УДК 004.91

Є.Р. КОВИЛІН
Університет імені Альфреда Нобеля

АДЕКВАТНІСТЬ ПРОЦЕСУ ПОБУДОВИ СЕМАНТИЧНОЇ МОДЕЛІ ДОКУМЕНТУ НА ОСНОВІ НЕСТРУКТУРОВАНОЇ БАЗИ ЗНАНЬ

Розробка прикладних програмних систем автоматичної обробки текстів має на увазі вибір того чи іншого механізму опису та реалізації моделі природної мови, доступної для обробки ЕОМ. Оскільки, мова є досить не формалізованою системою з нестабільністю і неоднорідністю власних правил, то головною проблемою при реалізації таких моделей є складність опису семантичних характеристик тексту на рівні алгоритмічного уявлення. У дисертаційній роботі [1] був реалізований підхід до побудови програмної семантичної моделі документа, яка базується на структурі гібридної семантичної мережі. Необхідність і важливість цієї моделі виходить насамперед із аналізу існуючих аналогів і алгоритмічних підходів до побудови семантичних мереж документа – усі вони або базуються на словниках, або не пророблені для флективно багатих груп мови. Розроблений підхід базується на алгоритмі латентно-семантичного аналізу, що дозволяє знаходити семантичні відповідності на основі вагових характеристик тексту і роботі з проекціями координат для базових текстових одиниць на двомірній площині. Використання такого підходу до роботи із семантичними характеристиками тексту є інноваційним не тільки тому, що сфера застосування латентно-семантичного аналізу в першу чергу стосується задач класифікації документів, тоді як у нашій моделі його використання було змінено, і ми наближаємо не документ до терміну, а речення з документів до термінів документа, а і тому, що алгоритм поєднує у собі багато специфічних додаткових етапів, що є нетиповими для підходів побудови семантичної моделі документу. Мова йде про використання алгоритмів кластеризації, із відповідними методами для визначення необхідних для неї параметрів, алгоритмів синтаксичного, морфологічно і просторового аналізу даних [1]. Отриманий підхід дозволяє будувати семантичні моделі наукових текстів без будь-якої попередньої семантичної розмітки або складання семантичних словників, які містять у своєму складі кількісні показники семантичних характеристик тексту, що значно спрощує процес побудови систем автоматичної обробки текстів. Дослідження, проведене у даній статті стосується перевірки спроможності моделі виконувати своє функціональне призначення, шляхом побудови її критеріїв адекватності і перевірки їх виконання шляхом проведення відповідних експериментів.

Ключові слова: семантична мережа, автоматична обробка тексту, система запит-відповідь, генерація тексту.

Y.R. KOVULIN
Alfred Nobel University

ADEQUACY OF THE PROCESS OF BUILDING THE SEMANTIC MODEL OF THE DOCUMENT BASED ON AN UNSTRUCTURED KNOWLEDGE BASE

The development of applied software systems for automatic text processing implies the choice of one or the other mechanism for describing and implementing a natural language model available for computer processing. Since language is a rather unformalized system with instability and heterogeneity of its own rules, the main problem in the implementation of such models is the difficulty of describing the semantic characteristics of the text at the level of algorithmic representation. An approach to building a programmatic semantic model of a document, which is based on the structure of a hybrid semantic network, was implemented in the dissertation [1]. The necessity and importance of this model comes primarily from the analysis of existing analogues and algorithmic approaches to the construction of semantic networks of a document. All of them are either based on dictionaries or have not been developed for language groups with a rich inflectional morphology. The developed approach is based on the algorithm of latent semantic analysis, which allows finding semantic correspondences based on weight characteristics of the text and working with coordinate projections for basic text units on a two-dimensional plane. The use of such an approach to work with the semantic characteristics of the text is innovative not only because the algorithm combines many specific additional stages that are atypical for approaches to building a semantic model of a document, but also because the scope of application of latent semantic analysis primarily concerns the tasks of document classification, while in our model, its use has been changed, and we approach not the document to the term, but the sentence from documents to document terms. It is a question of the use of clustering algorithms with appropriate methods for determining the necessary parameters for it, algorithms of syntactic, morphological and spatial data analysis [1]. The obtained approach makes it possible to build semantic models of scientific texts without any previous semantic marking or compilation of semantic dictionaries, which contain quantitative indicators of the semantic characteristics of the text, which greatly simplifies the process of building

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4

automatic text processing systems. The research carried out in this article concerns the verification of the model's ability to fulfill its functional purpose by constructing its adequacy criteria and the verification of their implementation by conducting relevant experiments.

Keywords: semantic network, automatic text processing, request-response system, text generation.

Постановка проблеми

Головною складністю впровадження процесу автоматичної обробки текстів у прикладні програмні продукти є необхідність побудови семантичної моделі документів, доступної для обробки комп'ютером. Окремо слід відзначити ускладнення, пов'язане із властивостями флексії та типології порядку слів мови документа, через що моделі, створені, наприклад, для англійської мови, не є застосовними для документів слов'янською мовою. Через це, головним підходом до створення семантичних моделей документів слов'янськими мовами є ручна побудова семантичної структури тексту, яка може бути представлена у вигляді онтологічної розмітки або семантичних словників, і вимагає використання великого обсягу ручної праці для їх складання і пошуку фахівців у прикладній лінгвістиці. Це значно ускладнює використання автоматичної обробки текстів та обмежує застосування моделі для довільної колекції семантично неструктурованих текстів. Питання розробки схожих моделей уявлення текстів для їх подальшої комп'ютерної обробки подані у багатьох роботах, які стосуються галузей штучного інтелекту, математичного моделювання та обробки природної мови. Серед вітчизняних вчених слід відзначити наукові розробки Н.Н. Леонтьевої, М.В. Мозгового, В.А. Тузова, І.О. Мельчука, Ю.Д. Апресяна, спрямовані на математичне моделювання семантичних властивостей тексту на основі онтологій та семантичних словників. Зарубіжні англомовні моделі представлення текстів відштовхуються від жорсткого порядку слів у мові, як наприклад в роботах N. Chomsky. Проведений аналіз наявної у відкритому доступі науково-технічної літератури і документації показав, що існуючі моделі представлення слов'яномовних текстів спрямовані саме на опис структурованих знань, і не дозволяють повністю автоматизувати процес опису семантичних властивостей та адаптивного додавання неструктурованих знань до пошукової системи. Усі вони мають на увазі залучення значної кількості лінгвістичних знань і використання попередньої семантичної розмітки, створеної лінгвістом-експертом вручну. З іншого боку, зарубіжні моделі орієнтовані в першу чергу на обробку англомовних текстів і не можуть бути застосовані для представлення слов'яномовних текстів. Таким чином, відсутність цілком автоматичних моделей представлення слов'яномовних текстів дозволяє зробити висновок про те, що на сьогоднішній день задачі розробки семантичної моделі документа на основі неструктурованої бази знань і перевірки її адекватності є актуальними і практично значимими.

Мета дослідження

Розробити критерії адекватності математичної моделі представлення семантичних властивостей текстів на основі неструктурованої бази знань, та довести її спроможність працювати із семантичними характеристиками тексту шляхом виконання відповідних експериментів.

Аналіз останніх досліджень і публікацій

Перевірка адекватності розробленої семантичної моделі тексту і розробленої на її основі прикладної програмної системи викликає ряд складнощів. Справа в тому, що класичні семантичні мережі, які використовуються для онтологізації бази знань тестуються порівнянням із деяким «золотим стандартом», відносно якого і вираховуються критерії успішності побудови семантичної мережі [2]. Проте, у нашому випадку, такий підхід може використовуватися лише при безпосередньому застосуванні отриманої моделі, що відсуне питання її семантичних властивостей в цілому на другий план. Важливішим же питанням, що постає на цьому етапі, є перевірка залежності моделі саме від семантичних характеристик тексту, а не від його статистичного

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4

частотного портрету. Тому було прийнято рішення про проведення ряду експериментів, спрямованих, в першу чергу, на спробу ввести систему в оману, будуючи моделі семантично невірних документів і порівнювати їх з отриманими результатами для коректних текстів.

Виходячи з цього, найбільш показовою перевіркою є зіставлення із автоматично згенерованим текстом, а саме, обробка тексту, побудованого за допомогою автоматичного SEO-генератора тексту. Генератор тексту – це комп'ютерна програма, що створює тексти, коректні з точки зору більшості норм мови, але, як правило, позбавлені сенсу [3]. Іноді у читача, який працює із згенерованим такою програмою текстом, може скластися враження, що цей текст є осмисленим і семантично пов'язаним, особливо якщо текст має тематику, з якою читач слабо ознайомлений. Існують різні види генераторів тексту, що розрізняються своїми можливостями (наприклад, деякі з них можуть самостійно формувати нові слова). Генерація тексту шляхом складання з повністю випадкових слів дає сміттєвий результат, безглуздий для розуміння людиною, тому зазвичай застосовується генерація за вручну написаними фразами-шаблонами.

Викладення основного матеріалу дослідження

Загальний процес побудови моделі має лінійну структуру, яка поділена на три рівня обробки документу: ***синтаксичний рівень***, який об'єднує у собі деякі типові задачі обробки тексту, такі як виділення слів і речень із вхідного тексту, визначення стем, зважування елементів тексту, автоматичне визначення частини мови, тощо; ***семантичний рівень***, який реалізує у собі компонент латентно-семантичного аналізу і додаткові засоби інтелектуальної обробки даних, завдяки яким відбувається формування фінальної семантичної моделі вхідного тексту; ***рівень виведення***, який відповідає за відображення і збереження отриманої семантичної моделі у базу знань. Його детальний опис наведено у роботі [1]. Результатом виконання процесу над неструктурованим науковим тестом є семантична структура, представлена сукупністю множин термінів T і сукупністю множин речень S, зв'язок між якими визначається матрицею семантичних ваг W:

$$T = \left\{T_1\{t_1 \dots t_{w_{T_1}}\} \cdots T_n\{t_n \dots t_{w_{T_n}}\}\right\},$$
$$S = \{S_1\{s_1 \dots s_{l_1}\}\} \cdots S_1\{s_1 \dots s_{l_n}\}\}, \quad (1)$$
$$W = \begin{vmatrix} \{w_{T1}, w_{S1}\}, & \{w_{T2}, w_{S1}\} & \cdots & \{w_{Tn}, w_{S1}\} \\ \vdots & \vdots & & \vdots \\ \{w_{T1}, w_{Sm}\}, & \{w_{T2}, w_{Sm}\} & \cdots & \{w_{Tn}, w_{Sm}\} \end{vmatrix},$$

де $t_1 \dots t_{w_{T_,}}$ – множина координат стем, що формують кластер-стему, $s_1 \dots s_{S_l}$ – множина координат речень, що формують кластер-речення, $w_{T_,}$ – кількість точок у кластері-стемі, S_l – розмір кластера речення, w_S – вага зв'язку між кластером-стемою і кластером-реченням, n – кількість кластерів-стем, $m = n$ – кількість кластерів-речень.

Графічний результат роботи етапу виведення зображений на рис. 1. Результуюча семантична модель складеться із кластерів-стем (позначено на рисунку як WordCluster), до яких прив'язане значення кількості стем у кластері (зображене на рисунку у квадратних дужках), та кластерів-речень (позначено на рисунку як SentenceCluster), які пов'язані із кластерами-стемами семантичним відношенням, вага якого позначена у круглих дужках.

Для оцінки адекватності отриманої моделі система була перевірена на текстах, створених в результаті генерування на основі шаблонів синонімізації. Для цього був використаний сервіс [4], що дозволяє створювати тексти за обраною тематикою та розміром.

Особливістю згенерованого документу є задовільна для систем статичної обробки тексту частотна картина. Наприклад, розглянемо частотні портрети двох текстів однакового розміру, які наведені в таблиці 1. Ґрунтуючись на частотних портретах документів, за умови їх однакового розміру, можна зробити висновок, що обидва тексти мають тематичну спрямованість «космос» і перший текст присвячений темі галактичного радіовипромінювання, а другий – астероїдів,

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4

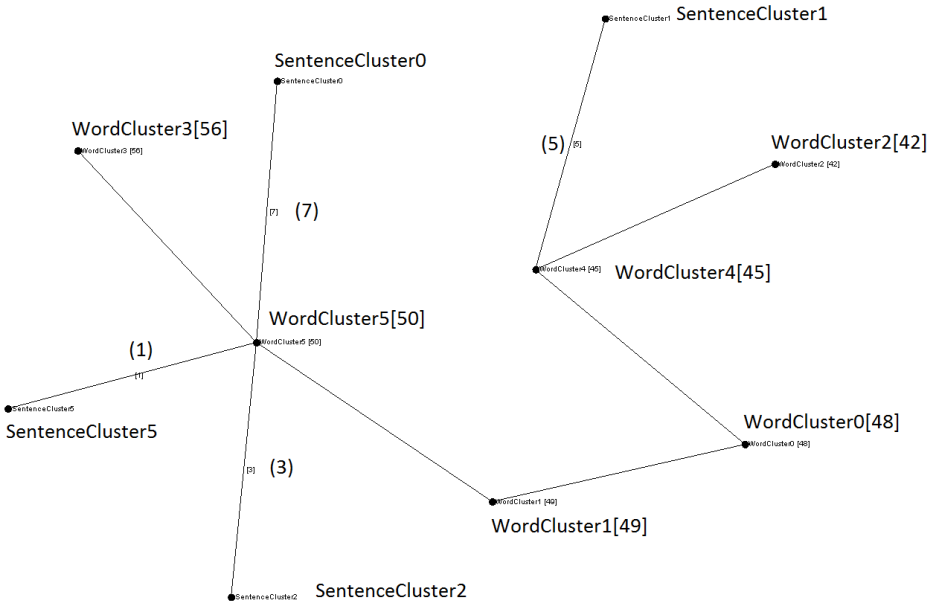

Рис. 1. Семантична модель для технічного тексту зі сильними семантичними зв'язками

однак це не зовсім так, оскільки перший текст є автоматично згенерованим текстом, а другий – семантично важливим знанням, що дійсно описує тему астероїдів. З цього витикає гіпотеза про перший критерій адекватності моделі, який можна сформулювати наступним чином: якщо модель орієнтується не тільки на синтаксичний і частотний, а й на семантичний рівень розуміння тексту, то у процесі роботи із згенерованим текстом, на відміну від семантично зв'язного тексту, повинні відбутися зміни кількісних характеристик семантичних властивостей створеної моделі, за умови однакової стилістичної спрямованості і близьких фізичних об'ємів обох текстів. Задля її перевірки було проведено ряд експериментів із згенерованим текстом і порівняння отриманих результатів із схожими семантично нормальними текстами.

Таблиця 1

Частотні портрети згенерованого тексту і знання

Документ1		Документ2	
Слово	Вага	Слово	Вага
Радіовипромінювання	88	астероїдів	92
галактичної	87	існують	23
джерела	83	планетні	22
об'єктів	53	поверхню	18
оптичне	44	земної	18
званих	42	зіткнення	17
зірки	42	залишалися	18

Роздивимось деякі приклади отриманих результатів. Результат обробки автоматично згенерованого документу наведено на рис. 2. Незважаючи на те, що не тільки обсяг автоматично згенерованого тексту співпадав з прикладом тексту із рис. 1, наведеним раніше, а і його технічна спрямованість, отримана семантична модель має значні яскраво виражені відмінності.

Перш за все, мова йде про кількість кластерів-стем, кластерів-речень, що мають зв'язки із кластерами-стемами, та вагу цих зв'язків – кількість «вільних» кластерів стем і їх загальна кількість значно перевищує аналогічні показники у еталонного тексту. Оскільки у нашій моделі,

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4

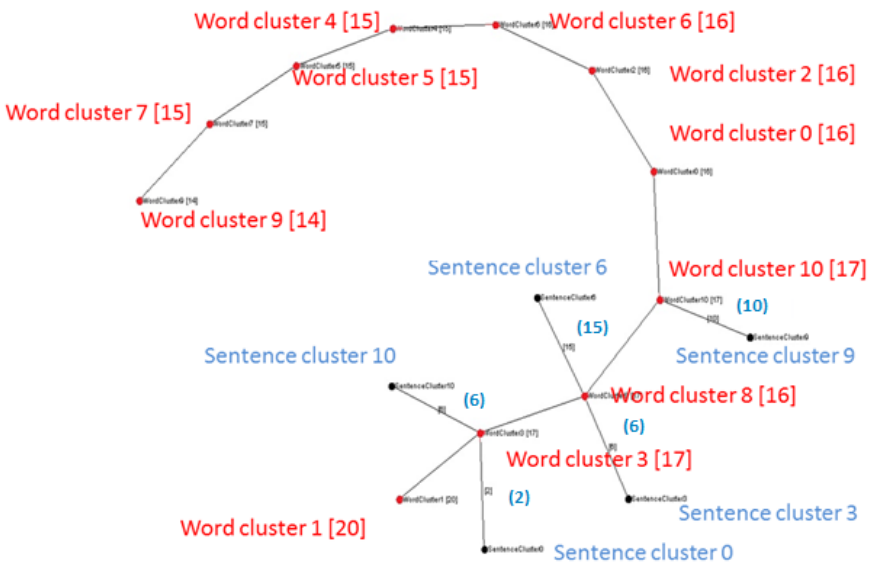

Рис. 2. Семантична модель для автоматично згенерованого технічного тексту зі слабкими семантичними зв'язками

саме ці данні і виражають кількісні характеристики семантичних властивостей тексту, то для оцінки роботи моделей на згенерованих текстах було виконано попарне порівняння числових значень семантичних моделей згенерованого тексту і знання щодо значення середнього коефіцієнта семантичної значущості моделі μ (2):

$$\mu(D) = \frac{\sum_{i=0}^{N} \frac{f_{\min}(D_i, D_{i+1})}{f_{\max}(D_i, D_{i+1})}}{N},$$

$$f_{\min}(D_1, D_2) = \min\left\{\frac{\sum w_{SD_1}}{N_{D_1}}, \frac{\sum w_{SD_2}}{N_{D_2}}\right\}, \quad (2)$$

$$f_{\max}(D_1, D_2) = \max\left\{\frac{\sum w_{SD_1}}{N_{D_1}}, \frac{\sum w_{SD_2}}{N_{D_2}}\right\},$$

де $\sum w_{SD}$ – вага всіх семантичних зв'язків в моделі знання D, N_D – кількість кластерів в моделі знання D, $f_{\min}(D_1, D_2), f_{\max}(D_1, D_2)$ – функції, які повертають мінімальне і максимальне значення відносин, N – кількість експериментів.

Дане значення характеризує розподіл семантичних міток документа (кластерів-стем) щодо семантичних зв'язків, і показує залежність цього значення для пар текстів в корпусі знань $D = \{d_1 \ldots d_N\}$. Коефіцієнт семантичної значущості моделі був розрахований для D_g – корпусу автоматично створених документів, D_I – корпусу знань і D_{gI} – корпусу, в якому зіставляється згенерований текст і знання між собою. Дослідження показали, що зміни коефіцієнта семантичної значущості для корпусу згенерованого тексту і корпусу знань при порівнянні текстів однакових розмірів практично не відбувається: $\mu_g(D_g) = 0{,}911$ та $\mu_I(D_I) = 0{,}932$ відповідно. Однак, якщо порівнювати між собою знання і згенерований текст, очевидне значне падіння значення коефіцієнта до $\mu_{gI}(D_{gI}) = 0{,}335$. Це відбувається через те, що на відміну від згенерованого тексту, семантичні мітки знання мають значно більшу вагу і відповідно об'єднують більшу кількість семантично пов'язаних термінів, підвищуючи таким чином семантичну значущість моделі. Через значно більшу кількість кластерів-стем в моделях згенерованого тексту, підтверджується головна ідея використання моделі в задачах автоматичної обробки знань в системах генерації відповідей – терміни не об'єднуються в семантичні мітки, оскільки не

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4

існує ніякого семантичного зв'язку знань в початковому тексті, що наочно показує адекватність розробленої моделі.

Окремою і важливою перевіркою адекватності алгоритму побудови семантичної моделі тексту є оцінка сенсової місткості семантичних міток документу. Сформульована гіпотеза має на увазі, що утворені семантичні мітки документа мають найбільшу кількість перетинів із семантичними контурами тоді, коли вони містять найбільшу кількість семантично значимих стем у своєму складі. Якщо це так, то отримана модель спроможна автоматично складати семантичні тезауруси понять, встановлюючи відповідності між наборами речень та семантичною міткою документу, реалізуючи при цьому схему розуміння абстрактного поняття завдяки набору значимих слів. Розглянемо прилад обробки технічного тексту, тематика якого стосується теми космосу, а саме – космічного реліктового випромінювання (рис. 3).

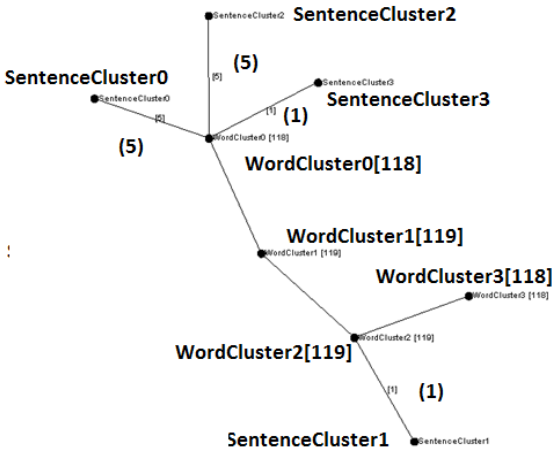

Рис. 3. Семантична модель тексту «Реліктове випромінювання»

На рис. 3 зображена семантична модель тексту, на основі якої видно, що кластер-стема із номером «0» має найбільшу сумарну вагу зв'язків із семантичними контурами, ніж інші, а кластер-стема із номером «3» не має перетинів із семантичними контурами стем. Тоді, відповідно до гіпотези що перевіряється, нульовий кластер має бути найінформативнішим і однозначно вказувати своїм змістом на тематику тексту, а третій кластер – мати у своєму складі неінформативні і семантично неоднозначні стеми. Дійсно, порівнявши кластери між собою, можна побачити значущу різницю – нульовий кластер містить стеми (всесвіт, реліктове, гравітаційних, супутникових, випромінювання, космічне та ін.), що не тільки пов'язані із головною тематикою тексту, а і семантично пов'язані між собою галузевою спрямованістю. З іншого боку, кластер, що не потрапив у семантичну мережу, не представляє ніякої сенсової цінності, на весь кластер присутня лише одна стема, що має відношення до теми космосу (телескоп), інші – являють собою семантично узагальнені слова, що не визначають ніякої предметної галузі.

Описана оцінка сенсової місткості семантичних міток документу не обмежується кількома тестами, і була проведена для окремої тестової бази знань, оскільки саме вона вказує на залежність структури семантичної мережі і семантичних властивостей документу. Для цього було перевірено 65 семантичних моделей, для яких було розраховано коефіцієнт середньої семантичної ємності ε (3) по кожній з тем – економіка, філософія, інформаційні технології і астрономія:

$$\varepsilon = \frac{\sum_{i=0}^{N} \frac{N_H}{N_L}}{N}, \qquad (3)$$

де N_H – кількість термінів, семантично пов'язаних з тематикою знання в кластері-стемі з найбільшою загальною вагою зв'язків з кластерами-реченнями (сильні кластери), N_H – кількість термінів,

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4

семантично пов'язаних з тематикою знання в кластері-стемі з нульовою загальною вагою зв'язків з кластерами-реченнями (слабкі кластери), N – загальна кількість проведених тестів для заданої теми.

Даний коефіцієнт показує, у скільки разів щільність семантично важливих термінів в сильних кластерах перевищує щільність семантично важливих термінів в слабких кластерах.

Отримані результати розрахунку сенсової місткості для кожного тематичного набору текстів показують, що у проведених тестах кількість семантично значних термінів у семантично сильних кластерах значно перевищує кількість термінів у семантично слабких кластерах, незалежно від кількості, розміру та тематичної спрямованості документів: для тематики «економіка» (15 документів) кількість термінів у сильних кластерах у середньому у 2,98 рази більша ніж у слабких, для тематики філософія (21 документ) у 3,37 рази, для тематики «астрономія» (24 документа) у 3,47 рази, для тематики «інформаційні технології» (5 документів) у 4,44 рази. При проведенні досліджень не враховувалась вага стем або їх будь-яке синтаксичне співвідношення із текстом – пов’язані терміни оцінювалися лише з точки зору належності до тематики документу, тому отримані результати вказують на цілком адекватне формування семантичних міток документу – кількість семантично значимих термінів у кластері-стемі прямо пропорційна до кількості та ваги пов’язаних із ним семантичних контурів речень у побудованих моделях документів, що доводить залежність структури семантичної мережі саме від семантичної складової тексту, і доводить її адекватність.

Висновки

Для оцінки отриманих за допомогою семантичної моделі результатів була сформульована та виконана поетапна перевірка семантичних властивостей розробленої моделі для визначення залежності саме від семантичних, а не від частотних характеристик документу. Виконаний ряд порівняльних тестів довів, що отримана модель є адекватною і може застосовуватися задля безсловникового універсального підходу до кількісного опису семантичних властивостей тексту. Встановлено, що через значно більшу кількість кластерів-стем в моделях хаотично згенерованого тексту підтверджується головна гіпотеза використання семантичної моделі в задачах автоматичної обробки знань в системах генерації відповідей – терміни не об'єднуються в семантичні мітки, оскільки не існує ніякого семантичного зв'язку знань в початковому тексті, що наочно показує адекватність розробленої моделі. Було встановлено, що кількість семантично значимих термінів у кластері-стемі прямо пропорційна до кількості та ваги пов’язаних із ним семантичних контурів речень у побудованих моделях документів, що доводить залежність структури семантичної мережі саме від семантичної складової тексту [5]. Завдяки цьому стає можливим подолати обмеження тлумачно-комбінаторного словника, що накладав строгі рамки використання лінгвістичних знань на створення прикладних програмних моделей, і застосувати парадигми м’якого розуміння Леонтьевої для автоматичної генерації текстів, автоматичної класифікацій документів і інших завдань автоматичної обробки текстів [5].

Список використаної літератури

1. Ковилін Є.Р. Модель генерації відповідей в пошукових системах на основі неструктурованої бази знань : дис ... канд. техн. наук : 01.05.02. Національна металургійна академія України. Дніпро, 2020. 233 с.
2. Усталов Д.А. Моделі, методи та алгоритми побудови семантичної мережі слів для задач обробки природної мови : дис … канд. фіз.-мат.наук : 05.13.17. ФГБУМ. Челябінськ, 2017. 129 с.
3. Болдас М.В., Соколова Є.Г. Генерація текстів на природній мові – теорії, методи, технології. *HTI. Сер. 2. Інформаційні процеси і системи*, 2006. С.1-15.
4. Генератор тексту. URL: https://online-generators.ru/text

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4

5. Volkovsky O.S., Kovylin Y. R. Computer system of intellectual semantic search with the text generation using. *Bulletin of the Kherson National University*. 2018. №3 (66). P. 238-245.

References

1. Kovylin, Y.R. (2020). Model' heneratsii vidpovidej v poshukovykh systemakh na osnovi nestrukturovanoi bazy znan'. Diss. kand. tekhn. nauk [Model of answers generating in the search engines based on an unstructured knowledge base] Dnipro.
2. Ustalov, D.A. (2017). Modeli, metody ta alhorytmy pobudovy semantychnoi merezhi sliv dlia zadach obrobky pryrodnoi movy. Diss. kand. fiz.-mat. nauk [Models, methods and algorithms for constructing a semantic network of words for natural language processing tasks] Chelyabinsk.
3. Boldas, M.V., & Sokolova, Ye.H. (2006). Heneratsiia tekstiv na pryrodnij movi – teorii, metody, tekhnolohii [Generation of texts in natural language - theories, methods, technologies]. *NTI "Informatsijni protsesy i systemy"* [NTI Information processes and systems], **2**, 1-15.
4. Henerator tekstu. URL: Available at: https://online-generators.ru/text (accessed 15 September 2022).
5. Volkovsky, O.S., & Kovylin, Y. R. (2018). Computer system of intellectual semantic search with the text generation using. *Bulletin of the Kherson National University*. **3** (66), 238-245.

Ковилін Єгор Романович – к.т.н., старший викладач кафедри інформаційних технологій Університету імені Альфреда Нобеля. e-mail: kovylin.yehor@gmail.com, ORCID: 0000-0002-2734-4095.

https://doi.org/10.32782/mathematical-modelling/2022-5-1-4