Формування тематичних лінгвістичних моделей для розпізнавання злитого мовлення з телевізійних новин

Н.Б. Васильєва, В.В. Пилипенко, В.В.Яценко

Міжнародний науково-навчальний центр інформаційних технологій та систем
просп. Академіка Глушкова 40, Київ 03680
{n.vassilleva; valeriy.pylypenko; yatsenko.valya} @gmail.com

Анотація

В статті описується інформаційна технологія розпізнавання мовлення з телевізійних новин за допомогою тематичних лінгвістичних моделей. При побудові системи розпізнавання спочатку проводиться розділення текстів для лінгвістичної моделі на теми та побудова окремих специфічних словників для кожної теми. При розпізнаванні використовується двохпрохідний підхід, де на першому кроці шукаються ключові слова для всіх тем та визначається тема повідомлення. Після цього на другому проході відбувається розпізнавання мовлення з використанням визначеної тематичної лінгвістичної моделі. Проведені експерименти по розпізнаванню мовлення з телевізійного каналу новин *NewsOne*.

1. Вступ

Розпізнавання мовлення з каналів новин потребує використання словників значно більших ніж 100 тис слів. Але для розміщення таких словників потрібні надвеликі ресурси комп'ютерів, зокрема величезні обсяги оперативної пам'яті. Звичайно, новини природно розподіляються на теми, які, на нашу думку, мають свої особисті словники, и таким чином, загальна задача розпізнавання може бути розділена на окремі задачі, в яких застосовуються менші, але тематично специфічні словники.

Для побудови лінгвістичних моделей для розпізнавання необхідні великі обсяги текстів. Пропонується використати ресурси Інтернет для формування корпусів текстів та відібрати корисні тексти за тематикою за допомогою алгоритмів кластеризації документів [1].

При розпізнаванні використовується двохпрохідний підхід, де на першому кроці шукаються ключові слова [2] для всіх тем. Відповідь розпізнавання розглядається як документ, і для нього визначається тематика за допомогою алгоритмів кластерізації документів. Після цього на другому проході відбувається розпізнавання мовлення з використанням визначеної тематичної лінгвістичної моделі.

У другому розділі описуються тестові навчальні вибірки, які застосовувалися для створення файлів ключових слів. У третьому розділі приведений алгоритм автоматичного віднесення текстів до теми, наводяться приклади роботи алгоритму. Четвертий розділ описує процес побудови ЛМ та словників для текстів, розбитих на тематики. У п'ятому розділі приведені результати експериментів по розпізнаванню мовлення з телевізійного каналу новин NewsOne.

2. 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 автоматичної розбиття корпусу текстів, завантажених із Інтернет сайтів, за допомогою файлів ключових слів (ФКС).

На Рисунку 1 представлена блок-схема структури тематичного розпізнавання мовлення.

3. Створення файлів ключових слів для тематик

Для створення лінгвістичної моделі теми, а також для пошуку ключових слів на першому етапі роботи були створені файли ключових слів (ФКС).

ФКС для кожної з обраних тем створювався експертним шляхом. Ці файли мали відповідати деяким умовам. Першою умовою було використання текстів сайту *NewsOne*. Детальніший опис дивись нижче у пункті 2.1. Другою умовою була обмежена лематизація: використовувалися лише ті словоформи, що потрапляли у текстову навчальну вибірку, та застосовувалася лематизація лише для слів, що потрапили в словник (пункт 2.2).

3.1. Текстові корпуси

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

1. 512 текстових файлів, взятих на сайті *NewsOne* [3];
2. 60 файлів, взятих із *AKUEM* [4];
3. 592 файлів, взятих на інших сайтах новин із Інтернету.

Перші два набори текстів мали відповідні їм набори звукових файлів. Частина цього матеріалу була використана в якості контрольної вибірки. Інша частина була використана для формування ФКС. Останній набір текстів доповнював 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 [5]. Таких категорій було обрано 12 – узагальнених та 50 – більш детальних.

У Таблиці 1 представлено назви та відношення загальних та детальних тем. Таблиця наглядно ілюструє, що деякі теми, які в пресі обговорюються частіше, мають

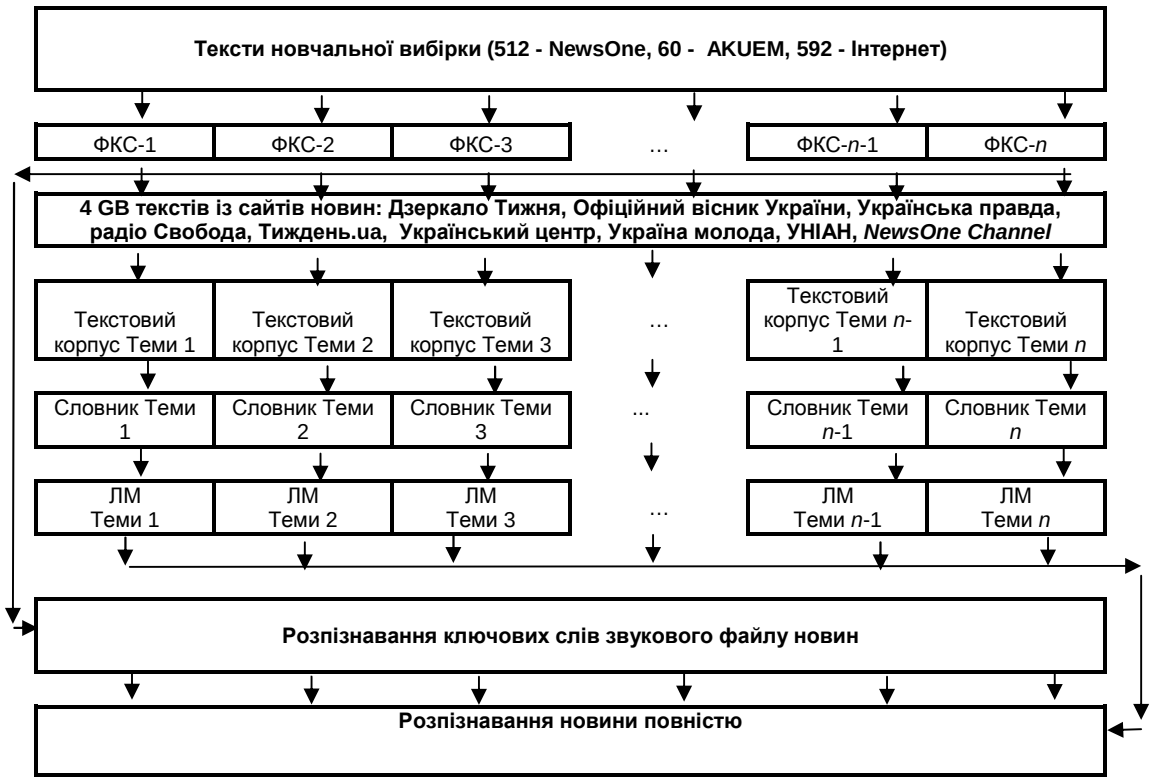


Рисунок 1. блок-схема структури тематичного розпізнавання мовлення

більше текстів для створення ФКС та, відповідно, більші самі словники ключових слів.

Таблиця 1. Входження детальніших тем до більш загальних.

№	Назва теми	Довжина словника	Назви підтем
1	досягнення	431	IT технології, наука США, відкриття-загадки-космос
2	економіка	801	благодійність, банки, виробництво, газові відносини, гроші, нерухомість, Київ, Євро-2012, економіка США
3	культура	920	книга, козаки, культура, мистецтво, освіта, ам'ятники, таблоїд, туризм, культура США, втрата
4	Київ	127	Київщина
5	події	1024	демонстрація, дороги, козаки, нещасні випадки, кримінальні події України, шахти, таблоїд, теракти, події США, війна та військові (армія), втрати, ВВВ
6	політика	297	політика світу, політика України, політика США, втрати
7	релігія	92	релігія
8	спорт	454	автоспорт, баскетбол, бокс, Євро-2012, футбол, хокей, теніс, велоспорт, інші види спорту
9	суд	103	суд над Тимошенко
10	тварини	238	гороскоп, тварини
11	явища	580	забруднення довкілля, погода, природні катаклізми
12	здоров'я	205	здоров'я

Словники ключових слів із 50 тем вже не мають стільки слів, але більш конкретно відповідають певній темі.

3.2. Обмежена лематизація

Слова, які були представлені в словниках ключових слів обмежено лематизувалися: процес переведення словоформи до леми — її нормальної (словарної) форми.

У Талиці 2 наведено приклад словника ключових слів (словник спорт)

Таблиця 2. Приклад словника ключових слів (словник спорт).

аварія аварії аварію
автогонка автогонок
автогонка-формула автогонок-формула
автодром автодромом
автомобіль авто автівки автомобілем автомобілю автомобілі
автомобільний автомобільної автомобільну
автоспорт автоспорту
автоспортивний автоспортивної
автоспрінт autosprint автоспрінт
аеробіка
азовмаш азовмашу азовмашем

4. Процедура автоматичного розбиття текстів на теми

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 кіот кість 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.

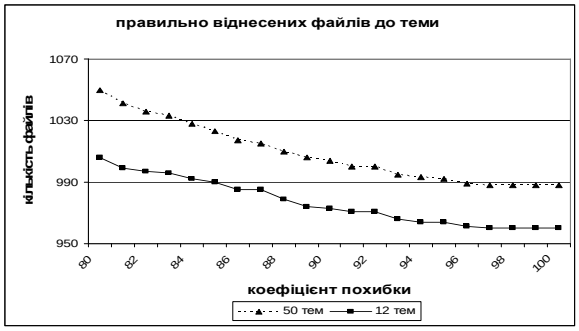
Слід зауважити, що якщо файл не має слів хоча б з однієї теми (тобто їх 0), то такі файли записуються у кошик. Щоб уникнути потрапляння випадкового файлу у тему, можна збільшити поріг від 0 до тієї мінімальної кількості слів, які точно можуть відповідати за тему.

У таблиці 1 відображені результати роботи програми автоматичного розбиття текстів на одну з 12 тем (для 50 тем результат роботи програми буде схожим), де можна побачити різницю при зупинці вищеописаного алгоритму на етапі 3 та на етапі 4. Необхідно також відмітити, що вказавши коефіцієнт 1,0 маємо такий самий результат, якщо зупинитися на етапі 3.

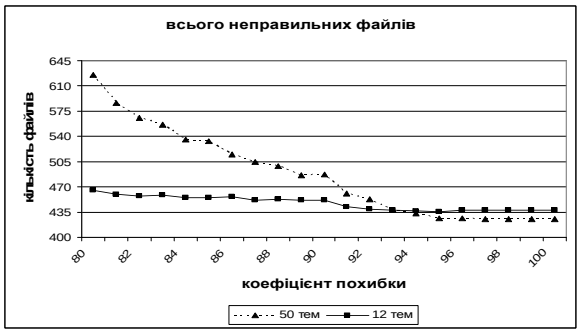
Таблиця 3. Відображення результатів роботи алгоритму віднесення файлу до певної теми при коефіцієнтах похибки 1,0 та 0,75

Назва теми	Словник ключових слів теми або ФКС	Кількість слів із даного файлу, які потрапили у словник відповідної теми	Тема, яка перемогла на етапі 3 (значення коефіцієнт у 1,0)	Тема, яка перемогла на етапі 4 (значення коефіцієнт у - 0,75)
досягнення	43	4		
економіка	801	25	+	+
культура	920	19		+
київ	127	2		
події	1024	1		
політика	297	2		
релігія	91	0		
спорт	453	2		
суд	103	0		
тварини	238	0		
явища	580	0		
здоров'я	205	0		

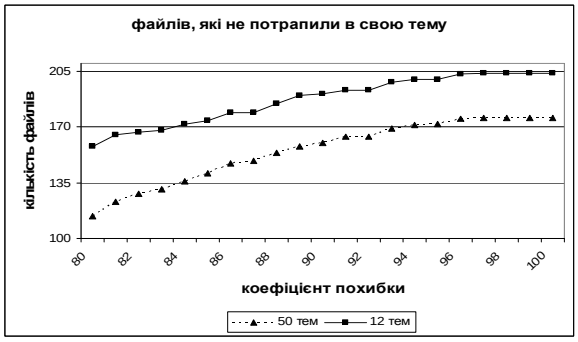
Щоб визначити, який коефіцієнт необхідно взяти, було проведено ряд експериментів з кроком 0,05 (від 0,5 до 1,0) і з кроком 0,01 (від 0,8 до 1,0). Файли, на яких проводилися тестування, це ті ж 1164 файлів, які використовувалися для створення ФКС. Результати експериментів, приведених на Рисунках 2 (а-в), порівнювалися з початковим експертним віднесенням файлів до тем (як 12 так і 50).



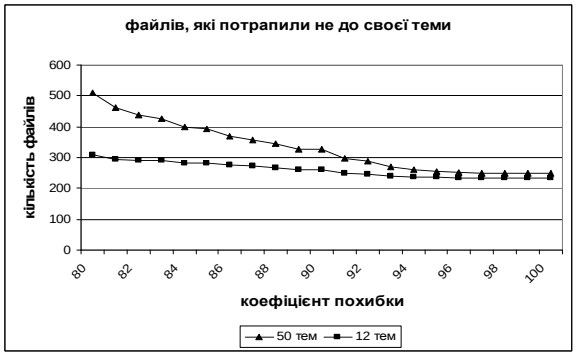
а) файли, які співпали з експертним;



б) разом файлів, яких не потрапили в свою тему та тих, які потрапили не до свої теми;



в) файли, які не потрапили в свою тему порівняно з експертним;



г) файли, які потрапили не в свою тему порівняно з експертним.

Рисунок 2. Результати віднесення файлів контрольної вибірки до теми з кроком коефіцієнта похибки 0,01 (від 0,8 до 1,0), порівняні з експертним розбиттям

5. Процес побудови ЛМ та словників для текстів, розбитих на тематики

Як було зазначено, на другому проході відбувається розпізнавання мовлення з використанням визначеної лінгвістичної моделі тематики.

Формально поняття лінгвістичної моделі (ЛМ) можна описати наступним чином. По навчальній вибірці будуються послідовності слів, які називаються *N*-грамами (*N* – кількість слів в послідовності). *N*-грами лінгвістичної моделі використовуються для прогнозування кожного символу в послідовності, зазначеної її *N*-1 попередником, де кожне слово розглядається в поєднанні з попередніми і наступними словами як вони зустрілися в навчальній вибірці. Це робиться для того, щоб потім новий розпізнаний текст можна було оцінити, порівнюючи послідовно з побудованими N-грамами, і по максимальній ймовірності вивести результат розпізнавання.

Хоча основний принцип *N*-грам ЛМ дуже простий, на практиці існують, як правило, набагато більше потенційних *N*-грам, ніж може бути зібрано при підготовці тексту в достатній кількості для отримання надійної частотної оцінки.

Крім того, для будь-яких реальних додатків, таких як розпізнавання мови, використання статистичних і підготовлених кінцевих текстів ускладнює генерування однієї ЛМ, яка добре відповідала б тому чи іншому тестовому матеріалу. Наприклад, ЛМ навчена на газетних текстах буде мати гарний результат для звітів диктування новин, але та ж ЛМ буде мати поганий результат для особистого листування або мовного інтерфейсу для системи резервування квитків. Остаточна складність полягає в тому, що словник *N*-грам ЛМ скінчений і фіксується на час побудови.

Таким чином, кожна тематика, для якої проводиться розпізнавання, повинна мати свою ЛМ і свій словник, на якому ця модель будується.

5.1. Формування словників.

Розглянемо як формуються словники для побудови ЛМ на визначених тематиках.

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.

Сумарний обсяг тематичних словників становив біля 300 тис. слів.

Далі, використовуючи таким чином сформований словник і попередньо розмічені тексти навчальної вибірки, будується ЛМ тематики.

5.2. Побудова лінгвістичної моделі

Лінгвістична модель будується в декілька етапів.

На етапі підготовки у вхідних текстах визначеної тематики розмічаються позиції початку і кінця речення відповідними позначеннями, тобто відбувається автоматична попередня розмітка текстів і їх об'єднання.

Потім, на першому етапі, використовуючи розмічені тексти та сформовані словники для кожної тематики, будується файл карти слів, в якому підраховуються *N*-грами і зберігаються в базі даних грам файлів. При побудові *N*-грам кожному слову із тексту ставиться у відповідність числовий унікальний ідентифікатор та кількість входжень цього слова в текст. Паралельно формується ще один файл *N*-грам, який містить трійки слів в ідентифікаторах та кількість входжень цих трійок в текст.

Отриманий файл карти слів треба порівняти із словником розпізнавання для того, щоб з'ясувати, які слова із карти слів співпали (тобто знайшлися в словнику розпізнавання) – вони залишаються. Ті слова із карти слів, для яких не знайшлося слово в словнику розпізнавання, замінюються на «*UNK*».

А потім на заключному етапі розраховуються результуючі грам файли, які використовуються для обчислення ймовірностей *N*-грам, вони зберігаються у файлі лінгвістичної моделі.

Усі тексти розподілялися на 12 кластерів. Окрім цього був сформовано 13-й кластер (кошик), в якій потрапили тексти не належні до базових кластерів.

6. Результати експериментальних досліджень

Інструментарій *HTK* [6] 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 ЛМ та словники і проводилися експериментальні дослідження. В таблиці 5 наведено об'єми текстових матеріалів, які були сформовані для кожної тематики.

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

Таблиця 5 показує середню точність розпізнавання звукових файлів, які належать до тематик, з використанням належних лінгвістичних моделей. Курсивом виділені кластери файлів, для яких досягнута не найкраща точність серед інших тематик.

Вважаємо, що крім тематики СУД всі інші, для яких досягнуто не найкраща точність, мають малу статистику як для текстів ЛМ, так і серед звукових файлів КВ. Всі ці тематики необхідно додатково поповнювати текстами або віднести до кошику, якщо не буде знайдено необхідний обсяг текстів.

Тематику СУД необхідно переглянути, тому що окремі новини більш належать до політики ніж до суду.

Таблиця 4. Середня точність розпізнавання звукових файлів, які належать до теми ПОДІЇ, з використанням різних лінгвістичних моделей.

№	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 5. Розпізнавання звукових файлів, належних до тематик, за використанням відповідної тематичної ЛМ

Тема, до якої належать файли КВ	Об'єм текстів для ЛМ, *MBt*	Довжина вибірки розпізнавання, слів	Послівна точність, %	Найкращий результат серед всіх ЛМ, %
досягнення	8	2012	62.77	74.26
економіка	288	5851	80.02	80.02
культура	254	8963	69.79	69.79
київ	9,3	979	68.64	77.32
події	298	14880	81.44	81.44
політика	429	4099	83.27	83.27
релігія	17,1	-		
спорт	38	11185	72.81	72.81
суд	29	4546	80.80	84.36
тварини	3	2087	51.41	74.32
явища	55	1464	79.03	77.94
здоров'я	9,5	141	63.83	79.43

7. Висновки

Розроблено та експериментально перевірено технологію автоматичного розбиття текстів на тематики для побудови специфічних лінгвістичних моделей. Для найбільш представлених тематик досягнута найкраща точність розпізнавання мовлення. Сумарний обсяг словників системи розпізнавання становив біля 300 тис. слів.

8. Література

[1]. 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.

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

[3]. Електроний ресурс: *http://newsone.ua*

[4]. 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.

[5]. Електроний ресурс: *http://ru.wikipedia.org*

[6]. Young S.J. HTK Book, version 3.1 / Young S.J. [et al]. – Caridge University, 2002. – 355 p.