International Science Journal of Engineering & Agriculture
2024; 3(4): 87-92
https://isg-journal.com/isjea
doi: 10.46299/j.isjea.20240305.08
ISSN: 2720-6319

INTERNATIONAL SCIENCE GROUP

Автоматизоване формування синонімічних рядів для елементів логіко-лінгвістичної моделі речення природньої мови

Андрій Динько
кафедра комп'ютерних систем та мереж, Національний авіаційний університет, Київ, Україна
ORCID 0009-0002-5414-4301

Для цитування цієї статті:
Динько Андрій. Автоматизоване формування синонімічних рядів для елементів логіко-лінгвістичної моделі речення природньої мови. International Science Journal of Engineering & Agriculture. Vol. 3, No. 5, 2024, pp. 87-92. doi: 10.46299/j.isjea.20240305.08.
Надійшла до редакції: 10 вересня 2024 р.; **Схвалено:** 30 вересня 2024 р.;
Опубліковано: 01 жовтня 2024 р.

Анотація: У статті описано процес автоматизованого формування списків синонімів для кожного елемента логіко-лінгвістичної моделі речення української мови на основі раніше сформованих предикатів, процес формування яких описано в проаналізованих наукових публікаціях, для подальшого використання як компонент у системах порівняння електронних текстів за змістом. Для опису даного процесу складено блок-схему алгоритму і програмно реалізовано за допомогою середовища розробки програмного забезпечення *Visual Studio 2019*, з використанням засобів *.NET Windows Forms* та прикладу бази даних синонімів слів української мови. Для заповнення таблиць у базі даних запропоновано використовувати три поля для ідентифікації вмісту: ідентифікаційний номер, слово та масив слів синонімів до даного слова, розділений символом крапка з комою. На виході роботи програмного модуля отримано об'єкт, як екземпляр класу, що містить елементи логіко-лінгвістичної моделі речення та списки їхніх синонімів як приватні поля і публічні властивості даного класу.

Ключові слова: логіко-лінгвістичні моделі, текстові аналізатори, порівняння текстів за змістом.

1. Вступ

Кількість інформації, котра подана у вигляді електронних текстів, зростає щохвилини і разом із цим з'являється необхідність у їхньому аналізі та пошуці серед даного масиву даних необхідного для кінцевого користувача об'єму. Вседоступність інформації також зачіпає поняття інтелектуальної власності, що досить часто порушується або через недостатню обізнаність, або ж навмисно. Разом із появою все новіших системи для виявлення повторень, з'являються і нові методи обходу виявлення використання чужої праці. На основі аналізу публікацій, пропонується посприяти визначенню повторень змісту тексту, попри використання перефразування і словозамінювання аналогами слів, шляхом використання логіко-лінгвістичного моделювання.

2. Об'єкт і предмет дослідження

Об'єктом даного дослідження є процес автоматичної побудови логіко-лінгвістичних моделей простих речень природної мови. Предметом дослідження є системи аналітичної обробки текстової інформації за змістом.

3. Мета та задачі дослідження

Метою статті є покращення технології автоматизованого формування логіко-лінгвістичних моделей шляхом додавання синонімічних рядів елементів моделі для збільшення рівня виявлення повторень текстів за змістом. Для досягнення поставленої мети необхідно проаналізувати попередні дослідження щодо автоматизованої побудови логіко-лінгвістичних моделей речень природньої мови та описати процес автоматичного формування синонімічних рядів елементів моделі речення. Написати програмний модуль що виконуватиме автоматизоване формування списків синонімів для кожного елемента моделі.

4. Аналіз літератури

В наукових працях [1-2] описуються загальна проблематика аналізу текстів за змістом та методи що намагаються вирішити ці питання. Одним із способів подати текст у зрозумілому для комп'ютерного пристрою вигляді є використання логіко-лінгвістичного моделювання як засобу представлення у вигляді предикатів складові тексту, що описується в [3-6]. Поданий у вигляді логіко-лінгвістичної моделі текст сприяє полегшенню виявлення повторень за змістом та інших операцій пов'язаних із аналізом електронного тексту, як описано в [7-11].

Для формування логіко-лінгвістичної моделі речень української мови слід виконати наступні кроки: зчитати текст, виконати парсинг тексту на речення, виконати поділ речення на слова, провести морфологічний аналіз слів, провести синтаксичний аналіз речень, сформувати словосполучення на основі проведеного раніше аналізу, побудувати моделі речень, як описано в [1, 12-13]. Виконавши дані кроки і отримавши сукупність моделей речень, можна провести співставлення із моделями, сформованими на основі іншого тексту, таким чином можна виявити збіги за змістом попри перефразування слів речення. Для покращення порівняння пропонується додати списки синонімів для сформованих моделей, щоб збільшити кількість виявлення збігів попри використання синонімів при переписуванні тексту. Пошук синонімів описаний в[14-15].

В даній статті пропонується на основі вихідних даних дослідження, що описано в [12], тобто масиву логіко-лінгвістичних моделей речень української мови, описати і програмно реалізувати автоматизований процес формування синонімічних рядів для елементів цих даних.

5. Методи досліджень

Метод порівняння логіко-лінгвістичних моделей[7], методи обробки текстової інформації[1, 2], методи представлення знань[5-6, 8], методи синтаксичного аналізу[1-2, 13]

6. Результати досліджень

В [1] речення природньої мови подається у вигляді простого предикату, що описує частину цього речення, яке має закінчений зміст та відображає у реченні S p-е відношення з h-ю характеристикою між суб'єктом x з характеристикою g і об'єктом y з характеристикою q, предмет якого z володіє характеристикою r:

$$L_p^S(x, g, y, q, z, r, h) \quad (1)$$

Технологія автоматизованої побудови логіко-лінгвістичної моделі (1) простого речення української мови описана в [12]. Пропонується для кожного елемента сформованої моделі речення додати за наявності його синонімічний ряд, тобто масив синонімів. Для програмної реалізації доцільно створити клас, що міститиме елементи моделі та списки для подальшого зберігання синонімів як поля класу як показано на рис. 1.

```
class LLM
{
    string p, x, g, y, q, z, r, h; //елементи моделі
    List<string> synonimsP, synonimsX, synonimsZ, synonimsG, synonimsY, synonimsQ, synonimsR, synonimsH; //списки що міститимуть синоніми для елементів
    //конструктор
    1 reference
    public LLM(string init_p, string init_x = "0", string init_g = "0", string init_y = "0", string init_q = "0", string init_z = "0", string init_r = "0", string init_h = "0")
    {
        this.p = init_p;
        this.x = init_x;
        this.g = init_g;
        this.y = init_y;
        this.q = init_q;
        this.z = init_z;
        this.r = init_r;
        this.h = init_h;
        synonimsP = new List<string>();
        synonimsX = new List<string>();
        synonimsG = new List<string>();
        synonimsY = new List<string>();
        synonimsQ = new List<string>();
        synonimsZ = new List<string>();
        synonimsR = new List<string>();
        synonimsH = new List<string>();
    }
```

Рис. 1. Поля та конструктор класу для зберігання логіко-лінгвістичної моделі речення.

Для формування масиву синонімів кожного елемента логіко-лінгвістичної моделі слід до описаного в [12] програмного модуля автоматизованої побудови логіко-лінгвістичної моделі простих речень української мови додати можливість зчитувати електронний словник синонімів, тобто підключити базу даних з таблицями синонімів українських слів. Оскільки одним і вихідних даних програмного модуля є модель що має вигляд як описано в (1), то для даного дослідження ці дані використовувалися як вхідні – тобто елементи моделі вводяться вручну. Після підключення бази даних слід провести пошук за кожним з елементів моделі і зберегти видобуті за наявності синоніми у список або динамічний масив. Блок-схема алгоритму автоматизованого формування синонімічного ряду для логіко лінгвістичної моделі речення української мови зображено на рис. 2.

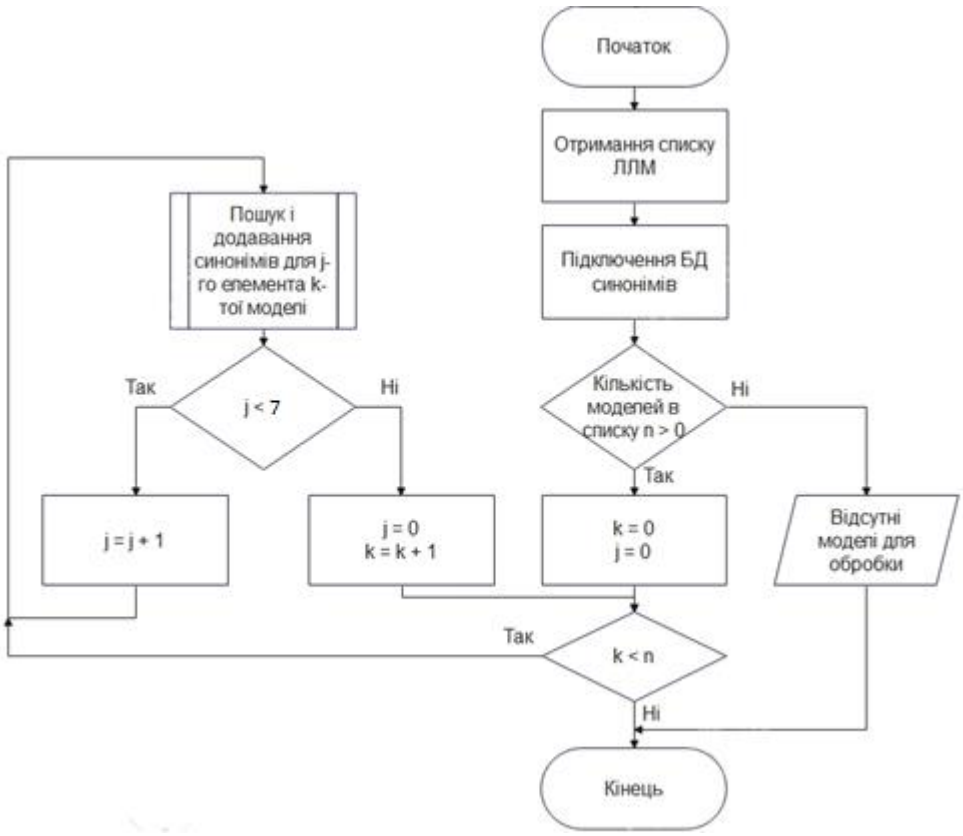

Рис. 2. Блок-схема формування синонімічних рядів для елементів логіко-лінгвістичної моделі.

Пошук і додавання синонімів для j-го елемента k-тої логіко-лінгвістичної моделі речення, як частина схеми на рисунку 2, подано на рис. 3.

Рис. 3. Блок-схема отримання списку синонімів з бази даних синонімів.

Запис у базі даних синонімів має вигляд як показано на рис. 4, де наявні три стовпці в таблиці: ідентифікатор рядка, слово, список синонімів для даного слова розділених символом крапка з комою.

Рис. 4. Приклад запису у таблиці синонімів.

Для речення "*Дозрілу бульбу регулярно подають на обідній стіл*", логіко-лінгвістична модель матиме вигляд: *подають(0, 0, бульбу, дозрілу, стіл, обідній, регулярно)*. Відповідно результат автоматизованого формування синонімічних рядів для елементів даної моделі зображено на рис. 5.

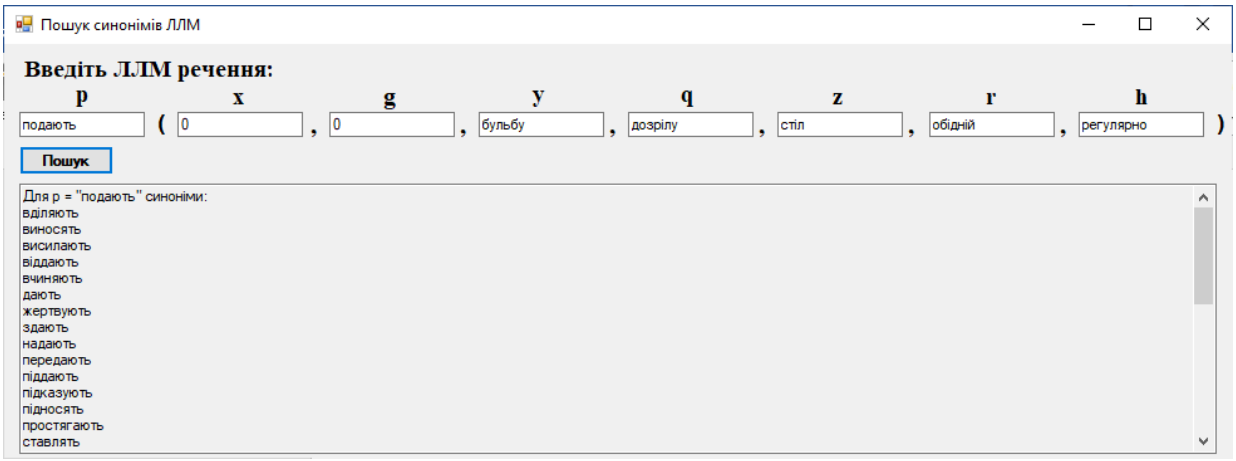

Рис. 5. Сформовані синоніми для елементів моделі

Результатом пошуку і формування списків синонімів моделі є об’єкт класу(рис. 1) що містить в собі значення елементів логіко-лінгвістичної моделі речення та значення синонімічних рядів цих елементів як подано на рис. 6.

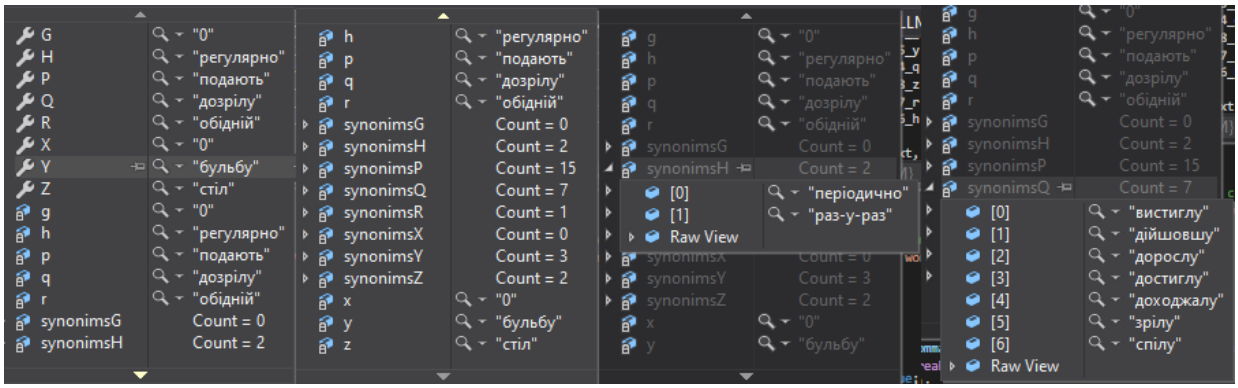

Рис. 6. Вміст об’єкта після обробки моделі.

7. Перспективи подальшого розвитку досліджень

Результати даного дослідження можна використовувати як елемент для систем виявлення повторень текстів за змістом, а також для іншого роду текстових аналізаторів. Алгоритм формування синонімічних рядів можна удосконалити шляхом видозміни елементів моделі під час пошуку синонімів у вигляді початкової форми слова, що сприятиме більшій варіативності результатів пошуку в базі даних. Пропонується використовувати даний компонент для формування масиву об’єктів класу, що містить в собі опис логіко-лінгвістичної моделі речення та списків синонімів цих елементів, кожного речення тексту, таким чином подаючи його у вигляді зручному для подальшого порівняння та обробки.

8. Висновки

Описаний в даному дослідженні процес автоматизованого формування синонімічних рядів для елементів логіко-лінгвістичної моделі речення природньої мови на виході генерує масив даних що сприятиме виявленню повторень електронних текстів за змістом незважаючи на використання синонімів як аналогів для заміни вмісту тексту з метою приховати дані збіги від програмних засобів для аналізу текстів. Було створено програмний модуль, що виконує описані в статті кроки для побудови списків синонімів кожного елемента моделі. За можливості використання бази даних синонімів, що містить дані подані у вигляді відмінному від того яке використовувалося у даній роботі, загальний алгоритм формування синонімічних рядів не зазнає змін, проте для коректного функціонування програмного модуля слід замінити частину коду що стосуватиметься запитів до нового компонента. Вихідні дані програмного модуля можуть слугувати основою для нових ідей у текстовому аналізі.

Список літератури:

1) Вавіленкова, А. І. (2016). *Теоретичні основи аналізу електронних текстів: монографія*. Київ: ТОВ «СІК ГРУП УКРАЇНА».

2) Вавіленкова, А. І. (2013). Аналіз методів обробки текстової інформації. *Вісник Національного технічного університету ХПІ. Серія: Інформатика та моделювання*, (39), 35-40.

3) Вавіленкова, А. І. (2017). Правила синтезу логіко-лінгвістичних моделей речень природної мови. *Наукоємні технології*, (1), 3-7.

4) Вавіленкова, А. І. (2014). Алгоритм побудови логіко-лінгвістичної моделі текстового документу. *Электротехнические и компьютерные системы*, (16), 85-92.

5) Вавіленкова, А. (2015). Побудова змістовної моделі тексту на основі використання логіко-лінгвістичних моделей. *Вісник Національного університету Львівська політехніка. Комп'ютерні науки та інформаційні технології*, (826), 169-175.

6) Вавіленкова, А. І. (2017). *Аналіз і синтез логіко-лінгвістичних моделей речень природної мови: монографія*. Київ: ТОВ «СІК ГРУП УКРАЇНА».

7) Вавіленкова, А. І. (2012). Логіко-лінгвістичні моделі речень як засіб порівняння текстових документів за змістом. *Математические машины и системы*, *1*(1), 166-173.

8) Вавіленкова, А. І. (2015). Методологічні основи автоматичного аналізу логіко-лінгвістичних моделей текстових документів. *Математичні машини і системи*.

9) Вавіленкова, А. І. (2015). Система порівняльного аналізу текстових документів. *Вісник Житомирського державного технологічного університету. Серія: Технічні науки*, (4), 94-100.

10) Вавіленкова, А. І. (2015). Умови тотожності логіко-лінгвістичних моделей простих речень природної мови. *Вісник Національного технічного університету ХПІ. Серія: Інформатика та моделювання*, (32), 27-35.

11) Вавіленкова, А. І. (2012). Алгоритм порівняння логіко-лінгвістичних моделей речень природної мови. *Системи підтримки прийняття рішень. Теорія і практика: зб. доп. наук.-практ. конф. з міжнар. участю.–Київ: ІПММС НАНУ* (с. 132-135).

12) Динько, А. Ю. (2020). Технологія автоматизованої побудови логіко-лінгвістичних моделей.[магістерська дипломна робота, Національний авіаційний університет]. Репозитарій Національного авіаційного університету. http://er.nau.edu.ua/handle/NAU/38241.

13) Динько, А. Ю. (2024). Автоматизована побудова словосполучень як елемент синтаксичного аналізу для побудови логіко-лінгвістичної моделі. *The 10th International scientific and practical conference "Problems and prospects of modern science and education" Stockholm, Sweden. International Science Group. 2024. 381 p.* (с. 333).

14) Вавіленкова, А. І. (2014). Аналіз методів пошуку синонімів в електронних документах. *Вісник Чернігівського державного технологічного університету. Серія: Технічні науки*, (2), 119-128.

15) Динько, А. Ю. (2024). Використання логіко-лінгвістичних моделей для визначення збігів за змістом у текстових документах. *The I International scientific and practical conference "Innovative scientific research: theory, methodology, practice". Boston, USA. International Science Group. 2024. 289 p.* (с. 274).

Automated formation of synonym sets for elements of the logical-linguistic model of a natural language sentence

Andrii Dynko
Faculty of Computer Science and Technologies, National Aviation University, Kyiv, Ukraine
ORCID 0009-0002-5414-4301

Abstract: The article describes the process of automated formation of synonym lists for each element of the logical-linguistic model of a sentence in the Ukrainian language, based on previously created predicates. The process of their formation is detailed in the analyzed scientific publications, for further use as a component in content-based text comparison systems. A flowchart of the algorithm describing this process has been created and implemented in software using the Visual Studio 2019 development environment, utilizing .NET Windows Forms and an example of a synonym database for Ukrainian words. To populate the database tables, it is proposed to use three fields for content identification: identification number, word, and an array of synonym words for the given word, separated by a semicolon. The output of the program module is an object, as an instance of a class, containing the elements of the logical-linguistic model of a sentence and their synonym lists as private fields and public properties of the class.

Keywords: logical-linguistic models, text analyzers, comparison of texts by content.