МЕТОДИКА ВИОКРЕМЛЕННЯ КЛЮЧОВИХ СЛІВ І СЛОВОСПОЛУЧЕНЬ ТА ПОБУДОВИ НАПРАВЛЕНИХ ЗВАЖЕНИХ МЕРЕЖ ТЕРМІНІВ ІЗ ЗАСТУВАННЯМ PART-OF-SPEECH TAGGING

Д.В. Ланде [0000-0003-3945-1178] О.О. Дмитренко [0000-0001-8501-5313]
Інститут проблем реєстрації інформації НАН України, Київ, Україна
dwlande@gmail.com, dmytrenko.o@gmail.com

У цій роботі запропонований новий метод виокремлення ключових слів і словосполучень з тематичних інформаційних потоків та новий метод встановлення напрямків зв'язків між вузлами у ненаправлених мережах термінів із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Представлено ідею встановлення вагових значень зв'язків між вузлами у направленій мережі термінів. Також представлена цілісна методика комп'ютерної обробки текстових корпусів та побудови направлених зважених мереж термінів (ключових слів та словосполучень), виокремлених за допомогою попереднього процесу класифікації слів за частинами мови та відповідним маркуванням – Part-of-Speech tagging, та подальшого статистичного зважування. Апробацію запропонованої методики було проведено на прикладі алегоричної повісті-казки "Маленький принц" (англ. "The Little Prince") Антуана де Сент-Екзюпері. Застосовуючи запропонований метод було виокремлено ключові терміни та побудовано направлену зважену мережу зі слів та словосполучень, які відповідають окремим ключовим поняттям у досліджуваному творі.

Ключові слова: текстовий корпус, обробка природньої мови, Part-of-speech (PoS) tagging, термінологічна онтологія, мережа термінів.

Постановка проблеми

Ця стаття присвячена вирішенню актуального науково-практичного завдання, що стосується концептуалізації та подальшої формалізації у вигляді мережі термінів неструктурованих текстових даних, що містяться у тематичних інформаційних потоках розподілених в мережі Інтернет.
Враховуючи той факт, що багато задач, що виникають під час роботи з текстовими інформаційними потоками, лежать на перетині між математичними науками та лінгвістикою, то це відкриває широкі можливості для застосування потужного математичного апарату та лінгвістичної теорії.

Метою цієї роботи є запропонувати новий метод визначення напрямків зв'язків між вузлами ненаправленої мережі, побудованої із ключових слів та словосполучень тематичного текстового масиву, щоб будувати термінологічні онтології у вигляді направлених мереж термінів для того, щоб у подальшому робити конструктивні висновки щодо мережевої структури та її параметрів, та на основі цього приймати ефективні рішення у

відповідно розглянутих проблемних предметних галузях, з якими змістовно пов'язані тексти.

Методика

Побудова направленої мережі термінів здійснюється в межах кожного окремого речення текстового корпусу.

У цій роботі для автоматичного розбиття на токени та розмічування тексту й присвоєння тегів кожному слову застосовуються відповідно окремі функції "word_tokenize" та "pos_tag" спеціалізованої надбудови – модуля NLTK (Natural Language Toolkit), що розроблений на мові програмування Python [1].

Також у цій роботі окрім стандартних наборів стоп-слів, що доступні за посиланнями [2], [3] пропонується використовувати список стоп-слів, що сформований експертами в межах досліджуваної предметної галузі.

Запропонований у цій роботі метод визначення ключових слів та словосполучень, а також напрямків зв'язків базується на використанні результатів отриманих за допомогою процесу класифікації слів за частинами мови та відповідним маркуванням – розмічування частин мови (Part-of-Speech tagging) [4]. Виходячи із практичних досліджень [1] можна помітити, що найбільш вживаними членами речення у англійській мові є артиклі (DT – determiner), іменники (NN – sing or mass noun, NNS – plural noun), займенники (PR – personal pronoun), дієслова (VB – verb base form), означення (JJ – adjectives) та прислівники (RB – adverb). Загалом ключовими словами являються окремі іменники, що зазвичай стосуються людей, місць, речей чи концептів, та іменники у парі з означеннями – словосполучення виду "JJ NN". Також в цій роботі вважається, що важливими можуть бути словосполучення виду "NN_1 NN_2", "JJ_1 JJ_2", "JJ_1 JJ_2 NN", "JJ_1 JJ_2 NN_1 NN_2". Хоча артиклі, прийменники (IN – preposition), сполучники (CC – conjunction, coordinating), окремі дієслова, прислівники та займенники являються стоп-словами, проте словосполучення виду "VV_1 to VV_2" "NN_1 IN/CC NN_2", "JJ_1 IN/CC JJ_2", "JJ NN_1 IN/CC NN_2", "JJ_1 IN/CC JJ_2 NN", "JJ_1 JJ_2 NN_1 IN/CC NN_2", "JJ_1 IN/CC JJ_2 NN_1 IN/CC NN_2" можуть бути ключовими. Після формування вищеназваних термінів та упорядкування їх у певному порядку (формується послідовність, де словосполучення з більшою кількістю слів розташовуються перед словосполученнями та словами, які є їх частиною) здійснюється видалення одиничних стоп-слів (окремих артиклів, прийменників, сполучників, деяких дієслів, прислівників та займенників).

Далі за допомогою глобальної частоти терміна – GTF [5], що визначається відношенням загальної кількості появи терміна у всіх документах корпусу до загальної кількості термінів у документах корпусу, здійснюється статистичне зважування слів та словосполучень, що входять у сформовану на попередньому етапі послідовність.

Для кожного слова у порядку його зустрічання у тексті формується так званий кортеж. Кожен елемент кортежу складається з трьох значень: перше – термін (слово або словосполучення); наступне – тег, який присвоюється

слову в залежності від його приналежності до певної частини мови; останній елемент такого набору – числове значення GTF. Важливо зазначити, що GTF обчислюється з урахуванням двох попередніх значень – слова або словосполучення та частини мови, до якої воно належить. Кількість таких однакових кортежів у всьому тексті, що нормована на загальну кількість сформованих термінів, і визначає значення третього елемента.

На наступному кроці пропонується визначити ненаправлені зв'язки між термінами у тексті. Для досягнення цієї мети застосовується алгоритм графа горизонтальної видимості для часових рядів (Horizontal Visibility Graph algorithm – HVG) [6]. Часовим рядом у нашому випадку є послідовність числових значень GTF, що сформована на попередньому етапі. Ідея алгоритму полягає у тому, що два вузли t_i та t_j, які відповідають елементам часового ряду x_i і x_j, знаходяться у горизонтальній видимості тоді й тільки тоді, коли $x_k < \min(x_i; x_j)$ для всіх t_k таких, що $t_i < t_k < t_j$. У нашому випадку послідовність t_i, $i=1,..,n$ – це послідовність слів у межах речення (n – кількість слів, що залишились у реченні після вищеописаної попередньої обробки). HVG дозволяє будувати мережеві структури на основі текстів, в яких окремим словам або словосполученням деяким чином поставлені у відповідність числові вагові значення.

Якщо між вузлами t_i до t_j часового ряду існує ненаправлений зв'язок, встановлений за вищеописаним алгоритмом, то:

- напрямок зв'язку пропонується встановлювати від вузла t_i до t_j, якщо у реченні слово (не словосполучення), якому відповідає вузол t_i зустрічається раніше ніж термін (слово або словосполучення), якому відповідає вузол t_j;
- напрямок зв'язку пропонується встановлювати від вузла t_j до t_i, якщо у реченні словосполучення (не слово), якому відповідає вузол t_j зустрічається раніше ніж термін, якому відповідає вузол t_i.

Беручи до уваги принцип формування послідовності із термінів, що описаний вище, та запропоновані правила встановлення зв'язків можна помітити, що слова та словосполучення будуть входити у відповідні словосполучення, мають більшу кількість слів. Тобто значна частина словосполучень з більшою кількістю слів є розширенням відповідних їм словосполучень та слів. Подібний принцип побудови направлених мереж зі слів, побудова мереж природніх ієрархій термінів, запропонований у роботі [7], де направлена мережа зі слів та словосполучень будується за принципом входження терміна у відповідне йому словосполучення.

Вагові значення зв'язків між вузлами направленої мережі визначаються за запропонованим у роботі [8] принципом, який полягає у тому, що вузли, які відповідають однаковим термінам побудованої на попередньому етапі направленої мережі, об'єднуються ("склеюються"), а кількість однаково-направлених зв'язків між відповідними вузлами і визначає вагове значення зв'язку між цими вузлами.

Результати досліджень

Запропонована методика обробки текстових корпусів та побудови направлених зважених мереж термінів була апробована на прикладі найвідомішого твору Антуана де Сент-Екзюпері алегоричної повісті-казки “Маленький принц” (англ. “The Little Prince”).

Відповідно до методики, що запропонована вище, було здійснено обробку обраного текстового документу й виокремлено ключові терміни та побудовано направлену зважену мережу зі слів та словосполучень, які відповідають окремим ключовим поняттям у досліджуваному творі (рис. 1). Для побудованої мережі було видалено всі зв’язки, які мають вагове значення рівне 1 та вузли, вихідна та вхідна степінь яких дорівнює нулю.

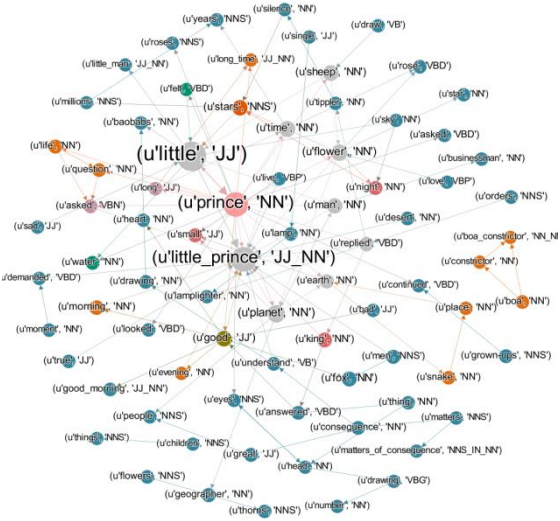

Рис. 10. Направлена зважена мережа термінів побудована для тексту “The Little Prince” (мітки вузлів містять термін та відповідний йому тег).

Висновки

У цій роботі було запропоновано новий метод виокремлення ключових термінів та новий метод встановлення напрямків зв’язків із застосуванням більш широкої обробки природної мови, що базується на розбитті на частини мови (Part-of-speech tagging). Також представлена цілісна методика, що дозволяє будувати направлені зважені мережі з ключових слів та словосполучень текстового корпусу.

Апробацію запропонованої методики було проведено на прикладі алегоричної повісті-казки “Маленький принц” (англ. “The Little Prince”) Антуана де Сент-Екзюпері. Проаналізувавши результати дослідження було виявлено найбільш вагомі зв’язки між відповідними вузлами у мережі

термінів, що відповідають окремим ключовим поняттям у досліджуваному творі. У межах запропонованої онтологічної моделі ключовими виявились терміни “little”, “prince” і “little_prince”, що відповідають назві твору, а найбільш вагомими, як і очікувалось, зв’язки між цим ж термінами “little → little_prince” та “little → prince”.

Літературні джерела

1. Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. O'Reilly Media (2009). ISBN 0-596-51649-5
2. Google Code Archive: Stop-words. https://code.google.com/archive/p/stop-words/downloads/. Accessed 24 Oct 2020
3. Text Fixer: Common English Words List. http://www.textfixer.com/ tutorials/commonenglishwords.php. Accessed 24 Oct 2020
4. Extract Custom Keywords using NLTK POS tagger in python. https://thinkinfi.com/extract-custom-keywords-using-nltk-pos-tagger-in-python/. Accessed 24 Oct 2020
5. Lande, D., Dmytrenko, O., Radziievska, O.: Determining the Directions of Links in Undirected Networks of Terms. In: CEUR Workshop Proceedings (ceur-ws.org). Vol-2577 urn:nbn:de:0074-2318-4. Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019), vol. 2577, 132-145. (2019). ISSN 1613-0073 [http://ceur-ws.org/Vol-2577/paper11.pdf]
6. Luque, B., Lacasa, L., Ballesteros, F., & Luque, J.: Horizontal visibility graphs: Exact results for random time series. Physical Review E, 80(4), (2009). DOI: doi.org/10.1103/PhysRevE.80.046103.
7. Lande, D. V., Snarskii, A. A., Yagunova, E. V., & Pronoza, E. V.: The use of horizontal visibility graphs to identify the words that define the informational structure of a text. In: 2014 12th Mexican International Conference on Artificial Intelligence, pp. 209-215 (2014).
8. Lande D.V., Dmytrenko O.O.: Creating the Directed Weighted Network of Terms Based on Analysis of Text Corpora. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC) (Kyiv, 5-9 Oct. 2020). DOI: doi.org/10.1109/SAIC51296.2020.9239182.