DOI: https://doi.org/10.36910/6775-2524-0560-2021-44-22
УДК 004.912
Рябоконь Тетяна Олексіївна
Петрашенко Андрій Васильович, к.т.н., доцент
https://orcid.org/0000-0003-0239-1706
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ПРОГРАМНІ ЗАСОБИ ФОРМУВАННЯ ТА ОБРОБКИ БАЗИ ДАНИХ СЛОВОСПОЛУЧЕНЬ УКРАЇНСЬКОЇ МОВИ

Рябоконь Т. О., Петрашенко А. В. Програмні засоби формування та обробки бази даних словосполучень української мови. Дана стаття присвячена опису створення бази даних словосполучень та виявленню найбільш ефективних методів знаходження словосполучень у тексті. Проведено аналіз існуючих досліджень статистичних методів виділення колокацій з текстових даних та запропоновано критерій для порівняння їх ефективності в роботі з текстами саме українською мовою. Також описано архітектуру автоматизованої генерації бази даних словосполучень та можливі способи її прискорення. Проведено експеримент та визначено найбільш ефективний метод знаходження словосполучень для обраного корпусу текстів українською мовою.
Ключові слова: статистичні методи знаходження словосполучень, база даних словосполучень, колокація, текстовий корпус.

Рябоконь Т. А., Петрашенко А. В. Программные средства формирования и обработки базы данных словосочетаний украинского языка. Данная статья посвящена описанию создания базы данных словосочетаний и выявлению наиболее эффективных методов поиска словосочетаний в тексте. Проведен анализ существующих исследований статистических методов выделения коллокаций из текстовых данных и предложен критерий для сравнения их эффективности в работе с текстами именно на украинском языке. Также была описана архитектура автоматизированной генерации базы данных словосочетаний и возможные способы ее ускорения. Проведен эксперимент и определен наиболее эффективный метод поиска словосочетаний для выбранного корпуса текстов на украинском языке.
Ключевые слова: статистические методы нахождения словосочетаний, база данных словосочетаний, коллокация, текстовый корпус.

Riabokon T., Petrashenko A. The software system for generation and processing of a database of collocations of the Ukrainian language. This article is devoted to the description of creating a database of collocations and identifying the most effective methods of collocation extraction from the text. Existing researches of statistical methods of collocation extraction from text data were analyzed and the criterion for comparison of their efficiency in work with Ukrainian language texts is offered. The architecture of automated generation of the database of collocations and possible ways of its acceleration is also described. An experiment was conducted and the most effective method of finding collocations for the selected corpus of texts in the Ukrainian language was determined.
Keywords: statistical methods of collocation extraction, the database of collocations, colocation, text corpus.

Вступ. На сьогоднішній день існує безліч різноманітних ресурсів для вивчення та дослідження майже будь-якої живої мови. Люди досліджують та вивчають мови не тільки задля потреб філології, а й для розвитку NLP додатків та алгоритмів, для вдосконалення пошукових систем та голосових асистентів, інструментів, що домагають редагувати та підсумовувати тексти, та алгоритмів, що генерують цілі тексти автоматично. Такими ресурсами можна вважати словники, перекладачі та майже будь-які тексти в мережі Інтернет. Адже мова постійно трансформується та змінюється, тому для ефективної роботи алгоритмів, що працюють з живою мовою, потрібен постійних аналіз справжнього мовлення. Як вже було сказано, одним з видів мовних ресурсів є різноманітні словники, у тому числі й словники словосполучень.

Аналіз словосполучень є важливим розділом NLP досліджень. Вміння аналізувати, класифікувати та знаходити словосполучення дає можливість оперувати контекстом та змістом, що закладені у речення, а не окремими словами. Це допомагає значно вдосконалити системи, що працюють з натуральними мовами. Існує багато алгоритмів та підходів, які дозволяють аналізувати окремі слова, але аналіз та пошук групи слів, що зв'язані між собою є більш складним завданням.

Словосполучення важливі для ряду застосувань: генерація природної мови – щоб переконатися, що вихідні дані звучать природно і уникнути помилок; обчислювальна лексикографія – для автоматичного визначення важливих словосполучень, які мають потрапити до словника та корпусні лінгвістичні дослідження, наприклад, вивчення суспільних та культурних явищ через мову.

Дана робота присвячена пошуку словосполучень в текстах, що написані українською мовою, з подальшим упорядкуванням та морфологічним аналізом слів. Виділення словосполучення - це задача, що передбачає використання комп'ютера для автоматичного виділення словосполучення з корпусу. Традиційний метод виконання виділення словосполучення полягає у знаходженні формули на основі статистичних величин для обчислення оцінки пов'язаної з кожною парою слів.

Також в роботі досліджені основні підходи до пошуку словосполучень: відбір словосполучень за частотою, вибір за середнім значенням та дисперсія відстані між головним та залежним словом.

Пов'язані роботи. Для виділення словосполучень з тексту можна використовувати різні методи навчання асоціативних правил. Наприклад, застосовують метод точкової взаємної інформації, t-критерій Стьюдента, індекс Соренсена, логарифмічна правдоподібність для обчислення ступеня близькості між складовими словосполучення у текстовому корпусі.

Відомості про t-критерій можна знайти у багатьох підручниках загальної статистики. Зокрема найвідоміша праця це «Statistical methods» Дж. Снедекора та В. Кокрена [1]. Однією з перших публікацій про дослідження словосполучень була робота К. Черча та П. Хенкса 1989, в якій автори вказують на ефективність методу взаємної інформації для ідентифікації словосполучення в лексикографії. Також автори зосередили увагу на новому типі словнику, що побудований на основі корпусу і розробили програму обчислювальної лексикографії, яка поєднує корпусні данні, обчислювальні методи та людське сприйняття мови для побудови більш точних словників, які краще відображають дійсне використання мови [2, 3].

Серед багатьох загальних методів, представлених в роботі К. Маннинга та Г. Шютце 1999, найкращі результати можна досягти шляхом виділення словосполучення на основі як мовної, так і статистичної моделі [4]. Ф. Смаджа 1993 представив метод під назвою Xtract, на першому етапі якого парні лексичні відношення одержуються лише за допомогою статистичної інформації, на наступному етапі ідентифікуються поєднання кількох слів і складні вирази та нарешті, шляхом поєднання методів синтаксичного аналізу та статистичних методів, відмічаються та фільтруються словосполучення, які були отримані на першому етапі [5].

Т. Данінг (1993) вказав на слабкість методу взаємної інформації і показав, що логарифмічна правдоподібність є більш ефективним методом при виявленні одномовних словосполучень, особливо коли їх кількість дуже мала [6].

Ще одним методом, що використовується для аналізу словосполучень є z-критерій, близький до t-критерію [7, 8]. Він використовується в деяких пакетах програмного забезпечення для аналізу тексту.

Отже, можна зробити висновок, що проблема сумісності слів або виділення словосполучень з текстових даних є дійсно актуальною та активно досліджуваною проблемою у корпусній лінгвістиці.

Методи знаходження словосполучень у тексті

Дане дослідження посвячене побудові словника словосполучень. Такий словник можна створити, обробивши велику кількість текстів з використанням профільних електронних словників та методів виділення словосполучень з корпусу. Існує декілька методів виділення словосполучень з тексту і кожен з них має певні переваги та недоліки. Головним недоліком статистичних методів є ігнорування синтаксичних співвідношень між словами на великій відстані.

В процесі дослідження були порівняні між собою такі статистичні методи: підрахунок, точкова взаємна інформація, t-критерій Стьюдента, критерій хі-квадрат та логарифмічна правдоподібність. Всі вони застосовуються для обчислення ступеня близькості між складовими словосполучень у текстовому корпусі.

Безумовно, найпростіший метод пошуку словосполучень у текстовому корпусі – це підрахунок. Якщо декілька слів зустрічаються разом багато разів, це є свідченням того, що існує математичне відношення, яке не може бути просто описане математичною функцією комбінацій цих слів.

Для ефективного пошуку словосполучень в корпусі корисно визначити, чи є між словами деякий зв'язок і вони зустрічаються разом частіше, ніж один раз. Оцінка того, чи є щось випадковою подією — це одна з класичних проблем статистики. Зазвичай така проблема визначається з точки зору перевірки гіпотез. Треба сформулювати нульову гіпотезу H_0 про відсутність зв'язку між словами поза ймовірнісними випадками, обчислити ймовірність p, що подія відбудеться, якщо гіпотеза H_0 істинна, а потім відкинути H_0, якщо p занадто мала, або зберегти H_0 в іншому випадку.

Один з методів перевірки гіпотез, що широко використовується для виявлення словосполучень називається t-критерій Стьюдента. Перевірка враховує різницю між спостережуваним та очікуваним середнім значенням, масштабованим за дисперсією даних, і означає, наскільки ймовірно буде отримана вибірка цього середнього значення та дисперсії, якщо припустити, що вибірка береться з нормального розподілу із середнім значенням. Обчислити ймовірність отримання такої вибірки можна за формулою

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}},$$

де $\bar{x}$ – середнє значення вибірки, s^2 – дисперсія вибірки, N – розмір вибірки, а μ – середнє значення очікуваного розподілу.[2]

Використання t-критерію Стьюдента часто критикують за те, що даний метод передбачає нормально розподілені ймовірності, що найчастіше не відповідає дійсності. Альтернативним методом, який не має такої проблеми, є критерій хі-квадрат. Повна назва даного методу – це критерій узгодженості Пірсона, але для спрощення його називають просто хі-квадрат, тому що він найвідоміший з цієї групи. Суть методу полягає в порівнянні спостережуваних подій з очікуваними подіями відповідно до припущення, що спостережувані події не залежать одна від одної. Якщо різниця між спостережуваними та очікуваними значеннями велика, треба відхилити нульову гіпотезу незалежності. [9]

Даний метод підсумовує квадратичні відмінності між очікуваними та спостережуваними частотами, масштабованими на очікувані частоти, у всіх комбінаціях вживання розглянутих слів у корпусі разом або з іншими словами.

Ще один відомий метод, який відноситься до методів перевірки гіпотез – це логарифмічна правдоподібність. Даний метод більше підходить для розріджених даних, ніж критерій хі-квадрат. Застосовуючи логарифмічну правдоподібність до виявлення словосполучень, ми вивчаємо наступні два альтернативні пояснення частоти виникнення біграму w^1w^2:

$$H1\colon P(w^1|w^2) = p = P(w^2|\neg w^1)$$

$$H2\colon P(w^1|w^2) = p_1 \neq p_2 = P(w^2|\neg w^1)$$

$H1$ – перша гіпотеза – формалізація незалежності, тобто поява w2 не залежить від w1, а $H2$ – друга гіпотеза – формалізація залежності, що є свідченням гарного зразка словосполучення. [5]

Інший розглянутий метод – це точкова взаємна інформація, що є статистичним методом, який порівнює ймовірність знайти два слова поруч одне з одним із ймовірністю того, що ці два слова незалежні одне від одного. Якщо x'і y' є словосполученням, ймовірно, що $P(x'y')$ буде значно більшим, ніж $P(x') * P(y')$. Таким чином, чим вище значення, тим більша ймовірність, що два слова є словосполученням. [5]

$$I(x'y') = log_2 \frac{P(x'y')}{P(x') * P(y')}$$

Однак взаємна інформація, як правило, дає велике значення для рідкісних подій. Наприклад, коли деякі слова найчастіше вживаються разом, метод взаємної інформації поверне вище значення, якщо частота їх знаходження в тексті буде меншою. Також точкова взаємна інформація погано вловлює інтуїтивне уявлення про хороше словосполучення, тому даний метод часто не обирають для практичного застосування. [7]

Програмні засоби формування та обробки бази даних словосполучень української мови. Дане дослідження проводиться з практичною метою – за результатами аналізу текстів та знаходження в них всіх словосполучень, які вдасться виділити за допомогою обраного алгоритму – побудувати словник словосполучень української мови. Створена база словосполучень дозволить шукати слова, які найчастіше вживаються разом, також можна буде переглянути різні тип словосполучень та речення-приклади для того, щоб зрозуміти контекст, в якому найчастіше вживається шукане слово. Такий словник може бути корисним, як для використання людьми, наприклад, у навчанні, так і для потреб NLP, адже кількість україномовних ресурсів, які можна використати для досліджень у даній сфері дуже обмежена.

Подібні системи вже існують для інших мов, вони також вміють працювати зі словосполученнями, але особливість системи, що буде створена в результаті даного дослідження буде в тому, що вона автоматично оновлюватиметься використовуючи декілька джерел, кількість яких буде зростати з часом. Тобто даний словник не буде обмежений в кількості джерел, він буде весь час поповнюватися новими текстами різних жанрів та авторів.

Через те що система повинна буде оброблювати велику кількість даних, а алгоритми нормалізації слів та виділення словосполучень працюють недостатньо швидко, планується використати систему Apache Spark. Вона дозволяє швидше обробляти набори великих даних, розділяючи роботу на частини та розподіляючи ці фрагменти між обчислювальними ресурсами. [10] Таким чином можна обробляти велику кількість текстів за порівняно невелику кількість часу та постійно оновлювати словник.

Виділення словосполучень з тексту є ключовою частиною системи, для неї буде використаний один з алгоритмів пошуку, що був розглянутий вище. Вибір алгоритму та його обґрунтування на основі експерименту будуть розглянуті у наступному пункті даної статті.

Ще одною важливою частиною даної системи є пошук. Описана база словосполучень буде корисною тільки якщо можна буде зручно отримати із неї слова, що найчастіше вживаються із шуканим словом та приклади вживання даного слова. Ця частина системи буде реалізована за допомогою Elasticsearch, тому що він дозволяє швидко зберігати великі обсяги даних та шукати інформацію в них. Він дозволяє легко писати складні запити для пошуку за будь-якими критеріями, шукати під час введення тексту, а також збирати статистику за великими обсягами даних, і що не менш важливо він добре масштабується. [11]

Розроблювана система має такий алгоритм роботи:

1. завантаження текстів в систему,
2. попередня обробка текстових даних (нормалізація, фільтр за стоп-словами та ін.),
3. застосування методу або декількох методів знаходження колокацій,
4. фільтрація колокації,
5. індексування даних та завантаження до пошукової системи.

При чому виконання пунктів 2–4 саме будуть прискорені за допомогою Spark, а в якості системи пошуку буде використаний Elasticsearch.

Отже, розроблювана система вирізняється серед інших подібних автоматизованим наповненням словника та прискореним аналізом текстів, на основі яких саме будується база словосполучень.

Експериментальна оцінка. В ході дослідження були проведені порівняльні оціночні експерименти наведених вище методів виділення словосполучень з використанням корпусу текстів української мови UA-GEC, який є першим анотованим корпусом української мови. Цей корпус розроблений і підтримується компанією Grammarly, та знаходиться у вільному доступі. Даний корпус складається з текстів різних авторів, жанрів, сфер наукової та суспільної діяльності. Всього корпус складається з 328771 слів, 20715 речень та зібраний за участі 492 авторів. [12] Даний корпус має зручний Python пакет. Усі корпусні дані та метадані знаходяться в директорії *./data*. Ця директорія має дві вкладені папки для розділів *навчальний* та *тестовий*. При проведенні експерименту були використані текстові дані з обох розділів та всього у вибірку для експерименту попали 166330 слів.

Перед початком експерименту була проведена попередня обробка тексту. Текст був завантажений, розбитий на окремі токени, очищений від стоп-слів, символів пунктуації, посилань та цифр. Опціональним етапом попередньої обробки також є лематизація слів – таким чином можна збільшити кількість слів, що зможе підрахувати алгоритм, оскільки слова, що мають однакову лему, але вжиті у тексі в іншій формі, наприклад, в іншому часі, або роді, не будуть рахуватися як різні слова. Але експериментальним шляхом було визначено, що результати отримані з попередньою лематизацією, менш лексично вірні, ніж ті, що отримані без неї.

Для проведення дослідження була обрана Python бібліотека Natural Language Toolkit, яка є провідною платформою для проведення досліджень та побудови додатків, що працюють з текстами та натуральними мовами. [13] Дана бібліотека має імплементацію описаних методів виділення словосполучень та дозволяє провести дослідження ефективності даних методів для української мови.

Всі знайдені словосполучення були відфільтровані за принципом – в словосполучення не повинні входити слова менше трьох символів, оскільки найчастіше, це службові частини мови і вони не представляють лексичного інтересу у складі словосполучень, а також результати повинні підпадати під певну структуру утворення словосполучення.

За допомогою бібліотеки py-morphy ми можемо визначити частину мови до якої відноситься певне слово та перевірити чи словосполучення утворене правильно. [14] Тобто перевіряється чи є словосполучення іменним, дієслівним або прислівниковим. Для цього кожне слово перевіряється на відповідність певним частинам мови згідно із правилами утворення даних типів словосполучень.

Отже, під час дослідження тексти корпусу UA-GEC були проаналізовані на предмет наявності біграм та триграм за допомогою різних методів виділення словосполучень в тексті.

На Рис. 1. Можна побачити результати 20 найкращих біграм виділених за допомогою обраних методів, а на Рис. 2 можна побачити результати аналогічного експерименту для словосполучень з трьох слів.

Ми бачимо, що методи точкової взаємної інформації та критерій хі-квадрат дають хороші результати. Їх результати також досить схожі, але метод взаємної інформації працює краще для виділення біграм в даному корпусі. Методи підрахунку та t-критерію також подібні один до одного.

	Frequency	PMI	T-test	Chi-Sq Test	Likeihood Ratio Test
0	(мою, думку)	(потерпимо, зневаги)	(мою, думку)	(поцілувався, страшненькою)	(мою, думку)
1	(штучного, інтелекту)	(потужно, інформувати)	(штучного, інтелекту)	(посідала, територіальної)	(штучного, інтелекту)
2	(точки, зору)	(потужним, ультрафіолетом)	(точки, зору)	(посіви, знищуються)	(точки, зору)
3	(наступного, дня)	(потугу, руїнної)	(наступного, дня)	(посунулась, милю)	(ніна, іванівна)
4	(має, право)	(потрошки, поверталось)	(має, право)	(посуваються, встаючи)	(молекулярної, динаміки)
5	(ніна, іванівна)	(потребую, старту)	(ніна, іванівна)	(постійною, міграцією)	(сталого, розвитку)
6	(сталого, розвитку)	(потребувало, гнучкої)	(сталого, розвитку)	(постійних, відрядженнях)	(державною, мовою)
7	(державною, мовою)	(потрапляю, висловлюю)	(державною, мовою)	(постулати, недооцінені)	(надання, впевненості)
8	(останнім, часом)	(потраплю, ідеальну)	(останнім, часом)	(постулат, виправданий)	(наступного, дня)
9	(дає, змогу)	(потонули, земному)	(дає, змогу)	(пострілу, снайперської)	(своєю, чергою)
10	(теорії, ігор)	(потомства, розрада)	(надання, впевненості)	(постригся, ченці)	(маркіяна, шашкевича)
11	(своєю, чергою)	(потицяти, віртуалці)	(своєю, чергою)	(поставлять, крапельницю)	(суспільно, орієнтоване)
12	(іншого, боку)	(потилиця, стиснена)	(теорії, ігор)	(посприяло, активній)	(суспільно, орієнтованого)
13	(надання, впевненості)	(потенціали, згладжують)	(іншого, боку)	(посприяли, зникненню)	(останнім, часом)
14	(сказав, скрудж)	(пострілу, снайперської)	(сказав, скрудж)	(посполитих, посполиті)	(другої, світової)
15	(теорія, ігор)	(потемніли, примарилось)	(молекулярної, динаміки)	(посольстві, культурний)	(дає, змогу)
16	(пів, години)	(потаповича, скупого)	(другої, світової)	(послідовний, скований)	(теорії, ігор)
17	(другої, світової)	(потайний, замкнутий)	(теорія, ігор)	(послуговуватися, транспортним)	(усунення, неоднозначності)
18	(молекулярної, динаміки)	(посіяло, насіння)	(пів, години)	(послана, филипа)	(ямної, культури)
19	(двадцять, років)	(посідала, територіальної)	(маркіяна, шашкевича)	(поскакушками, радували)	(ніни, іванівни)

Рис.1. Порівняння списків з двадцяти найкращих результатів для всіх розглянутих методів виділення словосполучень, що складаються з двох слів

	Frequency	PMI	T-test	Chi-Sq Test	Likeihood Ratio Test
0	(суспільно, орієнтоване, навчання)	(*навчання, провадити, кейсовим)	(суспільно, орієнтоване, навчання)	(*навчання, провадити, кейсовим)	(суспільно, орієнтоване, навчання)
1	(звітності, сталого, розвитку)	(поповнилося, залізницею, тунелем)	(суспільно, орієнтованого, навчання)	(поповнилося, залізницею, тунелем)	(суспільно, орієнтованого, навчання)
2	(суспільно, орієнтованого, навчання)	(порішав, затащили, катки)	(звітності, сталого, розвитку)	(порішав, затащили, катки)	(мою, думку, людина)
3	(двадцять, років, довгий)	(порізав, електричну, проводку)	(двадцять, років, довгий)	(порізав, електричну, проводку)	(мою, думку, бурхлива)
4	(аудиту, звітності, сталого)	(поріжте, кубиками, ребром)	(всередині, вуглецевих, нанотрубок)	(поріжте, кубиками, ребром)	(мою, думку, прояви)
5	(сказав, високий, чоловік)	(порядне, ґаздівство, остапову)	(аудиту, звітності, сталого)	(порядне, ґаздівство, остапову)	(мою, думку, озвучкою)
6	(другої, світової, війни)	(поршні, заскрипіли, штовхнули)	(другої, світової, війни)	(поршні, заскрипіли, штовхнули)	(мою, думку, найефективнішим)
7	(всередині, вуглецевих, нанотрубок)	(поршневих, шиберних, відцентрової)	(води, всередині, вуглецевих)	(поршневих, шиберних, відцентрової)	(мою, думку, порушується)
8	(води, всередині, вуглецевих)	(порцію, демотивації, вивчи)	(сказав, високий, чоловік)	(порцію, демотивації, вивчи)	(лінивих, мою, думку)
9	(ресурсів, черпаєш, корисну)	(портів, танкє, горєл)	(ресурсів, черпаєш, корисну)	(портів, танкє, горєл)	(єпатажу, мою, думку)
10	(черпаєш, корисну, інформацію)	(портупеї, меча, зашнурованому)	(користувацького, інтерфейсу, державною)	(портупеї, меча, зашнурованому)	(дидероти, мою, думку)
11	(одному, міських, департаментів)	(портативними, лікарнями, медиками)	(черпаєш, корисну, інформацію)	(портативними, лікарнями, медиками)	(акторами, мою, думку)
12	(користувацького, інтерфейсу, державною)	(порода, німецька, вівчарка)	(таблиці, каскадних, стилів)	(порода, німецька, вівчарка)	(мою, думку, ключовий)
13	(яких, ресурсів, черпаєш)	(поразкам, захоплюватися, подвигами)	(бобе, чикаго, думає)	(поразкам, захоплюватися, подвигами)	(романтика, мою, думку)
14	(старі, добрі, часи)	(порадитися, тамтешніми, інженерами)	(мер, сидячи, ліжку)	(порадитися, тамтешніми, інженерами)	(символічні, мою, думку)
15	(інтерфейсу, державною, мовою)	(попереджає, пластиковій, тарі)	(одному, міських, департаментів)	(попереджає, пластиковій, тарі)	(мою, думку, наступна)
16	(бобе, чикаго, думає)	(посиджу, втиснулась, покривало)	(інтерфейсу, державною, мовою)	(посиджу, втиснулась, покривало)	(мою, думку, правильного)
17	(учасників, бойових, дій)	(поодиноких, беріз, чагарників)	(старі, добрі, часи)	(поодиноких, беріз, чагарників)	(творчість, мою, думку)
18	(часи, двоє, чоловіків)	(пообідати, відкритій, терасі)	(часи, двоє, чоловіків)	(пообідати, відкритій, терасі)	(досвід, мою, думку)
19	(таблиці, каскадних, стилів)	(помішувала, курячий, бульйон)	(яких, ресурсів, черпаєш)	(помішувала, курячий, бульйон)	(філософське, мою, думку)

Рис.2. Порівняння списків з двадцяти найкращих результатів для всіх розглянутих методів виділення словосполучень, що складаються з трьох слів

Аналіз результатів та критерій порівняння

В даній статті були описані різні методи знаходження колокацій у тексті, в попередньому розділі також були наведені результати роботи даних методів, але все ще не зрозуміло який з них найкраще справляється з поставленим завданням.

В даному дослідженні був використаний підхід, описаний Евертом [15], суть якого полягає в виділенні невеликої випадкової вибірки позитивних і негативних прикладів (тобто n-грамів, що є колокаціями та ні відповідно), які будуть використані для обчислення влучності та повноти серед n-кращих кандидатів для кожної міри. Позитивні приклади були отримані шляхом ручного фільтрування людиною випадково вибраних n-грамів з корпусу.

Випадкова вибірка для біграм складається з 229 колокацій (вважаються позитивними прикладами) та 229 неколокацій (вважаються негативними прикладами), так само й для триграмів. Звісно, в дійсності вибріка буде відрізнятися, адже зазвичай буде більше негативних, ніж позитивних прикладів. Але це дає нам надійну основу для порівняння розглянутих мір асоціації, оскільки можна очікувати, що відносна ефективність вимірювання не залежить від тестової вибірки.

Результати для біграм наведені на Рис 3. вони були отримані після застосування POS-фільтра та фільтру за списком стоп-слів. На Рис 3. видно, що всі перевірені показники працюють краще, ніж простий підрахунок частоти, що виправдовує використання мір асоціації. Також видно, що метод точкової взаємної інформації (PMI) працює найкраще, за ним слідують хі-квадрат та логістична правдоподібність, а найгірше себе показав t-критерій Стьюдента.

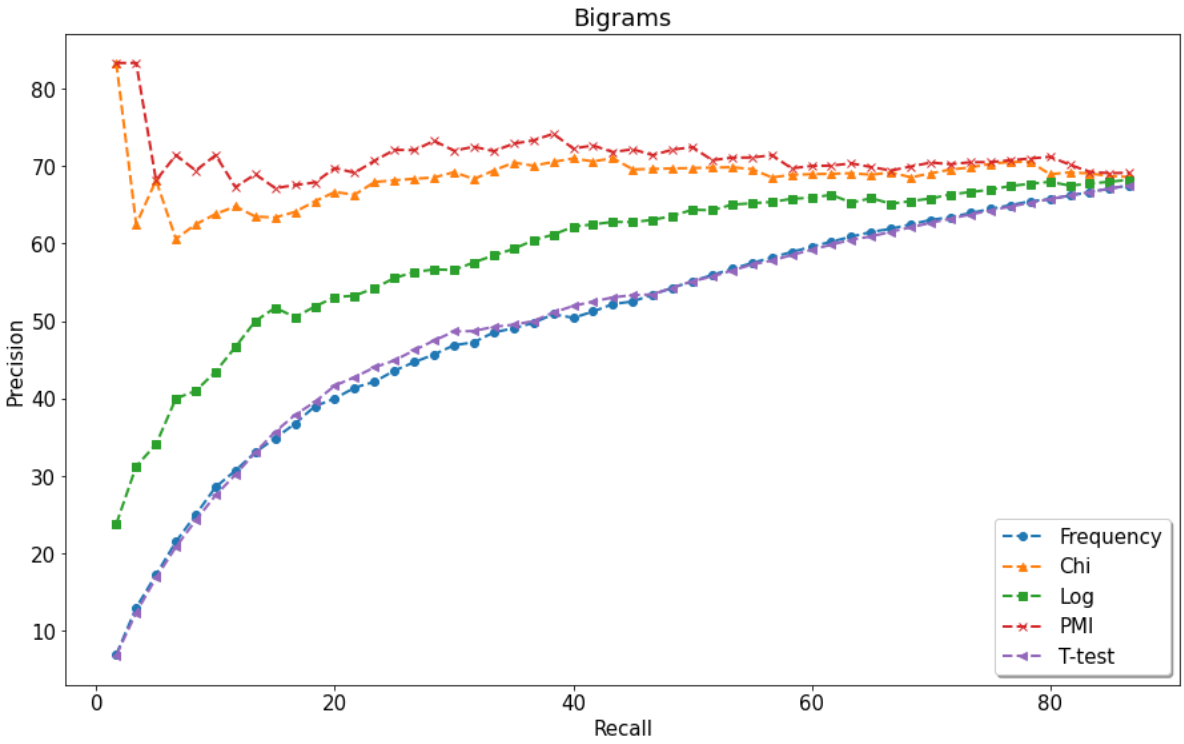

Рис. 3. Результати обчислення влучності та повноти для біграм

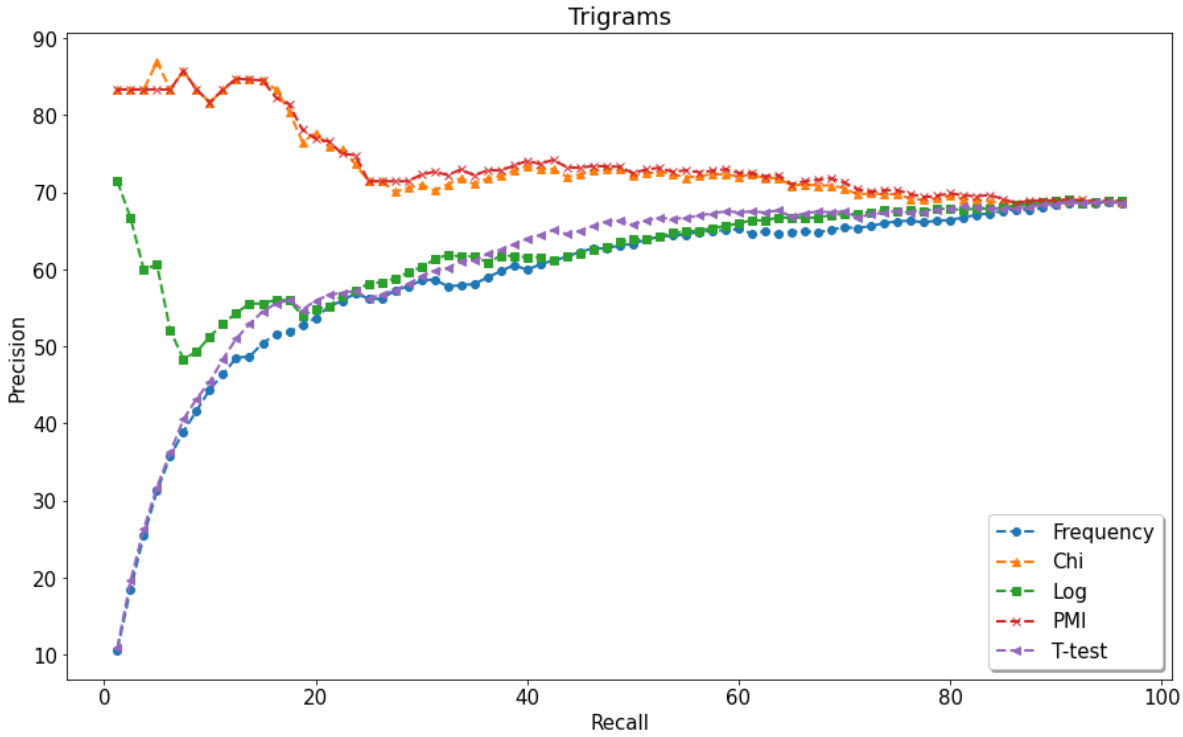

Рис. 4. Результати обчислення влучності та повноти для триграм

Результати для триграм показані на Рис 4. Метод точкової взаємної інформації перевершив всі інші розглянуті методи, на другому місці знову хі-квадрат, а от логарифмічна правдоподібність цього разу спрацювала так само, як і t-критерій та простий підрахунок, отже можна зробити висновок, що для різних типів колокацій методи працюють по-різному.

Отже, для виділення словосполучень у текстах з метою побудови словника словосполучень української мови може бути обраний метод точкової взаємної інформації. Також можна проводити тести на різних корпусах, щоб дізнатися, який метод працює найефективніше для обраного набору даних. Крім того, можна об'єднати результати роботи декількох методів, що ефективні для досліджуваного корпусу. В даному випадку на думку автора було б ефективно поєднати списки словосполучень виділених методом точкової взаємної інформації та методом логарифмічної правдоподібності.

Висновок

Були досліджені деякі з найпоширеніших методів статистичного аналізу текстів з метою виділення словосполучень. Дані методи були описані та проаналізовані з теоретичного та експериментального боку. За допомогою виділення біграм та триграм із текстів корпусу UA-GEC та порівняння результатів між собою було досліджено, який з наведених методів найкраще справляється з поставленою задачею.

Даний аналіз був проведений для побудови статистичної моделі для подальшого використання разом із іншими засобами обробки природної мови з метою подальшого використання для побудови бази словосполучень української мови. Розроблювана система буде постійно оновлюватися, джерела для пошуку словосполучень будуть додаватися, а знайдені словосполучення вже автоматично потраплятимуть до словника.

Основною перевагою такого словника буде відносна автономність, тобто його не потрібно буде наповнювати людям, наповнення відбуватиметься автоматично. Дана система також матиме пошук найчастіше вживаних слів разом зі словом, що ввів у систему користувач. Даний словник може бути використаний як допоміжний матеріал для вивчення української мови та як ресурс для досліджень у сфері NLP.

Майбутні дослідження будуть зосереджені на виділенні словосполучень, що складаються більше ніж з трьох слів, або є досить незалежними, щоб бути знайденими за допомогою простих статистичних методів. Також дослідження будуть зосереджені на збільшенні швидкості обробки текстових корпусів для більш швидкого та ефективного наповнення бази словосполучень. Остання та не менш важлива сфера для майбутніх досліджень – це аналіз накопичених даних, наприклад, словосполучення можуть бути використані в системах рекомендацій, в текстових редакторах, тощо.

References

1. Snedecor, George Waddel, and William G. Cochran. 1989. *Statistical methods. Ames:* Iowa State University Press. 8th edition. 53 c.
2. Church K. and Hanks P., 1990. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics.
3. Sinclair, John ted. 1995. ColIins COBUILD English dictionary. London: Harper Collins. New edition, completely revised.
4. Manning C. and Schütze H., 1999. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.
5. Smadja F., 1993. Retrieving Collocations from text: Xtract, Computational Linguistics, 19: 143-177.
6. Dunning T., 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics.
7. Fontenelle, Thierry, Walter Briils, Luc Thomas, Tom Vanallemeersch, and Jacques Jansen. 1994. DECIDE, MLAP-Project 93-19, deliverable D-la: survey of collocation extraction tools. Technical report, University of Liege, Liege, Belgium.
8. Hawthorne, Mark. 1994. The computer in literary analysis: Using TACT with students. Computers and the Humanities.
9. Church, Kenneth W., and Robert L. Mercer. 1993. Introduction to the special issue on computational linguistics using large corpora. Computational Linguistics. 20 c.
10. Apache Spark - Unified Analytics Engine for Big Data. URL: https://spark.apache.org/ (дата звернення 06.09.2021)
11. Free and Open Search: The Creators of Elasticsearch, ELK & Kibana | Elastic. URL: https://www.elastic.co/ (дата звернення 06.09.2021)
12. UA-GEC: перший анотований GEC-корпус української мови вже у вільному доступі! URL: https://ua-gec-dataset.grammarly.ai/ (дата звернення 05.09.2021)
13. Natural Language Toolkit — NLTK 3.6.2 documentation. URL: https://www.nltk.org/ (дата звернення 05.09.2021)
14. Морфологический анализатор pymorphy2 — Морфологический анализатор pymorphy2. URL: https://pymorphy2.readthedocs.io/en/stable/ (дата звернення 05.09.2021)
15. S. Evert, B. Krenn, Using small random samples for the manual evaluation of statistical evaluation measures. Computer speech and language, 19: pp. 450–466; 2005.