УДК 681.3.

РОЗРОБКА СЕМАНТИКО-СИНТАКСИЧНОЇ МОДЕЛІ ПРИРОДНОЇ МОВИ ЗА ДОПОМОГОЮ МЕТОДІВ НЕВІД'ЄМНОЇ ТЕНЗОРНОЇ І МАТРИЧНОЇ ФАКТОРИЗАЦІЇ

О.О. Марченко

Київський національний університет імені Тараса Шевченка, факультет кібернетики,
03680, Київ, проспект Академіка Глушкова, 2, корпус 6
Тел.: (044) 259 04 27; Факс: (044) 259 04 39; E-mail: rozenkrans@yandex.ua

Стаття описує методику розробки структурної моделі опису синтаксису і семантики природної мови. Дані про семантико-синтаксичні відношення мови, представлені у вигляді керуючих просторів синтаксичних структур речень, записуються у багатовимірних масивах. Після факторизації масиви даних служать основою для створення процедур семантичного та синтаксичного аналізу текстів.

A method of developing a structural model of natural language syntax and semantics is proposed. Syntactic and semantic relations between parts of a sentence are presented in a form of a recursive structure called a control space. Numerical characteristics of these data are stored in multidimensional arrays. After factorization, the arrays serve as the basis for the development of procedures for natural language semantic and syntactic analyses.

Вступ

Невід'ємна тензорна факторизація (NTF) останнім часом – це дуже популярна технологія в таких галузях як інформаційний пошук, обробка зображень, машинне навчання, обробка природної мови, та в інших суміжних напрямах. Даний підхід є одним з найбільш перспективних для виявлення й аналізу зв'язків і відношень у масивах даних, де описуються взаємопов'язані об'єкти N різних типів. У комп'ютерній лінгвістиці N-мірний тензор реалізується як багатовимірний масив даних, отриманих при частотному аналізі великих корпусів текстів. Факторизація N-мірного тензора при ранзі розкладання k формує N матриць, що складаються з k стовпців, які представляють відображення кожного окремого виміру тензора на k фактор-вимірів латентного семантичного простору. Це служить унікальним засобом для моделювання та виявлення взаємозв'язків лінгвістичних змінних у масиві N-мірних даних.

Метод невід'ємної факторизації тензорів можна назвати n-мірним узагальненням латентного семантичного аналізу [1], який використовується для обробки двомірних масивів даних. Структуру, отриману в результаті факторизації тензора, можна порівняти з багатошаровою нейронною мережею, що складається з N шарів, які представляють множини об'єктів N типів, та з прихованого комутаційного шару, що складається з множини комутаційних вузлів з різними ваговими коефіцієнтами. Даний шар моделює взаємозв'язок між об'єктами N типів і пов'язує N шарів в єдину нейронну мережу.

На даний час невід'ємна тензорна факторизація є перспективним методом у вирішенні задач комп'ютерної лінгвістики, про що свідчать численні роботи в цьому напрямку [2–5].

Особливий інтерес представляють роботи [2, 3], в яких описуються моделі тензорного представлення даних про частоту різних типів синтаксичних сполучень слів у реченнях, наприклад 3-вимірних сполучень типу subject – verb – object, або 4-вимірних сполучень типу subject – verb – direct_object – indirect_object або інших синтаксичних сполучень довжини, що не перевищує розмірність тензора N.

У тензорі кожний вимір представляє вісь деякого фіксованого члена речення – підмета, присудка, додатка, означення, обставини і т. д. N-мірні тензори містять оцінки частоти вживання сполучень певних наборів слів у реченнях в корпусах текстів. При цьому враховуються синтаксичні позиції слів. Після обробки великих текстових корпусів та накопичення значного обсягу даних у тензорі, формується N-вимірний масив опису комутаційних властивостей лексичних одиниць в реченнях даної мови, тобто для множини слів, представлених у тензорі, дано опис, в які синтаксичні відношення вони мають властивість вступати, з якими словами встановлюються дані відношення і з якою частотою.

Причому відношення ці є багатовимірні (N – максимальна розмірність для запису в тензор). Після цього йде етап невід'ємної факторизації отриманого тензора. Факторизація призводить до значного перетворення моделі представлення даних. Спочатку багатовимірний тензор є розрідженим і величезним за обсягом. Кожна з N-вісей синтаксичного простору містить десятки тисяч або сотні тисяч точок-слів. Після факторизації тензора його дані представляються у вигляді N матриць, що складаються з k стовпців (де значення k набагато менше, ніж число точок-слів у будь-якому з N вимірів тензора). Параметр k – ступінь факторизації, розмірність латентного семантичного простору, число ознакових вимірів у ньому. Крім значно більш компактного представлення масиву даних, надається можливість швидкого обчислення оцінки ймовірності будь-якого можливого сполучення слів у різних синтаксичних конструкціях речень. Це можна виконати шляхом обчислення суми

добутків компонент N k-вимірних векторів, що відповідають цим словам, вибраних з матриць, які відповідають їх синтаксичним позиціям.

Наприклад, щоб перевірити, наскільки ймовірним є використання словосполучення «Повар смажить качку», потрібно знайти в матриці SUBJECT k-вимірний вектор s, який відповідає іменнику «повар», потім знайти в матриці VERB k-вимірний вектор v, який відповідає дієслову «смажить». Після цього – знайти в матриці DIRECT_OBJECT k-вимірний вектор do, який відповідає іменнику «качка»; далі обчислюється сума добутків відповідних компонент цих трьох векторів:

$$x_{svdo} = \sum_{i=1}^{k} s_i v_i do_i \quad \text{(для 3-вимірного тензора N=3)},$$

де s_i – i-ий елемент вектора s, v_i – i-ий елемент вектора v, do_i – i-ий елемент вектора do.

Якщо результат суми x_{svdo} перевищує деякий пороговий рівень, то робиться висновок про можливість використання в мові такої послідовності слів у реченні. Обчислення даної оцінки для сполучення «Качка смажить повара» приводить до висновку про малу ймовірність такого словосполучення.

Дана модель дозволяє досить успішно автоматично виділяти з корпусів текстів такі лінгвістичні структури, як селекційні преференції (selectional preferences) [2] та субкатегоріальні фрейми дієслів (Verb SubCategorization Frame) [3], які поєднують у собі дані про синтаксичні та семантичні властивості взаємозв'язків між дієсловами та їх аргументами-іменниками у реченнях.

Очевидною проблемою цієї перспективної і потужної моделі є певна негнучкість та обмеженість представлення синтаксису речень природної мови. Розмірність тензору обмежує максимальну довжину речень-словосполучень, що описуються даною моделлю. Кожній осі відповідає конкретна синтаксична позиція. У роботі [2] описується 3-вимірний тензор для моделювання одного синтаксичного сполучення – підмет-присудок-додаток. У роботі [3] автор описує тензори розмірністю 9 та 12 для моделювання двох десятків різних типів синтаксичних відношень-сполучень. Просте збільшення розмірності тензору для обробки більшої кількості типів синтаксичних відношень розширеної арності не виглядає дуже переконливим засобом вдосконалення моделі. Актуальним і затребуваним напрямком досліджень у цьому контексті є пошук універсальних засобів представлення синтаксичних структур речень природної мови. Доцільно використати таку формальну модель представлення синтаксису, яка за допомогою рекурсії могла би виразити синтаксичні відношення речень довільної довжини і будь-якого ступеня складності структури. Така модель дозволила би записати багатовимірний структурний зв'язок між словами в реченнях будь-якої довжини у масивах фіксованої розмірності. У якості моделі представлення синтаксису мови пропонується використати керуючий простір синтаксичних структур природної мови [6]. Існує ряд класичних перевірених часом формальних моделей представлення синтаксису мови. Вибір саме керуючих просторів обумовлений тим, що в цій моделі за допомогою рекурсії описуються довільні складні конструкції через суперпозиції двох базових синтаксичних відношень – предикативних та синтагматичних. Запропонована лексико-синтаксична тензорна модель складається з одного 3-вимірного тензора для предикативних відношень та однієї матриці для синтагматичних відношень. Застосування керуючих просторів виявилося ефективним засобом редукції довільних n-арних синтаксичних відношень до суперпозиції бінарних та 3-арних відношень.

Тензорні моделі містять дані про семантико-синтаксичні комунікаційні властивості лише тих слів, які містяться в оброблених текстових корпусах та лише в рамках тих речень і словосполучень, в яких дані слова зустрічалися. Іншими словами тензорна модель відтворює лише ті речення та словосполучення, які містяться в оброблених текстових корпусах. У роботі запропоновано використовувати ієрархічні лексико-семантичні бази типу WordNet [7] для узагальнення описів комунікаційних властивостей слів із застосуванням неявних механізмів наслідування по гілкам дерева таксономії. Припустимо, що якщо певна властивість є у слова A, то з великою імовірністю ця властивість може бути у всіх слів синсету, в якому міститься A. Також з великою імовірністю ця властивість є присутньою у слів синівського синсету, а також у слів батьківського синсету. Саме ці припущення стали основою для реалізації механізму узагальнення опису комунікаційних семантико-синтаксичних властивостей слів по принципу таксономічного наслідування.

Загальновідомо, що для володіння природною мовою потрібні знання безпосередньо про мову (лексика, морфологія, синтаксис) та знання про оточуючий світ (мовні реалії, семантика). Тензорні моделі містять дані, в яких інтегровані семантичні та синтаксичні комунікаційні властивості слів. Застосування лексико-семантичних баз типу WordNet посилює семантичну складову моделі. В роботі як навчальні тексти разом з корпусом The Wall Street Journal також були використані тексти статей English Wikipedia та Simple English Wikipedia, як такі, що містять визначення понять та основну інформацію про них, для поглиблення семантики в моделі.

1. Керуючий простір синтаксичних структур природної мови

Основні синтаксичні конструкції описуються в класичних схемах граматики мови, які належать до періоду античності і мало змінилися до теперішнього часу.

Досить тонкі відношення керування між словами виражаються у лінгвістичних моделях дерев підпорядкування і систем складових. Очевидною перевагою перерахованих моделей є їх коректність − адекватне відображення специфічних характеристик синтаксичної структури речення. Дані моделі не позбавлені недоліків. Модель дерев підпорядкування орієнтована на керуючі зв'язки між словами, а модель систем складових враховує ієрархічне відношення вкладеності словосполучень в лінійній структурі тексту. Ці моделі лише наближено описують дійсні комунікативні властивості синтаксичних структур.

Спроби побудови більш зручних для машинної обробки моделей, узагальнюючих властивості дерев підпорядкування і систем складових призвели до створення моделі системи компонент А.С. Нариньяні [8] та синтаксичних груп А.В. Гладкого [9]. В цих моделях відбувається переміщення кута зору на синтаксичні структури з лінійного порядку, нав'язаного послідовністю запису тексту, до складного простору, утвореного синтаксично зв'язаними групами об'єктів. У роботі [6] запропоновано перейти до простору представлення, не залежного від порядку запису тексту, а значить і від національної мови. Простір виражає всі предикативні та синтагматичні відношення, що містяться в синтаксичних структурах. Цей простір назвали *керуючим*.

Розглянемо запропоновану алгоритмічну модель речення природної мови. На відміну від суто лінгвістичного підходу, речення розглядається як деякий динамічний обчислювальний рекурсивний процес, який розвивається в керуючому просторі, що пов'язує синтаксично згруповані частини речення інформаційними каналами. Структура керуючого простору відображає семантику синтагматичних і предикативних конструкцій мови.

Крім властивості давати імена об'єктам навколишнього світу, мова володіє фундаментальною властивістю виражати динамічні відношення, в які вступають об'єкти. Так, дієслово пов'язує у відношення об'єкти, що беруть участь у схемі дії цього дієслова, прикметник задає відношення об'єкта з самим собою. Синтаксична модель має містити опис, які частини речення пов'язані між собою через відношення, і якого типу ці відношення. Існують два види синтаксичних відношень – предикативні і синтагматичні. Предикативне відношення виражає залежність між синтаксичними об'єктами через поняття, що означає дію і зазвичай виражається за допомогою присудка – дієслова. Синтагма − це поєднання двох синтаксичних об'єктів, з яких один є визначенням іншого, тому в моделі мають повністю виражатися саме ці види відношень. Крім того, в широкому розумінні синтагми мають утворювати синтаксичні групи.

Адекватна модель синтаксичної структури має також відображати основну властивість рекурсивності мови – здатність розгортати власні визначення, тобто давати уточнення, характеристики, коментарі до своїх частин, а також будувати визначення визначень.

Навмисно порушується традиційний лінгвістичний підхід, при якому присудок вважається головним членом речення, від якого ідуть керуючі зв'язки. Це успадкувалося від звички вважати ім'я функції головнішим, аніж її аргументи. Для побудови даної моделі зручніше задавати синтаксичні відношення зв'язками *генерації* і *передачі відношень*. При цьому досягається більш точна характеристика керуючих зв'язків.

Якщо два об'єкти A і B вступають у відношення C, то ми виділяємо об'єкт (припустимо A), що викликає (ініціює, породжує) це відношення C і об'єкт, на який передається це відношення − B. Таким чином, виділяємо два види спрямованих зв'язків: від *об'єкта-генератора відношення* до *відношення* і від *відношення* до *підпорядкованого об'єкту*. Перший вид зв'язку називаємо α-зв'язком (зв'язок генерування), другий − β-зв'язком (зв'язок розповсюдження). Об'єкти A, B і відношення С розміщуються в точках керуючого простору, і тому графічне представлення відношення C, що зв'язує A і B, має вигляд, зображений на рис. 1.

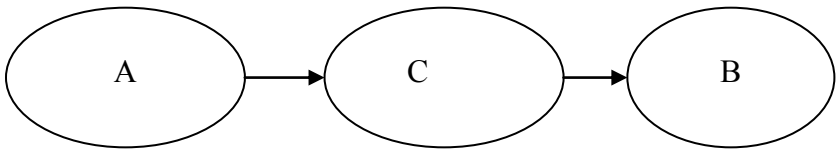

Рис. 1. Об'єкт A генерує відношення C, яке передається на об'єкт B

Дієслова визначають відношення між об'єктами. Тому в стандартній схемі простого речення: «іменник – дієслово – іменник» α-зв'язок спрямований від першого іменника до дієслова, а β-зв'язок спрямований від дієслова до іменника-визначення. Розглянемо приклад: *Дівчинка спекла торт*. Об'єкт *дівчинка* генерує відношення *спекла* і направляє його на об'єкт *торт*. Тому α-β-структура цього речення має вигляд як показано на рис. 2.

Рис. 2. Структура речення Дівчинка спекла торт

Розглянемо фразу: *Талановитий студент*. Тут об'єкт *студент* генерує унарне відношення *талановитий* і передає це відношення собі, як показано на рис. 3. Виникає кільцевий зв'язок, що характеризує визначення.

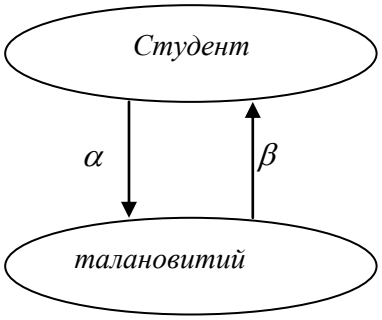

Рис. 3. Структура словосполучення Талановитий студент

Аналогічно міркуючи, для фрази *Талановитий студент швидко розв'язує рівняння* отримуємо структуру, яка показана на рис. 4.

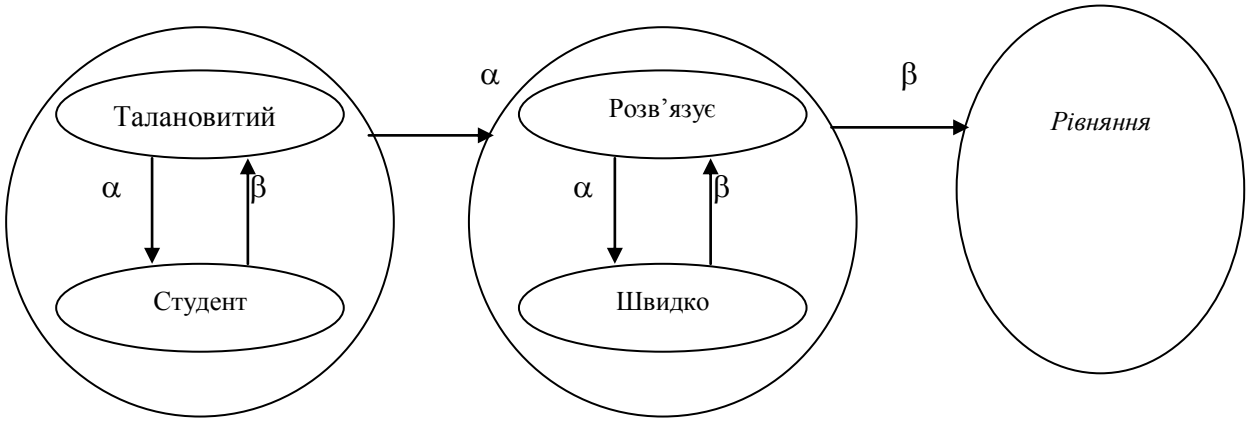

Рис. 4. Структура речення Талановитий студент швидко розв'язує рівняння

Речення мають два типи α-β-зв'язків: строго лінійна залежність і замкнута кільцева залежність. Першу називають лінійної конструкцією, другу − визначенням. Перша відповідає предикативним конструкціям мови, друга синтагматичним.

Формальна модель, орієнтована на завдання складних структур необхідного виду у формі керуючих просторів, будується наступним чином.

Дано клас базових об'єктів U. З кожним об'єктом асоціюється певний тип. Всього різних типів скінчене число. Типи можна виразити числами з інтервалу $[0, N]$. Припускаємо неоднозначність при зіставленні об'єктів типам, тобто функція приписування типів φ, взагалі кажучи, відображає U у множину всіх підмножин, утворених числами з інтервалу $[0, N]$. Конструкціями є або об'єкти з U, або конструкції, що отримані з інших конструкцій за допомогою підстановки останніх у точки лінійної або визначальної залежності. Правила обчислення типів конструкцій мають наступний вигляд:

1. Якщо в лінійній залежності об'єкт типу ***i*** з'єднується α-зв'язком з об'єктом типу ***j***, а останній β-зв'язком з об'єктом типу ***k***, то тип такої конструкції дорівнює ***f* (*i, j, k*)**, де ***f*** – задана частково-визначена функція.

2. Нехай ***d* (*i, j*)** – задана частково-визначена функція, тотожно рівна 1 в точках свого визначення. Функція ***d*** називається функцією узгодження. Якщо об'єкт типу ***i*** у визначальній конструкції уточнюється за допомогою об'єкта типу ***j***, то всій конструкції приписується тип ***i***, якщо тільки значення функції узгодження дорівнює 1. В іншому випадку значення типу не визначено.

Так як множина базових типів є скінченою, то функції ***f*** і ***d*** можуть бути задані таблицями. Правило 2 дозволяє легко обчислювати тип будь-якої складної конструкції. Легко довести, що тип будь-якої конструкції збігається з типом однієї з базових конструкцій, що задаються функціями ***f*** або ***d.***

Якщо неможливо обчислити тип конструкції, то вона вважається некоректною. Всі коректні конструкції утворюють керуючі простори класу U.

Стосовно синтаксичних структур дане визначення уточнюється наступним чином.

Базові об'єкти – це слова і прості словосполучення, що представляють собою частини мови (іменники, прикметники, дієслова, частки і т. д.) з відповідними морфологічними ознаками, а також складносурядні відношення і корелятори, призначені для з'єднання підпорядкованих речень з головними. Тип слова – це повна його граматична характеристика. Наприклад, тип слова *книга* дорівнює (іменник, неживе, однина, називний відмінок). Можливе розширення поняття типу додаванням деяких семантичних атрибутів. Неоднозначність завдання типу проявляється в неоднозначності розуміння значення деяких слів у відриві від контексту. Наприклад, слово ***мати*** може бути іменником або дієсловом. Функція ***f*** задає типи простих речень, а також тип складного речення залежно від його конструкції верхнього рівня. До базових об'єктів відносимо корелятори, що представляють собою або скріпу, або пару службових слів – (скріпа, співвідносне слово). Скріпа – службове слово в підпорядкованому реченні, що служить для прив'язування цього речення до головного. Співвідносне слово знаходиться в головному реченні (якщо воно є) і служить для зв'язку з відповідною скріпою. Функція ***d*** задає умови узгодження типів об'єкта, що визначається, та об'єкта, який визначає. Наприклад, визначеннями до іменника можуть бути прикметники, прийменникові групи або підпорядковані речення, до дієслова – прислівник, дієприслівник або підпорядковане речення; дієслово не може бути визначенням для іменника і т. д.

Таким чином, функції ***f*** і ***d*** виконують роль фільтру, що виділяє допустимі конструкції. Побудова таблиць значень функцій ***f*** і ***d*** представляється трудомісткім, але цілком реальним завданням. Все необхідне для цього є в класичній граматиці мови. Так як у визначальній конструкції роль підпорядкованої частини зводиться до коментарю або до уточнення головної частини, то значення типу всієї синтагматичної конструкції вибрано рівним значенню головного об'єкта − генератора властивості.

В роботі [6] показано, як елементарними перетвореннями можна конвертувати керуючий простір довільного речення як у дерево підпорядкування, так і у дерево виведення. В цьому сенсі структура керуючого простору одночасно узагальнює як дерева підпорядкування, так і дерева виведення. Керуючі простори можуть виразити синтаксичну структуру довільної складності та арності у вигляді набору бінарних та 3-арних відношень, що дозволяє точно записати всі дані про семантико-синтаксичні зв'язки всередині речення за допомогою однієї матриці D та одного тривимірного тензору F.

2. Побудова лексико-синтаксичної моделі природної мови

Для побудови семантико-синтаксичної моделі природної мови розроблена система автоматичного заповнення тривимірного тензору F та матриці D в процесі синтаксичного аналізу та пост-обробки синтаксичних структур речень великого текстового корпусу. Система має виконувати наступну послідовність дій:

- система послідовно приймає на вхід речення з великого текстового корпусу та виконує їх синтаксичний аналіз за допомогою модуля граматичного розбору Stanford Parser, який генерує синтаксичні структури речень у вигляді дерев підпорядкування та дерев виведення [10, 11];
- система аналізує дерево підпорядкування та дерево виведення поточного речення, збираючи керуючий простір його синтаксичної структури, перебираючи зв'язки між словами для виявлення предикативних сполучень довжиною 3 (підмет-присудок-додаток), а також синтагматичних сполучень довжиною 2 (іменник-прикметник, дієслово-прислівник і т. п.);
- після генерації керуючого простору синтаксичної структури поточного речення для кожної трійки точок (***i, j, k***), зв'язаних лінійною предикативною послідовністю α-β-зв'язків, в тензорі F у комірці ***F[I, J, K]*** значення збільшується на одиницю***: F[I,J,K]= F[I,J,K]+1***. Координати комірки тензору ***I, J, K*** відповідають парам (w_i, A_i), (w_j, A_j) та (w_k, A_k), де w – це слова, що є лексичними значеннями відповідних точок (***i, j, k***), а A – закодований опис характеристик цих лексем (частина мови, рід, число даної лексичної одиниці і т. д.);
- аналогічно у керуючому просторі синтаксичної структури поточного речення для кожної пари точок (***i, j***), зв'язаних між собою кільцевим синтагматичним α-β-зв'язком, у матриці D в комірці ***D[I, J]*** значення збільшується на одиницю: ***D[I,J]= D[I,J]+1***. Координати ***I, J*** відповідають парам (w_i, A_i) та (w_j, A_j), де w – це слова, що є лексичними значеннями відповідних точок (***i, j***), а A – закодований опис характеристик цих лексем.

Після обробки великого обсягу текстів у матриці D та у тривимірному тензорі F накопичується достатньо інформації про семантико-синтаксичні комунікативні властивості набору лексем для ефективної реалізації лексико-синтаксичної моделі природної мови. Надвелика розмірність та розрідженість утвореної матриці D та побудованого тензору F вимагають трансформації структур даних з метою більш економного та зручного представлення для збереження і обробки. Для оптимізації отриманих величезних масивів даних найкраще підходять методи невід'ємної матричної та тензорної факторизації.

3. Факторизація матриці D

Для розкладання матриці великої розмірності $D(N \times M)$ у вигляді добутку двох матриць $W(N \times k) \times H(k \times M)$, де ($k << N, M$), доцільно використати алгоритм невід'ємної матричної факторизації NMF, що був запропонований Лі та Суном [12]. У цільовій функції використовується норма Фробеніуса, як описується формулою.

$$\min_{W,H} \|D - WH\|_{\mathrm{F}}^2, \tag{1}$$

причому елементи матриць W та H повинні бути невід'ємними.

Для такої цільової функції, та для двох початкових матриць W_0 і H_0, NMF алгоритм складається з ітераційного виконання двох кроків:

$$(H_k)_{i,j} = (H_{k-1})_{i,j} \times \frac{(W_{k-1}^T D)_{i,j}}{(W_{k-1}^T W_{k-1} H_{k-1})_{i,j}}, \tag{2}$$

$$(W_k)_{i,j} = (W_{k-1})_{i,j} \times \frac{(D H_{k-1}^T)_{i,j}}{(W_{k-1} H_{k-1} H_{k-1}^T)_{i,j}}. \tag{3}$$

На практиці, кроки алгоритму повторюються, доки не буде досягнута нерухома точка або не буде виконана максимальна кількість ітерацій. Лі та Сун довели дві основні властивості цього алгоритму: по-перше, цільова функція є монотонно спадною під час застосування правил; по-друге, матриці W і H стають постійними тільки у випадку досягнення стаціонарної точки цільової функції.

4. Факторизація тензору F

Для розкладання тензору використовується невід'ємна тензорна факторизація [13]. Він подібний до паралельного факторного аналізу з обмеженням, що всі дані мають бути невід'ємними. Паралельний факторний аналіз – це мультилінійний аналог сингулярного розкладання матриць, що використовується в латентному семантичному аналізі. Головна ідея методу − мінімізація суми квадратів різниць між оригінальним тензором і факторизованою моделлю тензору. Для 3-вимірного тензору $T \in R^{D_1 \times D_2 \times D_3}$ визначається цільова функція:

$$\min_{x_i \in R^{D_1}, y_i \in R^{D_2}, z_i \in R^{D_3}} ||T - \sum_{i=1}^{k} x_i \circ y_i \circ z_i||_F^2, \tag{4}$$

де k − розмірність факторизованої моделі, а $\circ$ − зовнішній добуток (outer product).

Для невід'ємної факторизації додаються обмеження щодо невід'ємності значень елементів:

$$\min_{x_i \in R_{\geq 0}^{D_1}, y_i \in R_{\geq 0}^{D_2}, \ldots, z_i \in R_{\geq 0}^{D_N}} ||T - \sum_{i=1}^{k} x_i \circ y_i \circ \ldots \circ z_i||_F^2. \tag{5}$$

Результат роботи алгоритму − представлення тензору у вигляді трьох матриць, які описують відображення кожної з розмірностей тензору на k фактор-вимірів латентного семантичного простору. NTF модель підганяється методом найменших квадратів. На кожній ітерації дві з розмірностей фіксуються, а третя розмірність підганяється методом найменших квадратів. Процес триває до моменту збіжності.

5. Властивості лексико-синтаксичної моделі природної мови

Факторизацією матриці D та тензору F система формує потужну базу, яка містить у собі дані про будову синтаксичних структур речень природної мови, в які інтегрований опис лексико-семантичних відношень між словами. Окрім загального синтаксису, що задає структуру речень в загальному абстрактному вигляді, база містить лексико-семантичні обмеження, які визначають, які слова можуть утворювати зв'язок певного синтаксичного типу. Для того, щоб визначити, чи можуть два слова **a** та ***b*** утворити кільцевий синтагматичний зв'язок, треба взяти з матриці W вектор-строку $\boldsymbol{W_a}$, що відповідає слову ***a***, з матриці H – вектор-стовпчик $\boldsymbol{H_b}$, що відповідає слову ***b***, та обчислити скалярний добуток векторів ($\boldsymbol{W_a}$, $\boldsymbol{H}_b^T$). Якщо значення добутку перевищує певний пороговий рівень, то даний зв'язок є визначеним. Для того, щоб визначити, чи можуть три слова ***a***, ***b*** та ***c*** утворювати предикативний зв'язок (***a*** → ***b*** → ***c***), потрібно з першої матриці X розкладеного тензору F взяти вектор X_a, що відповідає слову ***a***, з другої матриці Y розкладеного тензору F взяти вектор Y_b, що відповідає слову ***b***, з третьої матриці Z розкладеного тензору F взяти вектор Z_c, що відповідає слову ***c***, та обчислити значення

$$S_{abc} = \sum_{i=1}^{k} X_a[i] * Y_b[i] * Z_c[i].$$

Якщо значення S_{abc} перевищує певний пороговий рівень, то даний зв'язок є визначеним. Всі зв'язки, які не є визначеними, вважаються невизначеними.

Отримані матриці в неявному вигляді задають множину визначених речень мови, що задається текстами вхідного корпусу. Вектори слів з отриманих матриць є неявним описом їх «структурної поведінки» − вони визначають, в які синтаксичні відношення ці слова мають властивість вступати, та з якими словами вони вступають в ці відношення. За допомогою отриманих матриць можливо виконувати синтаксичний аналіз речень з побудовою керуючого простору їх синтаксичних структур, використовуючи висхідні алгоритми аналізу типу Кока–Янгера–Касамі [14–16].

6. Програмна реалізація

Як навчальний текстовий корпус використані статті English Wikipedia та Simple English Wikipedia, а також тексти корпусу The Wall Street Journal. Тексти послідовно обробляються парсером та блоком побудови керуючих просторів їх синтаксичних структур. Спочатку речення аналізуються Стенфордським парсером. Його вихід – дерево виведення речення (parse tree) та дерево підпорядкування (dependency tree). Для побудови керуючих просторів речень був розроблений алгоритм конвертації дерева підпорядкування та дерева виведення речення у керуючий простір синтаксичної структури речення [17]. Алгоритм представляє собою рекурсивний обхід дерева виведення речення зліва – направо з породженням точок керуючого простору в кожному вузлі дерева виведення та з конвертацією відповідних цим вузлам зв'язків з дерева підпорядкування у α-β-зв'язки керуючого простору – предикативні або синтагматичні. За кожною точкою простору закріплюється певне лексико-семантичне значення (слово чи словосполучення) та його характеристики (частина мови, рід, число і т. д.). На початку роботи алгоритму кожне слово представляє собою незв'язану точку керуючого простору. Коли точки A та B з'єднуються і утворюють нову точку простору C, яка представляє α-β-зв'язок між A та B, ця нова точка отримує своє власне лексико-семантичне значення. Це значення може бути наслідуване від головного елемента пари (A, B). Наприклад, у випадку словосполучення *зелений паркан* у парі (*зелений*, *паркан*) головним є іменник, тому нова утворена точка унаслідує значення *паркан*. Або в результаті об'єднання двох точок їх лексичні значення утворюють стале словосполучення, про що можна зробити висновок із наявності даного словосполучення у спеціальній базі – базі назв статей Вікіпедії. Наприклад, якщо об'єднуються точка A із значенням *теорема* та точка B із значенням *Вейєрштрасса* – тоді утворюється стале словосполучення *теорема Вейєрштрасса*, яке стає лексико-семантичним значенням нової утвореної точки C.

Після побудови керуючого простору синтаксичної структури речення для всіх кільцевих синтагматичних α-β-зв'язків у матриці кільцевих зв'язків D нарощується значення $d[I, J]$ (I – індекс першого слова, J – індекс другого слова) $d[I, J] = d[I, J] + 1$. Для всіх трійок лінійних предикативних зв'язків ***A***-α-***B***-β-***C*** у тривимірному тензорі лінійно-предикативних зв'язків F нарощується значення $f[I, J, K]$ (I – індекс слова *A*, J – індекс слова *B*, K – індекс слова *C*). $f[I, J, K] = f[I, J, K] + 1$.

Оброблено 800 тисяч статей англомовної English Wikipedia та Simple English Wikipedia. Також було оброблено корпус статей The Wall Street Journal. За рахунок того, що даний корпус є розміченим вручну і містить коректні синтаксичні структури наявних у корпусі речень, які напряму переконвертовано в керуючі простори, для навчальної вибірки отримано велику кількість керуючих просторів синтаксичних структур високої якості (коректних майже на 100%).

Таким чином, згенеровано велику розріджену матрицю кільцевих зв'язків D (розмір приблизно 2,3 млн. слів × 2,3 млн. слів, біля 57 млн. ненульових елементів) та великий тривимірний тензор лінійно-предикативних зв'язків F (розмір приблизно 2,3 млн. слів × 152 тис. слів × 2.3 млн. слів, близько 78 млн. ненульових елементів). Дані масиви факторизовані за допомогою алгоритму невід'ємної матричної факторизації Лі та Суна та алгоритму паралельної факторизації тривимірного тензору PARAFAC [13]. Алгоритми факторизації реалізовані із застосуванням паралельних обчислень на графічних картах, як у роботах [18, 19].

Факторизовані масиви даних дозволяють елементарно обчислювати значення ймовірності утворення кільцевих синтагматичних зв'язків між двома будь-якими словами за допомогою простого скалярного добутку двох відповідних їм векторів. Також аналогічно просто можна обчислювати значення ймовірності утворення лінійних предикативних зв'язків між трьома будь-якими словами.

На основі отриманих масивів лексико-синтаксичної сполучності реалізований синтаксичний аналізатор, який по реченню англійською мовою напрямку будує керуючий простір його синтаксичної структури. Як базовий метод застосовано алгоритм Кока–Янгера–Касамі.

Запропонована модель містить опис лише тих зв'язків між словами, які фактично мали місце у відповідних реченнях навчального корпусу. Якщо для пари слів A та B кільцевий синтагматичний зв'язок прописаний, так як у навчальних текстах він є присутнім, то для пари A_1 та B_1 (де A_1 – синонім A, B_1 – синонім B) такого зв'язку може і не бути. Для трійки слів A, B, та C, які пов'язані лінійно-предикативним зв'язком, це твердження також має місце. З використанням словників синонімів ця проблема досить легко розв'язується. В розробленій системі у якості такого словника використовується WordNet та його синсети. Система робить припущення, що якщо зв'язок між A та B існує, то він може існувати також між довільною парою A_i та B_i, де A_i –

довільне слово з синсету, який містить A, B_i – довільне слово з синсету, який містить B. Але тут постає проблема омонімії, коли одному слову в WordNet відповідає декілька синсетів, – яким чином визначити пару чи трійку коректних синсетів в кожному конкретному випадку під час синтаксичного аналізу речення.

Існує декілька підходів для розв'язання цієї класичної проблеми неоднозначності слів WSD. Найбільш придатними у даному випадку можуть виявитися методи, що розроблялися спеціально для інтеграції сторінок Wikipedia в якості нових вузлів у WordNet [20–23].

З іншого боку, отримані в результаті невід'ємної факторизації матриці D дві матриці W та H – є потужним інструментарієм для визначення міри семантичної близькості між словами згідно методики латентного семантичного аналізу.

Для розв'язання проблеми неоднозначності слів розроблений наступний алгоритм.

Для визначення наявності кільцевого синтагматичного α-β-зв'язку між ***a*** та ***b***:

(*A*): для того щоб визначити, чи можуть два слова ***a*** та ***b*** утворити кільцевий синтагматичний зв'язок, треба взяти з матриці W вектор-строку $\boldsymbol{W_a}$, що відповідає слову ***a***, з матриці H – вектор-стовпчик $\boldsymbol{H_b}$, що відповідає слову ***b***, та обчислити скалярний добуток векторів ($\boldsymbol{W_a}$, $\boldsymbol{H_b^T}$). Якщо значення ($\boldsymbol{W_a}$, $\boldsymbol{H_b^T}$)$> T$ (де T – пороговий рівень, оптимальне значення якого було визначено експериментальним шляхом), то даний ***α-β***-зв'язок є визначеним. Інакше:

(*B*): за словами ***a*** та ***b*** переходимо до їх синсетів у лексико-семантичній базі WordNet. Отримаємо набір синсетів-вузлів $\{A_i\}$, на які посилається слово ***a***, та набір синсетів-вузлів $\{B_i\}$, на які посилається слово ***b.*** Перевіряємо попарно $\{A_i\}$ та $\{B_i\}$, чи існує якісь значення k та j, що у синсетах A_k та B_j містяться відповідно слова a'_k та b'_j, для яких скалярний добуток векторів $(W_{a'_k}, H^T_{b'_j})> T$. Якщо такі k та j знайдено, то даний зв'язок між ***a*** та ***b*** є визначеним. Інакше:

(*C*): множина $\{A_i\}$ розширяється синсетами, що сполучаються з вузлами $\{A_i\}$ зв'язками гіпонімії та гіпернимії, так само розширяється множина $\{B_i\}$; після цього відбувається перевірка, чи існує для розширених $\{A_i\}_{ext}$ та $\{B_i\}_{ext}$ якісь значення k та j, що у A_k та B_j містяться відповідно слова a'_k та b'_j, для яких скалярний добуток векторів $(W_{a'_k}, H^T_{b'_j})> T$. Перевірка виконується лише для тих пар синсетів, що до того не перевірялися. Якщо такі k та j знайдено, то даний зв'язок між ***a*** та ***b*** є визначеним.

Інакше робимо ще раз розширення множин $\{A_i\}$ та $\{B_i\}$ та пошук таких A_k та B_j, для яких $(W_{a'_k}, H^T_{b'_j})> T$.

Якщо за 2-3 ітерації попереднього кроку не знайдено таких синсетів, то такого зв'язку не існує.

При розширенні $\{A_i\}$ та $\{B_i\}$ треба уникати додавання синсетів із списку концептів найбільш загальних значень з верхньої частини ієрархії WordNet. При залученні подібних понять швидко втрачається смислова близькість між синсетами в наслідуванні властивостей та відношень по зв'язках гіпонімії/гіпернімії.

Для лінійного предикативного α-β-зв'язку даний алгоритм працює аналогічним чином.

Таксономічна ієрархія лексико-семантичної бази WordNet разом із механізмом наслідування дають можливість узагальнення описаної моделі представлення синтаксичних зв'язків та лексико-семантичних відношень мови. Це робить побудовану систему універсальним засобом аналізу синтаксису та семантики природної мови.

7. Експерименти

Ключовим елементом для формування бази синтакичних та лексико-семантичних відношень мови є наявність великого корпусу коректно розмічених текстів. Використання корпусу The Wall Street Journal відчутно вплинуло на якість отриманої моделі. Для отримання розмічених текстів з English Wikipedia та Simple English Wikipedia використовується Стенфордський парсер, який породжує дерева виведення та дерева підпорядкування речень текстів. Оцінка точності побудови дерев виведення – біля 87%. Оцінка точності побудови дерев підпорядкування приблизно дорівнює 84%. Частина некоректно сформованих дерев виведення та дерев підпорядкування, звісно, призводить до формування певної відповідної долі некоректних описів керуючих просторів синтаксичних структур речень. Алгоритм конвертації дерев виведення та дерев підпорядкування у керуючий простір синтаксичних структур речень на коректних входах не виявив власних помилок при побудові відповідних керуючих просторів. При формуванні та факторизації матриці кільцевих зв'язків D та тривимірного тензору лінійно-предикативних зв'язків F не було виявлено ніяких втрат та спотворення первісної інформації.

Після розробки системи синтаксичного аналізу та генерації керуючих просторів для речень природної мови на основі створених лексико-синтаксичних баз проведені експерименти вимірювання коректності побу-

дови керуючих просторів синтаксичних структур. Сформовані тестові виборки – 1500 речень зі статей Simple Wikipedia, 1500 речень зі статей англомовної Wikipedia (з текстів інших, аніж 800 тисяч статей, оброблених для побудови матриці D та тензору F). Тестові текстові набори з Wikipedia та Simple Wikipedia оброблені Стенфордським парсером, їх синтаксичні дерева автоматично трансформовані у керуючі простори розробленим алгоритмом конвертації. Після того отримані керуючи простори вручну перевірені та виправлені за допомогою команди експертів-лінгвістів. Таким чином сформовано анотований тестовий текстовий корпус для перевірки якості роботи системи синтаксичного аналізу та генерації керуючих просторів синтаксичних структур на текстах Simple Wikipedia та English Wikipedia.

Система синтаксичного аналізу та генерації керуючих просторів побудувала керуючі простори синтаксичних структур для речень з анатованого корпусу. Ці побудовані керуючі простори речень співставлені з еталонними керуючими просторами з анотованого тестового корпусу. Перевірка відбувалася автоматично по кожному знайденому кільцевому синтагматичному α-β-зв'язку та по кожному знайденому лінійному предикативному α-β-зв'язку.

Також проведено тестування на текстах корпусу Wall Street Journal методом крос-валідації (коли в процесі перевірки якості роботи системи на окремих частинах корпусу із масивів бази тимчасово вилучалися дані, отримані безпосередньо при обробці цих частин). Перевірка якості роботи системи на корпусі Wall Street Journal відбувалася в автоматичному режимі.

Перевірка здійснювалася із врахуванням алгоритмічного випадку, в якому знайдено той чи інший синтаксичний зв'язок (випадок ***A*** – пряме знаходження імовірності наявності зв'язку між словами через скалярний добуток векторів слів, випадок ***B*** – підключення синонімів даних слів для перевірки імовірності наявності зв'язку, випадок ***C*** – підключення гіпонимів та гіперонимів даних слів для перевірки імовірності наявності зв'язку). Перевірка здійснювалася лише для речень, які успішно оброблені системою з повною побудовою керуючого простору їх синтаксичної структури (успішно опрацьовано 82,9 % з 3000 речень тестового набору текстів, та 96,2% на текстах Wall Street Journal). Результати тестування представлені у табл. 1 та 2.

Таблиця 1. Оцінки точності визначення кільцевих синтагматичних α-β-відношень на корпусах текстів статей Simple Wikipedia, Wikipedia та Wall Street Journal

	Simple Wikipedia	Wikipedia	WSJ corpus
Випадок A	96,88 %	94,48 %	95,23 %
Випадок B	94,62 %	91,89 %	92,80 %
Випадок C	92,21 %	85,71 %	86,58 %

Таблиця 2. Оцінки точності визначення лінійних предикативних α-β-відношень на корпусах текстів статей Simple Wikipedia, Wikipedia та Wall Street Journal

	Simple Wikipedia	Wikipedia	WSJ corpus
Випадок A	97,38 %	95,41 %	96,12 %
Випадок B	95,21 %	92,29 %	93,72 %
Випадок C	94,11 %	88,71 %	91,59 %

Треба зазначити, що оцінки точності визначення лінійних предикативних α-β-відношень є вищими за оцінки точності визначення кільцевих синтагматичних α-β-відношень. Це виглядає природним з точки зору відносної позиційної стійкості відношень типу *підмет-присудок-додаток* у структурі речень. Певний незначний відсоток помилок, що присутній навіть у найпростішому випадку A, свідчить про наявність помилок у навчальному масиві керуючих просторів речень, на основі якого складалися матриця кільцевих зв'язків D та тривимірний тензор лінійно-предикативних зв'язків F. Додатковою перевіркою та виправленням даних навчального масиву можна вдосконалити побудовану модель. Найкращі показники відповідають оцінкам роботи системи на реченнях Simple Wikipedia, що є цілком зрозумілим через просту та чітку синтаксичну структуру речень у Simple Wikipedia. Речення English Wikipedia по структурі є набагато складнішими і через те виникає значно більше можливостей для різноманітних інтерпретацій граматичних структур. Обробка речень із корпусу The Wall Street Journal за оцінками точності переважає результати роботи системи на реченнях English Wikipedia, що свідчить про те, що якісні навчальні дані з розміченого корпусу The Wall Street Journal призвели до відчутного покращення моделі.

Висновки

Рекурсивність керуючих просторів синтаксичних структур природної мови дозволяє точно виразити структуру речень довільної складності та довжини. Це дає можливість при розробці семантико-синтаксичної **тензорної** моделі природної мови замість нарощування мірності лінгвістичних масивів сполучності лексичних одиниць обмежитись лише побудовою одного тривимірного тензору та однієї матриці. Розроблена на основі факторизованих масивів система аналізу та побудови керуючих просторів синтаксичних структур речень під час тестування продемонструвала високу якість та точність роботи, що доводить коректність та ефективність запропонованої моделі. Із цього випливає її актуальність як в теоретичному плані, так і в аспекті застосування на практиці в прикладних лінгвістичних системах.

1. *Deerwester S., Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman.* Indexing by Latent Semantic Analysis. // In Journal of the American Society for Information Science. – 1990. – P. 391–407.
2. *Tim Van de Cruys.* A Non-negative Tensor Factorization Model for Selectional Preference Induction // In Journal of Natural Language Engineering. – 2010. 16(4):417–437.
3. *Tim Van de Cruys, Laura Rimell, Thierry Poibeau, and Anna Korhonen* Multi-way Tensor Factorization for Unsupervised Lexical Acquisition // In Proceedings of COLING – 2012. – P. 2703–2720.
4. *Cohen S.B., Michael Collins.* Tensor Decomposition for Fast Parsing with Latent-Variable PCFGs // In NIPS. – 2012. – P. 2528–2536.
5. *Peng Wei, Li Tao.* On the equivalence between nonnegative tensor factorization and tensorial probabilistic latent semantic analysis // Applied Intelligence, Springer Journals. – 2011. October, Vol. 35, Issue 2, P. 285–295
6. *Anisimov A.V.* Control space of syntactic structures of natural language // Cybernetics. – 1990. – N 3, P. 11–17.
7. *Miller G.A., Beckwith R., Fellbaum C.D., Gross D., Miller K.* WordNet: An online lexical database // Int. J. Lexicograph. – 1990. – 3, 4. – P. 235–244.
8. *Нариньяни А.С.* Формальная модель: общая схема и выбор адекватных средств. Препр. № 400/ВЦ СО АН СССР. – Новосибирск, 1978. – 19 с.
9. *Гладкий А.В.* Синтаксические структуры естественного языка в автоматизированных системах общения. – М.: Наука, 1985. – 144 с.
10. *Klein D. and Manning C.D.* Accurate Unlexicalized Parsing // In Proceedings of ACL. – 2003. – P. 423–430.
11. *Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning.* Generating Typed Dependency Parses from Phrase Structure Parses // In Proceedings of LREC. – 2006.
12. *Lee D.D. and Seung H.S.* Algorithms for Non-Negative Matrix Factorization // In Proceedings of NIPS. – 2000.– P. 556–562
13. *Cichocki A., Zdunek R., Phan A.-H., Amari S.-I.* Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation // J. Wiley & Sons, Chichester. – 2009.
14. *Kasami T.* An efficient recognition and syntax-analysis algorithm for context-free languages // Scientific report AFCRL-65-758, Air Force Cambridge Research Lab, Bedford, MA. –1965.
15. *Cocke J. and Jacob T.* Schwartz Programming languages and their compilers: Preliminary notes // Technical report, Courant Institute of Mathematical Sciences, New York University, 1970
16. *Younger D.H.* Recognition and parsing of context-free languages in time n3 // In Information and Control – 1967. 10(2). – P. 189–208.
17. *Марченко О.О.* Алгоритм конвертації дерева залежностей у керуючий простір синтаксичної структури речення // Вісник Київського національного університету імені Тараса Шевченка. Серія: фізико-математичні науки. – 2013. – № 5.
18. *Antikainen J., Havel J., Josth R., Herout A., Zemcík P., Hauta-Kasari M., Zemcík P.* Nonnegative Tensor Factorization Accelerated Using GPGPU // In TPDS. – 2011. – P. 1135–1141.
19. *Kysenko V., Rupp K., Marchenko O., Selberherr S., Anisimov A.* GPU-Accelerated Non-negative Matrix Factorization for Text Mining // In Lecture Notes in Computer Science. – 2012. – Vol. 7337. – P. 158–163.
20. Ponzetto S.P., Navigli R. Knowledge-rich Word Sense Disambiguation rivaling supervised systems // In Proceedings of ACL. – 2010. – P. 1522–1531.
21. *Ponzetto S.P., Navigli R.* Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia // In *Proceedings of IJCAI*. – 2009. – P. 2083–2088.
22. *Ponzetto S.P., Navigli R.* BabelNet: Building a Very Large Multilingual Semantic Network // In Proceedings of ACL. – 2010. – P. 216–225.
23. *Ruiz-Casado M.* Enrique Alfonseca and Pablo Castells // Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In Proceedings of AWIC. – 2005.