Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)
(рівень вищої освіти)

Дослідження методів семантичного аналізу текстів для інтелектуалізації Web-сайтів.
(тема)

Виконав:
студент 2 курсу, групи СШІмзд-18-1
Войтенко Я.С.
(прізвище, ініціали)
Спеціальність 122 – Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо -наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва освітньої програми)
Керівник проф. Удовенко С.Г.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
Кафедра Штучного інтелекту
Рівень вищої освіти другий (магістерський)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо -наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові Войтенку Ярославу Станіславовичу
(прізвище, ім'я, по батькові)

1. Тема роботи «Дослідження методів семантичного аналізу текстів для інтелектуалізації Web-сайтів»
затверджена наказом по університету від 30.03.2020 р. № 44Стз
2. Термін подання студентом роботи до екзаменаційної комісії 20 травня 2020 р.
3. Вихідні дані до роботи Науково-технічні публікації, дані Інтернет-джерел та відомих наукових проектів щодо розробки та дослідження семантичного аналізу текстів
4. Перелік питань, що потрібно опрацювати в роботі Аналіз науково-технічної літератури з питань обробки природномовних текстів, вивчення метрик міжнародних рейтингів, аналіз контенту сайту університету, проведення експериментального моделювання та навчання моделі, розробка тематичної моделі сайту університету

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) _____
_____
_____
_____
_____
_____
_____
_____
_____
_____

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Аналіз предметної галузі	проф. Удовенко С.Г.		12.04.2020
Методи і алгоритми обробки текстів	проф. Удовенко С.Г.		20.04.2020
Експериментальне моделювання	проф. Удовенко С.Г.		30.04.2020
Проектування системи помічника університету	проф. Удовенко С.Г.		08.05.2020

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Аналіз предметної галузі	11.04.2020	виконано
2	Аналіз методів обробки текстових документів	20.04.2020	виконано
3	Експериментальне моделювання	25.04.2020	виконано
4	Проектування системи помічника університету	08.05.2020	виконано
5	Написання пояснювальної записки	14.05.2020	виконано
6	Попередній захист	15.05.2020	виконано
7	Захист перед ЕК	20.05.2020	

Дата видачі завдання ____30 березня_2020 р.

Студент _____
(підпис)

Керівник роботи _____ _____проф. Удовенко С.Г._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Записка пояснювальна: 79 с., 20 рис., 53 джерела.

СЕМАНТИЧНИЙ МЕТОД, ШТУЧНИЙ ІНТЕЛЕКТ, ЛАТЕНТНО-СЕМАНТИЧНИЙ АНАЛІЗ, КОНТЕНТ-АНАЛІЗ, НЕЙРОННІ МЕРЕЖІ, ІНТЕРНЕТ, МАШИННИЙ ПЕРЕКЛАД, WEB-САЙТ

Об’єкт дослідження – методи семантичного аналізу даних.

Предмет дослідження – методи семантичного аналізу природномовних текстів.

Мета роботи – дослідження методів семантичного аналізу текстів для інтелектуалізації Web-сайтів. Згідно з метою роботи необхідно виконати такі завдання, як аналіз існуючих наукових джерел за темою дослідження, огляд існуючих систем семантичного аналізу, аналіз алгоритмів семантичного аналізу, дослідження проблем в сучасних методах інтелектуалізації Web-сайтів, дослідження найперспективніших шляхів використання семантичного аналізу для інтелектуалізації сайтів, висновки та рекомендації з приводу опрацьованого матеріалу.

Методи дослідження – методи семантичного аналізу природномовних текстів, латентно-семантичний аналіз, методи лексичного аналізу природномовних текстів, методи семантико-синтаксичного аналізу природномовних текстів.

РЕФЕРАТ

Пояснительная записка: 79 с., 20 рис., 53 источника.

СЕМАНТИЧЕСКИЙ МЕТОД, ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ, ЛАТЕНТНО-СЕМАНТИЧЕСКИЙ АНАЛИЗ, КОНТЕНТ-АНАЛИЗ, НЕЙРОННЫЕ СЕТВ, ИНТЕРНЕТ, МАШИННЫЙ ПЕРЕВОД, WEB-САЙТ

Объект исследования – методы семантического анализа данных.

Предмет исследования – методы семантического анализа естественно-языковых текстов.

Цель работы – исследование методов семантического анализа текстов для интеллектуализации Web-сайтов. Согласно с целью работы необходимо выполнить такие задачи, как анализ существующих научных источников по теме исследования, обзор существующих систем семантического анализа, анализ алгоритмов семантического анализа, исследования проблем в современных методах интеллектуализации Web-сайтов, исследования перспективных путей использования семантического анализа для интеллектуализации сайтов, выводы и рекомендации по поводу обработанного материала.

Методы исследования – методы семантического анализа естественно-языковых текстов, латентно-семантический анализ, методы лексического анализа естественно-языковых текстов, методы семантико-синтаксического анализа естественно-языковых текстов.

ABSTRACT

Explanatory note: 79 p., 20 fig., 53 sources.

SEMANTIC METHOD, ARTIFICIAL INTELLIGENCE, LATENT SEMANTIC ANALYSIS, CONTENT ANALYSIS, NEURAL NETWORKS, INTERNET, MACHINE TRANSLATION, WEBSITE

The object of the study is the methods of semantic data analysis.

The subject of the study is the methods of thematic modeling of natural-text cases.

The purpose of the work is to study methods of semantic analysis of texts to intelligentize Web sites. According to the purpose of the work, it is necessary to perform such tasks as analysis of existing scientific sources on the topic of research, review of existing systems of semantic analysis, analysis of semantic analysis algorithms, research of problems in modern methods of intellectualization of Web sites, research of promising ways of using semantic analysis to intelligent sites, conclusions and recommendations about processed material.

Research methods – methods of semantic analysis of natural language texts, latent-semantic analysis, methods of lexical analysis of natural language texts, methods of semantic-syntax analysis of natural language texts.

ЗМІСТ

ВСТУП

Лінгвістична обробка природномовних текстів є однією з центральних проблем інтелектуалізації інформаційних технологій. Цій проблемі приділяється значна увага в розвинутих країнах світу, доказом чого є задіяння вагомих ресурсів на розробку лінгвістичного програмного забезпечення. У зв'язку з бурхливим розвитком Інтернету та інших комп'ютерно-комунікаційних технологій ця проблема набуває ще більшої популярності та значимості у світі технологій.

Тема семантичного аналізу набула популярності ще в минулому столітті. Значні зусилля науковців були спрямовані на розробку математичних алгоритмів та комп'ютерних програм для обробки текстів природною мовою. Були створені різноманітні алгоритми та структури представлення даних для автоматизації аналізу та синтезу текстів. Традиційно аналіз текстових документів поділявся на послідовність процесів морфологічного, синтаксичного та семантичного аналізів. Були розроблені моделі та алгоритми, які відповідають кожному з цих етапів. Зокрема для семантичного аналізу тексту були запрпоновані та розвинуті класичні семантичні мережі та фреймові моделі Мінського, для синтаксису речення – граматики Хомського, системні граматики Холідея, дерева підпорядкування та системи складових Гладкого, розширенні мережі переходів; для морфологічного аналізу розроблено багато різних моделей, орієнтованих на конкретні групи мов.

Найбільшу складність обробки природномовних текстів викликають явища полісемії та омонімії, які привносять у мовні структури неоднозначність і значно ускладнюють задачу коректного відображення семантично-синтаксичної структури тексту в його формальне логічне представлення. Всі ці проблеми вирішуються на рівні семантичного аналізу.

З іншого боку, функції логікосемантичного аналізу призводять до того, що програми обробки тексту є складними та повільно працюють. Людина в

реальному процесі розуміння тексту не так часто застосовує логіку – лише по мірі виникнення логічних задач, а в решті випадків відбувається застосування інших механізмів інтелектуального аналізу (більш автоматичних), у першу чергу – пошук за асоціацією, за формою чи контекстом.

Метою даної атестаційної роботи є дослідження методів семантичного аналізу текстів для інтелектуалізації Web-сайтів.

Згідно з метою роботи були поставлені наступні задачі:

- аналіз існуючих наукових джерел за темою дослідження;
- огляд існуючих систем семантичного аналізу;
- аналіз алгоритмів семантичного аналізу;
- дослідження проблем в сучасних методах інтелектуалізації Web-сайтів;
- дослідження найперспективніших шляхів використання семантичного аналізу для інтелектуалізації сайтів;
- висновки та рекомендації з приводу опрацьованого матеріалу.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Огляд методів семантичного аналізу текстів

Крім знань про структуру мови, семантика тісно пов'язана з філософією, психологією і іншими науками, так як неминуче стосується питань про походження значень слів, їх ставлення до буття і мислення. При семантичному аналізі необхідно враховувати соціальні та культурні особливості носія мови. Процес людського мислення, як і мова, яка являється інструментом вираження думок, дуже гнучкий та важко піддається формалізації. Тому семантичний аналіз по праву вважається найскладнішим етапом автоматичної обробки текстів.

На даний момент існує чимало методів аналізу сенсу висловлювань, проте жоден з них не є універсальним. Над співвідношенням сенсу тексту працювали багато науковців. Так, І.А. Мельчук в роботі [1] ввів поняття лексичних функцій, розвинув поняття синтаксичних і семантичних валентностей і розглянув їх в контексті толково-комбінаторного словника, який представляє мовну модель. Він показав, що значення слів співвідносяться безпосередньо не з навколишньою дійсністю, а з уявленнями носія мови про цю дійсність. В.Ш. Рубашкін і Д.Г. Лахути ввели ієрархію синтаксичних зв'язків для більш ефективної роботи семантичного аналізатора [2]. Найважливішими є обов'язкові рольові зв'язки, далі йдуть зв'язки кореференції, потім факультативні рольові зв'язки і тільки потім предметно-асоціативні.

Відомий лінгвіст Є.В. Падучева пропонує розглядати тематичні класи слів, зокрема дієслів, оскільки вони несуть основне смислове навантаження [3]. Суттєвою в даному підході є ідея поділу понять мови на деякі семантичні

групи з урахуванням того, що ці поняття мають певний нетривіальний загальний смисловий компонент. Елементи таких груп схильні мати один і той же набір залежних понять.

Універсальна мова представлення знань має бути зручною для здійснення виведення нових знань із уже наявних, а значить, необхідно створити апарат для перевірки правильності висловлювань. Для цього доцільно використовувати логічні моделі подання знань. Наприклад, семантична мова містить у собі формалізми логіки предикатів, де присутні атомарні поняття, функції над цими поняттями і правила виводу, за допомогою яких можна описувати нові поняття. Не виключено, що в напрямку створення подібних семантичних мов буде розвиватися наукова думка в майбутньому.

1.1.1 Концептуальний і прецедентний аналіз

На етапі морфологічного і семантико-синтаксичного аналізу текстів основними одиницями, які позначають поняття, є слова. При такому підході вважається, що сенс словосполучень і фраз може бути виражений через сенс присутніх в них слів. Такий підхід базується на припущенні, що словосполучення, які зустрічаються в текстах, можна розділити на вільні і невільні. Інший підхід ґрунтується на тому, що неподільними одиницями сенсу є категорії і поняття, що складаються не з самостійних слів, а з словосполучень [4]. Такі категорії і поняття називаються концептами. Ідея концептуального аналізу як невід'ємної складової семантичного аналізу зустрічається в дослідженнях [2, 5, 6]. Розглянемо завдання, що доцільно вирішувати засобами концептуального семантичного аналізу.

З точки зору використовуваних методів і засобів семантичний аналіз повинен передбачати два етапи: етап інтерпретації граматично виражених (синтаксичних і анафоричних) зв'язків і етап розпізнавання зв'язків, які не

мають граматичного вираження. Неоднозначності повинні вирішуватися самим процесом аналізу за критерієм ступеня смислової задовільності одержуваного в будь-якому вигляді результату. Ключовим пунктом системи семантичного аналізу є ефективна словникова підтримка. У цьому сенсі будь-яка система семантичного аналізу є тезаурусно орієнтована. Процедури семантичного аналізу в усіх без винятку випадках спираються на функціональність понятійного словника.

Словник для підтримки семантичного аналізу має оперувати смислами і, отже, описувати властивості і відносини понять, а не слів, тому його можна назвати концептуальним словником [2]. У певному сенсі роль концептуального словника можуть виконувати семантичні мережі. У семантичному інтерпретаторі насамперед слід уточнювати, як розрізняються типи семантичних відносин в тексті: рольові (зв'язки з валентністю предиката), предметно-асоціативні (відносини між об'єктами, процесами, що є значимі в предметної області) тощо. Приймаються наступні основні правила інтерпретації синтаксичних зв'язків:

- тип встановлюваного семантичного відношення визначається семантичними класами і в певних випадках більш детальними семантичними характеристиками синтаксичного «господаря» і «слуги»;

- прийменник розглядається не як самостійний об'єкт інтерпретації, а як додаткова (семантико-граматична) характеристика зв'язку між синтаксичним «господарем» прийменника і керованим ним словом;

- для вирішення лексичної і синтаксичної омонімії, що фіксується синтаксичним аналізатором, семантичний інтерпретатор використовує систему емпірично встановлюваних переваг. На рівні типів семантичних відносин встановлюється наступний порядок переваг (відповідає зменшенню пріоритету зв'язку): функціональні зв'язки і зв'язки, що встановлюють факт смисловий надмірності; рольові зв'язки, які визначаються як обов'язкові, при наявності семантично узгодженого актанта; зв'язки кореференції; рольові зв'язки, які визначаються як факультативні; специфіковані предметно-

асоціативні зв'язки; неспецифіковані предметно-асоціативні зв'язки.

Все більшого значення набуває аналіз «за зразком» (прецедентний аналіз), заснований на використанні корпусу попередньо розмічених текстів [7]. Система аналізу повинна забезпечувати не тільки отримання знань з конкретного тексту, а й накопичення результатів як на синтаксичному, так і на семантичному рівнях для використання їх далі в якості прецедентів.

Суттєвою в даному підході є ідея поділу понять мови на деякі семантичні групи з урахуванням того, що ці поняття мають певний нетривіальний загальний смисловий компонент. Елементи таких груп схильні мати один і той же набір залежних понять. В такому випадку словник для підтримки семантичного аналізу має оперувати смислами і, отже, описувати властивості і відносини понять, а не слів. Залишається питання, як правильно структурувати і представляти інформацію в подібних словниках, щоб пошук по ним був зручним і швидким, а крім того, можна було б враховувати зміни в природній мові (зникнення старих і виникнення нових понять).

1.1.2 Латентно-семантичний аналіз (ЛСА)

Латентно-семантичний аналіз (ЛСА) – це метод обробки інформації на природній мові, що аналізує взаємозв'язок між колекцією документів і термінами, які в них зустрічаються, для зіставлення деяких тематичних чинників до всіх документів і термінів.

В основі методу латентно-семантичного аналізу лежать принципи факторного аналізу, зокрема виявлення латентних зв'язків досліджуваних явищ або об'єктів. При класифікації / кластеризації документів цей метод використовується для отримання контекстно-залежних значень лексичних одиниць за допомогою статистичної обробки великих корпусів текстів.

Програма LSA (спочатку відома як «Латентние семантичне індексування» («Latent Semantic Indexing», LSI)) розроблялася для вирішення завдань пошуку і вилучення інформації (information retrieval) і являє собою виділення з великої бази даних невеликої кількості документів, релевантних заданому запиту. Попередні підходи до вирішення цього завдання включали в себе пошук за ключовими словами (keyword-matching), оцінювання ваги цих ключових слів і побудову векторної основи, яка зображує наявність слів в документах. LSA поширив векторну основу на декомпозицію на сингулярні значення перебудови бази (Singular Value Decomposition (SVD)).

Загальний алгоритм роботи LSA полягає в наступному:

- збір великого масиву релевантного тексту та поділ його на «документи». У більшості випадків кожен параграф обробляється як окремий документ. Такий підхід заснований на тому, що інформація всередині параграфа має тенденцію бути логічно пов'язаною (когерентною) і послідовною;

- створення суміжної матриці документів і термів. Клітка в цій матриці відповідає документу і терму і містить кількість випадків, коли зустрічається в . Терм визначається як слово, яке зустрічається більш ніж в одному документі, і при морфологічному пошуку або іншому морфологічному аналізі не дає спроби комбінувати різні форми того ж слова. Якщо є термів і документів, то така матриця може бути розглянута як репрезентація, в якій існують -мірний вектор для кожного документа і -мірний вектор для кожного терма;

- зменшення значення кожної клітини за рахунок ефекту від загальних слів, які зустрічаються у всьому корпусі (тобто від слів, які найбільш часто зустрічаються в загальному масиві текстової інформації). Метод загального зважування – «логарифм ентропії» – базується на теорії інформації (Information Theory), в якій значення підвищується при отриманні інформації;

- SVD за допомогою параметра точно визначає бажане число вимірювань. (В принципі, SVD розраховується з усіма вимірами і створює

три матриці, які при перемножуванні дають вихідні дані, але відповідну кількість пам'яті, яка потрібна для такої операції, занадто велика. Тому для вирішення даного завдання використовують алгоритми, оптимізовані для розрідженого простору даних, і підраховують тільки найбільш значущі вимірювання матриць.) Результат описаного вище процесу – три матриці. Одна має -мірний вектор для кожного документа, інша – -мірний вектор для кожного терма в корпусі, третя – -сингулярні значення. Перші дві матриці визначають два різних векторних простору, які також відрізняються від простору, що визначається вихідної матрицею.

Таким чином, кожен терм і документ представляються за допомогою векторів в загальному просторі розмірності *k* (так званому просторі гіпотез). Близькість між будь-якою комбінацією термів і / або документів легко обчислюється за допомогою скалярного добутку векторів.

Зазвичай, вибір *k* залежить від поставленого завдання і здійснюється емпірично. Якщо вибране значення *k* занадто велике, то метод втрачає свою потужність і наближається за характеристиками до стандартних векторних методів. Занадто мале значення *k* не дозволяє вловлювати відмінності між схожими термами або документами [8].

1.2 Використання методів семантичного аналізу текстів в мережі Інтернет

Зазвичай, під терміном «семантичний аналіз» в лінгвістичних системах мається на увазі певне перетворення структури вхідного тексту у внутрішню модель представлення даних певної системи. Такі внутрішні моделі мають назви «семантична мережа», «семантичний граф», «семантичне представлення» тощо.

Подальша робота типових семантичних алгоритмів базується на порівнянні «семантичної структури» нових документів зі структурою, що отримана системою під час навчання і знаходиться в її базі знань[9].

Зрозуміло, що даний підхід не можна вважати строго семантичним та націленим на розумну, інтелектуальну обробку тексту, коли автоматизована система насправді розуміє контекст, а не проводить порівняння двох структур, побудованих на базі статистично зібраної інформації про документи. Незважаючи на певну обмеженість згаданих методів, проведемо огляд найбільш практично успішних з них.

1.2.1 Огляд існуючих систем семантичного аналізу текстів та їхнього впливу на Інтернет

Первинний семантичний аналіз. В роботі [10] зазначено, що назва «семантичний» є умовною, адже так називається метод, в якому використовується валентна структура, описана в словнику РОСС [11]. Семантичний аналіз будує семантичну структуру речень російською мовою. Семантична структура складається із семантичних вузлів і семантичних відносин. Семантичний вузол – це такий об’єкт текстової семантики, в якого заповнені всі валентності: як експліцитно виражені в тексті, так і імпліцитно (тобто ті, які виходять із екстралінгвістичних джерел). Семантичне відношення – це універсальний зв’язок, який визначається носієм мови у тексті. Цей зв’язок бінарний, тобто він йде від одного семантичного вузла до іншого. Семантичні вузли утворюються зі слів вихідного речення. Головне джерело гіпотез про склад семантичного вузла дає синтаксичний аналіз. Більшість синтаксичних груп можуть перейти в семантичні вузли, деякі повинні перетворитися в атрибути вузлів. Крім самого тексту, джерелами гіпотез виступають словник тимчасових груп, словник РОСС та інші тезауруси [11].

Синтактико-семантичний підхід. В основі підходу лежить лінгвістична модель, відповідно до якої основу семантичної структури висловлювання представляє так званий пропозиційний компонент плану змісту. Цей компонент відбиває позамовну ситуацію, що описується реченням, і

характеризує його об’єктивний зміст, на відміну від інших компонентів (модального, комунікативного), які так чи інакше характеризують відношення до ситуації або співвідносять ситуацію з якимось моментом часу або умовами її реалізації, і тому відносяться до сфери суб’єктивного. Таким чином, синтактико-семантичний підхід до отримання знань припускає виділення зі структури фрази її семантичного ядра – об’єктивного опису ситуації, і абстрагування від несуттєвих, суб’єктивних компонентів плану змісту. З цією метою використається синтаксичний аналізатор тексту, що працює на підставі знання загальних правил граматики мови, а також словник моделей керування, що описує для кожного предиката способи вираження в мові його аргументів (прийменників та відмінків актантів) [12].

Семантичний аналіз в системі «Мінерва». Виконується у вигляді перетворення графа речення (отриманого після роботи синтаксичного аналізатора) у вирази на внутрішній мові «Мінерва». Функціонує на базі аналізатора та бібліотеки шаблонів синтаксичних конструкцій російської мови, для яких уже створено опис формальною мовою подання знань [13].

Семантика в пошуковій машині. В даному випадку [14] семантичний аналіз тексту має своєю метою витяг змісту з тексту та відображення його у формальну модель, що дозволяє знаходити змістовну близькість двох текстів (для задачі пошуку – близькість запиту та документу). При семантичному аналізі тексту множина синтаксем кожного речення відображається в неоднорідну семантичну мережу, запропоновану Г.С. Осиповим, з синтаксемами у вершинах та семантичними зв’язками як ребрами.

Семантичний аналіз тексту оперує в основному іменними синтаксемами, які виділяються в результаті морфологічного та синтаксичного аналізу. Іменна синтаксема представляється в тексті іменною або прийменниковою групою – словосполученням з іменником або прийменником як керуючим словом. Іменна синтаксема характеризується морфологічною формою – прийменником, відмінком і категоріально-семантичним класом іменника, від якого вона утворена. Морфологічна форма

синтаксеми й категоріально-семантичний клас визначаються за допомогою лінгвістичного аналізатора тексту. Синтаксема характеризується також синтаксичною функцією, що вона може виконувати в реченні, і синтаксичним значенням. У ході семантичного аналізу тексту необхідно встановити значення іменних синтаксем, які є носіями змісту тексту. Морфологічна форма та категоріально-семантичний клас іменної синтаксеми не однозначно задають її значення, тому для вирішення неоднозначностей використовується контекст – дієслово або віддієслівний іменник, при якому іменна синтаксема входить в речення. Для розбору таких випадків використовується спеціальний словник, що описує найбільш часті сполучення певного дієслова з можливими синтаксемами.

Проект «Відкрите пізнання» (Open Cognition). У рамках проекту «Відкрите пізнання» використовується аналізатор Link Grammar Parser, який відповідає за обробку естетичної мови [15]. Цей аналізатор почали розроблювати у 1990-х рр. в університеті Карнеги – Меллона [16]. Даний підхід відмінний від класичної теорії синтаксису. Система визначає запропоновану синтаксичну структуру, яка складається з безлічі відмічених зв’язків (коннекторів), з'єднуючих пари слів. Link Grammar Parser використовує інформацію про тип пам'яті між різними словами. На даний момент він підтримує словники для іврита, англійської, німецької, російської, турецької, перської, арабської, латиської і в'єтнамської мов. Головною причиною, за якою аналізатор називають семантичною системою, можна вважати унікальний по повноті набір зв'язків (близько 100 основних, причому деякі з них мають 3-4 варіанти). У деяких випадках ретельна робота над різними контекстами привела авторів системи до переходу до майже семантичних класифікацій, побудованим на синтаксичних засадах. Проект Open Cognition, в рамках якого розвивається Link Grammar Parser, відкритий і безкоштовний, що є великою перевагою для проведення досліджень. Досить докладний опис і вихідний код можна знайти на сайті [17].

Open Cognition продовжує розвиватися, що також важливо, оскільки є можливість взаємодіяти з розробниками. Нарівні з Link Grammar ведеться розробка аналізатора RelEx, який дозволяє витягувати відносини семантичної залежності у висловлюваннях на природній мові і в результаті представляти пропозиції у вигляді дерев залежностей [18]. Він використовує кілька наборів правил для перестроювання графа з урахуванням синтаксичних зв'язків між словами. Після кожного кроку, згідно набору правил зіставлення, в отриманому графі додаються теги структурних характеристик і відносин між словами. Однак деякі правила, навпаки, можуть скорочувати граф. Таким чином відбувається перетворення графа. Цей процес застосування послідовності правил нагадує метод, який використовується в обмежувальних граматиках. Головна його відмінність полягає в тому, що RelEx працює з графовим поданням, а не з простими наборами тегів (позначають відносини). Ця особливість дозволяє застосовувати більш абстрактні перетворення при аналізі текстів. Іншими словами, основна ідея полягає в тому, щоб використовувати розпізнавання образів для перетворення графів. На відміну від інших аналізаторів, які повністю спираються на синтаксичну структуру пропозиції, RelEx більше орієнтований на представлення семантики, зокрема, це стосується сутностей, порівнянь, питань, вирішення анафор і лексичної багатозначності слів.

Система «Діалінг». Це автоматична система російсько-англійського перекладу розроблялася в 1999-2002 рр. в рамках проекту «Автоматична обробка тексту». В різний час в роботі над нею брали участь двадцять два фахівці, більшість з яких відомі вчені-лінгвісти. За основу системи «Діалінг» була взята система французько-російського автоматичного перекладу, розроблена в ВЦП спільно з МГПИИЯ ім. М. Тореза в 1976-1986 рр., і система аналізу політичних текстів російською мовою «Політекст», розроблена в Центрі інформаційних досліджень в 1991-1997 рр.

Система «Політекст» була спрямована на аналіз офіційних документів російською мовою і містила повний ланцюжок аналізаторів тексту:

графематичний, морфологічний, синтаксичний і частково семантичний. В системі «Діалінг» був частково запозичений графематичний аналіз, адаптований під нові стандарти програмування. Програма морфологічного аналізу була написана заново, оскільки швидкість роботи була низькою, але сам морфологічний апарат не змінився [19]. На графематичному рівні константами є графематичні дескриптори: ЛЕ (лексема) – присвоюється послідовностям, що складається з кириличних символів; ЦК (цифровий комплекс) – присвоюється послідовностям, що складається з цифр, та т.д. На морфологічному рівні для позначень використовуються граммеми: ов – орудний відмінок, мн – множина, нп – неживий предмет, дв – досконалий вид, пд – перехідність дієслова і т.д. Можливі типи фрагментів на етапі фрагментаційного аналізу: головні речення, підрядні речення в складі складного, дієприкметникові, дієприслівникові та інші відокремлені звороти. Про кожен фрагмент відомо, які фрагменти в нього безпосередньо вкладені і в які він безпосередньо вкладений. Основними складовими застосовуваного в «Діалінгі» семантичного апарату є семантичні відносини і семантичні характеристики. Приклади семантичних відносин: ИНСТР – «інструмент», ЛОК – «локація, місце розташування», ПРИНАДЛ – «приналежність» та ін. Вони досить універсальні і мають схожість з предикатами і семантичними ролями. Семантичні характеристики дозволяють будувати формули з використанням логічних зв'язок «і» та «чи». Кожному слову приписується деяка формула, складена з семантичних характеристик. У семантичному словнику «Діалінга» міститься близько 40 семантичних характеристик. Приклади семантичних характеристик: ГЕОГР – географічний об'єкт; ДВИЖ – дієслова руху; ІНТЕЛ – дії, пов'язані з розумовою діяльністю; НОСІНФ – носії інформації; ЭМОЦ – прикметники, які виражають емоції, і т.д. Деякі характеристики є складовими, так як їх можна виразити через інші. Семантичні характеристики нарівні з граматичними характеристиками забезпечують перевірку узгодження слів при інтерпретації зв'язків в тексті.

В даний момент всі інструменти, розроблені в рамках проекту «Автоматична обробка тексту» (в тому числі система «Діалінг»), є вільним кросплатформним програмним забезпеченням (ПЗ).

Існують і інші системи, що містять компоненти семантичного аналізу. Однак вони мають істотні недоліки для досліджень: складно знайти опис, не є безкоштовними і вільно поширюваними або не працюють з текстами російською та українською мовами. До таких систем відносяться OpenCalais[20], RCO [21], Abbyy Compreno [22], SemSin[23], DictaScope [24] тощо.

Слід згадати систему вилучення даних з неструктурованих текстів Pullenti [25]. Вона посіла перше місце на доріжках T1, T2, T2-m і друге місце на T1-l на конференції «Діалог 2016» в змаганні Fact-RuEval. На сайті розробників системи Pullenti є також демоверсія семантичного аналізатора, що дозволяє за пропозицією будувати семантичну мережу. Інструментальне середовище «Декла» розроблено в кінці 90-х років і використано для побудови експертних систем, оболонок для експертних систем, логіко-аналітичних систем, лінгвістичних процесорів, що забезпечують обробку і автоматичне вилучення знань з потоків неформалізованих документів на природній мові [26].

Система машинного перекладу «ЕТАП-3» призначена для аналізу і перекладу текстів російською та англійською мовами. Система використовує перетворення текстів на природній мові в їх семантичне подання на мові Universal Networking Language. Як вже говорилося раніше, розмітка синтаксичного корпусу «Національний корпус російської мови» виконується лінгвістичним процесором ЕТАП-3, заснованим на принципах теорії «Сенс ⇔ Текст» [27].

Останнім часом з'являється все більше систем уявлення баз знань у вигляді графів. Оскільки обсяги інформації постійно збільшуються з неймовірною швидкістю, такі системи повинні підтримувати побудову та поповнення баз знань в автоматичному режимі. Автоматична побудова баз

знань може здійснюватися на основі структурованих джерел даних. Прикладами таких систем є Yago [28] , DBpedia [29], Freebase [30], Google's Knowledge Graph [31], OpenCyc [32].

Інший підхід дозволяє отримувати інформацію з відкритих ресурсів в Інтернеті без участі людини: ReadTheWeb [33], OpenIE [34], Google Knowledge Vault [35]. Подібні системи є експериментальними, кожна з них має свої особливості. Наприклад, Knowledge Vault намагається враховувати невизначеності, кожному факту ставляться у відповідність коефіцієнт довіри і походження інформації. Таким чином, всі твердження діляться на ті, які мають високу ймовірність бути істинними, і ті, які можуть бути менш імовірними. Передбачення фактів і їх властивостей здійснюється методами машинного навчання на основі дуже великої кількості текстів і вже наявних фактів. В даний момент Knowledge Vault містить 1,6 млрд фактів. Система NELL, що розробляється в рамках проекту ReadTheWeb університетом Карнегі-Меллона, містить більше 50 млн тверджень з різними ступенями довіри. Близько 2 млн 800 тис. Фактів мають високу ступінь довіри. Процес навчання NELL також ще не завершений.

1.2.2 Перспективи інтелектуалізації Web-сайтів за допомогою семантичного аналізу текстів

В даний час ведуться активні дослідження в області розробки алгоритмів аналізу текстів. Результатом цих досліджень є десятки моделей і готових алгоритмів, яким необхідна перевірка. При цьому до цих пір не існує інструмента, який надає зручні засоби для розробки в цій галузі. Це змушує розробника-лінгвіста зосереджувати увагу не тільки на написанні алгоритму, але і на створенні системи, здатної запустити цей алгоритм, забезпечити його взаємодію з іншими і надати необхідну інформацію про його роботу.

Семантичний Web – це надбудова над існуючим Web-простором, яка покликана зробити розміщену в ній інформацію зрозумілішою для комп'ютерів і інтелектуальних агентів [36, 37]. Машинна обробка можлива в семантичній павутині завдяки двом її найважливішим характеристикам [36].

Використання уніфікованих ідентифікаторів ресурсів (URI), широко відомих як адреси. В Інтернеті ці ідентифікатори використовуються для установки посилань на адресуємий об'єкт (наприклад, Web-сторінку, файл). У семантичній павутині URI використовуються також для іменування об'єктів. Свої URI в семантичній павутині є не тільки у сторінок, але і у об'єктів реального світу (людей, міст, художніх творів і так далі), і навіть у абстрактних понять (наприклад, у властивостей «ім'я», «посада», «колір»). Оскільки URI глобально унікальні, вони дозволяють називати одні й ті ж предмети в різних місцях в семантичній павутині.

Використання семантичних мереж і онтологій. Сучасні методи автоматичної обробки даних, доступних в Інтернеті, засновані на частотному і лексичному аналізі текстового вмісту, яке призначене для сприйняття людиною. У семантичній павутині замість цього використовується стандарт RDF, що описує семантичні мережі, в яких вузли і дуги мають URI. Твердження, які кодуються за допомогою RDF, в подальшому можна інтерпретувати за допомогою онтологій, створених за стандартами RDF Schema і OWL, щоб отримувати з них логічні висновки. В основі онтологій лежать математичні формалізми, звані дескрипціонними логіками. Технічну частину семантичної павутини становить сімейство стандартів на мови опису, що включає XML, XML Schema, RDF, RDF Schema, OWL [36, 38].

Онтології призначені для: уявлення метаданих, що описують семантичну структуру предметної області; обміну інформацією та знаннями для забезпечення можливості взаємодії між інтелектуальними агентами, щоб синхронізувати терміни і поняття, що описують прикладну предметну область.

Ці інтелектуальні агенти виконують складний пошук за кількома критеріями в пошукових системах, здійснюють збір, аналіз, обробку даних, обмін з іншими агентами, даними і онтологіями, а також здатні самонавчатися.

Web-сервіс – це програмне забезпечення, що надає доступ до даних і певної функціональності в розподіленому середовищі. Значно полегшує вирішення складних завдань для користувачів. На концептуальному рівні ми можемо розглядати Web-сервіси як одиниці додатка, кожна з яких займається виконанням певного функціонального завдання. Якщо піднятися на рівень вище, то ці завдання можна об'єднати в бізнес-орієнтовані завдання для виконання певних бізнес-операцій, дозволяючи технічно непідготовленим людям розглядати додатки як обробники завдань в рамках потоку робіт додатків Web-сервісів. Таким чином, після того як технічні фахівці розробили Web-сервіси, користувачі бізнес-процесів можуть об'єднати їх для вирішення конкретних виробничих завдань [39].

Інтеграційне рішення на основі Web-сервісів забезпечують [39,40]:

- можливість взаємодії додатків, реалізованих на різних програмно-апаратних платформах;

- можливість підтримки гнучких змін в додатках;

- інтеграцію додатків за допомогою Web-сервісів у відповідність з бізнес-процесом.

Основні обмеження Web-сервісів – це статичність інтеграційного рішення і необхідність перезапису при виникненні змін у бізнес-процесах управління. Рішенням є використання семантичних Web-сервісів. Властивість динамічності семантичних Web-сервісів забезпечується можливістю модифікації параметрів його виклику в реальному часі, а високий ступінь автоматизації – онтологією, що дозволяє однозначно ідентифікувати програму-агенту призначення, зміст, технічні деталі виклику конкретного сервісу. Для забезпечення розуміння програмами-агентам

призначення Web-сервісу необхідно супровід його онтологією або семантичним описом.

Застосування вищевказаних технологій семантичного представлення даних і інтелектуальних програм-агентів сприяє інтелектуалізації розподіленої обробки даних в Інтернеті, і за рахунок цього їх використання в Інтернет підвищить ефективність обробки управлінських бізнес-процесів зі зменшенням людського фактору.

Поява в просторі Web необхідної кількості онтологій, що покриває всі сфери людської діяльності, і розвиток програмно-інформаційних засобів семантичного Web (системи логічного висновку, мови семантичних запитів, сховища знань на основі мережевої моделі даних RDF) забезпечать поширення семантичних Web-сервісів і їх впровадження на підприємствах. Це дозволить відшукувати і комбінувати Web-сервіси, що задовольняють вимогам бізнес-процесів, сформульованим на мові високого рівня.

Інтелектуальні агенти виконують складний пошук за кількома критеріями в пошукових системах, здійснюють збір, аналіз, обробку даних, обмін з іншими агентами, даними і онтологіями, також здатні самонавчатися.

Термін Web Mining можна перевести як "видобуток даних в Web". Web Intelligence (або Web-Інтелект) сприяє стрімкому розвитку електронного бізнесу. Здатність визначати інтереси і переваги кожного відвідувача, спостерігаючи за його поведінкою, є серйозною і критичною перевагою конкурентної боротьби на ринку електронної комерції.

WebMining з'явився з таких дисциплін як виявлення знань в базах даних, ефективний пошук інформації, штучний інтелект, машинне навчання і обробка природних мов.

Через різноманіття і надлишок інформації користувачі мережі Інтернет часто стикаються з проблемами аналізу і пошуку необхідної інформації. Можна виділити деякі проблеми роботи з інформацією у Всесвітній павутині:

- пошук значимої інформації (далеко не всі представлені користувачеві посилання несуть потрібну інформацію і вкрай важким є пошук

неіндексованої інформації);

- виявлення нових знань (серед усієї отриманої інформації складно витягти корисні знання);

- персоналізація інформації (виникає складність з осмисленням отриманих знань, поняття ідей, вкладених автором);

- вивчення споживача або індивідуального користувача (користувач не завжди отримує саме ту інформацію, яку хоче отримати).

Для вирішення цих проблем використовуються різні технології. До них відносяться: бази даних, інформаційний пошук, обробники природних мов та ін. Технологія WebMining спрямована як на пряме, так і на непряме рішення перерахованих проблем.

Системи Web Mining можуть відповісти на багато питань, наприклад, хто з відвідувачів є потенційним клієнтом Web-магазина, яка група клієнтів Web-магазину приносить найбільший дохід, які інтереси певного відвідувача або групи відвідувачів.

Технологія Web Mining охоплює методи, які здатні на основі даних сайту виявити нові, раніше невідомі знання і які в подальшому можна буде використовувати на практиці. Іншими словами, технологія Web Mining застосовує технологію Data Mining для аналізу неструктурованої, неоднорідною, розподіленої і значної за обсягом інформації, що міститься на Web-вузлах.

У Web Mining можна виділити наступні етапи:

- вхідний етап (input stage) – отримання «сирих» даних з джерел (логи серверів, тексти електронних документів);

- етап попередньої обробки (preprocessing stage) – дані подаються у формі, необхідної для успішної побудови тієї чи іншої моделі;

- етап моделювання (pattern discovery stage);

- етап аналізу моделі (pattern analysis stage) – інтерпретація отриманих результатів.

Це загальні кроки, які необхідно пройти для аналізу даних мережі Інтернет. Конкретні процедури кожного етапу залежать від поставленого завдання. У зв'язку з цьому виділяють такі категорії Web Mining.

- аналіз використання веб-ресурсів (Web Usage Mining);
- витяг веб-структур (Web Structure Mining);
- витяг веб-контенту (Web Content Mining).

Пошук знань в мережі Інтернет є непростим і трудомістким завданням. В значній мірі його вирішує напрям Web Mining, що здійснює витяг веб-контенту. Він заснований на поєднанні можливостей інформаційного пошуку, машинного навчання та Data Mining. Крім того, Web Content Mining має на увазі автоматичний пошук і витяг якісної інформації з різноманітних джерел Інтернету, перевантажених "інформаційним шумом". Тут також йдеться про різні засоби кластеризації і анотування документів.

У цьому напрямку, в свою чергу, виділяють два підходи: підхід, заснований на агентах, і підхід, заснований на базах даних.

Підхід, заснований на агентах (Agent Based Approach), включає такі системи:

- інтелектуальні пошукові агенти (Intelligent Search Agents);
- фільтрація інформації / класифікація;
- персоніфіковані агенти мережі.

Приклади систем інтелектуальних агентів пошуку:

- Harvest (Brown і ін., 1994);
- FAQ-Finder (Hammond та ін., 1995);
- Information Manifold (Kirk і ін., 1995);
- OCCAM (Kwok and Weld, 1996), and ParaSite (Spertus, 1997);
- ILA (Information Learning Agent) (Perkowitz and Etzioni, 1995);
- ShopBot (Doorenbos і ін., 1996).

Підхід, заснований на базах даних (Database Approach), включає системи: багаторівневі бази даних; системи web-запитів (Web Query Systems). Приклади систем web-запитів:

- W3QL (Konopnicki і Shmueli, 1995);

- WebLog (Lakshmanan і ін., 1996);
- Lorel (Quass і ін., 1995);
- UnQL (Buneman і ін., 1995 and 1996);
- TSIMMIS (Chawathe і ін., 1994).

Згідно з цим підходом аналізується зміст документів: знаходяться схожі за змістом слова та їх кількість. Потім вирішується завдання кластеризації та класифікації, де документи групуються за смисловим близькості. Цей напрямок може бути використано для оптимізації пошуку індексованих документів.

Для знаходження необхідної інформації користувачі зазвичай користуються пошуковими ресурсами. При цьому часто використовуються прості запити за ключовими словами. Результатом виконання запиту є список сторінок, відсортований за деякий індексом релевантності, що описує ступінь збігу результату із запитом. Однак існуючі пошукові механізми мають недоліки, основним з яких є низька точність результату, викликана недостатнім урахуванням семантичних зв'язків і контексту знайдених в тексті виразів [41]. Індексація сегментів мережі, які цікавлять користувача, з використанням інтелектуального аналізу даних, що застосовує алгоритми математичної лінгвістики і обробки природних мов, є перспективним напрямком Web Mining в області пошуку інформації.

2 АЛГОРИТМИ МАТЕМАТИЧНОЇ ЛІНГВІСТИКИ І ОБРОБКИ ПРИРОДНИХ МОВ

Внаслідок створення різноманітних моделей процесів обробки тексту, а також відповідних алгоритмів та структур представлення даних, сформувалася така міждисциплінарна галузь науки , як обробка природної мови (Natural Language Processing – NLP).

NLP – це наука на перетині комп’ютерних наук, штучного інтелекту та обчислювальної лінгвістики. NLP містить набір інструментальних засобів для вилучення змістовної та корисної інформації. Саме тому, проблемне поле цієї науки пов’язане з забезпеченням взаємодії між комп’ютерами та природними мовами.

Головною метою NLP є підвищення якості машинного аналізу і синтезу повідомлень на природній мові. Якість машинного аналізу залежить від здатності комп’ютера отримувати сенс з вхідних повідомлень природною мовою, а якість синтезу залежить від здатності комп’ютера генерувати адекватні вихідні повідомлення на природній мові. Згідно з дослідженнями Карен Спарк Джонс можна виділити наступні головні завдання NLP:

- видобування даних, яке полягає у вивченні даних, пошуку зв'язків та закономірностей між ними;

- синтез мовлення займається питаннями озвучування/прочитання тексту (документ, повідомлення і т. д.) голосом, який є наближеним до природного;

- розпізнавання мови. Дослідження в цієї проблемної галузі NLP стосуються виведення/розпізнавання тексту з малюнків, відсканованих документів або файлів у PDF форматі, а також розпізнавання мовлення, продукованого людським голосом;

- генерування природної мови (конвертування комп'ютерних даних у

природну мову людини);

- машинний переклад. Оскільки комп’ютер не володіє тими знаннями, що володіє людина, які формують «розуміння» тих чи інших фраз, автоматичний переклад вважається надзвичайно складним;

- розпізнавання/визначення теми. Це завдання вирішується шляхом поділу тексту на логічні частини, визначаючи провідні теми кожної з цих частин.

- інформаційний пошук (окрім пошуку відбувається розпізнавання та вилучення змістовної інформації);

- отримання інформації. Це завдання корелюється з попереднім. Різниця в тому, що воно більш зосереджене на вилученні семантичної інформації з тексту;

- проблема лексичної багатоманітності (вирішується наданням списку можливих значень конкретного багатозначного слова, серед яких можна вибрати відповідник згідно контексту);

- демодуляція окремих лінгвістичних одиниць (відбувається шляхом перетворення окремих термінів (медичних, публіцистичних, технічних) у зрозумілу форму);

- забезпечення функціонування питально-відповідальної системи. Питання, поставлені людською мовою, не завжди бувають конкретизованими. Існують абстрактні питання, які створюють перешкоди питально-відповідальним системам.

Відповідно до цих завдань можна виокремити чотири загальні класи завдань NLP: аналіз мови, аналіз текстів, синтез мови, синтез текстів.

Важливими прикладами застосування NLP є:

- віртуальний помічник Windows Cortana, який розпізнає мову. За допомогою Cortana можна створювати нагадування, відкривати додатки, відправляти листи, грати в ігри, дізнаватися погоду і т.д.;

- поштовий сервіс Gmail (вміє визначати спам та запобігати його потраплянню до поштових скриньок;

- платформа від Google – Dialogflow, яка дозволяє створювати NLP ботів (наприклад, можна створити бота для замовлення піци).

Отже, ми бачимо, що обробка природної мови використовується для вирішення різноманітних типів завдань. У процесі дослідження обробки природної мови було досягнуто значних результатів, серед яких розробка потужних лексикографічних систем, програм для машинного перекладу, електронних словників та ін. Однак, існують проблеми, які досі не знайшли свого вирішення, вони коріняться у самій природі людської мови.

Дослідження автоматичного аналізу тексту в комп'ютерній лінгвістиці впливає на розвиток не тільки теоретичних знань лінгвістичних основ, необхідних для створення штучного інтелекту, а й на реалізацію практичних потреб людини, наприклад, створення систем машинного перекладу.

Автоматичний аналіз тексту складається з ряду складних трансформацій природної мови відповідно до заданого алгоритму. Лінгвістичні алгоритми можна класифікувати за певними критеріями, а саме: спосіб комунікації, форма мовлення, рівень інтелектуальності, рівень мовної системи.

Відповідно до способу комунікації розрізняють лінгвістичні алгоритми письмового тексту та усного мовлення. Лінгвістичні алгоритми аналізу тексту почали розробляти в межах NLP ще в 50-х роках XX століття для розробки інформаційно-пошукових систем та систем автоматичного реферування. Алгоритми аналізу усного мовлення почали розробляти пізніше (на початку 90-х років). Зараз вони активно застосовуються в автовідповідачах, системах розпізнавання індивідуальних характеристик людини (вік, стать, рівень алкогольного сп'яніння), в системах, що керуються голосом тощо.

За формою мовлення розрізняють алгоритми обробки монологічного та діалогічного мовлення. Тривалий час об'єктом автоматичного аналізу слугували монологічні тексти, переважно наукові роботи. Завдяки розвитку мережі Інтернет почали аналізувати жанр діалогічного мовлення (чати,

блоги, форуми). Обробка діалогічних текстів вимагає застосування специфічних алгоритмів, враховуючи паралінгвістичні особливості цього жанру.

Алгоритми, які дозволяють встановити імпліцитну або нову інформацію, яка не міститься в тексті, мають вищий рівень інтелектуальності, ніж алгоритми інформаційного пошуку та реферування, які просто узагальнюють найважливішу інформацію тексту.

Алгоритми автоматичного аналізу тексту можуть застосовуватись на різних рівнях мовної системи, починаючи з окремого символу, котрий виступає об’єктом аналізу оптичних систем розпізнавання тексту та закінчуючи дискурсивним рівнем, на якому відбувається моделювання структури зв’язного тексту. Системи аналізу текстів складаються з компонентів (так званих «лінгвістичних процесорів»), які один за одним обробляють вхідний текст. Вхід одного процесору, як правило, є виходом іншого.

2.1 Моделі і алгоритми контент-аналізу текстів

Передача змісту в процесі обробки тексту складається з двох компонентів – змістового і граматичного. Сенс речення виникає як результат поєднання елементарних семантичних одиниць відповідно до визначених правил. Інакше кажучи, речення представляє код окресленого змісту, а синтаксис – умовну форму послідовності змістових одиниць, що дає змогу структурувати цей зміст. Головною умовою правильної семантичної інтерпретації тексту є контекст. Будь-яка галузь діяльності відображається конкретними термінами і взаємозв’язками між ними. Таку множину можна розбити на певну кількість типів термінів і типів взаємозв’язків (семантичних одиниць). Кожне речення можна перевести в текст, який складається з ряду термінів і типів зв’язків, без урахування граматичних особливостей, відображаючи кожен термін або зв’язок у певний тип. Цей процес має назву

канонізації тексту, а зміст, який при цьому виникає, – канонічним сенсом тексту. Проте повна відмова від граматики не завжди виправдана. Іноді зміст речень визначається прийменниками, відмінками слів, і тому їхнє врахування може значно полегшити семантичний аналіз. Тому вводиться додаткова семантична одиниця – граматична роль лексем або їх частин для зазначення відповідних граматичних ознак мови.

Під час процесу аналізу тексту слід враховувати, що зміст також залежить від специфіки сфери, до якої він належить. Цим викликано впровадження третього типу семантичних одиниць – спеціальних ролей лексем. У загальному випадку канонічний зміст тексту визначається за допомогою значень семантичних одиниць усіх вказаних типів.

Останнім часом набув поширення контент-аналіз (аналіз змісту) текстів. Він передбачає пошук у тексті мовних індикаторів (одиниць аналізу, символів), певних змістових понять (категорій аналізу, термінів), визначення частоти їх уживання, оцінювання співвідношення з іншими одиницями і зі змістом усього твору. Основними процедурами контент-аналізу текстів є (рис. 2.1):

- процедура символьного аналізу;
- процедура морфологічного аналізу;
- процедура термінологічного аналізу.

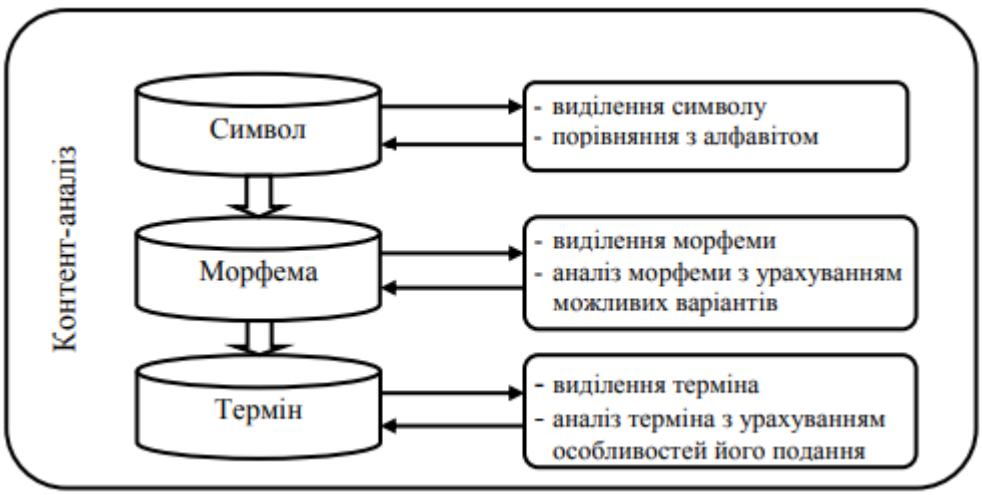

Рисунок 2.1 – Структурна схема алгоритму контент-аналізу

Процедура символьного аналізу полягає у порівнянні послідовності символів текстового фрагменту з символами алфавіту. Результат такої оброки переходить на етап морфологічного аналізу. Він проводиться у декілька кроків. На першому компонується підмножина морфем, а на другому виділяються їх основні характеристики:

- один з символів у морфемі може бути замінений;
- символ може випадати;
- може бути доданий додатковий або зайвий символ;
- символи можуть бути змінені місцями.

Під час етапу термінологічного аналізу необхідно врахувати дві умови. Перша з них визначає такі особливості подання терміну в тексті: слова та словосполучення можуть мати різні рід, відмінок, множину; між словами у терміні, який складається з декількох слів, можуть стояти інші слова; відсутність строгого порядку слів у термінологічному словосполученні.

Друга умова пов’язана з тим, що в мовознавстві часто не визнається наявність у мові багатоосновних термінів, які ідентифікуються як концептуальні об’єднання. За кількістю компонентів виділяють такі типи термінологічних словосполучень: двокомпонентні, трикомпонентні, чотири-, п’яти- і шестикомпонентні. Аналіз словесних конструкцій термінів дає підставу вважати, що більшість із них двокомпонентні (рис. 2.2). Привертає увагу той факт, що нині з’явилася тенденція до збільшення багатокомпонентних структур термінології. Трикомпонентні словесні конструкції мають структуру, відображену на рис. 2.3.

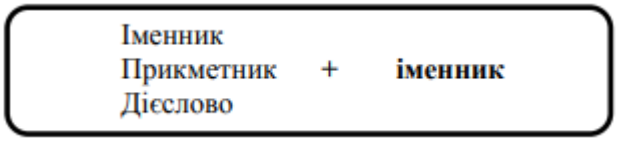

Рисунок 2.2 – Типові лексико-граматичні моделі двокомпонентних термінологічних словосполучень

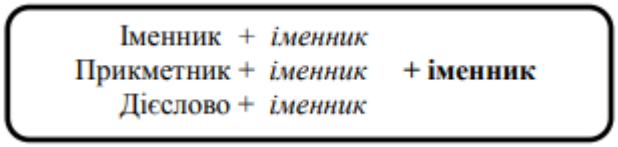

Рисунок 2.3 – Типові лексико-граматичні моделі трикомпонентних термінологічних словосполучень

На практиці в багатьох текстах формуються багатослівні словосполучення, тобто ті, які складаються з чотирьох і більше слів. Вони вживаються для вираження складних понять, кожному з яких відповідає свій термін. Сьогодні спостерігається тенденція до їх збільшення (рис. 2.4).

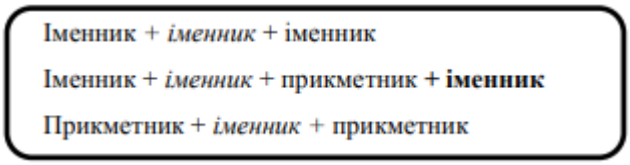

Рисунок 2.4 – Типові лексико-граматичні моделі мультикомпонентних термінологічних словосполучень

На основі цих факторів можна розробити характеристику рівня термінів, яка несе в собі інформацію про те, у якій позиції і з якими словами наявні у тексті терміни. А. Я. Шайкевич висунув гіпотезу, що слова, які зв’язані між собою за змістом, у тексті повинні траплятися недалеко один від одного. І навпаки, слова, які часто трапляються разом у тексті, пов’язані між собою за змістом. Цю гіпотезу він використав для виділення семантичних рівнів при аналізі віршованих текстів. Але, якщо під час аналізу віршованих

текстів за «одиницю аналізу» можна взяти один рядок, то виділення семантичного поля у звичайних прозових текстах є проблемою. Зрозуміло, що чим більша частина тексту вибирається як «одиниця», тим більшою може виявитися відстань між термінами.

Слід зауважити, що семантична подібність слів буде залежати не лише від їх відстані відносно один одного, а й від їхнього граматичного розташування. Щоб спростити процедуру аналізу взаємозв'язку термінів, пропонується виділяти один термін як «домінанту», а інший, який буде траплятися разом з ним, умовно називати «супровідним» (рис. 2.5). При цьому не слід забувати, що один і той самий термін може виступати як у ролі «домінанти», так і «супровідним». Після того, як зв'язки між окремими парами термінів будуть установлені, слова, які тісно пов'язані один з одним за сенсом, можна об'єднати у семантичні підгрупи тощо доки у підгрупі не з'являться цілі змістові фрази.

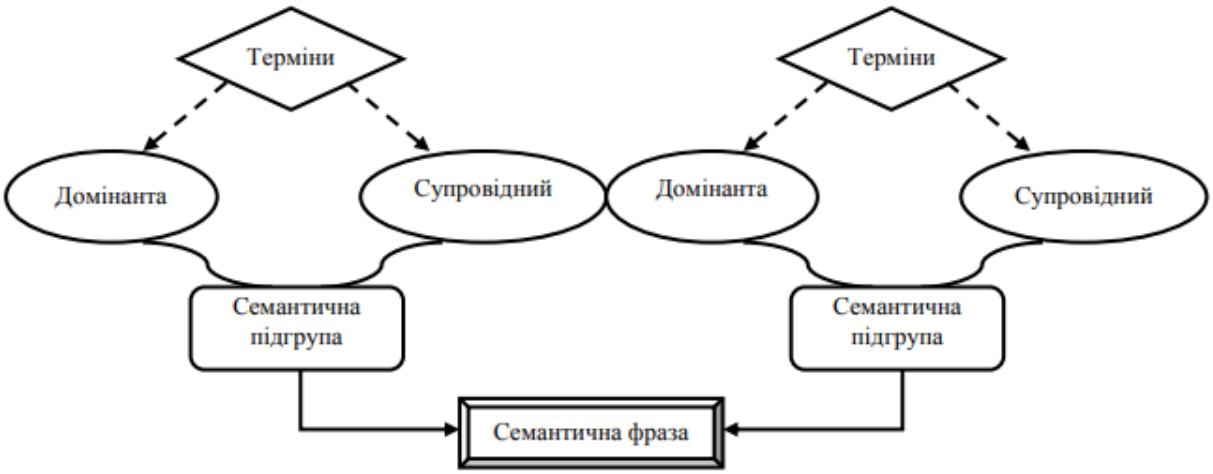

Рисунок 2.5 – Схематична модель термінологічного процесу аналізу тексту

Завдяки контент-аналізу можна отримати достатньо об'єктивний результат та зробити зміст тексту вимірюваним і придатним для точного обчислення.

2.2 Алгоритми лексичного аналізу

Головне завдання лексичного аналізу – розпізнати лексичні одиниці тексту. На вході у програму цього типу – текст, на виході – список лексичних одиниць тексту (рис. 2.6)

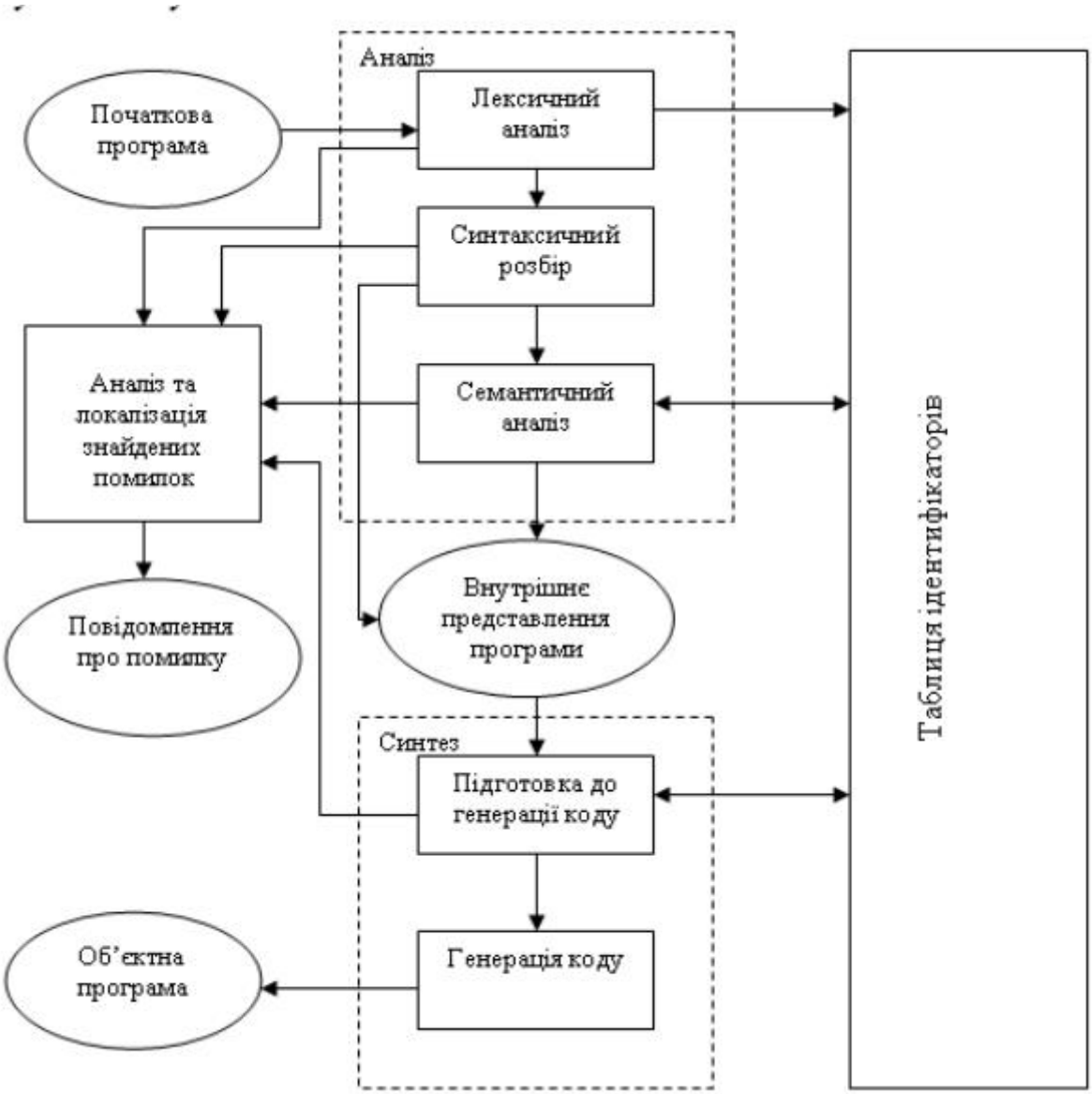

Рисунок 2.6 – Лексичний аналіз схема

Одним з фундаментальних алгоритмів лексичного аналізу є лексична декомпозиція, яка передбачає розбивку тексту на токени; відповідно, програми, що виконують лексичну декомпозицію, називаються

токенайзерами. Як правило, токени співпадають зі словоформами, однак для позначення лексичних одиниць тексту використовується термін «токен», а не «слово», так в ряді випадків під токеном розуміють одиницю меншу, ніж слово (окремі морфеми, клітики), або більшу, ніж слово (словосполучення).

Англомовні токенайзери виконують декомпозицію на основі пробілів між словами і зазвичай розпізнають в якості окремих токенів апостроф та символи, що йдуть за ним [44].

Програми цього типу містять проблему розпізнавання словосполучень і абревіатур. Очевидно, що скорочення типу e.g. слід розпізнавати як один токен; те ж саме відноситься до дат, наприклад 11.11.2111.

Окрему проблему становлять ініціали та скорочення перед особовими іменами, наприклад, J. Smith, Dr. Smith, J.B. Smith. Якщо ці імена кореферентні, то потрібно розглядати ініціали як окремі маркери: це дозволить розпізнати Smith як ім'я одного з персонажів і нарахувати адекватні вагові коефіцієнти залежно від частотності його вживання. Якщо ж маються на увазі різні люди, слід розглядати прізвище та ініціали як один токен.

Зазвичай лексична декомпозиція проводиться на основі списків скорочень [45]. Крім того, в окремому файлі збираються стійкі словосполучення та ідіоми, які розпізнаються як один токен. Наприклад, because of доцільно розглядати як один токен, оскільки це союзне словосполучення має одне значення.

Лексична декомпозиція має фундаментальне значення для проведення автоматичного аналізу тексту, оскільки лежить в основі цілого ряду інших алгоритмів. Зрозуміло, що для проведення стемінгу слід спочатку розбити текст на токени; на основі списку токенів зазвичай виконується синтаксична декомпозиція, зважування, нарешті, анотування, також виконується на лексичному рівні.

Анотування проводиться за допомогою тегерів, на вході в яких – список токенів, на виході – список, в якому кожному токену привласнено

умовне позначення (тег), що вказує на його лінгвістичні характеристики. Найпоширенішим видом тегерів є тегер частин мови (*POS taggers*), який розпізнає частину мови токена і приписує йому відповідний тег. Окрім інформації про частини мови зазвичай вказується інформація про лексикограматичні і семантичні характеристики слова, наприклад, *NN* – загальний іменник в однині, *NNS* – загальний іменник у множині, *AJC*-прикметник у порівняльному ступені і т. д. [46]. Списки тегів частин мови відрізняються ступенем дробності. Більш подрібнена класифікація дозволяє видавати користувачу більший обсяг інформації, однак обумовлює більшу кількість помилок, знижуючи швидкодію програми [47].

Тегери частин мови послідовно виконують три основні операції: токенізацію, морфологічну класифікацію та зняття неоднозначності. Морфологічна класифікація передбачає зіставлення кожного токена вхідного тексту зі словником і приписування йому тегів частин мови. У словнику зазвичай містяться словоформи з можливими тегами частин мови. Досить велика кількість слів співвідноситься тільки з однією частиною мови (прийменники, артиклі, займенники), проте цілий ряд слів може використовуватися в якості різних частин мови.

Якщо якесь слово з вхідного тексту відсутнє в словнику, застосовуються спеціальні правила для розпізнавання частини мови. У тому випадку, якщо неможливо застосувати правила, токену приписується тег, який використовується за умовчанням, зазвичай – тег іменника. Іменники – найбільш частотна частина мови, і саме вони позначають нові об'єкти, імена яких можуть бути відсутніми в словнику. Якщо всім словам в тексті приписати тег іменника, то можна правильно проанотувати 14,6% слів.

Токени, яким приписано більше одного тега, а також статистична інформація про них передаються для подальшої обробки в модуль зняття неоднозначності.

Зняття неоднозначності передбачає вибір одного з двох або більше тегів, приписаних даному токену. Залежно від алгоритмів, що

застосовуються для зняття неоднозначності, тегери частин мови класифікують на стохастичні і засновані на правилах.

В стохастичних тегерах проводиться аналіз імовірнісних параметрів кожного з тегів, в результаті якого вибирається один тег з найбільшим імовірнісним значенням.

У тегерах, заснованих на правилах, аналіз імовірнісних характеристик не проводиться, хоча враховується частотність використання тегів з тим чи іншим токеном. Такий тегер навчається на досить великому анотованому корпусі, запам'ятовуючи найбільш частотні теги морфологічно омонімічних словоформ, далі, для підвищення якості анотування застосовуються спеціальні правила автоматичного виправлення помилок.

Загалом стохастичні технології істотно знижують кількість помилок, однак негативно впливають на швидкодію системи. Їх можна успішно використовувати для анотування статичних корпусів. Для динамічного анотування краще застосовувати тегери, засновані на правилах, оскільки вони забезпечують більшу швидкодію.

2.3 Семантико-синтаксичний алгоритм стиску

Цей алгоритм належить до групи алгоритмів, які виконують абстрагування, з опорою на зовнішні джерела інформації. У наведеному вигляді він не здатний прореферувати весь текст, або стиснути його до малого об’єму. Проте він є корисним, оскільки дозволяє отримати для ряду випадків стиск там, де простий вибір буде змушений втратити інформацію.

Передбачається, що є ряд додаткових механізмів, а саме: синтаксичний та семантичні аналізатори, синтаксичний синтезатор для речень. (рис. 2.7)

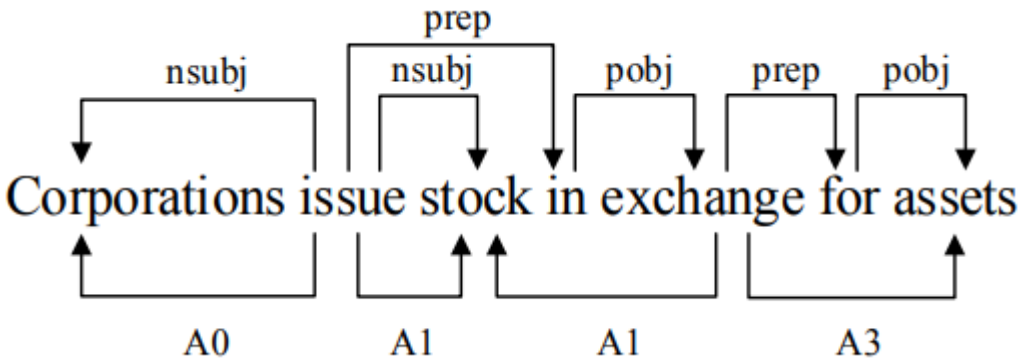

Рисунок 2.7 – Приклад результату роботи синтаксичного аналізатора (Зверху) і результату роботи аналізатора, який визначає рольову структуру

У межах областей між двома маркерами «Тема почалася» і «Тема скінчилася», які належать одній темі, застосовується перший алгоритм узагальнення. Він працює переважно з онтологією, використовуючи зв'язок «бути» (*is_a*). У процесі узагальнення цей алгоритм пробігає по онтології від понять нижчого рівня до понять вищого рівня у пошуках поняття, яке є водночас допустимими та досить абстрактними, для можливості здійснення узагальнення. Для нього є обов'язковою синтаксична передобробка, оскільки він інтенсивно використовує синтаксичні дані.

Позначимо тематичну область як . Розглянемо алгоритм, що назвемо алгоритм стиску 1.

Для кожного є :

скласти предикатну базову структуру відповідно до підмета і присудка;

для кожного є (від даного і до кінця області) :

скласти предикатну структуру відповідно до підмета і присудка;

порівняти предикатну структуру з базовою;

якщо для підметів присудків або і підметів і присудків виконуються «Умови групування» на відстань 2, то з та будується одне , більш поширене і, використовуючи більш загальні поняття, вилучається.

Оцінка складності роботи алгоритму стиску 1: , де – довжина блоку тексту в реченнях, – найбільша кількість сенсів слова, – найбільша довжина речення в тексті.

Недоліком цього алгоритму є його чутливість до синтаксичних неоднорідностей та «короткозорість», оскільки він не реагує на зв'язки між поняттями у WordNet, що мають довжину більшу за 2. Проте, якщо збільшити відстань до 3х або 4х, часто відбувається надлишкове узагальнення, що негативно впливає на якість реферату.

Попри очевидні перевагу такого алгоритму, а саме здатність до складання висновків, хоч би і обмежену, необхідно зауважити його малу частоту очікуваного використання.

2.4 Семантичний алгоритм стиску

Аналогічно до розглянутого алгоритму 1, цей алгоритм належить до групи алгоритмів, які виконують абстрагування, з опорою на зовнішні джерела інформації. У наведеному вигляді він не здатний прореферувати весь текст, або стиснути його до малого об'єму. Проте він є корисним, оскільки дозволяє отримати для ряду випадків стиск там, де простий вибір буде змушений втратити інформацію (рис. 2.8).

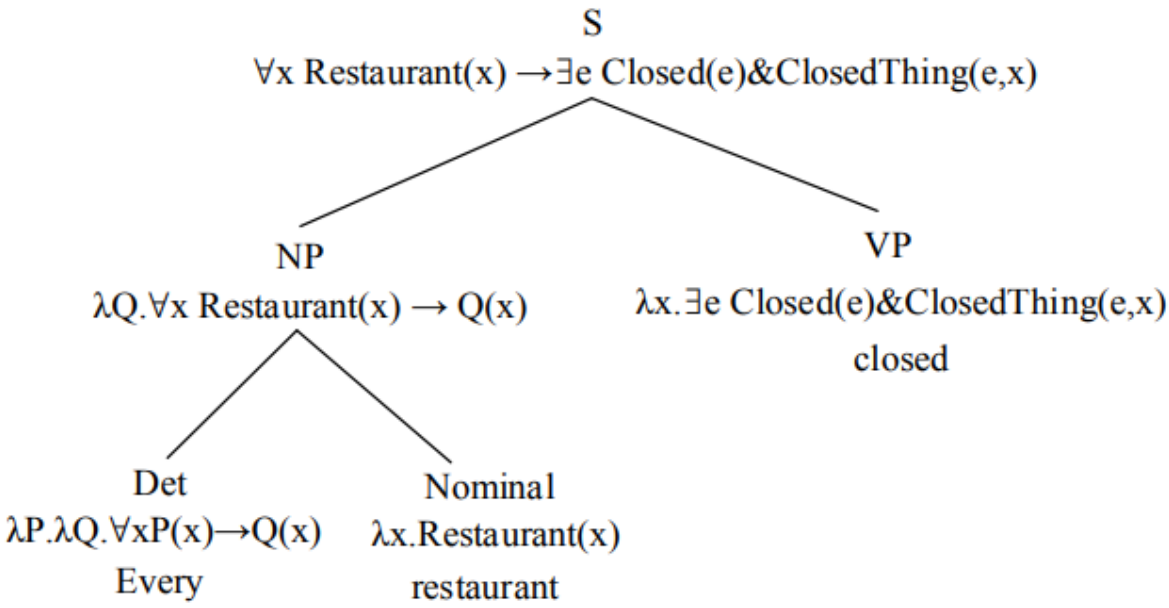

Рисунок 2.8 – Результат розбору за формальною синтаксичною і семантичною граматикою речення «Every restaurant closed»

Передбачається, що є ряд додаткових механізмів, а саме: синтаксичний та семантичні аналізатори, синтаксичний синтезатор для речень.

У межах областей між двома маркерами «Тема почалася» «Тема скінчилася», які належать одній темі, також застосовується другий алгоритм узагальнення. Його основою є пошук у ширину в орієнтованому графі онтології. Умови зупинки:

- як тільки зустрічається вершина (концепт), що є забороненою (зайвою), алгоритм припиняє обчислювати ваги для цієї вершини та всіх вершин, для яких вона є нащадком в ієрархії WordNet. До заборонених сенсів належать загальні поняття, якщо вони не представлені явно в лексичному ланцюжку;

- ящо вершина поза бажаною тематикою, алгоритм припиняє обчислювати ваги для цієї вершини та всіх вершин, для яких вона є нащадком в ієрархії WordNet;

- якщо досягнуто довжину шляху, рівну 5.

Позначимо тематичну область як . Повнозначне слово – – сенс слова

Алгоритм узагальнення:

Створити список

Для кожного є

Для кожного є

Для кожного

Додати у

/// Для кожної ітерації

Створити список

Якщо не виконуються умови зупинки

Для є

Визначити кількість шляхів, що проходять через в напрямку більш загального у WordNet

Занести у

Встановити вагу, рівну сумі всіх шляхів від нижніх синсетів через нього

Замінити

///

Для кожного є

Для кожного є

Для кожного є

Для кожного

Виконати розмітку за WordNet.

///

Для кожного є

Створити список (речень кандидатів)

Для кожного є

Порівняти маркери та

Якщо маркери співпадають,

Додати у

Інакше:

Для кожного є (речення зі списку)

Порівняти предикатну структуру зі .

Якщо немає відповідностей -вилучити

Опрацювати , створивши більш загальне речення

Вилучити використані з тектсу

Зробити список порожнім

Оцінка складності роботи алгоритму на одному проході: , де – довжина блоку тексту в реченнях, – найбільша кількість сенсів слова, – найбільша довжина речення в тексті, – кількість сенсів у списку сенсів. Очевидно, що чим більше різних змістовно наповнених елементів є в межах теми, то повільніше працює алгоритм.

Таким чином можна визначити ті концепти з WordNet, що не представленні в тексті явно, проте сильно пов’язані з його змістом.

Це дозволяє, наприклад, узагальнити „стіл, стілець, ліжко” до „меблі” але не до „об’єкт”. Проблема полягає у тому, що як і попередній алгоритм, цей також чутливий до синтаксичних структур. Так само, як і у попередньому випадку, необхідно зауважити малу ефективність (частоту очікуваного вживання) даного алгоритму.

2.5 Машинний переклад

Машинний переклад передбачає виконання комп’ютером перекладу тексту з однієї природної мови на іншу без участі людини та результат такої роботи. Задача машинного (автоматичного) перекладу потребує морфологічного аналізу, аналізу і перекладу лексики, синтаксичного аналізу і синтезу семантичних трансформацій, які б забезпечували смислову рівність введеної і виведеної текстової інформації. Звідси випливає, що задача

машинного(автоматичного) перекладу – це задача штучного інтелекту, який би зміст не вкладався в поняття штучного інтелекту [48,49].

На даний час серед найбільш досліджуваних є системи з використанням мови-посередника (проміжної мови), змішаного підходу, на основі навчання машин, на основі корпусів, на основі прикладів та гідридних методів на основі статистик та прикладів. Працюють як одномовні, так і багатомовні системи, не лише для письмових текстів, але й для усного мовлення.

Окремо стоїть задача автоматизованого перекладу, яка передбачає тісну взаємодію перекладача та системи на всіх етапах перекладу. Завдяки успіхам в розробці систем автоматизованого перекладу професійні перекладачі масово користуються такими системами, особливо для перекладів з повторюваною тематикою або перекладів, де треба узгоджувати роботу багатьох перекладачів.

2.5.1 Автоматичний переклад

Для першого етапу розвитку машинного перекладу було характерне так зване «кодування-декодування».

Цей підхід називається прямим методом, у ньому переклад розглядається як звичайний аналог тексту оригіналу. Відповідно до методу прямого перекладу, вихідний і цільової тексти повинні бути схожі і за своєю формою, і за концептуальним змістом.

Ця ідея виявилась обмеженою вузьким колом текстів спеціалізованої тематики (наприклад, прогнозом погоди).

Схему прямого методу наведено на рис. 2.9.

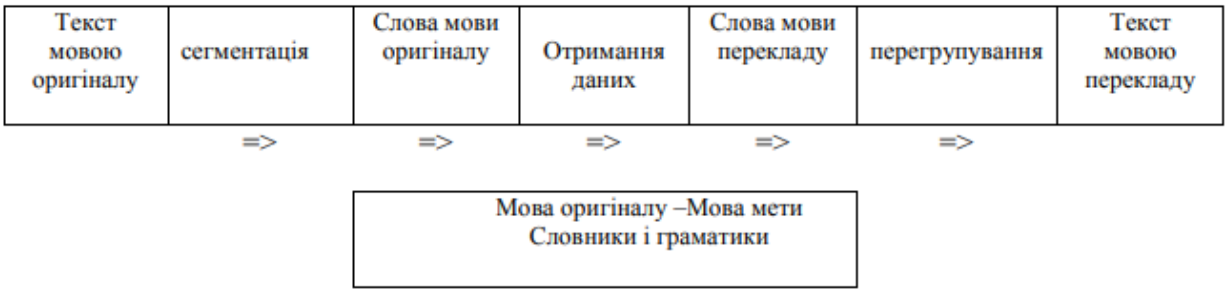

Рисунок 2.9 – Прямий метод

Протягом 1970х-80х рр. відбувався розвиток так званих систем «другого покоління», побудований на правилах, спрямованих на лінгвістичну обробку, як правило, у три етапи: синтактико-семантичний аналіз вихідного тексту, застосування правил перетворень з більш-менш абстрактним рівнем представництва та генерування цільового тексту з синтаксичного представлення вхідного тексту. У той же час точилися дебати про те, як можна використовувати системи, побудовані за цієї архітектурою, для забезпечення прийнятного рівня перекладу для реальних користувачів. Найпопулярнішими були ідеї обмеження входу (підмова і контрольовані мови), або переклад за участю користувача в перед- і пост-редагуванні [50]. Схему оптимізованого прямого методу наведено на рис. 2.10.

Рисунок 2.10 – Оптимізований прямий метод

Щоб поліпшити якість прямого перекладу, застосовуються два наступні методи, а саме: синтаксичні фільтри і статистичне ранжування перекладних еквівалентів, які б дозволили вибрати найбільш ймовірні з них для конкретного документа, що перекладається.

Синтаксичні фільтри мають форму логічних фреймів, де слоти заповнені синтаксичними структурами з зазначенням функції. Зазвичай в системах машинного перекладу на основі прямого методу досить багато фільтрів для «згладжування» сирого перекладу.

Другий основний метод машинного перекладу – це спосіб переносу інформації, заснований на правилах перетворень (першим вважається прямий метод) [51].

У системі на основі перетворень процес перекладу включає наступні стадії обробки: морфологічний та синтаксичний аналіз, власне перенос інформації у проміжному представленні, синтез синтаксичних структур, морфологічний синтез (побудова тексту перекладу). Досить часто системи на основі переносу містять семантичну складову. Мережа семантичних описів і відносин накладається на синтаксичні структури вихідного тексту і тексту мети (тобто власне перекладу). Метою семантичної компоненти є підвищення точності перекладу. Схему методу, заснованого на правилах перетворень, наведено на рис. 2.11.

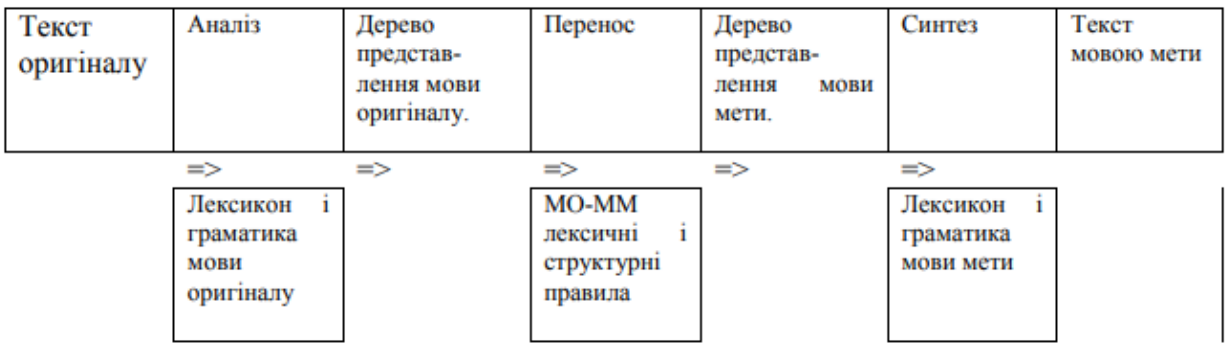

Рисунок 2.11 – Метод, заснований на правилах перетворень

Третім основним методом є використання проміжної мови [52]. У певному сенсі він схожий на попередній метод, однак існує декілька важливих відмінностей. На відміну від процедур переносу, які застосовуються здебільшого на синтаксичному рівні з деякими коригуванням семантики, представлення інформації через проміжну мову включає всю доступну лінгвістичну інформацію.

Крім того, системи на основі проміжної мови претендують на універсальність, тобто поширюються на будь-які мови.

Проміжна мова є формальним описом морфологічних, синтаксичних і семантичних характеристик мовної одиниці у вигляді співвідношення одиндо-одного. Кожна одиниця мови пов'язана з конкретним незмінним атомом у структурі проміжної мови і навпаки – кожен атом структури проміжної мови незмінно пов'язаний з одиницями різних мов.

В ідеалі, модель із застосуванням проміжної мови у машинному перекладі має включати наступні етапи обробки: морфологічний, синтаксичний та семантичний аналіз вихідного тексту, використовуючи інформацію зі словника мови оригіналу і парадигм; формування представлення мови вихідного тексту модулем проміжної мови; перетворення початкового представлення модулем проміжної мови у текст перекладу, використовуючи відповідні семантичні, синтаксичні, лексичні та морфологічні дані зі словника мови перекладу і парадигм.

Зазвичай формалізм проміжної мови має вигляд графічного мережі або її аналітичного еквівалента. Це дуже складна система морфологічних, синтаксичних та семантичних одиниць і відносини. Схему моделі на осноі проміжної мови наведено на рис.2.12.

Рисунок 2.12 – Модель на основі проміжної мови

На особливу увагу заслуговують системи на основі методу штучного інтелекту (ШІ) – artificial intelligence (AI), які спираються на енциклопедичні дослідження.

Основним компонентом моделі перекладу на основі ШІ є його так звана „база знань”. Відповідно до моделі перекладу, заснованій на ШІ, – основні результати лінгвістичного аналізу на всіх рівнях мови перевіряються за допомогою позамовної інформації, що міститься в базі знань.

У всіх трьох вищезгаданих способах моделювання перекладу усунення неоднозначності здійснюється тільки за допомогою контексту. Жодна з цих моделей, однак, не використовує двох інших інструментів усунення неоднозначності, тобто ситуації та довідкової інформації.

У моделях перекладу, заснованих на ШІ, процедури усунення неоднозначності радикально відрізняються і ґрунтуються перш за все на аналізі ситуації та довідкової інформації (бази знань), в той час як лінгвістичні методи аналізу контексту служать тільки в якості вторинних резервних засобів. Бази знань містять особливим чином впорядковані ієрархії фактів про реальний світ, а вербальна інформація відіграє підлеглу роль, лексично позначаючи факти і ситуації. Ще одним важливим компонентом моделювання перекладу на основі ШІ є модуль прийняття рішень, який включає структурну ієрархію логічних побудов з оцінкою імовірності.

Нинішній рівень складності моделювання перекладу на основі ШІ досить неоднозначний: з одного боку, результати розвитку моделей ШІ,

призначених для перекладу як такого вельми обмежені, а з іншого, однак, розробка моделей ШІ, призначених для інтерфейсу природною мовою, особливо для експертних систем, дуже ефективна.

Машино-орієнтовані статистичні методи складають наступну групу підходів [53]. У статистичних методах моделювання перекладу передбачається, що з певною ймовірністю кожне слово тексту перекладу може бути перекладом кожного слова вихідного тексту, але різні статистичні моделі відрізняються щодо подальших імовірнісних оцінок. Модель може оцінювати:

- імовірності узгодження порядку слів у тексті оригіналу і результуючому тексті перекладу;

- імовірності словосполучень у тексті оригіналу і результуючому тексті перекладу тощо.

Схему статистичної моделі наведено на рис 2.13.

Рисунок 2.13 – Статистична модель

На початку 1990-х рр. розвивається модель з використанням пар «приклад-переклад» для довгих конструкцій. Це стало можливим завдяки таким факторам розширенню можливості ЕОМ зберігати великі бази

прикладів та наявності великих двомовних корпусів текстів в електронних форматах. Схему відповідної моделі перекладу наведено на рис 2.14.

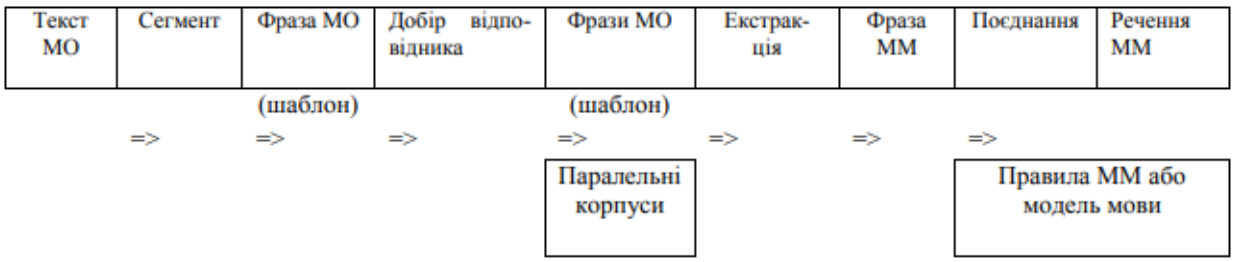

Рисунок 2.14 – Модель перекладу, побудована на прикладах

Слід зазначити, однак, що жоден з методів машинного перекладу не використовується в реальних системах у чистому вигляді.

У даний час розробляються також змішані або гібридні системи, що використовують як імовірнісні, так і лінгвістичні методи для отримання найкращого результат.

Про автоматичний переклад текстів вперше заговорили вже з моменту появи перших електронних обчислювальних машин. Потенційні сфери застосування машинного перекладу досить різноманітні. Наприклад, машинний переклад корисний – на побутовому рівні, спрощуючи комунікацію людей різних мовних груп, котрі не володіють належними мовними навичками. Машинний переклад актуальний і в бізнес-сфері, коли потрібен переклад значних обсягів даних.

На даний момент існують 3 технології машинного перекладу:

- аналітичний машинний переклад;
- статистичний машинний переклад;
- нейронний машинний переклад.

Аналітичний машинний переклад – це перша з технологій машинного перекладу. Метод має на увазі використання набору правил перекладу з вихідної мови в цільову, написаних лінгвістом, і двомовних словників

(набору лексичних елементів обох мов). Аналіз в даному методі часто сильно завищений, а процес перекладу проходить наступні етапи:

- морфологічний аналіз, в ході якого вказується рід, число, особа і інші морфологічні характеристики слів. При цьому виникає проблема багатозначності, коли одне і те ж слово може відноситися до різних частин мови;

- об'єднання окремих слів у групи;

- синтаксичний аналіз пропозицій, тобто визначення членів речення та їх місця в реченні. Спочатку програма шукає присудок. Потім перед знайденим присудком шукається підмет. Якщо його там немає, то алгоритм шукає підмет за присудком. Якщо підмета немає і там, то вважається, що він відсутній;

- синтез речень – узгодження знайдених частин речень та груп залежних слів;

До переваг аналітичного методу можна віднести:

- синтаксичну і морфологічну точність;

- стабільність і передбачуваність результату;

- можливість налаштування на предметну область.

З недоліків можна виділити:

- трудомісткість і тривалість розробки: для розробки лінгвістичних правил, яких в системі аналітичного перекладу може налічуватися десятки тисяч, необхідно залучення лінгвістів; процес розробки цих правил може займати від декількох місяців до декількох років;

- необхідність підтримувати і актуалізувати лінгвістичні БД;

- «машинний акцент» при перекладі – орієнтуючись виключно на правила, такі системи часто можуть ігнорувати контекст, підставляючи в цілому вірні, але не самі доречні варіанти перекладу окремих лексем.

Статистичний машинний переклад є підвидом методу, заснованого на корпусах тексту (corpus-based machine translation, CBMT). В основі CBMT є робота не з готовими правилами, тобто не раціоналістичний (аналітичний)

підхід, а емпіричний підхід, що використовує формування правил на основі паралельних двомовних корпусів текстів.

В основі такої технології дежить теорема Байеса: з речення виділяються окремі граматичні одиниці (слова і фрази), перебираються всі варіанти перекладу для кожного фрагмента і зважується ймовірність кожного з них.

Система статистичного перекладу зазвичай складається з трьох компонентів:

- модель перекладу, або таблиць перекладу – це таблиця-словник, в якій для всіх відомих системі слів і фраз на одній мові перераховані всі можливі їх переклади на іншу мову і вказана ймовірність цих перекладів;

- імовірнісна модель мови – це уявлення знань системи про мову, на яку потрібно перекласти текст. Вона використовується для того, щоб після вибору найбільш ймовірних варіантів перекладу окремих слів і фраз виходячи з моделі перекладу, вибрати з цих варіантів найбільш підходящі, виходячи з контексту;

- декодер – складова перекладача, яка є безпосередньо перекладом. Для кожного речення вихідного тексту він підбирає всі варіанти перекладу, поєднуючи між собою фрази з моделі перекладу, і сортує їх по спадаючій ймовірності. Потім всі отримані варіанти декодер оцінює за допомогою моделі мови.

Крім цього, може бути застосована таблиця зміни порядку, яка вказує, як можна змінювати порядок слів при перенесенні на мову перекладу. Іноді також необхідно включати додаткову лінгвістичну інформацію для мов з багатьма словозмінами, наприклад, російську або українську мови.

До переваг статистичних систем можна віднести:

- швидке налаштування: оскільки система навчається сама, лінгвісти необхідні тільки для допомоги в написанні алгоритму аналізу корпусів текстів; для подальшого навчання системи використовуються тексти, які можна знайти у вільному доступі;

- такі системи добре справляються з перекладом складних і рідкісних слів, термінів і стійких виразів;

- легко додавати нові напрямки перекладу: якщо мова почне змінюватися, система помітить це як тільки до неї потраплять відповідні тексти;

- відсутність глибокого аналізу тексту економить обчислювальні ресурси.

З недоліків можна відзначити наступне:

- статистичні системи набагато гірше працюють для сильно несхожих одна на одну мов без використання складних моделей типу tree-to-tree / tree-to-string (наприклад, при перекладі з англійської на японський;

- «дефіцит» паралельних корпусів: якість перекладу сильно залежить від кількості паралельних корпусів, для коректних перекладів статистичної системі необхідно як мінімум 500 тисяч, в ідеалі від декількох мільйонів паралельних текстів;

- нестабільність перекладу: незважаючи на здатність переводити стійкі вирази, ці самі вирази і терміни можуть переводитися по-різному виходячи з контексту;

- часто результат перекладу схожий на «зібраний пазл»: хоча загальний зміст речення зрозумілий, але частини речень існують окремо один від одного.

2.5.2 Нейронний переклад

Незважаючи на збереження актуальності аналітичної і статистичної технологій (в тому числі і гібридної), останнім часом активно розвиваються методи перекладу з використанням штучних нейронних мереж (neural machine translation, NMT).

На перший погляд нейронний переклад дуже схожий на статистичний, оскільки також використовує аналіз паралельних даних і формує на основі

цього аналізу певні залежності і закономірності. Однак в основі даного методу лежать зовсім інші принципи.

В основі нейронного перекладача лежить механізм двонапрямлених рекурентних нейронних мереж, побудований на матричних обчисленнях, який дозволяє будувати істотно більш складні імовірнісні моделі, ніж статистичні машинні перекладачі.

Хоча нейронний переклад також використовує для навчання паралельні корпуси, в процесі навчання він оперує не окремими фразами, але й цілими реченнями. Одна з головних проблем полягає в тому, що нейронній мережі потрібно набагато більше корпусів для навчання, ніж статистичної системі: щонайменше близько 100 мільйонів токенів для адекватної роботи, для перекладів ж належної якості – не менше близько 500 мільйонів. Також для навчання подібної системи потрібно набагато більше обчислювальних потужностей.

Однак головною причиною того, що нейронний переклад почав проявлятися порівняно нещодавно, є не стільки апаратні обмеження, оскільки тренувати подібні системи можна було і раніше, нехай ціною великих затрат часу, а скоріше розвитком теорії та практики нейронних мереж.

Провідні розробники систем нейронного перекладу вели свої дослідження вже досить давно, проте високі очікування від можливостей нейронних мереж і побоювання, що недостатньо досконалі системи нейронного перекладу не виправдають ці очікування, змушували розробників не оголошувати результати своїх розробок завчасно. Разом із тим у районі 2015-2016 років всі розробники подібних систем почали представляти свої варіанти нейронних перекладачів один за іншим. Так, Яндекс оголосив про поступовий перехід на нейронний переклад з вересня 2017 року, а Google зробив це березні 2017 року.

Що стосується переваг і недоліків нейронних мереж, вони багато в чому схожі з недоліками статистичних систем. З одного боку, нейронна

мережа, при належній кількості вхідних даних здатна видати практично ідеальний або близький до такого переклад, оскільки вона не просто навчається, вона «розуміє» принципи, за якими будується переклад.

З іншого боку, по-перше, вона вимагає куди більше обчислювальних потужностей для свого навчання, і, по-друге, в умовах нестачі паралельних даних, яких вона вимагає на порядок більше, ніж статистична мережа, нейронна мережа видасть переклад вкрай низької якості, тобто нейронні перекладачі ще більш вимогливі до обсягу вхідних даних, ніж статистичні системи.

Особливості нейронних перекладачів:

- в загальному випадку нейронний автоматичний переклад дає результат більш високої якості, ніж «чисто» статистичний підхід;

- автоматичний переклад через нейронну мережу краще підходить для вирішення завдання «універсального перекладу»;

- жоден з підходів до машинного перекладу сам по собі не є ідеальним універсальним інструментом для вирішення будь-якої задачі перекладу;

- для вирішення завдань з перекладу в бізнесі тільки спеціалізовані рішення можуть гарантувати відповідність всім вимогам.

Незважаючи на очевидні перспективи нвикористання нейронних мереж в якості основного інструменту машинного перекладу в майбутньому, на поточний момент недолік в першу чергу належного обсягу даних в різних предметних областях не дозволяє говорити про беззастережну перевагу нейронних мереж у всіх сферах застосування машинного перекладу. Тому на поточний момент як і раніше зберігають актуальність як статистичні системи, що вимагають менший обсяг даних для навчання, так і аналітичні системи, що дають стабільний і точний результат у вузьких областях.

3 ОБҐРУНТУВАННЯ РЕКОМЕНДАЦІЙ З ВИКОРИСТАННЯМ МЕТОДІВ СЕМАНТИЧНОГО АНАЛІЗУ ТЕКСТІВ ДЛЯ ІНТЕЛЕКТУАЛІЗАЦІЇ WEB-САЙТІВ

3.1 Порівняльний аналіз методів природної обробки мови

Суттєвою характеристикою є виділення динамічних і статичних алгоритмів. Динамічні алгоритми виконуються «на льоту», у відповідь на запит користувача, в той час як статичні алгоритми виконуються в процесі попереднього аналізу тексту, до того, як до нього звертається користувач. Відповідно, істотно розрізняються вимоги до швидкодії, що в свою чергу впливає на вибір архітектури і мови використовуваного програмного забезпечення.

Найбільша швидкодія досягається при застосуванні алгоритмів поверхневого рівня, до яких відносяться позиційно-статистичні алгоритми. Більш складні алгоритми семантичного рівня, що передбачають аналіз семантики мовних одиниць (наприклад, структурно-семантичних відносин), або аналіз структури зв'язного тексту, в тому числі моделювання його тематичної структури. Таким чином, можна виділити три групи алгоритмів: алгоритми поверхневого, семантико-синтаксичного і дискурсивного рівнів. Для підтримки цих алгоритмів використовуються і різні лексикографічні ресурси. Алгоритми поверхневого рівня виконуються на основі словників, що містять статистично-імовірнісні дані про розподіл мовних одиниць. Для виконання алгоритмів семантичного рівня потрібні словники-тезауруси, семантичні словники, онтології. Перспективним напрямком виступає розробка словників, в яких вказуються такі семантичні ознаки слів, що виділяються в компонентному аналізі , як «натхненність – неживий», «абстрактність – конкретність», завдяки яким можна істотно підвищити ефективність виконання, наприклад, дозволу анафори. Взагалі, актуальним є

більш широке застосування і алгоритмізація таких лінгвістичних методів, як компонентний, предікаціонний, відмінково-рольовий аналіз.

При розробці лінгвістичного програмного забезпечення використовуються найрізноманітніші мови програмування. Найбільш популярні мови групи С (С ++, С #). У США широко застосовується мова Python, зокрема в інструментальному ПО NLTK (Natural Language Toolkit), розробленому в Пенсильванському університеті [Mertz, 2004].

Статистичні методи обробки тексту

Ці методи засновані на підрахунку кількості слів в текстах. Як правило, вони досить прості і не вимагають глибоких теоретичних знань в лінгвістиці, іноді – вимагають глибоких математичних знань. Витоки цих методів лежать в математиці, в обчислювальній геометрії: текст представляється вектором, довжина вектора – кількість слів.

В цілому статистичні методи:

- вельми громіздкі;
- спираються на дуже великий досвід і розробки математиків;
- не враховують особливості самих текстів (рідко враховують);
- дають непогані результати;
- можуть дати не просто відповідь, а ще й певну кількість, яке є мірою впевненості алгоритму в цій відповіді (скажімо, 75%). Якщо оперувати цією впевненістю і самими результатами, можна досягнути дуже хороших результатів у практичних завданнях.

Точні методи обробки тексту

Якщо мова йде про видаляння інформації з тексту, «розумінні тексту», це точні методи. Їх точність при цьому вельми умовна: в якійсь мірі вони точні, в якийсь – ні. Точні алгоритми можуть «сказати», згадується в тексті якесь ім'я, дія чи ні.

Особливість статистичних методів в тому, що ми майже ніколи не можемо зрозуміти причину, по якій алгоритм зробив той чи інший висновок.

Людині доведеться просто повірити алгоритму на слово. З точним методом ми можемо зрозуміти причину набагато частіше: є конкретне правило, яке працює в конкретному випадку і дозволяє зробити такий висновок. Або алгоритм все показав правильно, або він все показав неправильно.

Штучний інтелект та автоматизація майже всіх процесів набувають стрімкого розвитку. Створення віртуальних помічників дозволяє перекласти чимало рутинних завдань на комп'ютер.

Практичні рекомендації з використання результатів наведеного вище аналізу, згідно з завданням, стосуються вдосконалення Web-сайту університету на основі семантичного аналізу текстових документів.

Кожного року в університеті з'являється тисячі нових студентів, роки йдуть, люди змінюються, але питання кожного року залишаються незмінні. Новачки розгублено намагаються знайти потрібну інформацію, абітурієнти хочуть знати актуальні дані з приводу вступу, а студенти потребують розкладу занять та екзаменів.

Створення системи технічної підтримки університету дозволить вирішити чимало проблем як студентів, так і викладачів без залучення додаткових кадрів.

Зокрема ця система має навчитися розпізнавати текст повідомлень, класифікувати їх та знаходити шляхи вирішення проблеми чи питання. Віртуальний помічник зможе підсказати розклад занять, екзаменів, документів для вступу до ВНЗ, до магістратури, строки оплати навчання, гуртожитку тощо, а при нестандартному питанню, яке потребує індивідуального рішення, буде рекомендовано звернутися до адміністрації факультету чи університету в цілому.

Це надасть змогу:

- зменшити навантаження робітників адміністрації;

- забезпечити студентів та працівників актуальною інформацією, яку вони зможуть миттєво отримати;

- підвищити рівень довіри та статус університету серед інших ЗВО

(закладів вищої освіти);

- допоможе адаптуватися іноземним студентам, які не володіють достатніми навичками комунікації на високому рівні;

- при використанні мобільних додатків надасть змогу швидкого сповіщення останніми новинами;

- створення робочих місць для студентів, які цікавляться ШІ та сучасними проектами.

3.2 Проектування системи помічника університету

Аналіз текстів на природних мовах, або Natural Language Processing, – це широкий план завдань, які можна розділити на 3 частини: Natural Language Understanding – розуміння текстів; Natural Language Generation – генерація текстів; усне мовлення – розпізнавання і синтез.

Підсистема Natural Language Understanding має самостійну цінність, коли ми робимо систему, яка автоматично аналізує текстовий контент, витягує необхідні факти про організації, продукти тощо (рис. 3.1).

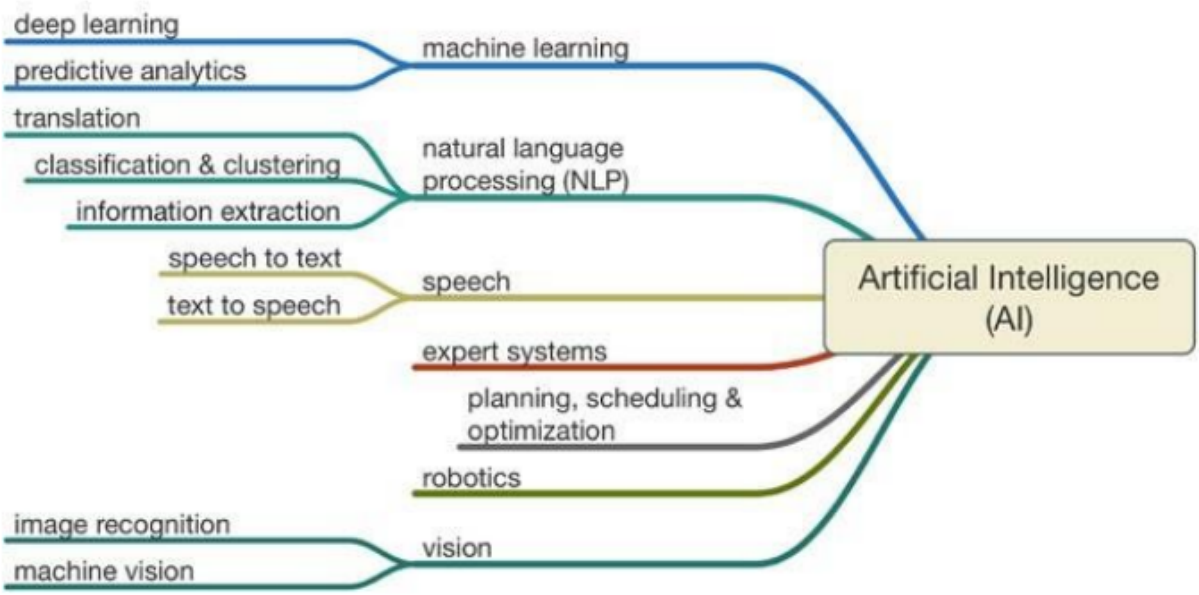

Рисунок 3.1 – Схема машинного навчання системи

Для розуміння тексту, перш за все, потрібно виділити семантику: визначити характер тексту, клас підписів, намір користувача, інтонацію, тональність і, нарешті, про що взагалі йде мова в тексті, про які об'єкти реального світу, людей, організацій, місць, даних. Завдання виділення об'єктів стосується розпізнавання іменованих сутностей (Natural Entity Recognition); визначення того, що хоче користувач, – це класифікація призначених для користувача намірів (Classification Intent); розпізнавання тональностей – це Sentiment Analysis.

Підходи до їх вирішення існують давно.

Розпізнавання іменованих сутностей з тексту, який згенерував клієнт, стосується необхідності виділення таких іменованей сутностей (Named Entities) (рис. 3.2) , як імена людей, назв організацій і геолокації, а також соціальних сутностей (більш докладних адрес, телефонів тощо).

Коли ми спілкуємося з чат-ботом, йому важливо знати, чого ми від нього хочемо. Розуміння намірів користувача дозволить визначити місце в графі діалогу і найбільш адекватну реакцію чат-бота. Інтенти (наміри користувача) можуть виражатися по-різному. Ось фрагменти реальних чатів з клієнтами бота: «Сімку, кажу, заблокуйте!» або «Можу я відключити на час номер?». Здавалося б, класифікацію інтентів можна робити за ключовими словами. Але навіть ці два простих приклади, що відносяться до одного і того ж інтенту (блокування сім-карти) взагалі не містять пересічних слів. Тому краще всього зробити систему на основі машинного навчання, яка б сама мала змогу класифікувати тексти, вибудувати та налаштовувати чат-бот на новий набір класів інтентів в залежності від того, де будуть цей чат-бот експлуатувати.

Для всіх цих завдань найефективніше будувати модель, яка навчається з учителем. Відповідно, якщо ми використовуємо людину для розмітки датасета, вона може виділити, які емоції в тексті згадано, до яких класів віднести ті чи інші призначені для користувача тексти, а до якій тональності

– запити до оператора call-центру. Після цього можна сформувати вектор ознак і на його основі отримувати вектора текстів або слів тексту і далі навчати систему.

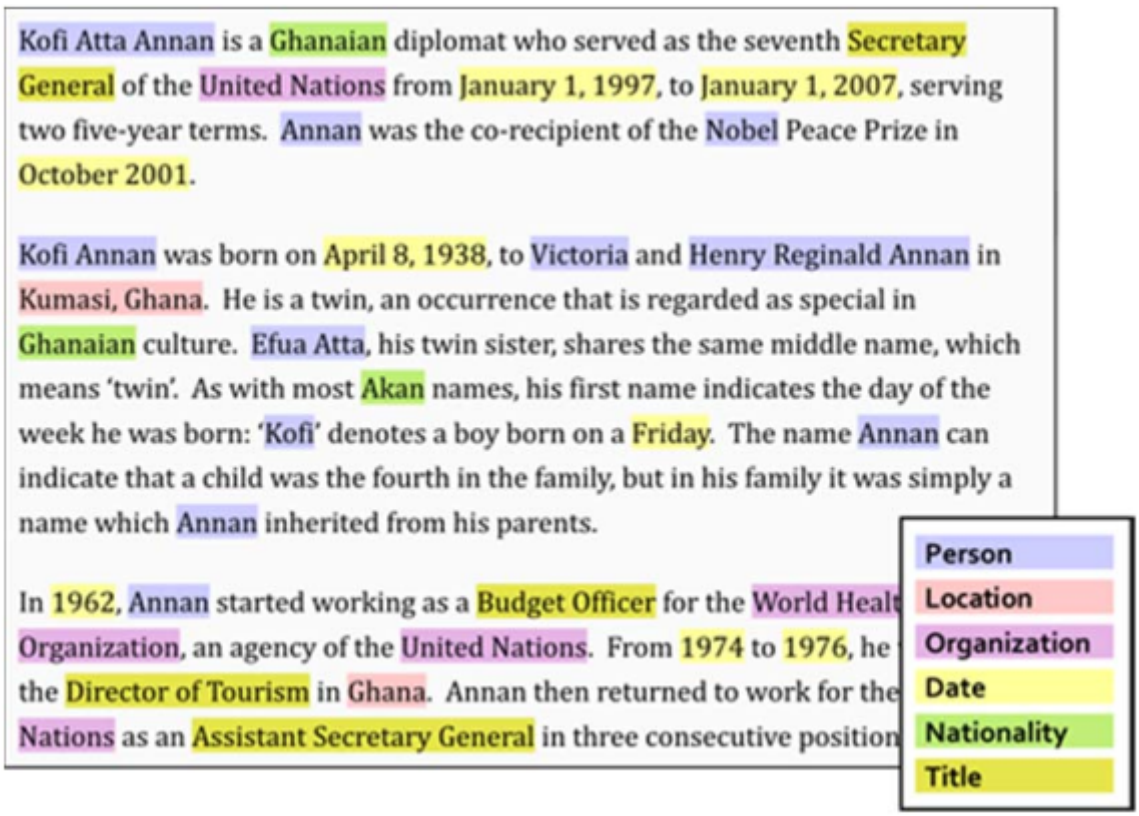

Рисунок 3.2 – Приклад іменування сутностей

Для підготовки різних текстових корпусів існує безліч інструментаріїв. Наприклад, один з найбільш популярних інструментаріїв – це brat an notation tool. Ми можемо завантажити в нього набір текстів, а самі вчителі за допомогою миші виділяють слова з тексту: наприклад, назва організації іменованої сутності класу org, іменована сутність класу money. Цей інтерфейс досить зручний та наочний, проте розрізняти великі обсяги текстів тут дуже довго і виснажливо.

Зазвичай в реальних задачах при проектуванні чат-ботів або при створенні систем аналізу контенту немає можливості використовувати різні великі корпуси. Стандартний обсяг корпусів – це 100, 200, інколи 1000, 2000 текстів, не більше. В той же час класична постановка задачі навчання з

учителем передбачає набагато більшу кількість текстів (десятки тисяч). Можна побудувати систему з хорошою навчальною здатністю і вирішувати завдання, але тут це зробити неможливо. На допомогу приходить сучасний підхід до машинного навчання, який називається Transfer Learning (перенесення навчання) – навчання великої нейронної мережі.

Нейронна мережа за рахунок своєї багатошарової структури є одночасно і класифікатором, і ілюстратором ознак. Молодші шари витягують елементарні особливості зображення: рисочки, графічні примітиви. І чим ближче до виходу з нейронної мережі, тим більше високорівневі, абстрактні елементи зображення витягуються. Якщо ми навчимо мережу розпізнаванню картинок з ImageNet (сайтом з більш ніж 14 мільйонами різних зображень, на яких можна побудувати дуже круту глибоку нейронну мережу), а потім приберемо кілька останніх шарів (голову) і залишимо тільки початкові шари (тіло), то знання, акумульовані у вигляді початкових шарів, вже дозволяють отримувати елементарні відповідні зображення. А оскільки будь-який шар складається з елементарних, ми вирішуємо завдання навчання нейромережі не з нуля. Беремо звичайне «тіло» для будь-якого завдання в цій же області, наприклад, в комп'ютерному зорі, пришиваємо нову «голову» і донавчаємо вже всю її структуру разом.

Починаючи з 2017 року в області комп'ютерної лінгвістики та аналізу тексту стала відбуватися свого роду революція. З використанням ефективних методів Transfer Learning, зокрема моделей ELMo (Embeddings from Language Models) та BERT (Bidirectional Encoder Representations from Transformers) TransferLearning дослідження в галузі комп'ютерної лінгвістики вийшли на новий рівень.

Модель ELMo дозволяє враховувати глибоку семантику в тексті, та здійснювати розпізнавання іменованих сутностей, класифікацію тексту, аналіз тональності (все ці завдання засновані на дослідженні семантики тексту та витяганні сенсу). Таким чином, можна навчити глибоку нейронну мережу мовному моделюванню та пошуку залежностей у текстах.

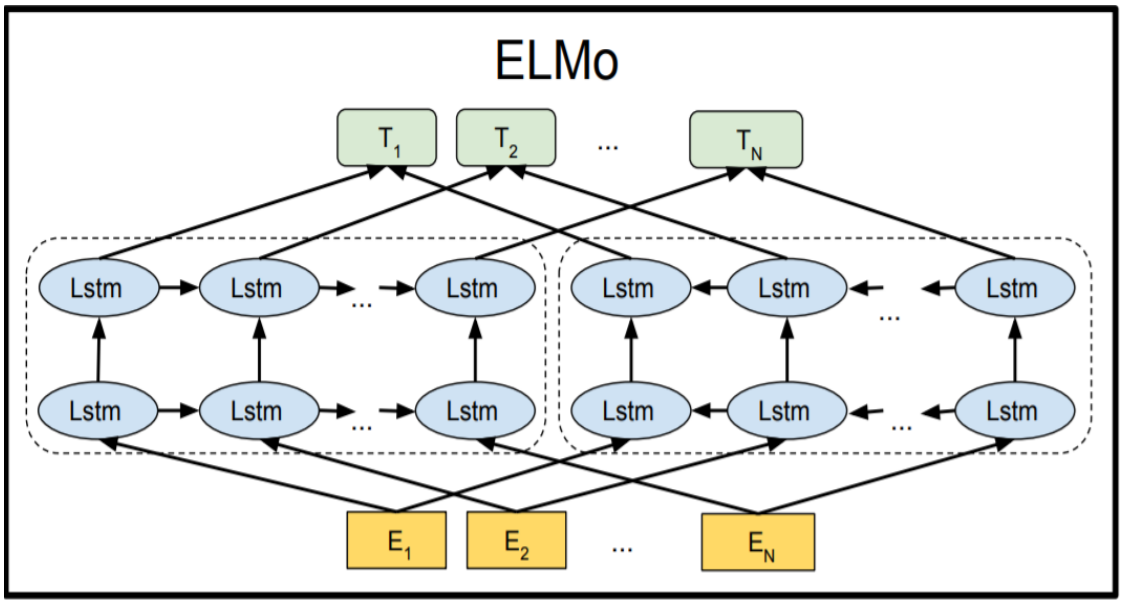

Рисунок 3.3 – архітектура моделі ELMo

Рекурентні нейронні мережі, завдяки зворотним зв'язкам, мають пам'ять і можуть враховувати аналізовану залежність скільки завгодно раз. На основі виявлених закономірностей можна вирішувати завдання мовного моделювання. Яким би довгим текст не був, ми намагаємося змусити нейронну мережу спрогнозувати наступне слово по, припустимо, двох, трьох, чотирьох словах, з цілого абзацу. Такого роду мовні моделі виявляються більш ефективними.

Таким чином, непотрібно довго робити розмітку для кожної мови, а можна генерувати мовні моделі для будь-яких мов дуже легко з використанням компьютера. Після навчання моделі в режимі експлуатації подається на вхід текст, і при обробці кожного слова в тексті модель намагається спрогнозувати наступне слово. З допомогою ELMo беруться приховані стани на всіх рівнях нейронної мережі, зворотної рекуррентной нейронної мережі, а потім формується векторне подання цього слова в тексті.

Модель ELMo враховує глибоку семантику, вирішує проблему омонімії, однак і ця модель не позбавлена недоліків. Рекурентні нейронні

мережі, з одного боку, можуть враховувати легкі залежності, а з іншого вони дуже важко навчаються і не завжди можуть добре аналізувати довгі послідовності-так чи інакше, проблема з забування у них є.

Модель BERT замість рекурентних нейронних мереж використовує модель Transformer, засновану не на так званому зворотньому зв'язку, а на механізмі уваги. При читанні тексту мимоволі виділяються ключові слова, які несуть найбільшу семантичну навантаженість.

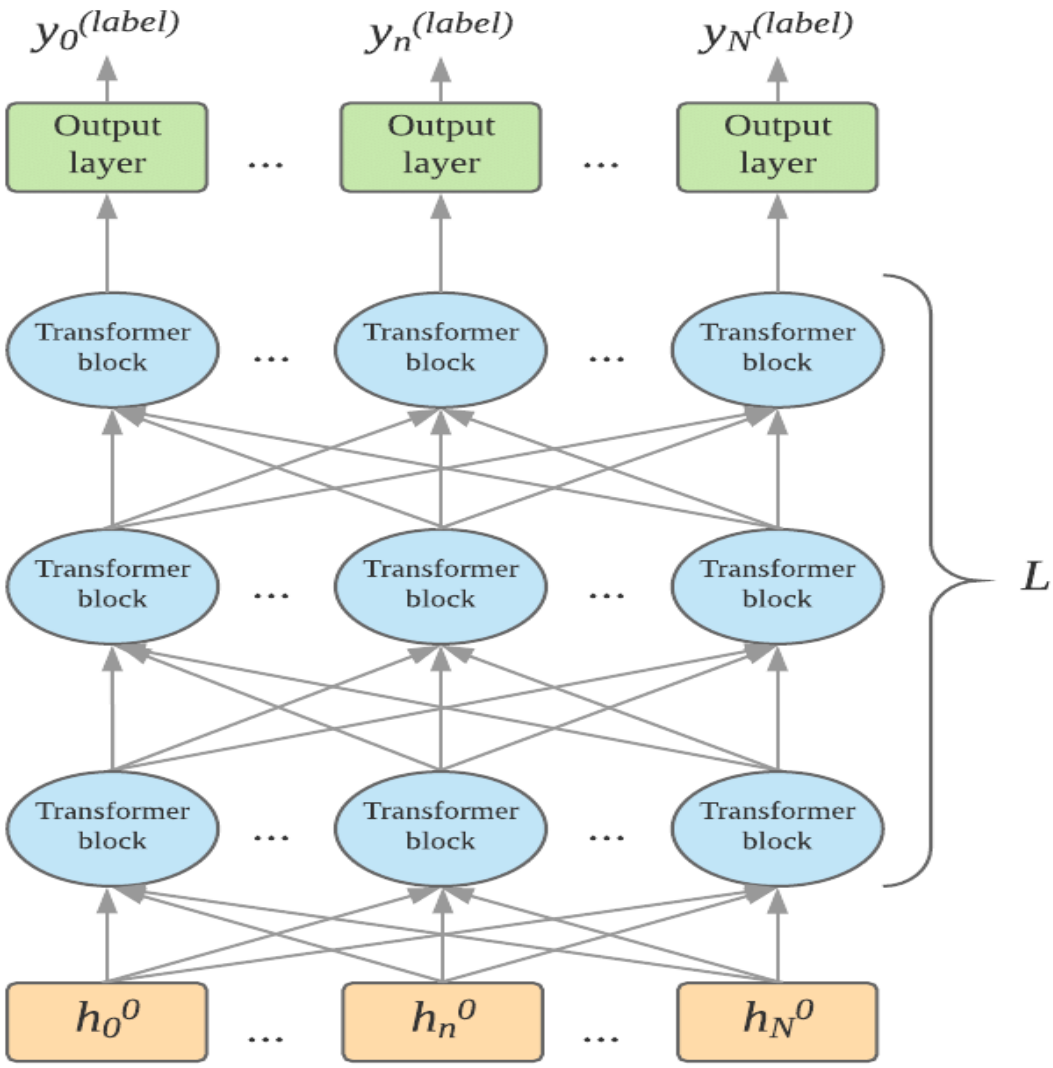

Рисунок 3.4 – Архітектура моделі BERT

Можна локалізувати якісь об'єкти на зображенні або в тексті. Механізм уваги присутній в нейронних мережах і заснований приблизно на тому ж

принципі. Певний шар зважує кожен елемент послідовності, потім – кожну послідовність в кожен момент часу. Виходить, що можна бачити, які елементи послідовності на даному етапі найбільш або найменш важливі. У Transformer використовується спеціальний варіант уваги (Multi-Head Attention). Більш того, Transformer цілком і повністю використовує не цілі слова, а квазіморфеми (BPE), тому що це виявилося більш ефективно і зручно. При цьому виділяються найбільш стійкі поєднання символів в тексті і створюються слова з символів. Проблема словника дуже гостро стоїть для мов типу української, у яких є безліч словоформ: «мама», «мамочка», «матуся», «мамі» (відмінки, зменшувально-пестливі форми та ін). Зберігання повного словника усіх словоформ є затратним та незручним. Компромісним рішенням є виділення найбільш стійких підслів в слові статистично, при цьому частотні біграми (повторення символів) замінюються на спецсимволи BPE. У підсумку формується скорочений словник BPE.

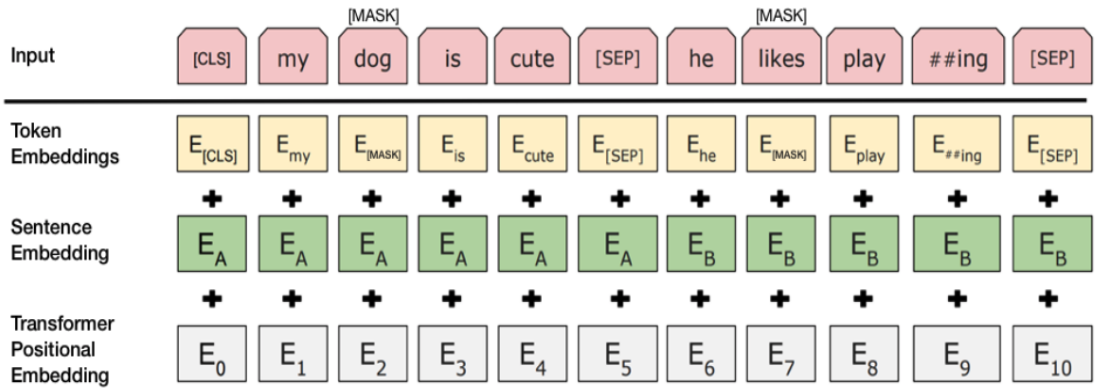

Рисунок 3.5 – Схема роботи MLM («Masked Language Modeling») у моделі BERT

Припустимо, є якийсь текст: «Let's stick to improvisation in this skit». Починається він з квазітокена, який означає початок, і в кожному тексті випадковим чином 10-15% слів замінюються на mask (спецсимвол), який є вхідним сигналом. Тобто ми змушуємо модель BERT вчитися відновлювати

пропущені слова, які замасковані під словом mask (ці слова повинні бути в тексті, в реченні, в абзаці по контексту). Для такого датасета не потрібна ніяка розмітка, адже текст сам є розміткою в даному випадку, і BERT вчиться моделювати семантику всередині тексту. Друга задача – це навчити BERT визначати, чи є друге речення логічним розвитком першого (Multi-task learning).

3.3 Етапи створення системи

Етап 1. Визначаємо тематику тексту за допомогою машинного навчання.

Спочатку потрібно побудувати дерево тематик звернень і натренувати класифікатор орієнтуватися в них. Як згадувалося вище, найкращий варіант для цього використати заздалегідь навчену модель на основі BERT. Тобто для класифікації тексту запиту потрібно представити його у вигляді векторів так, щоб схожі за змістом пропозиції лежали поруч в отриманому просторі.

Модель BERT заздалегідь навчається на двох завданнях з немаркованими текстами. Перші 15% токенів випадковим чином замінюються на [MASK], а мережа на підставі контексту прогнозує вихідні токени – це забезпечує моделі природну «двонаправленність». Друге завдання – навчити модель визначати зв'язок між реченнями (два поданих на вхід речення можуть розташовуватися поряд або бути розкидані по тексту).

Для отримання мережі, здатної передбачати тематику повідомлення з поправкою на специфіку нашого сервісу, необхідно довчити архітектуру BERT на вибірці запитів.

Однак частота тематик і самі тематики змінюються: щоб мережа оновлювалася разом з ними, окремо донавчаємо тільки нижні шари моделі на найсвіжіших даних (за останні кілька тижнів). Таким чином знання особливостей текстів підтримки будуть зберігаєтися, а ймовірності для можливих класів розподілятися адекватно поточного дня.

Етап 2. Працюємо з інформацією про питання: прописуємо бізнес-правила для кожного шаблону.

Співробітникам системи підтримки буде запропонований інтерфейс, де для кожного шаблону відповіді потрібно прописати деяке обов'язкове правило. Наприклад, для випадку з оплатою:

Шаблон: «Добрий день! Я все перевірив: навчання оплачене один раз. Гроші спершу «заморожуються» на вашій карті і тільки потім списуються, через це банк може двічі оповістити про одну операції. Будь ласка, перевірте виписку з банківського рахунку, щоб у всьому переконатися. Якщо побачите там дві списані суми, надішліть, будь ласка, скан або фото виписки »

Правило: payment_type is "card" and transaction_status is "clear_success" and transaction_sum == order_cost

Етап 3. Вибираємо відповідь: з'єднуємо відповідні тематики тексту і бізнес-правил для шаблонів

Кожній тематиці ставимо у відповідність відповідні шаблони відповідей: тематика визначається методами машинного навчання, а шаблони, які відгукуються на неї, перевіряються на істинність правилом з попереднього пункту. Користувач отримає відповідь, перевірка якого видала значення "True". Якщо таких варіантів декілька, буде обраний найбільш популярний варіант у співробітників підтримки.

Система спроектована, оптимізація прогнозує чудові результати. Автовідповіді повинні стабільно функціонувати без постійного втручання і легко масштабуватися – самостійно або в напівручному режимі.

Цього можливо досягти завдяки трьохступінчатої структури системи:

- офлайн-розробка – на цій стадії змінюються моделі, готуються правила;

- Production service – мікросервіс, який підхоплює оновлення, застосовує їх і відповідає користувачам в реальному часі;

- подальший аналіз результатів, щоб переконатися, чи нова модель працює коректно, а користувачі задоволені автовідповідачем.

Кожен з розглянутих методів має свої технології і технічні засоби. Їх розвиток дає змогу реалізувати у відповідних системах досконаліші функції та наділяє систему елементами інтелектуальної діяльності людини. Можна припустити, що найперспективнішим серед них є метод з використанням нейронних мереж, який дає змогу представити, проаналізувати семантичну інформацію тексту та працювати з нею в будь-якому ключі.

В атестаційній роботі подано модель семантичного аналізу текстів, яка може бути використана у комп'ютерно-лінгвістичній системі обробки текстової інформації. Запропонована модель дає змогу врахувати семантичні зміни (зміни порядку слів, зміни множини/роду/відмінку, вставляння слів у середину фрази), виразити характеристики тексту на рівні символів, морфем, термінів у вигляді послідовності множин.

Розглянута модель подання тексту ґрунтується на різних семантичних рівнях, що дає змогу зменшити розмір інформації, яка зберігається. Це збільшує швидкість подальшого аналізу за рахунок зменшення інформаційного навантаження тексту і кількості елементів, що обробляються.

Узагальнюючи проведені дослідження, можна зробити висновок, що лише чітке співвідношення контенту з окресленими особливостями мови дасть змогу перейти семантичному аналізу тексту на вищий рівень, більш наближений до елементів людського інтелекту.

Також варто зазначити, що перенесення навчання дозволяє ефективно вирішувати різні завдання, особливо коли немає можливості створювати великі аналітичні центри. Другий момент: при тонкій і глибокій семантиці в текстах складні моделі лідирують, в порівнянні з простими, з великим відривом.

Для програмної реалізації можна використати технологію Azure від Microsoft. Можливість розуміти, що користувач хоче сказати і який вкладає контекст, може бути складним завданням, але також може сприяти більш природній бесіді з ботом. API розпізнавання мови, так само зване LUIS,

дозволяє робити так, щоб бот міг розпізнавати наміри користувальницьких повідомлень, використовувати більш природню мову користувача і краще направляти потік спілкування. Схема створення бота для помічника університету на основі Azure (рис. 3.6).

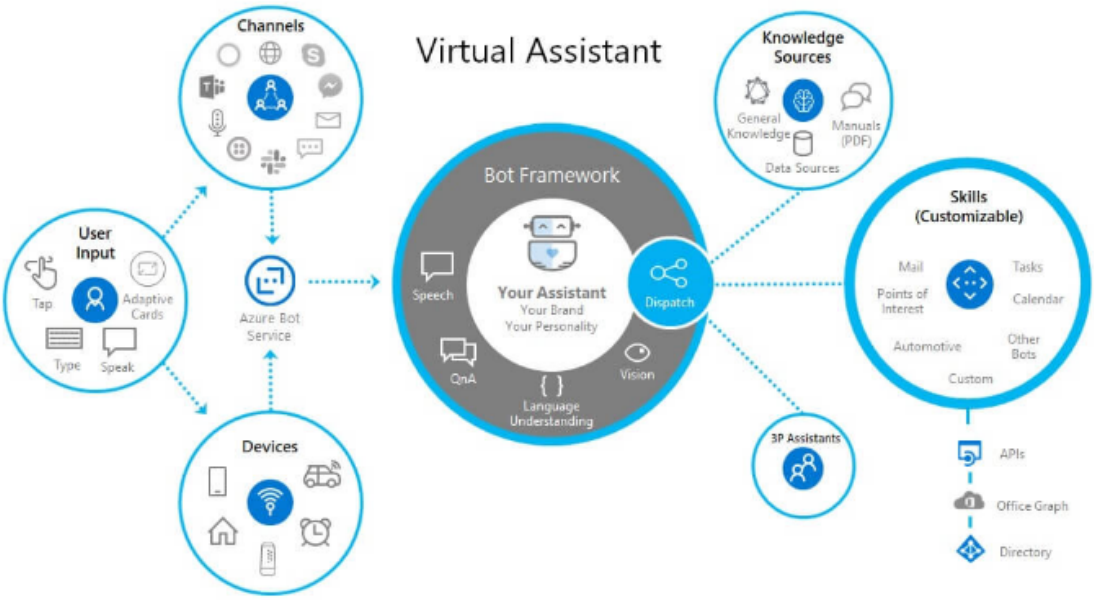

Рисунок 3.6 – Бот на основі Azure

Таким чином, на основі проведеного аналізу методів семантичного аналізу природномовних текстів у роботі запропоновано концепцію створення бота-помічника університету та визначені етапи його створення.

ВИСНОВКИ

В атестаційній роботі було розглянуто найпоширеніші методи семантичного аналізу тексту. Було проаналізовано та порівняно їх характеристики та всі недоліки і переваги для створення рекомендацій щодо інтелектуалізації Web-сайтів з використанням методів семантичного аналізу тексту.

Відповідно до поставленого завдання було виконано огляд існуючих методів семантичного аналізу тексту, системи їх використання та перспективи розвитку.

Вирішення задачі інтелектуалізації з використанням методів семантичного аналізу тексту є актуальною задачею, що стає популярнішою щодня, це є цілком виправдано, враховуючи переваги, які надають результати їх використання.

На основі проведення аналізу відомих методів та систем було створено ряд рекомендацій з побудови системи помічника університету на основі аналізу тексту та машинного навчання.

В даній роботі детально розглядалися не лише методи аналізу даних, але й їх використання в машинному перекладі, використанні Data mining та ін. Істотна проблема, з якою стикаються розробники сучасного лінгвістичного програмного забезпечення – погана якість текстів, що розміщуються в Інтернеті. До таких текстів, як чати, досить важко, а іноді неможливо, застосувати традиційні алгоритми аналізу в силу численних відхилень від норм орфографії, пунктуації та граматики. Разом з тим саме такі жанри текстів, як чати, блоги, форуми є цінним джерелом інформації і об'єктом аналізу багатьох галузей, в першу чергу, в програмах інтелектуального аналізу тексту. Розробка алгоритмів аналізу діалогічних текстів також є перспективним напрямком в рамках автоматичної обробки природної мови.

Продовженням цієї роботи може стати реалізація спроектованого онлайн помічника для загального використання в університеті, для підключення тематичного помічника до загальної системи необхідно створити нове дерево, шаблони та правили та підключити їх.

ПЕРЕЛІК ПОСИЛАНЬ

1. Семеріков С. О. Мобільне навчання: історія, теорія, методика/ Сергій Семеріков, Ілля Теплицький, Світлана Шокалюк // Інформатика та інформаційні технології в навчальних закладах. –2008. –No 6. –С. 72-82; 2009. –No 1. –С. 96-104.

2. Семеріков С. О. Мобільне навчання : історико-технологічний вимір / Семеріков С. О., Стрюк М. І., Моісеєнко Н. В. // Теорія і практика організації самостійної роботи студентів вищих навчальних закладів : монографія / кол. авторів; за ред. проф. О.А.Коновала. – Кривий Ріг : Книжкове видавництво Киреєвського, 2012. –С. 188-242.

3. Semerikov S. O. Computer Simulation of Neural Networks Using Spreadsheets: The Dawn of the Age of Camelot [Electronic resource] / Serhiy O. Semerikov, Illia O. Teplytskyi, Yuliia V. Yechkalo, Arnold E. Kiv // Augmented Reality in Education : Proceedings of the 1stInternational Workshop (AREdu 2018). Kryvyi Rih, Ukraine, October 2, 2018 / Edited by : Arnold E. Kiv, Vladimir N. Soloviev. –P. 122-147. –(CEUR Workshop Proceedings (CEUR-WS.org), Vol. 2257). [Electronic resource]. – URL: http://ceur-ws.org/Vol-2257/paper14.pdf.

4. Markova O. M. CoCalc as a Learning Tool for Neural Network Simulation in the Special Course “Foundations of Mathematic Informatics” [Electronic resource] / Oksana Markova, 68Serhiy Semerikov, Maiia Popel // ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer 2018 : Proceedings of the 13th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume II: Workshops (ICTERI, 2018). Kyiv, Ukraine, May 14-17, 2018 / Edited by : Vadim Ermolayev, Mari Carmen Suárez-Figueroa, Vitaliy Yakovyna, Vyacheslav Kharchenko, Vitaliy Kobets, Hennadiy Kravtsov, Vladimir Peschanenko, Yaroslav Prytula, Mykola Nikitchenko, Aleksander Spivakovsky. –P. 388-403. –(CEUR Workshop

Proceedings (CEUR-WS.org), Vol. 2104). –[Electronic resource]. – URL: http://ceur-ws.org/Vol-2104/paper_204.pdf.

5. Анисимов А.В.Система обработки текстов на естественном языке / А.В.Анисимов, А.А. Марченко // Искусственный интеллект. –2002. –No 4. –С. 157-163.

6. Анисимов А. В. Компьютерная лингвистика для всех: Мифы. Алгоритмы. Язык / А.В.Анисимов.–К.: Наук. думка,1991. –208 с.

7. Литвиненко О. Є.Інженерно-лінгвістичні принципи аналізу текстів / О.Є.Литвиненко, Д. А. Бурко // Наукоємні технології. –2009. –Том 3, No 3. –С.60-62.–DOI: 10.18372/2310-5461.3.5130

8. Заболеева-Зотова А. В. Латентный семантический анализ: новые решения в Internet / А. В. Заболеева-Зотова, А. Ю. Пастухов, П. В. Сердюков, Н. А. Козлова, С. А. Чернов // Информационные технологии. –2001. –No 6. –С. 67-82.

9. Ландэ Д. В. Поиск знаний в Internet / Д. В. Ландэ.–М.: Вильямс, 2005. –272 с.–(Профессиональная работа)

10. Сокірко А. В. Семантичні словарі в автоматичній обробці тексту: По матеріалам системи ДІАЛІНГ. М. 2001. 120 с.

11. Ліонтова Н.Н. Російський загальносемантичний словар (ROSS): структура, наполнение. NTI. Сер. 2. 1997. № 12. С 5-20

12. Єрмаков А.Й. Тезіси докладу міжнародного конгресу "Російська мова: історична доля та сучасність"

13. Проект “Minerva“ [Electronic resource]. – URL: http://www.inteltec.ru/publish/articles/textan/concept.shtml

14. Тихомиров І.А. Матеріальна міжнародна конференція "Діалог 2009"

15. Open Cognition – [Electronic resource]. – URL: http://opencog.org/

16. Link Grammar Parser. AbiWord, 2014 – [Electronic resource]. – URL: http://www.abisource.com/projects/link-grammar/

17. The CMU Link Grammar natural language parser – [Electronic resource]. – URL: https://github.com/opencog/link-grammar/

18. RelEx Dependency Relationship Extractor. OpenCog – [Electronic resource]. – URL: http://wiki.opencog.org/wikihome/index.php/Relex

19. Сокирко А.В. Семантичні словники в автоматичній обробці тексту (за матеріалами системи ДІАЛІНГ): дис. ... канд. тех. наук. М .: МГПИИЯ, 2001. 120 с.

20. OpenCalais – [Electronic resource]. – URL: http://www.opencalais.com/ opencalais-api /

21. RCO – [Electronic resource]. – URL: http://www.rco.ru/?page_ id = 3554

22. Abbyy Compreno – [Electronic resource]. – URL: https: // www.abbyy.com/ru-ru/isearch/compreno/

23. SemSin– [Electronic resource]. – URL: http://www.dialog-21.ru/media/1394/kanevsky.pdf

24. DictaScope – [Electronic resource]. – URL: http://dictum.ru/

25. Pullenti – [Electronic resource]. – URL: http://semantick.ru/

26. Інструментальне середовище «Декла» – [Електронний ресурс]: – URL: http://ipiran-logos.com/

27. Національний корпус російської мови – [Електронний ресурс]:– URL: http:// www.ruscorpora.ru/

28. Yago – [Electronic resource]. – URL: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

29. DBpedia [Electronic resource]. – URL: http: //wiki.dbpedia.org/

30. Freebase – [Electronic resource]. – URL: https: // developers.google.com/freebase/

31. Google's Knowledge Graph [Electronic resource]. – URL: https://developers.google.com/know-ledge-graph/

32. OpenCyc [Electronic resource]. – URL: http: // www. opencyc.org/

33. ReadTheWeb [Electronic resource]. – URL: http://rtw.ml.cmu.edu/rtw/

34. OpenIE – [Electronic resource]. – URL: http://nlp.stanford.edu/software/openie.html

35. Google Knowledge Vault [Electronic resource]. – URL: https://www.cs.ubc.ca/~murphyk/Papers/kv-kdd14.pdf

36. Berners-Lee T., Hendler J., Lassila O. // Scientific American. 2001. May. P.28-37.

37. Вишняков В.А., Бородаєнко Ю.В., Бородаєнко Д.С. Моделі і засоби інтеграції додатків, маркетингу, аутсорсингу, обробки знань в комп'ютерних мережах: монографія. Мінськ, 2011 року.

38. Allemang D., Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. USA, 2008.

39. Ньюкомер Е. Веб-сервіси. XML, WSDL, SOAP і UDDI. Для професіоналів. СПб, 2010 року.

40. Вишняков, В.А. // Економіка та управління. 2012. № 4 (32). С. 111-117.

41. Web Mining [Электронный ресурс] // Businessdataanalytics.ru. 2008, грудень. – URL: http://www.businessdataanalytics.ru/WebMining.html

42. URL: http://wiki.opencog.org/wikihome/index.php/Web [Електронний

43. Kilgarrif A. BNC database and word frequency lists / A. Kilgarrif – 1998 [Electronic resource]. – URL: http://www.kilgarrifF.co.uk/bnc-readme.html

44. Tokenizer : Opennlp [Electronic resource]. – SourceForgeNet, 2010. – URL: http://sourceforge.net/apps/mediawiki/opennlp/index.php?title=Tokenizer

45. BNC part of speech tags [Electronic resource]. – URL: http://pie.usna.edu/POScodes.html

46. Яцко, В.А. Симметричное реферирование : теоретические основы и методика [Текст] / В.А. Яцко // Научно-техническая информация. Сер.2., 2002. С. 18-28.

47. Яцко В.А. Алгоритмы предварительной обработки текста: декомпозиция, аннотирование, морфологический анализ [Текст] / В.А. Яцко, М.С. Стариков, Е.В. Ларченко [и др.] // Научно-техническая информация. Сер.2., 2009. – С. 8-18.

48. Марчук Ю.Н. Компьютерная лингвистика М.: Изд-во Восток-Запад, 2007г., – 317 с –[Електронний ресурс]: – URL: http://www.twirpx.com/file/398578/

49. Clark A. The Handbook of Computational Linguistics and Natural Language Processing /Clark A., Fox C., Lappin S.// Blackwell Publishing, 2010, – 775p [Electronic resource]. – URL: stp.lingfil.uu.se/~santinim/sais/ClarkEtAl2010_HandbookNLP.pdf

50. Волошин В.Г. Комп'ютерна лінгвістика: Навчальний посібник. – Суми: Університетська книга, 2004. –382 с.

51. Mitkov R. The Oxford handbook of computational linguistics /Oxford University Press, 2003 –786 p. [Electronic resource]. – URL: http://www.google.com.ua/books?hl=uk&lr=&id=yl6AnaKtVAkC&oi=fnd&pg=PP2&dq=5.%09The+Oxford+handbook+of+computational+linguistics

52. Nirenburg S. Ontological semantics / Nirenburg S., Raskin V. //MIT Press, 2004, – 420 p – [Electronic resource]. – URL: http://books.google.com.ua/books/about/Ontological_semantics.html?id=OPek3LpMIigC&redir_esc=y

53. Jurafsky D. Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition, / Jurafsky D., Martin J.// 2009 – [Electronic resource]. – URL: http://www.cse.iitk.ac.in/users/mohit/Speech-and-Language-Processing.pdf

URL: http://wiki.opencog.org/wikihome/index.php/Relex