про мультиагентні системи та інтелектуальну лінгвістичну лабораторію, а також наскрізну методологічну лінію посібника – питання відтворюваності, простежуваності й меж автоматизації в повному циклі наукового дослідження. Кожен розділ супроводжується запитаннями для самоперевірки та практичними завданнями, складність яких дібрано так, щоб їх можна було виконувати на обох рівнях із різною глибиною опрацювання.

Окремо відзначу одну особливість цього видання. Експерименти, описані в останніх розділах, виконано тими самими засобами, яким посібник присвячено, а частину дослідницьких процедур здійснено безпосередньо мовними моделями. Ми свідомо не приховували ані успіхів, ані невдач цих експериментів: описано і те, що система виконала добре, і те, чого вона виконати не змогла. Переконаний, що для навчального видання друге є не менш повчальним за перше, а подекуди й повчальнішим.

Вихід цього посібника став можливим завдяки багаторічній співпраці з видавництвом «Магнолія 2006», з яким двадцять років тому було укладено генеральну угоду, що започаткувала проєкт серії «Комп'ютинг» та підготовку її першого підручника – «Сховища даних». Щиро вдячний колективу видавництва

за незмінну підтримку серії впродовж усіх цих років і за готовність видавати книги з тематики, яка щойно формується.

Висловлюю також щиру вдячність колегам із Національного університету «Львівська політехніка» за наукове середовище, у якому ця праця постала, а рецензентам – за уважне прочитання рукопису та слушні зауваження, що допомогли зробити виклад точнішим.

Я вірю у світле майбутнє України, Батьківщини Ціолковського, Вернадського, Корольова, Глушкова, Патона, переконаний, що вона, спираючись на інтелектуальний потенціал, традиції наукових шкіл і здобутки вітчизняних учених, безсумнівно посідатиме чільне місце в майбутній високотехнологічній цивілізації, базованій на знаннях. Технології, яким присвячено цю книгу, дають для цього нові можливості – за умови, що ми навчимося користуватися ними розважливо й відповідально.

Щиро вдячний Тобі, мій Шановний Читачу, за увагу та відданість нашому спільному об'єкту професійної діяльності й наукових досліджень – тій масштабній і багатовимірній сфері, якою була, є і залишається галузь комп'ютингу.

З глибокою повагою
Науковий редактор серії КОМП'ЮТИНГ,
доктор технічних наук,професор,
Лауреат Державної премії України в галузі науки та техніки
Володимир ПАСІЧНИК

ВСТУП

В цивілізаційному вимірі розвитку історію інформаційних технологій зручно розглядати як певну послідовність революційних докорінних змін у способах накопичення, зберігання, опрацювання й передавання знання. Концептуальну основу такого революційного погляду закладено в працях видатного вченого, академіка НАН України, Віктора Михайловича Глушкова, який у фундаментальній праці «Основи безпаперової інформатики» (1982) сформулював теорію інформаційних бар'єрів і виокремив послідовні переломи в історії опрацювання інформації – появу мовлення, писемності, книгодрукування та електронно-обчислювальних машин. Продовжуючи цей ряд із позицій сьогодення, до нього додаються ще дві віхи – виникнення глобальної мережі Інтернет та широке впровадження методів і засобів штучного інтелекту. Саме шоста інформаційна революція визначає актуальний стан наукової, освітньої та професійної діяльності в галузі інформаційних технологій.

Слід зазначити, що всередині цієї шостої за ліком інформаційної революції відбулася подія, яка заслуговує на окреме осмислення й є центральним об'єктом та безпосереднім предметом цього навчального посібника. Ідеться про появу великих мовних моделей – систем, здатних опрацьовувати природномовний текст у всій його багатозначності. Впродовж півстоліття автоматизоване опрацювання мови спиралося на спроби формалізувати мовні правила, і ці спроби неодноразово наштовхувалися на те, що природна мова, на відміну від формальної, не має вичерпного опису. У парадигмі великих мовних моделей задачі розв'язуються принципово інакше. В них не описується мова правилами, а реконструюються закономірності її вживання зі статистики великих текстових масивів. Наслідком стало те, що операції, які донедавна вважалися винятково людськими – тлумачення значень, виокремлення концептів, побудова узагальнень, оцінювання якості наукового викладу, – раптово опинилися в зоні досяжності процесів автоматизації.

Швидкість цієї зміни створила розрив, який відчуває сьогодні кожен, хто працює з науковим текстом. Інструмент став масово доступним значно раніше, ніж було вироблено методологію його застосування. Дослідник має у своєму розпорядженні систему, здатну за хвилини виконати те, що донедавна потребувало годин ручної праці, але не має відповідей на питання, без яких наукове використання такої системи є не вичерпно осмисленим: що саме їй можна доручити, а що ні; як відрізнити обґрунтований результат від правдоподібного; звідки взялося те чи інше твердження в згенерованому тексті; хто несе відповідальність за його зміст; і чи залишається в традиційному розумінні науковою праця, у якій частину операцій виконано машиною. Саме цей розрив між доступністю інформаційної технології та зрілістю методології й зумовлює актуальність теми посібника.

Загострює проблему й те, що дана інформаційна технологія розвивається нерівномірно щодо різних мов. Переважна більшість досліджень, наборів даних, оцінювальних методик і практичних рекомендацій створюється на

англійськомовному матеріалі. Українськомовний науковий простір опиняється в становищі, коли готові рішення доводиться не просто застосовувати, а перевіряти на придатність: показники, відкалібровані для англійської мови, можуть давати систематично спотворені результати на мові з розвиненою морфологією й вільним порядком слів. Це стосується і якості генерування, і надійності автоматичного оцінювання, і навіть засобів виявлення машинно згенерованих текстів. Тому перенесення іншомовного досвіду на український ґрунт є не технічною, а власне науковою задачею, яка потребує окремої перевірки результатів на кожному кроці та етапі використання цієї, без перебільшення, інноваційної інформаційної технології.

Не менш істотним є термінологічний вимір проблеми. Галузь формується настільки швидко, що українські відповідники для її ключових понять здебільшого ще не усталилися. У фаховій літературі паралельно вживаються кілька варіантів того самого терміна, частина понять функціонує в англійській формі, а деякі – на кшталт prompt – не мають прийнятного відповідника взагалі. Ця обставина не є суто лексикографічною незручністю. Термін є мовною реалізацією концепту, і поки терміносистема не впорядкована, не може бути впорядкованим і понятєвий апарат галузі. Тому питання термінології проходить через увесь посібник як наскрізна лінія, а не як додаток до основного викладу.

Прикметно, що ця проблематика має в Україні тривалу й ґрунтовну традицію, на яку спираються автори посібника. Вітчизняна кібернетична школа, започаткована академіком В. М. Глушковим, від самого початку розглядала опрацювання інформації не як суто технічну задачу, а як питання подання знання. Саме в цій традиції сформувалися уявлення про інформаційні бар'єри та безпаперову інформатику, які й сьогодні лишаються методологічно чинними. У науковій школі академіка Анатолія Олександровича Стогнія ці ідеї було перенесено на ґрунт інформаційного моделювання систем баз даних та знань, звідки походить і наукова школа, у межах якої підготовлено це видання.

Мовний вимір названої традиції найповніше розвинувся в працях академіка НАН України, професора Володимира Анатолійовича Широкова та втілюється в роботах Українського мовно-інформаційного фонду НАН України. Її засадничою ідеєю є трактування словника не як переліку статей, а як системи знань із власною внутрішньою структурою. Лексикографічна праця спирається тут на формальні моделі мовних одиниць та відношень між ними, а результатом стають онтологокеровані лексикографічні платформи. Представники цієї школи послідовно обстоюють погляд, за яким майбутнє опрацювання мови пов'язане з інтеграцією мовних ресурсів, знань і штучного інтелекту в єдину інтелектуальну інфраструктуру. Під науковим керівництвом академіка Володимира Анатолійовича Широкова провадиться й системна праця над проблемами створення національної системи наукової термінології, до якої залучено Комітет наукової термінології НАН України та Національну комісію зі стандартів державної мови.

Ідеї, викладені в цьому посібнику, продовжують обидві лінії названої традиції. Від кібернетичної школи успадковано настанову на формальне подання знання й перевірюваність результату; від мовно-інформаційної – розуміння того,

що праця з мовою потребує власного апарату й не зводиться до застосування обчислювальних засобів. Великі мовні моделі не скасовують жодної з цих настанов, а лише додають до них новий інструмент, який сам собою природно потребує глибокого методологічного впорядкування.

Окремої уваги потребує вимір академічної доброчесності. Поширення генеративних технологій поставило перед академічною спільнотою питання, на які традиційні уявлення про авторство відповіді не дають. Якщо текст породжено машиною за завданням людини, кому він належить? Чи є використання таких засобів порушенням, і за яких умов? Чи можна покладатися на технічні засоби виявлення машинного тексту як на інструмент контролю? Досвід, описаний у завершальних розділах цього посібника, дає підстави для стриманої відповіді на останнє питання: надійність таких засобів залежить не лише від мови та жанру, а й від способу породження тексту, а отже, технічна експертиза не може бути єдиною опорою доброчесності. Значно надійнішим є шлях декларування використаних засобів і забезпечення простежуваності дослідницького процесу – тобто саме те, чому присвячено методологічну лінію цієї книги.

Нарешті, тема має виразно міждисциплінарний характер, і це отримує окрему інформаційно-методологічну складність. Робота з великими мовними моделями в наукових цілях лежить на межі мовознавства та інформаційних технологій, причому обидві сторони приходять на це спільне поле з різним рівнем підготовки. Фахівець з інформаційних технологій володіє достатньо глибокими знаннями, але часто не має інструментарію для того, щоб коректно сформувати та поставити мовознавчу задачу й оцінити якість її розв'язання. Філолог має такий апарат, але нерідко сприймає модель як «чорну скриньку», що унеможливлює обґрунтоване тлумачення її поведінки. Подолання цього розриву є однією з головних дидактичних цілей запропонованого навчального посібника.

Метою видання є формування в читача системного розуміння того, як великі мовні моделі можуть бути включені в наукову працю з текстом – не як універсальний «відповідач», а як компонент контрольованих інформаційних технологій, у яких модель виконує конкретну функцію, а її результат проходить багатоетапну, методологічно обґрунтовану та опрацьовану перевірку. Відповідно до цієї мети в посібнику послідовно вирішуються такі завдання, як розкриття базових засад та концепцій великих мовних моделей у формі, придатній для читача без спеціальної математичної підготовки; можливостей і меж їхнього застосування до класифікації, аналізу та інтерпретації текстів; викладу методики концептуального моделювання предметних галузей і побудови онтологій; опис технології онтологічно-керованого генерування наукових текстів; формування підходів до оцінювання якості одержаних результатів; і, нарешті, демонстрація побудови мультиагентних систем, здатних підтримувати повний цикл дослідницької праці.

Наскрізною для всього викладу є одна методологічна теза, яку варто сформулювати, оскільки вона складає базову цінність підходу, що сповідують автори цього навчального посібника. Наукова цінність тексту визначається не його стилістичною рівністю, а можливістю простежити, звідки взялося кожне твердження та надійно верифікувати його коректність та обґрунтованість. Великі

мовні моделі породжують стилістично рівний текст без труднощів – і саме тому головним завданням є не поліпшення якості викладу, а забезпечення простежуваності результату. Усі інформаційні технології, подані в посібнику, підпорядковані цьому базовому принципові.

Автори методологічно поєднують дві дослідницькі перспективи, кожна з яких окремо була б недостатньою для обраної теми. Один з авторів є вихованцем вітчизняної київської кібернетичної школи. Впродовж майже півстолітнього періоду наукової та викладацької праці в Національному університеті «Львівська політехніка» під його керівництвом сформувалось науково-методичне ядро високофахового осередку за профілем інформаційного моделювання систем баз даних та знань. Витоки цього процесу сягають ініціативи академіків Віктора Михайловича Глушкова та Анатолія Олександровича Стогнія, а його наукова проблематика пов'язана з формальним поданням, реляційним моделюванням та онтологіями предметних галузей.

Результативність науково-педагогічної діяльності професора В. В. Пасічника відображають такі показники. За майже півстолітній період сформувалася генерація дослідників, науковців, високофахових ІТ-професіоналів, до якої належать дванадцять докторів наук, а разом із наступним поколінням учнів – близько двадцяти; кандидатів наук підготовлено понад сорок, а з урахуванням учнів-учнів – понад шістдесят. П'ятнадцятьом вихованцям цього наукового осередку присуджено Премію Президента України для молодих учених – найвищу наукову відзнаку України за наукову діяльність молодого вченого.

Помітною складовою творчо-наукової діяльності стало розбудовування наукових осередків поза межами Львова: за профілем систем баз даних та знань наукові й педагогічні кадри готувалися на замовлення університетів Чернівців, Тернополя, Рівного, Луцька, Дрогобича та Ужгорода. У Львівській політехніці постав базовий осередок кафедри інформаційних систем та мереж, з якого згодом виросли кафедри соціальних комунікацій та інформаційної діяльності і кафедра систем штучного інтелекту.

Цей набутий колективний досвід, зокрема, визначає позицію, подану у цьому навчальному посібнику. Школа інформаційного моделювання систем баз даних та знань десятиліттями розробляла засоби, що надають знанню строгої форми й забезпечують перевірюваність. Великі мовні моделі підходять до тієї самої задачі з протилежного боку. Вони працюють із неструктурованим текстом, але не надають при цьому формальних гарантій. Поєднання цих двох способів опису, за якого онтологія відповідає за перевірюваність, а мовна модель – за здатність працювати з живим текстом, становить наскрізну ідею книги й прямо продовжує традицію названої школи.

Другий співавтор навчального посібника Максим Віталійович Яромич представляє інше покоління й іншу траєкторію входження в тему. Ступені бакалавра та магістра за спеціальністю «Комп'ютерні науки» він здобув на факультеті прикладної математики та інформатики Львівського національного університету імені Івана Франка, після чого понад десять років працює в провідних ІТ-компаніях, розробляючи та втілюючи численні ІТ-проєкти. Нині

він навчається в аспірантурі Національного університету «Львівська політехніка» на кафедрі прикладної лінгвістики Інституту комп'ютерних наук та інформаційних технологій за напрямом «Філологія», а темою його дослідження є генерування наукових текстів великими мовними моделями на основі онтологічного підходу.

Практичні результати завершальних розділів – програмна реалізація мультиагентної системи та експериментальна перевірка її можливостей – одержано саме в межах цього дослідження.

Матеріал видання побудовано за принципом поступового ускладнення: від базових понять до складних технологічних рішень, від окремих операцій з текстом до повного циклу дослідницької праці. Кожен наступний розділ спирається на поняття, введені в попередніх, тому послідовне опрацювання матеріалу є бажаним, хоча розділи допускають і вибіркове їх використання.

У першому розділі «Основні базові засади та концепції великих мовних моделей» закладається понятєвий фундамент усього посібника. Виклад починається з історичної перспективи. Розглядаються шість інформаційних революцій, концептуальну основу яких заклав В. М. Глушков, і паралельно подається погляд американського дослідника комунікації Дж. Д. Пітерса, який простежив історію не технологій, а самого поняття комунікації. Зіставлення цих двох перспектив – технологічної та смислової – задає рамку для подальшого розгляду. Перша пояснює, чому великі мовні моделі стали можливими саме тепер, друга нагадує, що зростання технічних можливостей ще не гарантує ані розуміння, ані істини.

Далі послідовно вводяться базові поняття галузі: співвідношення штучного інтелекту, машинного навчання та глибинного навчання; будова штучного нейрона й багатошарової мережі; векторні представлення слів і дистрибутивна гіпотеза; токенізація та контекстне вікно; промпт і режими роботи моделі; стадії створення мовної моделі від передтренування до донавчання. Особливу увагу приділено тим термінам, які функціонують у двох різних значеннях залежно від наукової традиції. Так, поняття токена подається окремо в інформатичному тлумаченні як підслівна одиниця, якою оперує модель, і в корпусно-лінгвістичному як окреме слововживання, протиставлене типові; поняття контекстного вікна – як обсяг тексту, що його модель утримує водночас, і як проміжок довкола ключового слова в корпусному аналізі. Таке розмежування є принциповим для міждисциплінарної аудиторії, оскільки саме змішування значень є найпоширенішим джерелом непорозумінь між фахівцями двох галузей.

Окремий підрозділ присвячено розмежуванню мови й мовлення та природної й формальної мови. Перше розмежування, запроваджене Ф. де Соссюром, пояснює, чому щодо мовної моделі коректно казати, що вона генерує текст, а не бере участь у мовленнєвій діяльності. У неї немає ані комунікативного наміру, ані адресата, ані ситуації спілкування. Друге пояснює, чому опрацювання формальної мови є або правильним, або хибним і перевіряється механічно, тоді як результат опрацювання природної мови завжди лишається ймовірнісним і потребує тлумачення. Саме з цієї відмінності випливає потреба

поєднувати формальні структури з мовними моделями, до чого посібник повертатиметься в подальших розділах.

Завершується розділ розглядом можливостей і меж великих мовних моделей. Тут вводиться означення галюцинації як твердження, поданого у формі фактичного, але позбавленого підстав, і показується, що це явище не є збоєм, а прямо випливає з улаштування моделі, навченої добирати найімовірніше продовження тексту, а не перевіряти його істинність. Розглядаються також способи пом'якшення цієї та інших вад, типові прикладні задачі опрацювання природної мови та вимоги відповідального використання, зокрема ті, що випливають із засад академічної доброчесності.

У другому розділі «Великі мовні моделі як інструмент опрацювання та аналізу текстів» здійснюється перехід від засад до практики й показується, що саме мовні моделі здатні робити з текстом. Виклад починається з операцій, які передують будь-якому аналізові: класифікації текстів, нормалізації, тобто зведення тексту до єдиного вигляду, та побудови числового подання. Розглядаються класичні підходи на кшталт TF-IDF у зіставленні з сучасними векторними моделями, а також показано, у яких випадках простіші статистичні методи лишаються доцільнішими за складніші.

Значну частину розділу присвячено жанровому аналізові. Показано, що ідея описувати жанр набором ознак не є винаходом доби мовних моделей. В українській лінгвостилістиці відповідний інструмент здавна відомий як паспорт мовного жанру. Різниця між традиційним описом і автоматичною параметризацією полягає радше в способі заповнення, ніж у задумі – те, що дослідник робив вручну на невеликому матеріалі, модель здатна виконати для сотень текстів. Тому лінгвістичний опис жанру доцільно розглядати не як попередника автоматичної класифікації, а як її змістову основу.

Далі розглядаються семантичний і стилістичний аналіз, робота з термінологією та побудова структур знання. Окремо обговорюються фразеологізми як найвразливіше місце мовних моделей. Значення сталого звороту не виводиться зі значень складників, тоді як модель обчислює ймовірність послідовності слів, а українська фразеологія подана в навчальних даних значно гірше за англійську. Наведено типологію помилок і практичні прийоми, що покращують результат. Так само розмежовано два значення терміна «тезаурус» – мовознавче та інформатичне, – оскільки нерозрізнення їх призводить до того, що на запит побудувати тезаурус предметної галузі модель повертає щось середнє, що не відповідає ані стандартові, ані лексикографічній традиції.

Завершальна частина розділу присвячена методології порівняння моделей. Описано застосування методу аналізу ієрархій Т. Сааті до багатокритеріального оцінювання мовних моделей у мовознавчих задачах, розглянуто критерії такого оцінювання та проблему відтворюваності результатів. Тут формулюється теза, істотна для всього подальшого викладу. Через недетермінованість моделей одна відповідь не є результатом – результатом є розподіл відповідей, а отже, методика має заздалегідь передбачати кількість повторів і спосіб узгодження розбіжностей.

У третьому розділі «Концептуальне моделювання предметних галузей та побудова онтологій засобами великих мовних моделей» здійснюється перехід від мовних одиниць до структурованого знання. Він починається з концептуально-термінологічного аналізу – виокремлення термінів, формування дефініцій, розмежування значень. Наводиться перелік вимог, які термінознавство висуває до терміна: однозначність у межах галузі, дефінітивність, системність, відсутність синонімії, мотивованість, стислість, деривaційна здатність, стилістична нейтральність, відповідність нормам мови та усталеність. Показано, що на практиці всі ці вимоги одночасно виконуються рідко, тож термінологічна праця здебільшого зводиться до пошуку компромісу між ними, а остаточне рішення не може бути суто автоматичним.

Центральним поняттям розділу є концепт, і йому, як і в попередніх випадках, надано два означення. У мовознавстві концепт є одиницею ментального рівня, що узагальнює знання й досвід про фрагмент дійсності та має понятєвий, образний і оцінний складники. В інформатиці концепт – це клас в онтології, формально описана сутність із набором властивостей і відношень. Різниця між ними не суто теоретична. Мовознавчий концепт відкритий і не має чітких меж, онтологічний замкнений і повністю визначений своїм описом. Побудова онтології для гуманітарної галузі завжди означає редукцію, і усвідомлення цієї втрати є умовою коректного тлумачення результатів.

Розділ послідовно розкриває процедуру концептуального моделювання: від виявлення концептів за термінами до встановлення типізованих зв'язків між ними й перевірки узгодженості одержаної структури. Розглянуто застосування підходу до різних за характером предметних галузей – технічної та гуманітарної, – і показано, у чому полягає відмінність. Окремої уваги надано темпоральним онтологіям, які дають змогу зафіксувати не лише структуру галузі, а й її зміну в часі: появу понять, зміщення значень, зміну фаз розвитку. Такий вимір є принциповим для галузей, що швидко змінюються, оскільки без нього виникає ризик подати різночасові поняття як одночасні.

У четвертому розділі «Генерування та аналіз наукових текстів» здійснюється перехід від аналізу до породження тексту. Розглядаються типові задачі наукової праці, які піддаються автоматизації: реферування, анотування, підготовка оглядів літератури, структурування викладу. Показано, у чому полягає різниця між переказом джерела й науковим рефератом і чому саме друге завдання є для моделі істотно складнішим.

Ключовою для розділу є ідея онтологічно-керованого генерування. Її суть полягає в тому, що текст породжується не безпосередньо із запиту, а на основі попередньо побудованої структури знання – онтології, термінологічної бази, плану. Завдяки цьому кожне твердження фінального тексту може бути зведене до конкретного елемента цієї структури, а отже, перевірене. Такий підхід протиставляється монолітному генеруванню, за якого текст постає з одного запиту й простежується походження окремих тверджень неможливо.

Значну увагу приділено проблемі галюцинацій у науковому тексті та способам їхнього запобігання. Показано, що найризикованішим місцем є бібліографія. Саме там модель найчастіше вигадує неіснуючі джерела або

приписує наявним працям твердження, яких вони не містять. Розглядаються методи контролю – від вимоги простежуваності до звіряння покликань із зовнішніми бібліографічними службами.

У п'ятому розділі «Оцінювання якості та практичне застосування згенерованих текстів» дається відповідь на питання, яке неминуче постає після будь-якого генерування: як визначити, чи є одержаний текст прийнятним. Розділ починається з тези про багатовимірність якості наукового тексту: фактографічна точність, термінологічна сталість, логічність побудови, повнота огляду, обґрунтованість висновків і якість викладу є різними вимірами, які не зводяться до єдиної оцінки й можуть перебувати в конфлікті між собою.

Розглядаються автоматичні та модельні метрики оцінювання, їхні можливості й обмеження, а також людська оцінка й проблема узгодженості експертів. Описано процедуру багатокритеріального оцінювання, яка дає змогу поєднати різні виміри якості в обґрунтовану підсумкову оцінку. Окремо розглянуто підхід, за якого в ролі оцінювача виступає сама мовна модель, і показано як його переваги, так і принципові обмеження.

Завершальна частина розділу присвячена практичному застосуванню розглянутих методів в освітній галузі: сценаріям використання, навчальним цілям та організації роботи тих, хто навчається. Наголошено, що навчальна цінність полягає не в одержанні готової відповіді, а в тому, щоб зробити видимими етапи наукового мислення – від корпусу до термінології, від термінології до структури знання, від структури до тексту.

Шостий розділ «Побудова оркестратора агентів штучного інтелекту для генерування наукових текстів» є першим із двох практично зорієнтованих і присвячується мультиагентному підходові. Виклад починається з теоретичних засад: поняття агента, його властивостей, архітектур і механізмів координації, зокрема з урахуванням доробку українських дослідників у галузі мультиагентних систем. Далі показано, що змінилося з появою великих мовних моделей. Агент отримав здатність приймати інструкції природною мовою, використовувати інструменти й повертати структурований результат.

Центральною частиною розділу є опис експериментальної реалізації семиагентного конвеєра, у якому кожен агент виконує окрему функцію: добір корпусу, термінологічне опрацювання, побудову онтології, темпоральне моделювання, планування, генерування тексту й валідацію. Показано головну перевагу такої побудови над одним запитом: кожен етап має власний вхід, вихід і критерій якості, а отже, помилку можна локалізувати там, де вона виникла, а не виявляти її в готовому тексті.

Окремо розглянуто дві методологічно істотні задачі. Перша – обґрунтований добір моделей для кожної ролі методом аналізу ієрархій, який показує, що поняття «найкращої моделі взагалі» позбавлене змісту. Різні етапи потребують різних профілів здатностей. Друга – вибір інструментальної платформи для оркестрації, з порівнянням кодових, візуальних та автоматизаційних середовищ за критеріями наочності, підтримки циклів, можливості інспекції проміжних станів і порогу входу.

У сьомому розділі «Інтелектуальна лінгвістична лабораторія: мультиагентна автоматизація наукового дослідження» узагальнюється весь попередній матеріал і доводиться до найскладнішої форми. Теоретична частина розглядає рівні організації агентних систем – від окремого агента через лінійний конвеєр і оркестратор до моделі віртуального підприємства, у якій поєднано рольову, процедурну, комунікаційну та артефактну структури з вбудованим контролем якості. Перенесення цієї моделі на наукову діяльність дає поняття інтелектуальної лінгвістичної лабораторії – середовища, що підтримує повний цикл праці з науковим знанням: від виявлення проблеми до підготовки публікації.

Значну увагу приділено трактуванню інтелектуальної лінгвістичної лабораторії як гібридної інтелектуальної системи, у якій людський і машинний інтелект не заміщують, а доповнюють один одного. Розглянуто принципи, без яких така система втрачає наукову цінність: прозорість, простежуваність і контрольована автономія. Окремим напрямом описано застосування лабораторії до впорядкування українських терміносистем, а також зіставлення з традицією онтологокерованих лексикографічних платформ, розвиненою в науковій школі академіка НАН України Володимира Анатолійовича Широкова.

Практична частина розділу описує експеримент, у якому лабораторія з дванадцяти агентів пройшла повний цикл підготовки наукової статті. Наведено архітектуру системи, системні інструкції всіх агентів, програмний код вузлів і схему руху даних із циклами повернення. Темою експериментального дослідження обрано виявлення згенерованих текстів – вона має рефлексивну властивість, оскільки описані в статті маркери можна застосувати до самої статті.

Результати експерименту викладено без пом'якшень. Показано, що система успішно виконала всі етапи, які є впорядкуванням наявного матеріалу, і збоїла саме там, де потрібна була дія, для якої в неї не було входу. Найцікавіший результат виявився незапланованим: рефлексивна перевірка засвідчила, що стилометричні маркери, відкалібровані на текстах, породжених одним запитом, не переносяться на тексти, породжені багатоагентною системою. Це спостереження має прямий наслідок для практики контролю академічної доброчесності, до якого посібник повертається у висновках розділу.

Посібник адресовано магістрам спеціальностей «Філологія» (B11) та «Інформаційні системи і технології» (F6), а також магістрам і докторам філософії спеціальності «Системний аналіз та науки про дані» (F4) галузей знань «Культура, мистецтво та гуманітарні науки» та «Інформаційні технології», і залежно від рівня доцільними є різні траєкторії опрацювання. Перші два розділи дають базові засади й формують навички свідомого користування мовними моделями в навчальній і фаховій роботі. Магістрам, крім того, будуть потрібні розділи про концептуальне моделювання, генерування наукових текстів та оцінювання їхньої якості – тобто те, що безпосередньо стосується виконання кваліфікаційної праці. Докторам філософії призначено насамперед завершальні розділи про мультиагентні системи та інтелектуальну лінгвістичну лабораторію, а також наскрізну методологічну лінію посібника: питання відтворюваності, простежуваності та меж автоматизації наукового дослідження.

Кожен розділ супроводжується запитаннями для самоперевірки та практичними завданнями. Складність завдань дібрано так, щоб їх можна було виконувати на обох рівнях із різною глибиною опрацювання. Те саме завдання може бути виконане як навчальна вправа або як самостійне дослідження. Означення ключових понять винесено у виділені врізки, а приклади подано в окремому форматі, що дає змогу користуватися посібником і як довідником.

Окремо варто наголосити на одній особливості видання. Експерименти, описані в завершальних розділах, виконано тими самими засобами, яким присвячено посібник, а частину дослідницьких процедур здійснено безпосередньо мовними моделями. Автори свідомо не приховували ані успіхів, ані невдач цих експериментів: описано і те, що система виконала добре, і те, чого вона виконати не змогла, разом із причинами кожного збою та способами його усунення. Для навчального видання невдалий результат із розібраною причиною є не менш повчальним за вдалий, а подекуди й повчальнішим, оскільки саме на межах технології найкраще видно її природу.

Насамкінець варто повернутися до тези, з якої починався цей вступ. Великі мовні моделі істотно розширюють можливості наукової праці з текстом, але не звільняють дослідника від обов'язку тлумачити й перевіряти. Машина здатна впорядкувати наявне; здобути нове знання й узяти за нього відповідальність може лише людина. Ця межа проходить не між простими й складними завданнями, а між процедурами перетворення даних та здобуттям нового знання. Усвідомлення цієї межі – не як тимчасового обмеження нинішніх технологій, а як принципової властивості наукового пізнання – і є, на переконання авторів, головною фаховою компетентністю дослідника нової доби. Саме її формуванню підпорядковано зміст цього посібника.

Начальне видання

В. В. Пасічник, М. В. Яромич

ГЕНЕРУВАННЯ, ОПРАЦЮВАННЯ ТА АНАЛІЗ ТЕКСТІВ:
інформаційні технології на основі великих мовних моделей

Навчальний посібник

Формат 70x100/16. Папір офсетний. Друк цифровий.
Гарнітура Times New Roman.
Умовн. друк. арк. 31,12.

Видавництво «Магнолія 2006»
м. Львів-53, 79053, Україна, тел.: +38 (050) 370-19-57
e-mail: magnol06@ukr.net
https://magnolia.lviv.ua

Свідоцтво про внесення суб'єкта видавничої справи
до Державного реєстру видавців, виготівників і розповсюджувачів
видавничої продукції: серія ДК № 2534 від 21.06.2006 року,
видане Державним комітетом інформаційної політики,
телебачення та радіомовлення України

Надруковано в друкарні видавця Марченко Т. В.

383