НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Факультет прикладної математики

Кафедра прикладної математики

«До захисту допущено»

Завідувач кафедри

_____ О. Р. Чертов

«___» _____ 2016 р.

Дипломна робота

на здобуття ступеня бакалавра

з напряму підготовки 6.040301 «Прикладна математика»

на тему: Метод побудови специфічних словників для розпізнавання мовленнєвих сигналів за тематикою

Виконав: студент IV курсу, групи КМ-21

Коцовський Роман Ігорович _____

Керівник асистент _____

Громова В. В.

Консультант із нормоконтролю старший викладач _____

Мальчиков В. В.

Рецензент проф., канд. техн. наук, проф. _____

Корнійчук В. І.

Засвідчую, що в цій дипломній роботі немає запозичень із праць інших авторів без відповідних посилань.
Студент_____

Київ — 2016

Національний технічний університет україни

«Київський політехнічний інститут»

Факультет прикладної математики

Кафедра прикладної математики

Рівень вищої освіти — перший (бакалаврський)

Напрям підготовки 6.040301 «Прикладна математика»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ О. Р. Чертов

«___» _____ 2016 р.

ЗАВДАННЯ

на дипломну роботу студенту

Коцовському Роману Ігоровичу

1. Тема роботи: «Метод побудови специфічних словників для розпізнавання мовленнєвих сигналів за тематикою»,

керівник роботи Громова Вікторія Вікторівна, асистент,

затверджені наказом по університету від «6» травня 2016 р. № 1499-с.

2. Термін подання студентом роботи: «8» червня 2016 р.

3. Вихідні дані до роботи: розроблювана система повинна працювати з текстовими файлами що закодовані в фоматі Windows 1251 та UTF-8. Розмір файлів не повинен перевищувати 100 КБ. Їх назви повинні бути латинськими літерами.

4. Зміст роботи: виконати аналіз існуючих методів розв'язання задачі, спроектувати автоматизовану систему побудови тематичних словників, здійснити програмну реалізацію розробленої системи, провести тестування розробленої системи. Результатом роботи є тематичні словники що представляють собою текстові файли з тематичними словами.

5. Перелік ілюстративного матеріалу: зображення, що відображають результати роботи нормалізації текстів, підрахунку метрик, створення ФКС та процедури

розбиття текстів по тематикам, схема взаємодії модулів системи, зображення, що відображають результати роботи розробленого методу побудови специфічних словників.

6. Дата видачі завдання: «22» лютого 2016 р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів роботи	Примітка
1	Аналіз стану проблеми розпізнавання природної мови в Україні	18.04.2016	
2	Огляд літератури за тематикою та збір даних	21.04.2016	
3	Проведення порівняльного аналізу математичних методів	25.04.2016	
4	Підготовка матеріалів першого розділу роботи	30.04.2016	
5	Розроблення математичного забезпечення для побудови специфічних словників	06.05.2016	
6	Підготовка матеріалів другого розділу роботи	10.05.2016	
7	Підготовка матеріалів третього розділу роботи	12.05.2016	
8	Підготовка матеріалів четвертого розділу роботи	15.05.2016	
9	Розроблення програмного забезпечення	17.05.2016	
10	Оформлення пояснювальної записки	19.05.2016	

Студент _____ Коцовський Р. І.

Керівник роботи _____ Громова В. В.

АНОТАЦІЯ

Дипломну роботу виконано на 45 аркушах, вона містить 2 додатки та перелік посилань на використані джерела з 26 найменувань. У роботі наведено 3 рисунки та 6 таблиць.

Метою даної дипломної роботи є створення математичного та програмного забезпечення для побудови специфічних словників для розпізнавання мовленнєвих сигналів за тематикою.

У роботі проведено аналіз існуючих систем розпізнавання злитого мовлення та словників, які ними використоуються. Виконано їх порівняння з погляду точності отримуваних розв'язків, вартості, призначення, а також об'єму словника.

Сформовані вимоги до нормалізації текстів. Для кожної кожної тематики побудовано відповідні словники. Розроблено автоматизовану систему, що будує тематичні словники. Виконано тестування розробленої системи.

Ключові слова: корпуса текстів, файли ключових слів, нормалізація, лематизація, лінгвістична модель, метрика TFIDF, кластеризація.

ABSTRACT

The thesis is presented in 45 pages. It contains 2 appendixes and bibliography of 26 references. Three figures and 6 tables are given in the thesis.

The goal of the thesis is to develop mathematical and software tools for solving the problem of building specific dictionaries for recognition of speech signals for various subjects.

In the thesis, existing systems Continuous Speech Recognition and related dictionaries are analyzed. Was done comparing them in terms of the accuracy of the resulting solutions, value, purpose and scope of the dictionary.

Formed requirements for normalization of texts. For each subject creating corresponding dictionaries. Developed an automated system that builds thematic dictionaries. Tests developed system.

Keywords: text corpus, Files keywords, normalization, Lemmatisation, linguistic model, TFIDF metric, clustering.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ 8

ВСТУП 9

1 ПОСТАНОВКА ЗАДАЧІ 11

2 ОГЛЯД ІСНУЮЧИХ РІШЕНЬ 13

2.1 Класифікація методів, моделей і систем розпізнавання мовлення 13

2.2 Існуючі засоби розпізнавання 15

2.3 Висновки до розділу 19

3 ЗАГАЛЬНА СТРУКТУРА СИСТЕМИ ТЕМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ 21

4 МАТЕМАТИЧНІ МЕТОДИ РОЗВ'ЯЗКУ ПОСТАВЛЕНОЇ ЗАДАЧІ 22

4.1 Кластеризація, основні поняття і цілі 22

4.1.1 Класифікація кластеризації 23

4.1.2 Міра відстані між об'єктами 24

4.1.3 Алгоритм кластеризації — k–середніх 25

4.1.4 Модифікація алгоритму K-середніх для відбору корисних текстів із великихкорпусів документів 27

4.2 Визначення основних етапів категоризації текстів 28

4.3 Метрика TF-IDF 29

4.4 Висновки до розділу 31

5 ПОБУДОВА ТЕМАТИЧНИХ СЛОВНИКІВ 32

5.1 Загальний принцип побудови специфічних словників 32

5.2 Текстові корпуси 33

5.3 Процес нормалізації 35

5.4 Використання морфології 35

5.5 Формування вектора ключів 36

5.6 Процедура автоматичного розбиття текстів по тематикам 37

5.7 Формування словників 38

5.8 Висновки до розділу 39

6 РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ 40

ВИСНОВКИ 42

ПЕРЕЛІК ПОСИЛАНЬ 43

Додаток А Лістинги програм 45

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

АМ – акустична модель.

АСРЗМ – автоматична система розпізнавання злитого мовлення

КВ – контрольна вибірка

Корпус текстів — це вид корпусу даних, одиницями якого є тексти або їх достатньо значні фрагменти, що включають, наприклад, якісь повні фрагменти макроструктури текстів даної проблемної області.

Лематизація – процес переведення словоформи до леми — її нормальної (словарної) форми.

Лінгвістична модель — побудована на основі лінгвістичного моделювання сукупність символьних (лінгвістичних) послідовностей за обраними параметрами лінгвістизації та відновлена на її основі формальна граматика.

ЛМ – лінгвістична модель

МННЦ ІТіС – Міжнародний науково-навчальний центр інформаційних технологій і систем

HTK(Hidden markov model ToolKit) – інструментарій для побудови прихованих марковських моделей.

ФКС – файл ключових слів

ФКТ – файл ключових текстів

ВСТУП

В даний момент існує дуже багато мовленнєвої інформації що зберігається в цифрових звукових форматах. Значним мінусом такого зберігання є трата значної кількості пам'яті та повільний доступ і обробка інформації. Наприклад, для пошуку того що вже було записано потрібно буде повністю прослухати запис і це за умови якщо відомо де саме міститься необхідна інформація. В загальному випадку прийдеться прослуховувати практично всі записи що займе дуже багато часу. Значно полегшити цю задачу можна виконавши перетворення мовленнєвої інформації зі звукового формату в текстовий за допомогою АСРЗМ.

Лінгвістична модель мови є необхідним етапом у побудові АСРЗМ оскільки вона має бути настроєна на словник та тему повідомлення що розпізнається. Сучасні алгоритми АСРЗМ використовують статистичні методи для побудови лінгвістичної моделі за текстами, які найбільш схожі до розпізнаваних мовленнєвих повідомлень. Загалом, можуть бути задіяні будь-які тексти, але необхідно вибрати саме ті, що містять корисну інформацію за схожою тематикою.

В Україні основною установою, що займається розпізнаванням мовлення, є МННЦ інформаційних технологій і систем НАН України та МОН України. На даному етапі розвитку мовленнєвих технологій ще є не вирішені задачі, які можна і потрібно вирішувати. Зокрема задача розпізнавання злитої української мови в достатній мірі не вирішена. Однією з причин є те, що українська природна мова є слабко структурованою. Крім того, не аби яку роль відіграє вміст словника системи розпізнавання, який лише на якийсь відсоток співпадає зі словником мовленнєвого сигналу. Програма ж може розпізнавати ті слова, які містяться в словнику системи розпізнавання.

Саме тому досить актуальним є створення тематичних словників які будуть орієтуватися на тематику мовленнєвих сигналів. Крім того такі словники по своєму

обєму значно менші ніж загальні словники внаслідок чого зменшується похибка та прискорюються результати обчислень.

Не зважаючи на значні результати при дослідженні та розробці систем диктування тексту [1], актуальною є розробка автомазованої системи побудови тематичних словників для МННЦ ІТіС.

1 ПОСТАНОВКА ЗАДАЧІ

В ході виконання дипломної роботи потрібно розробити специфічні словники для розпізнавання мовленнєвих сигналів за тематикою для існуючої АСРЗМ в МННЦ ІТіС.

Метою досліджень є вдосконалення процесу розпізнавання мовленнєвих сигналів, забезпечення точності і адекватності стенографування у відповідності до голосового запису.

Науково-практична задача, що розв'язується в даній бакалаврській роботі, включає наступні завдання:

1) аналіз стану проблеми розпізнавання природної мови в Україні та за кордоном;
2) приведення текстів до відповідних форматів даних та їх нормалізація;
3) обрахунок метрик і створення ФКС
4) обмежену лематизацію
5) автоматизацію розбиття текстів на кластери
6) автоматичне розбиття текстів на теми
7) формування словників
8) отримання та аналіз результатів розпізнавання з використанням запропонованих тематичних словників.

Створені тематичні словники повиненні інтегруватись з існуючою в МННЦ ІТіС системою розпізнавання злитого мовлення.

Також потрібно проаналізувати ефективність тематичних словників порівняно із загальними словниками та обґрунтувати доцільність їх використання.

Повинен бути розроблений програмний застосунок, для автоматизації вищезгаданих процесів.

2 ОГЛЯД ІСНУЮЧИХ РІШЕНЬ

2.1 Класифікація методів, моделей і систем розпізнавання мовлення

Кожна система розпізнавання має деякі завдання, які вона покликана вирішувати і комплекс підходів, які застосовуються для вирішення поставлених завдань. Далі представлені основні ознаки, за якими можна класифікувати всі системи розпізнавання природної мови та описано вплив цих ознак на роботу системи.

Однією з них є розмір словника. Словник голосового повідомлення (мовленнєвого сигналу) на якийсь відсоток співпадає зі словником системи розпізнавання. Стверджують, що чим більше розмір словника, який закладений в систему розпізнавання, тим більше частота помилок при розпізнаванні слів системою. Наприклад, словник з 10 цифр може бути розпізнаний практично безпомилково, тоді як частота помилок при розпізнаванні словника в 100 000 слів може досягати 45% [2]. З іншого боку, більший словник охоплює більше слів, які диктор може подати на систему розпізнавання, а розпізнавання невеликого словника може дати велику кількість помилок розпізнавання, якщо в цьому словнику багато подібних між собою слів або ж словник голосового повідомлення сильно відрізняється від словника системи.

Наступна ознака – в залежності від типу мовлення, – роздільна або злита мова. Якщо при мовленні кожне слово відділяється від іншого ділянкою тиші, то кажуть, що ця мова – роздільна. Злита мова – це природно виголошені речення. Розпізнавання злитої мови набагато важче у зв'язку з тим, що межі окремих слів не чітко визначені і їх вимова іноді сильно спотворена змазуванням вимовлених звуків [2].

Ще одна ознака – дикторозалежність або дикторонезалежність системи. За визначенням, дикторозалежна система призначена для використання одним користувачем (людиною, яка навчала цю систему), в той час як дикторонезалежна

система призначена для роботи з будь-яким диктором [2]. Дикторозалежні системи налаштовуються на параметри того диктора, на прикладі якого навчаються (тренуються). Налаштування на голос диктора дикторозалежних систем займає звичайно від 30 хвилин до декількох годин. Системи розпізнавання мови, яким властива відносна незалежність від диктора, дозволяють користувачу працювати без попереднього налаштування. Крім того, існують системи з адаптацією до голосу диктора. Надійність таких систем розпізнавання поліпшується після навчання. Незалежність від диктора в таких системах зазвичай досягається за рахунок збереження звукових еталонів для усіх найбільш типових голосів носіїв мови. Це, безумовно, вимагає в кілька разів більшої продуктивності й обсягу пам'яті.

За призначенням системи розпізнавання мови поділяються на системи диктування тексту, голосові інтерфейси (командні системи), системи розшифрування записів, які попередньо збережені на цифрових носіях, тощо [3]. Призначення системи визначає необхідний рівень абстракції, на якому відбуватиметься розпізнавання вимовленого людиною тексту. У командній системі (наприклад, голосовий набір в стільниковому телефоні) частіше за все, розпізнавання слова чи фрази відбувається як розпізнавання єдиного мовного елемента. А системи диктування тексту потребують більшої точності розпізнавання і тому при інтерпретації вимовленої фрази буде враховуватися і аналізуватися не тільки те, що було сказано в поточний момент, а й те, як воно співвідноситься з тим, що було вимовлено до цього. Також, в системі повинен бути вбудований набір граматичних правил, яким повинен задовольняти вимовлений і розпізнаний текст. Чим суворіші ці правила, тим простіше реалізувати систему розпізнавання і тим обмеженішим буде набір висловлювань, які вона зможе розпізнати [1].

2.2 Існуючі засоби розпізнавання

Системи диктування тексту є дуже привабливими на даному етапі розвитку суспільства в силу новизни наданих користувачу можливостей.

Такі системи дозволяють користувачам мовленнєвий сигнал, записаний у звуковому файлі, перетворювати в звичайний текст.

На сьогоднішній день на ринку можна знайти достатньо комерційних систем розпізнавання мови, від систем дискретного диктування тексту до систем, що здатні розпізнавати злите мовлення, наприклад:

a) Dragon NaturallySpeaking, ViaVoice, Voice_PE (Voice Personal Edition) – для англійської мови;
b) «Горинич» – для російської мови;
c) Існуюча в МННЦ ІТіС система для розпізнавання української мови.

Крім того, відомим і популярним на сьогоднішній день є голосовий пошук Google, який має можливість розпізнавати запити в пошукову систему українською мовою і показує хороші результати по якості розпізнавання, питально-відповідна система Siri, адаптована для iOS, що спілкується природною мовою, щоб відповідає на питання і дає рекомендації. Однак ці системи не призначені для введення великих текстів, а обмежені розпізнаванням ізольованих слів або невеликого речення.

Характеристики існуючих програмних та апаратно-програмних засобів, які призначені для диктування тексту чи комп'ютерного документування усних виступів на конференціях, нарадах та інших подібних заходах представлений в таблиці 2.1.

Таблиця 2.1 – Порівняльна таблиця існуючих засобів розпізнавання

Назва фірми / Назва системи	Для кого призначена	Словник	Точність розпізнава- ння	Вартість	Особливості (переваги та недоліки)
Scansoft / Dragon NaturallySpeaking [4]	Диктування будь-яких текстів англійською, французькою, німецькою, іспанською, японською мовами	Можна експортувати/ імпортувати будь-які списки слів до 230 000 слів.	94 – 98%	Від 150$ до 600$	Переваги: зручна, надійна система, має досить широкий функціонал. Недоліки: не зручне введення чисел.
IBM / IBM ViaVoice [5]	Диктування будь-яких текстів англійською, іспанською та французькою мовами	64 000 слів	92 – 96%	Від 149$	Переваги: хороша якість розпізнавання. Можна диктувати в будь-який текстовий редактор. Недоліки: складності з деінсталяцією
IBM / VoiceType [6]	Диктування будь-яких текстів англійською, іспанською та французькою мовами	25 000 слів	82 – 90%	7 днів безкоштовна версія, далі 45$ реєстрація	Переваги: хороше розпізнавання простих слів. Недоліки: низька точність розпізнавання власних назв і скорочених слів.

Продовження таблиці 2.1

Назва фірми / Назва системи	Для кого призначена	Словник	Точність розпізнава- ння	Вартість	Особливості (переваги та недоліки)
IBM / MedSpeak [7]	Для диктування звітів лікарів- радіологів	25 000 слів	95 – 98 %	4495$	Переваги: сама висока безпомилковість розпізнавання, зручність у використанні, дикторонезалежність. Недоліки: словник системи обмежений набором специфічних медичних термінів.
Kurzweil / Voice_PE (Personal Edition) [8]	Для диктування текстів англійською мовою.	30 000 слів	90 - 97%	295$ (ціна, рекомендована виробником)	Переваги: дикторонезалежність, високий відсоток розпізнавання навіть без навчання системи, простота використання. Недоліки: –
Lernout & Hauspie / Voice Xpress Professional [9]	Для диктування текстів англійською мовою.	30 000 слів	80%	150 $	Переваги: висока якість розпізнавання чисел, зручність використання. Недоліки: дикторозалежність, нерівномірна якість розпізнавання.
Sakrament / "Сакрамент" [10]	Для диктування текстів білоруською, російською, українською мовами	до 10 000 слів	95-98%	–	Переваги: дикторонезалежність, зручність використання. Недоліки: розпізнавання коротких фраз, додаткові словники створюються лише по замовленню.

Продовження таблиці 2.1

Назва фірми / Назва системи	Для кого призначена	Словник	Точність розпізнава- ння	Вартість	Особливості (переваги та недоліки)
VoiceLock / "Горинич" ПРОФ 3.0 [11]	Для диктування текстів російською мовою.	10 000 слів з можливістю поповнення	75 - 85%	–	Переваги: можливість поповнювати словник. Недоліки: для прийнятної якості розпізнавання мови необхідно тривале навчання (начитування мовної бази). Говорити потрібно чітко, монотонно.
РеалСпікер Лаб / RealSpeaker [12]	Для диктування текстів англійською, іспанською, українською та іншими мовами.	–	Близько 70% для українських слів	Від 5$ до 50$ в залежності від терміну дії ліцензії	Переваги: можна вводити тексти в будь- який текстовий редактор. Недоліки: розпізнавання ізольованих слів з явно вираженими паузами між словами.
Існуюча в МННЦ ІТіС система	Для диктування будь-яких текстів, зокрема, текстів на суспільно-політичну тематику.	100 000 слів	80 – 85 %	–	Переваги: дикторонезалежність, розпізнавання української мови.

Варто відзначити, що системи диктування тексту на Заході знайшли своє практичне застосування в медицині. Це в першу чергу пов'язано з тим, що наукові дослідження в цій галузі добре фінансуються. Крім того задача тут спрощується тим, що словники медичних термінів в вузькій предметній області мають менший об'єм.

Усі вищезгадані системи в принципі використовують одні і ті ж методи та алгоритми. Різниця в об'ємі словника, типі мовлення та інших характеристиках обумовлена лише специфікою конкретної задачі та обмеженнями на швидкість обчислень та об'єм необхідної пам'яті.

Більшість з цих систем зручні в експлуатації і мають досить непогані можливості, але не вміють працювати з українською мовою. А от розроблена в МННЦ ІТіС система розпізнаваня мовлення дозволяє користувачам мовленнєві сигнали записані українською мовою перетворювати в текст (створювати стенограми українською мовою). Точність розпізнавання цієї системи дещо відстає від провідних закордонних розробок, однак це пояснюється специфікою української мови та тим, що українська мова є слабко структурованою.

2.3 Висновки до розділу

У розділі здійснено огляд існуючих рішень для задачі розпізнавання природної мови. Для кращого розуміння області досліджень проведено класифікацію методів і систем розпізнавання за основними ознаками та встановлено вплив цих ознак на роботу системи.

Досліджено методи розпізнавання мовлення. Аналіз існуючих методів розпізнавання мовлення показав, що широке застосування отримали алгоритми, в

яких для моделювання застосовуються приховані марковські моделі (ПММ), оскільки використання такого методу дає кращі результати в рамках запланованих досліджень.

Досліджено існуючі на ринку системи диктування текстів в Україні та за кордоном. Аналіз основних характеристик цих систем показав, що в даний час не існує універсальної системи, яка була б здатна до самонавчання, була б дикторонезалежною, стійкою до шумів, розпізнавала б злите мовлення, була б здатна працювати зі словниками великих розмірів і при цьому мала б низьку частоту появи помилок. Крім того, було встановлено, що більшість існуючих систем не вміють працювати з українською мовою, оскільки вона має свої особливості та є слабко структурованою.

Існуючу в МННЦ ІТіС систему, яка працює з українською мовою, планується вдосконалити з метою підвищення точності та забезпечення адекватності стенографування у відповідності до голосового запису.

3 ЗАГАЛЬНА СТРУКТУРА СИСТЕМИ ТЕМАТИЧНОГО РОЗПІЗНАВАННЯ МОВЛЕННЯ

Заздалегідь створюються акустичні та лінгвістичні моделі. Лінгвістичні моделі будуються окремо для кожної тематики. Корпуси текстів тематик створюються за допомогою автоматичного розбиття корпусу текстів, завантажених із Інтернет сайтів, за допомогою файлів ключових слів [16].

На Рисунку 3.1 представлена структурна схема тематичного розпізнавання мовлення існуючої в МННЦ ІТіС АСРМ .

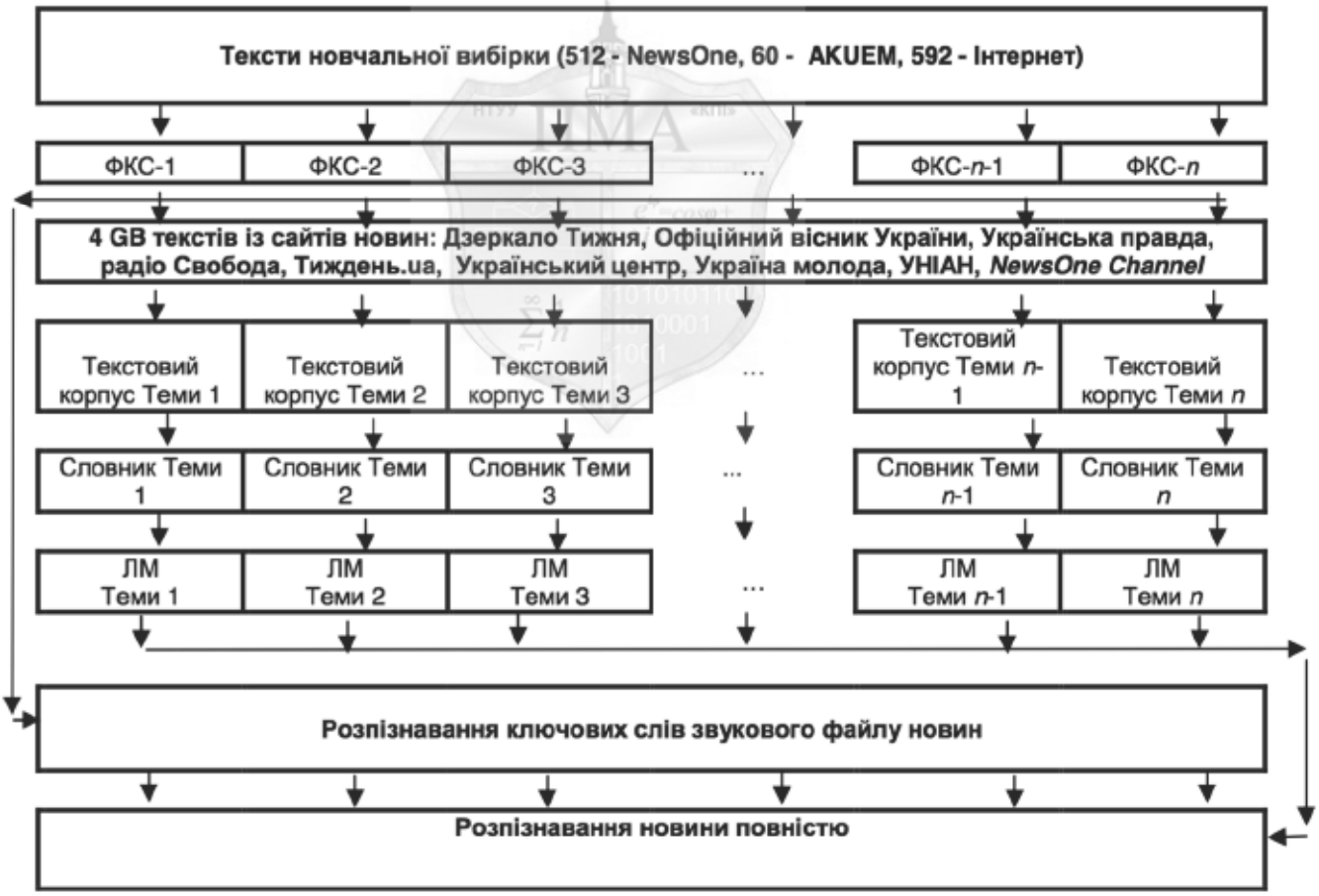

Рисунок 3.1 – Структурна схема тематичного розпізнавання мовлення

4 МАТЕМАТИЧНІ МЕТОДИ РОЗВ'ЯЗКУ ПОСТАВЛЕНОЇ ЗАДАЧІ

4.1 Кластеризація, основні поняття і цілі

Кластеризація (або кластерний аналіз) — це задача розбиття множини об'єктів на групи, звані кластерами. Усередині кожної групи повинні виявитися "схожі" об'єкти, а об'єкти різних груп повинні бути відмінні один від одного. Головна відміна кластеризації від класифікації полягає в тому, що перелік груп чітко не заданий і визначається в процесі роботи алгоритму [17].

Кластерний аналіз виконує такі основні завдання:

1. Розробка типології або класифікації.
2. Дослідження корисних концептуальних схем групування об'єктів.
3. Породження гіпотез на основі дослідження даних.
4. Перевірка гіпотез або дослідження для визначення, чи дійсно типи (групи), виділені тим чи іншим способом, присутні в наявних даних.

Незалежно від предмета вивчення, застосування кластерного аналізу припускає наступні етапи:

- Відбір вибірки для кластеризації.
- Визначення безлічі змінних, за якими будуть оцінюватися об'єкти у вибірці.
- Обчислення значень тієї чи іншої міри подібності між об'єктами.
- Застосування методу кластерного аналізу для створення груп схожих об'єктів.
- Перевірка достовірності результатів кластерного рішення.

Кластерний аналіз пред'являє наступні вимоги до даних:

- показники не повинні корелювати між собою.
- показники повинні бути безрозмірними.
- розподіл показників повинен бути близьким до нормального розподілу.

- показники повинні відповідати вимозі "стійкості", під якою розуміється відсутність впливу на їх значення випадкових факторів.
- вибірка повинна бути однорідна, не містити "викидів".

Після отримання та аналізу результатів можливе корегування обраної метрики і методу кластеризації до отримання оптимального результату.

Цілі кластеризації:

- Розуміння даних, шляхом виявлення кластерної структури. Розбиття вибірки на групи схожих об'єктів дозволяє спростити подальшу обробку даних і прийняття рішень, застосовуючи до кожного кластеру свій метод аналізу (стратегія "розділяй і володарюй").
- Стиснення даних. Якщо вихідна вибірка надлишково велика, то можна скоротити її, залишивши по одному найбільш типовому представнику від кожного кластера.
- Виявлення новизни (англ. Novelty detection). Виділяються нетипові об'єкти, які не вдається приєднати до жодного з кластерів.
- У першому випадку число кластерів намагаються зробити поменше. У другому випадку важливіше забезпечити високу ступінь подібності об'єктів усередині кожного кластера, а кластерів може бути скільки завгодно. В третьому випадку найбільший інтерес представляють окремі об'єкти, які не вписуються ні в один з кластерів [18].

4.1.1 Класифікація кластеризації

Класифікувати алгоритми кластеризації в широкому сенсі можна на два наступних класи:

1. Ієрархічні і плоскі. Ієрархічні алгоритми (також звані алгоритмами таксономії) будують не одне розбиття вибірки на непересічні кластери, а систему

вкладених розбиттів. Таким чином, на виході ми отримуємо дерево кластерів, коренем якого є вся вибірка, а листям — найбільш дрібні кластера. Так як основна вимога до системи це швидкодія, то розрахунок ієрархічними методами з вкладеними кластерами тільки ускладнить і сповільнить процес відображення просторових даних.Плоскі алгоритми будують одне розбиття об'єктів на кластери.Плоскі алгоритми вважаються досить швидкими та простими в дії, що повністю відповідає поставленим вимогам. До того ж одноразове розбиття дозволяє уникнути необхідності зберігати велику кількість проміжних даних.

2. Чіткі і нечіткі. Чіткі (або непересічні) алгоритми кожному об'єкту вибірки ставлять у відповідність номер кластера, тобто кожен об'єкт належить тільки одному кластеру.Так як основною ознакою об'єктів в кластерах є їх географічне положення, то чітка кластеризація найкращим чином вирішить проблему поділу об'єктів на чіткі групи.Нечіткі (або пересічні) алгоритми кожному об'єкту ставлять у відповідність набір речових значень, що показують ступінь відносини об'єкта до кластерів. Тобто, кожен об'єкт відноситься до кожного кластеру з деякою вірогідністю.При роботі з просторовими даними основною метою є наочність геоінформації, тому, якщо один і той же об'єкт буде входити в різні кластери, це значно ускладнить роботу і заплутає користувача.

4.1.2 Міра відстані між об'єктами

Існують основні етапи розподілу об'єктів по кластерам. Для початку необхідно скласти вектор характеристик для кожного об'єкта — як правило, це набір числових значень. Однак існують також алгоритми, що працюють з якісними (категорійними) характеристиками.В даному випадку вектор характеристик буде зберігати географічні координати просторових даних.Після того, як ми визначили вектор характеристик, можна провести нормалізацію, щоб всі компоненти давали однаковий внесок при розрахунку "відстані". У процесі нормалізації всі значення приводяться до

деякого діапазону, наприклад, [–1, –1] або [0, 1]. Для кожної пари об'єктів вимірюється "відстань" між ними — ступінь схожості.

4.1.3 Алгоритм кластеризації — k–середніх

Виходячи з вимог до системи і вибраних уподобань, можна визначитися з конкретним алгоритмом кластеризації.Задачу кластеризації можна розглядати як побудову оптимального розбиття об'єктів на групи. При цьому оптимальність може бути визначена як вимога мінімізації середньоквадратичної помилки розбиття.Алгоритми квадратичної помилки відносяться до типу плоских алгоритмів. Найпоширенішим алгоритмом цієї категорії є метод k–середніх. Цей алгоритм будує задане число кластерів, розташованих якнайдалі один від одного. Робота алгоритму ділиться на кілька етапів:

1. Випадково вибрати k точок, які є початковими "центрами мас" кластерів.
2. Віднести кожен об'єкт до кластеру з найближчим "центром мас".
3. Перерахувати "центри мас" кластерів згідно з їх поточним складом.
4. Якщо критерій зупинки алгоритму не задоволений, повернутися до п. 2.

Як критерій зупинки роботи алгоритму зазвичай вибирають мінімальну зміну середньоквадратичної помилки. Так само можливо зупиняти роботу алгоритму, якщо на кроці 2 не було об'єктів, що перемістилися з кластера в кластер.До недоліків даного алгоритму можна віднести необхідність задавати кількість кластерів для розбиття, але в даному випадку це можна вважати позитивною властивістю даного методу.Метод k–середніх також називають швидким кластерним аналізом, виходячи з назви, можна визначити, що даний метод найбільш відповідний для кластеризації просторових даних. Так обчислювальна складність методу k–середніх — O (nkl), де

k — число кластерів, l — число ітерацій. Даний метод повністю відповідає поставленим вимогам: швидкодія, чіткість розподілу по кластерам. Також завдяки

можливості спочатку ставити число кластерів можна розрахувати оптимальну кількість відображуваних об’єктів [19].

Тепер ми можемо розглянути кожен документ як вектор, розмірність якого дорівнює кількості термінів у словнику. Такий вектор документа будемо позначати $\vec{V}(d)$. Колекція документів може бути розглянута як безліч векторів в єдиному векторному просторі, в якому кожна вісь відповідає за один термін. Як вже було зазначено, таке уявлення "втрачає" порядок проходження термінів в документах.

Після того, як була побудована векторна модель, необхідно визначити спосіб обчислення міри близькості двох векторів, що належать векторному простору. Простою ідеєю є модуль різниці двох векторів, але цей метод не підходить з наступних причин: два документа з дуже схожим змістом можуть сильно відрізнятися в даній мірі тільки тому що один з них набагато довший за інший (відносні частоти термінів однакові, але їх абсолютні величини сильно розрізняються). Щоб компенсувати ефект довжини документа, стандартним рішенням є обчислення косинуса кута між векторами $\vec{V}(d1)$ і $\vec{V}(d2)$:

$$\cos(d_1, d_2) = \frac{\vec{V}(d_1)\cdot\vec{V}(d_2)}{\left|\vec{V}(d_1)\right|\left|\vec{V}(d_2)\right|} \tag{4.1}$$

Формула є спільною для всіх систем, що використовують векторне представлення документа. Вона не уточнює способів підрахунку компонент вектора.

Найпоширеніша косинусна міра дозволяє обчислювати відстань між документами (векторами) A_i та B_i:

$$\cos(\theta) = \left(\frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}}\right) \tag{4.2}$$

де $\sum_{i=1}^{n} A_i \times B_i$ – скалярний добуток;

$\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}$ – норма вектора A_i та B_i відповідно;

Для кластеризації спочатку обчислюється матриця відстаней між усіма документами. Потім агломераційний алгоритм кластеризації об'єднує два документи з

максимальною відстанню та переобчислює матрицю відстаней. Алгоритм зупиняється при досяганні деякого порогу в відстанях або при наявності необхідної кількості кластерів.

З метою запобігання занадто великих кластерів використовують поправочний коефіцієнт при обчисленні косинусної міри, залежний від обсягу кластера, наприклад, квадратний корінь або логарифм від кількості слів у кластері.

Агломераційний алгоритм потребує N2 пам'яті та O(N2) обчислень,
де N — кількість документів. Цей алгоритм використовується для кластеризації невеликої кількості документів, наприклад, для обчислення початкових кластерів для алгоритму K-середніх.

Таким чином формується декілька кластерів документів, які можна розглядати як однорідні за тематикою.

4.1.4 Модифікація алгоритму K-середніх для відбору корисних текстів із великихкорпусів документів

Алгоритм K-середніх потребує початкового завдання K центрів кластерів. Для цього використовується результат роботи агломераційного алгоритму кластеризації на відносно невеликому обсязі документів.

Алгоритм сканує всі документи з метою виявлення, до якого кластера належить поточний документ. Уводиться поріг відстані, який дозволяє не враховувати в подальшому розгляді документи з далекою відстанню від усіх центрів кластерів. Після сканування обчислюються нові центри кластерів і виконується декілька ітерацій алгоритму.

Після закінчення роботи алгоритму кластери містять тільки корисні документи за заданою тематикою.

4.2 Визначення основних етапів категоризації текстів

Першим етапом рішення задач автоматичної категоризації текстів є перетворення документі, що мають вид послідовності символів, до вигляду, що підходить для алгоритмів машинного навчання у відповідності із завданням категоризації. Зазвичай алгоритми машинного навчання мають справу з векторами в просторі (званому також простором ознак). Відображення документів у простір ознак також використовується і методами, заснованими на знаннях.

Другим етапом є побудова функції категоризації за допомогою навчання на прикладах.

Якість категоризації залежить і від того, як документи будуть перетворені в векторне подання, і від алгоритму, який буде застосований на другому етапі. При цьому важливо відзначити, що методи перетворення тексту в вектор специфічні для задачі класифікації текстів і можуть залежати від колекції документів, типу тексту (простий, структурований) і мови документа. Методи машинного навчання, застосовувані на другому етапі, не є специфічними для задачі класифікації текстів і застосовуються також в інших областях, наприклад, для задач розпізнавання образів.

Розглянемо класичний підхід для відображення тексту в вектор, який використовується багатьма системами автоматичної класифікації текстів. Цей метод ґрунтується на припущенні про те, що категорія, до якої належить цей документ, залежить від відносної частоти слів, що входять в текст. Це припущення, звичайно, є спрощенням. Існують приклади систем, які враховують більш складні фактори: порядок слів у тексті, структура тексту, що містить розмітку.

Базовий метод відображення тексту в вектор полягає в тому, що кожному слову, яке зустрічається в якомусь документі, відповідає певна координата в просторі ознак. Для слова, що зустрічається в документі, значення відповідної координати позитивно і пропорційно частоті слова в документі. Для слова, яке не зустрічається в документі, значення відповідної координати дорівнює нулю.

Є декілька причин, по яких слід прагнути зменшити розмір простору ознак. По-перше, облік всіх зустрінутих в документах слів призводить до занадто великої розмірності простору, хоча багато слова слабо впливають на результати категоризації (або взагалі не впливають). Висока розмірність простору ознак може призводити до високої обчислювальної похибки і низькій швидкості роботи алгоритмів навчання. По-друге, відображення декількох близьких за значенням слів в одну координату може поліпшити результати категоризації. Наприклад, різні морфологічні форми слова слід вважати еквівалентними.

Опишемо основні прийоми, застосовувані для перетворення текстів в вектори простору ознак.

4.3 Метрика TF-IDF

Тематика тексту визначається словами, які використовуються для передачі смислу, але різні слова мають бути враховані у різному ступені. Простіший погляд на частоту вживання слів звичайно модифікується з метою зменшення ролі загальнопоширених слів.

Вибір метрики істотно впливає на якість категоризації. У статті [22] наводиться докладне дослідження різних підходів до вибору метрик (ваг ознак). Результати експериментів, описаних в цій статті, показують, що однією з кращих формул обчислення ваг є метрика TF-IDF обчислюється для кожного слова зі словника документу

$$TFIDF_i = \frac{n_i}{\sum_k n_k} * \ln\left(\frac{|D_i|}{|t_i \in D_i|}\right), \tag{4.3}$$

де n_i - число входжень слова в документ;

$\sum_k n_k$ - загальна кількість слів в документі;

$|D_i|$ - кількість документів відповідної тематики;

$|t_i \in D_i|$ - кількість документів в яких зустрічається слово t_i;

де перший множник TF_i є частотність і-го слова у даному документі, а другий множник IDF_i враховує інверсію частотності появи і-го слова у всіх розглядуваних документах.

Такий вибір формули можна обґрунтувати теоретично такими міркуваннями:

1. Чим частіше слово зустрічається в документі, тим воно важливіше. Цей факт враховує множник tf_i.
2. Якщо слово зустрічається в багатьох або у всіх документах, то це слово не може бути істотним критерієм приналежності документа категорії і його вага слід знизити. Навпаки, якщо слово зустрічається в малій кількості документів, то його вага слід підвищити. Множник враховує це міркування і відповідає вазі слова ("контрастності") в даній колекції документів.
3. Для того щоб врахувати різну довжину текстів документів у колекції, ваги слів документів слід нормалізувати. У формулі (1.1) ваги нормалізуються так, щоб сума квадратів ваг кожного документа дорівнювала 1.

Існують також інші варіанти формули *tf*idf*, які дають близькі за якістю результати, наприклад можна використовували TF * IDF у формулюванні INQUERY [14]:

$$w_i = \beta + (1-\beta) \cdot \frac{tf_i}{tf_i + 0.5 + 1.5 \cdot \frac{dl}{avg\_dl}} \cdot \frac{\log\left(\frac{N+0.5}{n}\right)}{\log(N+1)} \qquad (4.4)$$

де dl - міра довжини документа, avg_dl - середня довжина документа, $\beta = 0.4$, де N - кількість документів у колекції, n - кількість документів, де зустрілося i-е слово.

У деяких випадках для обчислення ваги слова в тексті залучається також додаткову інформацію. Наприклад, можна враховувати інформацію про структуру тексту і словами, зустрінутим в заголовку, присвоювати більшу вагу [23].

4.4 Висновки до розділу

У зв'язку з зростанням необхідності обробки інформації необхідно спростити роботу людини і пришвидшити безліч завдань, таких як напівавтоматична перевірка творчих робіт, або систематизація і полегшення пошуку книг і наукових робіт в електронних бібліотеках, для всіх цих завдань можливо застосувати модуль категоризації, отже розробка такого модуля є актуальною. Проведено розгляд основних етапів, що необхідні для проведення автоматичної категоризації текстів.

5 ПОБУДОВА ТЕМАТИЧНИХ СЛОВНИКІВ

5.1 Загальний принцип побудови специфічних словників

Процес побудови тематичних словників зображено на Рисунку 5.1.

Рисунок 5.1 – Процес побудови тематичних словників

5.2 Текстові корпуси

ФКТ були сформовані та розподілені по тематикам експертами МННЦ ІТіС. Ці файли мали відповідати деяким умовам. Першою умовою є використання текстів з сайту *NewsOne*. Другою умовою є обмежена лематизація: використовувалися лише ті словоформи , що потрапляли у текстову навчальну вибірку, та застосовувалася лематизація лише для слів, що потрапили в словник.

Слова, які були вибрані для ФКС, були взяті з корпусів початкових текстових ресурсів:

1. 512 текстових файлів, взятих на сайті *NewsOne*;
2. 60 файлів, взятих із *AKUEM* [20];
3. 592 файлів, взятих на інших сайтах новин та з Інтернету.

Перші два набори текстів мають відповідні їм набори звукових файлів. Частина цього матеріалу буде використана в якості контрольної вибірки. Інша частина була використана для формування ФКТ. Останній набір текстів доповнює текстами та ключовими словами словники тематик.

Для кожного із 1164 файлів були виділені головні особливості експертним шляхом, таким чином, щоб віднести файл до тієї чи іншої теми, тобто класифікувати. Класифікацією текстів називають задачу інформаційного пошуку, в якій документ відноситься до однієї або декількох категорій на основі змісту документу. Таких категорій було обрано 12 – узагальнених та 50 – більш детальних.

У Таблиці 2 представлено назви та відношення загальних та детальних тем. наглядно ілюструє, що деякі теми, які в пресі обговорюються частіше, мають більше текстів для створення ФКС та, відповідно, будуть мати більші самі словники ключових слів.

Таблиця 5.1 – Входження детальніших тем до більш загальних.

№	Назва теми	Кількість текстів	Назви підтем
1	досягнення	65	IT технології, наука США, відкриття-загадки-космос
2	економіка	153	благодійність, банки, виробництво, газові відносини, гроші, нерухомість, Київ, Євро- 2012, економіка США
3	культура	126	книга, козаки, культура, мистецтво, освіта, ам'ятники, таблоїд, туризм, культура США, втрата
4	Київ	37	Київщина
5	події	270	демонстрація, дороги, козаки, нещасні випадки, кримінальні події України, шахти, таблоїд, теракти, події США, війна та військові (армія), втрати, ВВВ
6	політика	41	політика світу, політика України, політика США, втрати
7	релігія	1	релігія
8	спорт	315	автоспорт, баскетбол, бокс, Євро-2012, футбол, хокей, теніс, велоспорт, інші види спорту
9	суд	33	суд над Тимошенко
10	тварини	42	гороскоп, тварини
11	явища	76	забруднення довкілля, погода, природні катаклізми
12	здоров' я	6	здоров' я

5.3 Процес нормалізації

Процес нормалізації відбувався за наступними правилами:

1. Слова, що довжиною менше двох та цифри не враховувалися, були відкинуті
2. Слова що пишуться через дефіс, апостроф, містять перенос на наступний рядок – враховуються як одне слово
3. Усі слова приводились до малого регістру
4. Власні назви що складаються з кількох слів, наприклад Lotus Renault GP враховувалися як 3 різні слова

5.4 Використання морфології

Для того щоб об'єднувати різні морфологічні форми слова в одну координату простору ознак, кожне слово вихідного тексту приводиться до своєї нормалізованому формі (лемме). Для англійської мови зазвичай застосовується процедура нормалізації слів, яка полягає у відсіканні закінчення слова (stemming). Для української мови процедура нормалізації слів є більш складною, але на даний момент існують поширені методи її вирішення [14].

Слова, які будуть представлені в словниках ключових слів повинні обмежено лематизуватися. Для цього використовується словник українських слів об’ємом 1 809 362 слова. У Таблиці 3 наведена частина словника а якому перше слово – це основна форма слова, а друге – словофарма.

Таблиця 5.2 – Представлення словника форм українських слів

авіалінія	авіалінія
авіалінія	авіалінії
авіалінія	авіалінію
авіалінія	авіалінією
авіалінія	авіалінії
авіалінія	авіалініє
авіалінія	авіалінії
авіалінія	авіаліній
авіалінія	авіалініям
авіалінія	авіалінії
авіалінія	авіалініями
авіалінія	авіалініях

5.5 Формування вектора ключів

Для обчислення відстані між документами потрібно задати словник, для якого формується метрика TF- IDF. Такий словник називається вектором ключів. Вичерпну інформацію про відстань несе увесь словник для всіх розглядуваних документів, але обсяг обчислень стає дуже великим. Слід відзначити, що більша частина усього словника не є корисною для обчислень метрики — загальнопоширені та рідковживані слова не несуть ніякої корисної інформації.

Для формування вектора ключів обчислюються коефіцієнти TF-IDF для всіх розглядуваних документів, а всі слова документу впорядковуються за цими коефіцієнтами. Перші кілька десятків слів (приблизно 100) додаються до словника ключів. Потім видаляються найбільш частотні (з частотою появи більше ніж 0.001) та найменш частотні слова (з частотою появи менш ніж 0.00001). Достатньо

хорошим вважається вектор довжиною від 1000 до 10000 слів. У Талиці 4 наведено приклад словника ключових слів.

Таблиця 5.3 – Приклад словника ключових слів

аварія аварії аварію
автогонка автогонок
автогонка-формула автогонок-формула
автодром автодромом
автомобіль авто автівки автомобілем автомобілю автомобілі
автомобільний автомобільної автомобільну

5.6 Процедура автоматичного розбиття текстів по тематикам

Процедура автоматичного розбиття документів на теми зводиться до віднесення кожного файлу до однієї (чи більше) з вище визначених тем.

Етапи розбиття текстів на теми:

1. для кожного окремо із вхідних файлів кожне слово перевіряється, чи є це слово в будь-якому з ФКС;
2. якщо даного слова не має в словнику, воно не враховується (пропускається як сміття);
3. всі слова, які залишилися, підраховуються для кожної теми окремо. Та тема, в якій набирається більше всього слів – перемагає і файл відноситься до цієї теми. Такий спосіб розбиття відносить однозначно один файл до однієї теми. Та новини не завжди є таки однозначними. Деякі теми однаково стосуються 2–4 тем та різниця між ключовими словами цього файлу 10–15%. Для цього випадку введений ще один етап – етап порівняння відповідей.

4. кількість слів з теми, яка перемогла, множиться на коефіцієнт від 0 до 1. Цей коефіцієнт показує, яку похибку (в процентному відношенні) можна дозволити при розбитті файлів на теми. Наприклад: при підрахунку кожна тема для даного файлу отримала певну кількість слів: тема1 = 50, тема2 = 25, тема3 = 45, тема4 = 50, тема5 = 6. Вказавши коефіцієнт 1.0 (100%) перемагають тема1 та тема4, вказавши коефіцієнт 0.9 (90%), до теме1 та теме4, додається ще тема3.

Слід зауважити, що якщо файл не має слів хоча б з однієї теми (тобто їх 0), то такі файли записуються у кошик. Щоб уникнути потрапляння випадкового файлу у тему, можна збільшити поріг від 0 до тієї мінімальної кількості слів, які точно можуть відповідати за тему.

Щоб визначити, який коефіцієнт необхідно взяти, потрібно провести ряд експериментів з кроком 0,05 (від 0,5 до 1,0) і з кроком 0,01 (від 0,8 до 1,0).

5.7 Формування словників

ФКС для кожної з тематик недостатньо для створення словника. Тому необхідно виконати процедуру наповнення ФКС (Рисунок 5.2). Для цього необхідно розбити корпуса текстів по тематикам, після чого провести з ними ті ж маніпуляції що й з корпусами текстів для ФКС.

Рисунок 5.2 – Процес наповнення ФКС

Для кожної тематики потрібно сформувати свій словник. За основу береться словник визначеної тематики, тобто той словник, який найкраще (в повному об'ємі) характеризує цю тематику. Але не весь словник, а лише слова, які зустрілися з частотою більше 10 разів. Потім в загальному частотному словнику, побудованому на всіх файлах навчальної вибірки, слід взяти перші найчастотніші 10000 слів. Об'єднання цих двох підсловників і складає словник тематики.

5.8 Висновки до розділу

Відібрані тексти використовуються для побудови статистичної лінгвістичної моделі в системі розпізнавання злитого мовлення, що покращує точність розпізнавання. Крім алгоритму K-середніх можуть використовуватися більш сучасні алгоритми кластеризації, наприклад CLOPE [3,4].

6 РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

Інструментарій *HTK* на базі прихованих Марківських моделей (ПММ) використовувався для побудови акустичних та лінгвістичних моделей. Для розпізнавання мовлення розроблено систему яка сумісна з даними від HTK.

Акустична модель будувалась по звуковим файлам записанім з каналу *NewsOne* тривалістю 60 годин. Текстовий опис новин співпадає зі звуком на 85%. Для визначення фрагментів, де є точний збіг між звуком та текстовим описом, було використано автоматичне розпізнавання мовлення.

Для матеріалів, отриманих з Інтернету, були застосовані процедури автоматичного розбиття текстів на тематики. В результаті кожен текстовий файл було віднесено до тієї чи іншої тематики. Таким чином, була отримана текстова навчальна вибірка, на матеріалах якої будувалися словники і проводилися експериментальні дослідження.

Контрольна вибірка (КВ) для експериментів була сформована з файлів тривалістю 6 годин, які не належали до навчальної вибірки. Кожний файл розпізнавався за допомогою всіх 13 ЛМ. Всі файли КВ були розділені на тематики експертом та автоматично.

Таблиця 6.1 показує середню точність розпізнавання звукових файлів, які належать до теми ПОДІЇ, з використанням різних лінгвістичних моделей. Для ЛМ ПОДІЇ досягнута найкраща точність розпізнавання.

Таблиця 6.2 показує середню точність розпізнавання звукових файлів, які належать до тематик, з використанням тематичних ЛМ.

Врезультаті досягнуто не найкраща точність у звязку з тим, що тематичні словники досить мають досить малий об'єм. Всі ці тематики необхідно додатково

поповнювати текстами або віднести до кошику, якщо не буде знайдено необхідний обсяг текстів і повторно провести розпізнавання.

Таблиця 6.1 – Середня точність розпізнавання звукових файлів, які належать до теми ПОДІЇ, з використанням різних лінгвістичних моделей.

No	Тема ЛМ	Послівна точність, %
1	досягнення	54.22
2	економіка	76.04
3	культура	75.47
4	київ	56.06
5	інше	79.95
6	події	81.44
7	політика	77.71
8	релігія	57.73
9	спорт	64.40
10	суд	64.01
11	тварини	44.01
12	явища	68.04
13	здоров' я	55.21

Таблиця 6.2 – Розпізнавання звукових файлів, належних до тематик, за використанням відповідної тематичної ЛМ

Тема, до якої належать файли КВ	Довжина вибірки розпізнав ання, слів	Послівна точність, %
досягнення	1012	34.77
економіка	1851	18.02
культура	963	29.79
київ	979	38.64
події	3880	41.44
політика	4099	23.27
релігія	-	
спорт	1185	32.81
суд	3546	33.80
тварини	2087	21.41
явища	1464	22.03
здоров' я	11	30,87

ВИСНОВКИ

Було розглянуто задачу побудови специфічних словників для розпізнвання мовленнєвих сигналів за тематикою. Виконано аналіз існуючих методів розв'язання задачі.

Спроектовано та реалізовано автоматизовану систему побудови тематичних словників. Результатом роботи є тематичні словники, що інтегруються з існуючою в МННЦ ІТіС системою розпізнавання злитого мовлення

Результати випробувань реалізації на тестових прикладах показали, що за допомогою розроблених тематичних словників можна розпізнавати злите мовлення.

Для подальшого розвинення системи необхідно вдосконалювати процес побудови ФКС, та віднесення текстів по тематикам.

ПЕРЕЛІК ПОСИЛАНЬ

1. Пилипенко В.В. Распознавание ключевых слов в потоке речи при помощи фонетического стенографа. Искусственный интеллект. – Донецк, 2009. – No 4с. 220-224.

2. Федосин С.А., Еремин А. Ю.Классификация систем распознавания речи // электронное научное периодическое издание: электронный журнал / Электроника и информационные технологии / ГОУВПО «Мордовский государственный университет им. Н. П. Огарева», Саранск.

3. М.М.Биков, Т.В.Грищук. Моделювання процесу аналізу і класифікації голосових команд: Монографія – Вінниця: ВНТУ, 2009. – 128 с.

4. Система розпізнавання NaturallySpeaking [Електроний ресурс] — Режим доступу: http://www.dragonsys.com

5. Система розпізнавання IBM ViaVoice [Електроний ресурс] — Режим доступу: www.ibm.com/viavoice

6. Система розпізнавання VoiceType [Електроний ресурс] — Режим доступу: http://www-4.ibm.com/software/speech/

7. Система розпізнавання MedSpeak [Електроний ресурс] — Режим доступу: http://www.ibm.com/

8. Система розпізнавання Voice_PE [Електроний ресурс] — Режим доступу: http://www.voicerecognition.com/kurzweil/voicedes.html

9. Система розпізнавання Voice Xpress Professional [Електроний ресурс] — Режим доступу: www.lhs.com

10. Система розпізнавання Sakrament [Електроний ресурс] — Режим доступу: http://www.sakrament.com/

11. Система розпізнавання "Горинич" ПРОФ 3.0 [Електроний ресурс] — Режим доступу: http://www.upspecial.ru/gorynych-prof-3-0.html

12. Система розпізнавання RealSpeaker [Електроний ресурс] — Режим доступу: http://www.realspeaker.net/ua/

13. Lewis D. Feature Selection and Feature Extraction for Text Categorization. // Proceedings of the DARPA Workshop on Speech and Natural Language. —Harriman, New York, 1992. — pp. 212-217

14. 14. Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В., Штернов С.В. "Отправная точка" для дорожки по поиску в РОМИП (предварительный анализ). // Труды РОМИП'2003 (Российский семинар по Оценке Методов Информационного Поиска) — НИИ Химии СПбГУ / Под ред. И.С.Некрестьянова — Санкт-Петербург, 2003 — стр. 87-110.

15. Beuster G. MIC — A System for Classification of Structured and Unstructured Texts. Diploma Thesis. — University Koblenz, 2001.

16. Пилипенко В.В. Вибір текстів за тематикою для побудови лінгвістичної моделі мови в системах розпізнавання злитого мовлення. – Оброблення сигналів і зображень та розпізнавання образів: Десята Всеукраїнська міжнародна конференція. – Київ, 2010, С.63-64.

17. Мандель І.Д. Кластерний аналіз. — М.: Фінанси і Статистика, 1988.

18. Воронцов К.В. Алгоритми кластеризації та багатовимірного шкалювання. Курс лекцій. МДУ, 2007.

19. Котов А., Красильников Н. Кластеризація даних, 2006. 4-8 стор.

20. Н.Б. Васильєва, В.В. Пилипенко, О.М. Радуцький, В.В. Робейко, М.М. Сажок. Створення акустичного корпусу українського ефірного мовлення. Праці конференції УкрОбраз 2012, Київ, 2010, 55-58 стор.

21. Аношкина Ж.Г. Морфологический процессор русского языка. // Бюллетень машинного фонда русского языка / отв. редактор В.М. Андрющенко — М., 1996. — Вып.3, с.53-57.

22. Salton G, Buckley C. Term-Weighting Approaches in Automatic Text Retrieval. / Information Processing and Management, —1988 — pp. 513-523.

23. Beuster G. MIC — A System for Classification of Structured and Unstructured Texts. Diploma Thesis. — University Koblenz, 2001.