Автоматична Побудова Сематичної Мережі Тексту у Системах Запит-Відповідь

О. С. Волковський
кафедра автоматизованих систем обробки інформації
Дніпровський національний університет
імені Олеся Гончара
Дніпро, Україна
didivave@mail.ru

Є. Р. Ковилін
кафедра автоматизованих систем обробки інформації
Дніпровський національний університет
імені Олеся Гончара
Дніпро, Україна
kovilin.yegor@gmail.com

Automatic Construction of a Semantic Net of Text in the Request-Response Systems

O. Volkovskiy
Department of Automated information processing systems
Oles Honchar Dnipropetrovsk National University
Dnipro, Ukraine
didivave@mail.ru

Ye. Kovylin
Department of Automated information processing systems
Oles Honchar Dnipropetrovsk National University
Dnipro, Ukraine
kovilin.yegor@gmail.com

***Анотація*—Розглянуто структуру інтелектуальної системи запит-відповідь для отримання семантично зв'язних відповідей, генерованих на основі науково-технічних корпусів документів. Створено алгоритм автоматичної побудови семантичної мережі тексту на природній мові, розроблений в рамках завдання реалізації інтелектуальної системи запит-відповідь. Описано процес складання опорних міток для текстів, які характеризують семантичні зв'язки для тексту. На основі отриманого алгоритму була створена прикладна програмна система побудови графа смислових зв'язків науково-технічного тексту для мов слов'янської групи. Здійснено перевірку адекватності отриманої моделі на основі статистично нормалізованих текстів зі слабкими смисловими зв'язками.**

***Abstract*—The structure of the question-answer intellectual system for obtaining semantically connected answers which generated on the basis of scientific and technical corpus of documents is considered. An algorithm for the automatic construction of a semantic net of text in a natural language was developed for the task of implementing an intelligent request-response system. The process of compiling reference marks for texts characterizing semantic links for a text is described. On the basis of the obtained algorithm was created an applied software system for constructing a graph of the semantic links of the scientific and technical text for the languages of the Slavic group. The adequacy of the obtained model is checked on the basis of statistically normalized texts with weak semantic connections.**

Ключові слова—семантична мережа; система запит-відповідь; автоматична обробка тексту

Keywords—semantic network; question-answer system; automatic text processing

I. Вступ

Розробка прикладних програмних систем автоматичної обробки текстів (АОТ) має на увазі вибір того чи іншого механізму опису та реалізації моделі природної мови (ПМ), доступної ЕОМ. Оскільки, мова є досить не формалізованою системою з нестабільністю і неоднорідністю власних правил, то головною проблемою є складність опису семантичних характеристик тексту на рівні алгоритмічного уявлення. Оскільки ПМ це не просто набір слів, заснований на деяких граматичних складових (в задачах АОТ пріоритетним акцентом є отримання саме осмисленого тексту) це, в свою чергу, призводить багатьох розробників до необхідності врахування семантичних зв'язків не тільки між окремими словами, а й між реченнями і навіть між документами. Найчастіше, під семантикою і усвідомленням тексту машиною мається на увазі наступне: якщо ми ввели деякий текст в пам'ять ЕОМ і, скажімо, роздрукували його за допомогою принтера, про семантику годі й казати; але якщо над текстом виконується деяка обробка, в результаті якої користувач отримує зрозумілий і адекватний для нього новий текст (наприклад - переклад на іншу мову), то можна говорити про семантичне розуміння тексту комп'ютером. У цьому плані,

інтелектуальна генерація текстів є найбільш складним і цікавим завданням, оскільки при реалізації семантичної обробки шаблон синтезу тексту є більш нечітким, ніж, наприклад, при автоматичному перекладі. Тому в ході цієї роботи розглянута не тільки структура системи генерації тексту у вигляді відповіді на деякий запит, а й в першу чергу модель отримання семантичних зв'язків в тексті, доступна для прикладної програмної реалізації..

II. Система запит-відповідь

Загальна структура системи запит-відповідь що зараз розроблюється, зображена на рис.1. Найпершим вхідним параметром для системи є корпус документів, над яким проводиться операція розбиття на кластери, в результаті якої кожному отриманому кластеру відповідає набір семантичних міток, що характеризують семантичні властивості кожного документа, що міститься в кластері. Після цього виконується додавання розміченого таким чином кластера в загальний предметний корпус системи. При надходженні запиту користувача в систему, виконується його перетворення в набір опорних міток, на основі яких приймається рішення про формування визначаючого майбутню відповідь корпусу документів. По кожному з документів у визначальному корпусі виконується формування множини речень для включення їх в результуючу відповідь системи. Проаналізувавши отриману таким чином множину, система формує семантично пов'язану відповідь на запит, що надійшов.

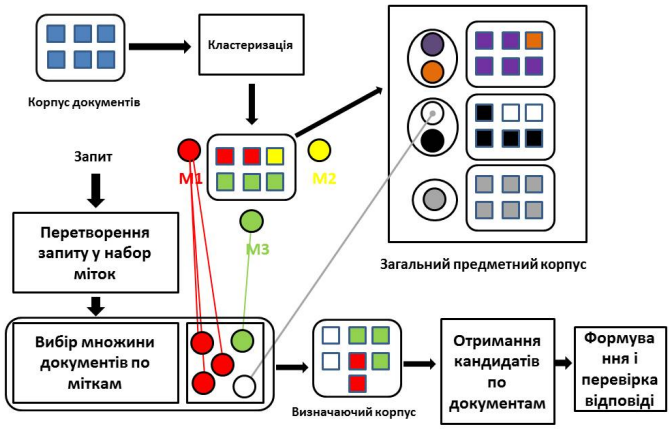

Рис. 1. Загальна структура роботи системи запит-відповідь

Одним з ключових моментів описаного алгоритму є отримання семантичних міток для документа, що має на увазі проведення семантичного аналізу всього документа і складання деякої опорної моделі для окремого тексту в корпусі. У нашому випадку, найкращим варіантом є побудова семантичної мережі документа, оскільки ця модель відображає смислові зв'язки між своїми елементами. Однак, програмна реалізація даної моделі викликає ряд складнощів, які є фундаментальними для більшості систем автоматичної обробки текстів. Йдеться і про відсутність чітких алгоритмів і математичних моделей, що описують саме поняття «семантична мережа документа» (і процес її отримання відповідно), і про самі властивості природної людського мови - встановлення зв'язків між елементами документа, написаного на мові флективно-багатої групи з вільним порядком слів (до якої відносяться всі слов'янські мови) значно складніше, ніж отримання аналогічних зв'язків, наприклад, в англомовному документі (де флексія набагато нижча і порядок слів жорстко фіксований граматичними правилами). Тому, складання семантичної мережі документа залишається відкритим та актуальним науковим питанням галузі автоматичної обробки текстів. В ході цієї роботи був складений і реалізований підхід до створення семантичних мереж документа.

III. Система побудови семантичної мережі документа

Алгоритм роботи системи зображений на рис.2. Першим етапом після завантаження документа і проведення синтаксичного аналізу (виділення речень і слів) стає визначення частини мови для кожного із знайдених слів. Для цього в систему була включена навчальна вибірка з 15 тис. слів і відповідних їм частин мови. На основі трьох типів закінчень (дві та три останні букви слова і закінчення отримане через стемінг за алгоритмом Портера) для кожного з елементів навчальної вибірки проводиться навчання наївного Баєсовського класифікатора, після чого на основі навченої моделі проводиться встановлення частини мови для кожного з виділених слів в тексті. Для підвищення точності в систему були включені кінцеві словники службових частин мови.

Рис. 2. Загальна структура роботи системи побудови семантичної мережі тексту

Після цього етапу проводиться визначення унікальних стем: для кожної пари слів проводиться відсікання приставки і закінчення за алгоритмом Портера - якщо довжина найбільшої загальної частини більше або дорівнює відстані Левенштейна для даної пари слів - то аналізоване слово замінюється на отриману стему. В результаті текст має такий вигляд (таб. 1) - для кожної із стем система визначає її частину мови та кількість входжень стеми в текст. На цьому етапі з тексту забираються всі слова помічені як службові частини мови.

ТАБЛИЦЯ I. РЕЗУЛЬТАТ СИНТАКСИЧНОЇ ОБРОБКИ

Текст до обробки	**Тест після обробки**
Сьогодні існують різні книги, відеокурси з програмування та інші способи швидко і відносно недорого навчитися писати програми і додатки.	|2|[годн{ADV}]|3|[існую{V}]|2|[різн{A}]|2|[книг{S}]|3|[відеокурс{S}]|22|программ{S}]|4|[інш{S}]7|[способ{S}]|2|[швидк{ADV}]|2|[орого{A}]|3|[навчити{V}]|8|[писа{V}]|22|[программ{S}] |2|[додатки{S}]

Для кожної із стем (загальною кількістю N) і кожного речення (загальною кількістю M) складається матриця $N*M$, значення якої визначаються кількістю входжень стем в речення. Над отриманої матрицею виконується операція сингулярного розкладання і проектування отриманих даних на площину. Оскільки сингулярне розкладання є стійким, ми можемо відкинути ті значення лівої і правої матриць, які відповідають низьким сингулярним значенням, залишивши тільки перші два, що представляють собою вектори координат двовимірної площині для стем і для речень. Отримана проекція зображена на рис.3.

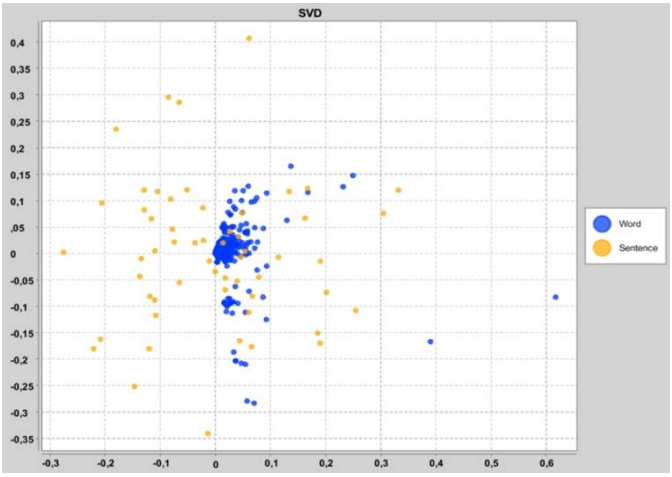

Рис. 3. Проекція сингулярного розкладання: Word – координати-стеми, Sentence – координати-речення

Наступним кроком стає кластеризація точок-координат для стем і для речень по алгоритму k-means. Кількість кластерів для стем і для речень cl визначається по формулі (1):

$$cl(W, W_U) = \frac{count(W)}{count(W_U)} \quad (1)$$

де W – слова, W_U – стеми. Центроїдами кластерів-стем є координати стем з найбільшою кількістю входжень в текст, що визначається за формулою (2):

$$Cst(W_U) = \max(W_0 ... W_{cl}) \quad (2)$$

де $W_0 \ldots W_{cl}$ – ваги стем. Центроїдами кластерів-речень є координати речень з найбільшим загальною вагою стем, яка визначається по формулі (3):

$$Cs(W_S) = \max\left(\sum_{i=0}^{SN} W_i\right) \quad (3)$$

де W_S – речення, W_i – вага стеми у речені, SN – кількість стем у реченні. Результат цих операцій можна побачити на рис. 4. (для стем) і на рис. 5 (для речень).

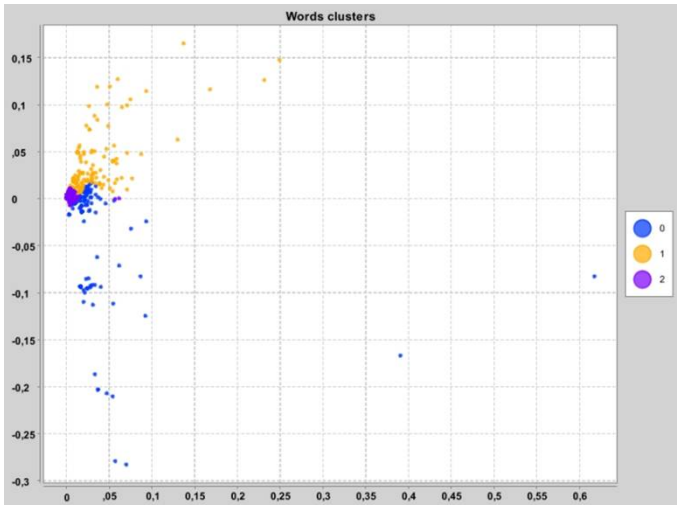

Рис. 4. Проекція кластерізації для стем. 0, 1, 2 – номера кластерів

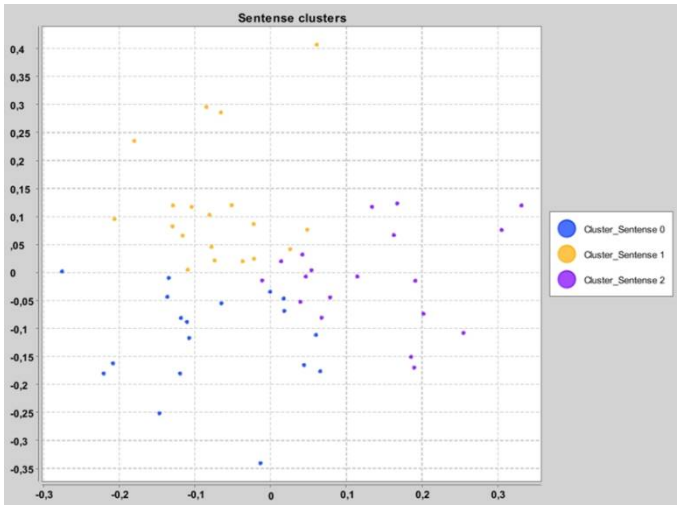

Рис. 5. Проекція кластерізації для реченнь. Cluster_Sentence 0, Cluster_Sentence 1, Cluster_Sentence 2 – номера кластерів

Заключним етапом стає формування семантичної мережі документа. На основі координат точок кожного кластера-стеми за алгоритмом Джарвіса будується контур опуклої фігури. Для кожного кластера-стеми визначається вага - кількість вхідних в нього стем, на основі чого будується семантичний граф зв'язків кластерів в порядку убування їх ваги. Для кожної фігури кластерів-стем, отриманої по алгоритму Джарвіса, перевіряється попадання точок що формують кожен кластер-речення.

Якщо такі точки знайдені - кластер речення приєднується в мережі до кластеру-стеми, де вага зв'язку - це кількість точок, що потрапили в контур кластера-стеми. Результат роботи системи зображений на рис. 6.

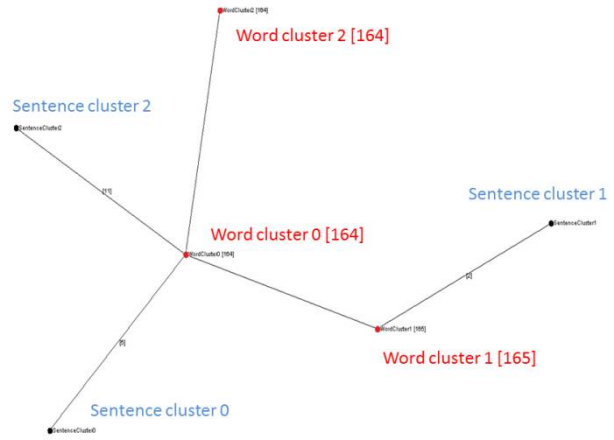

Рис. 6. Результат роботи системи – семантична мережа документа. WordCluster відповідають кластерам-стемам, SentenceCluste відповідаюсть кластерам-реченням.

IV. Аналіз отриманих результатів та висновки

Ґрунтуючись на отриманих результатах, стає можливим формування семантичної мітки документа в системах запит-відповідь. Так, мітка складається з тих кластерів-стем, які мають зв'язок з хоча б одним кластером-реченням, а для отримання кандидатів по мітці використовується перевірка входження мітки-запиту в мітку документа, і якщо таке входження знайдено - проводиться отримання кластера-речення з найбільшою вагою, пов'язаним зі знайденою міткою.

Для перевірки адекватності отриманої моделі система була перевірена на текстах, створених в результаті автоматичної генерації на основі патернів. Такі тексти є статично правильними, але мають слабкі смислові зв'язки між своїми частинами. Результат обробки такого тексту зображено на рис.7 (проекція кластерів стем та речень) та на рис.8 (результуюча семантична мережа).

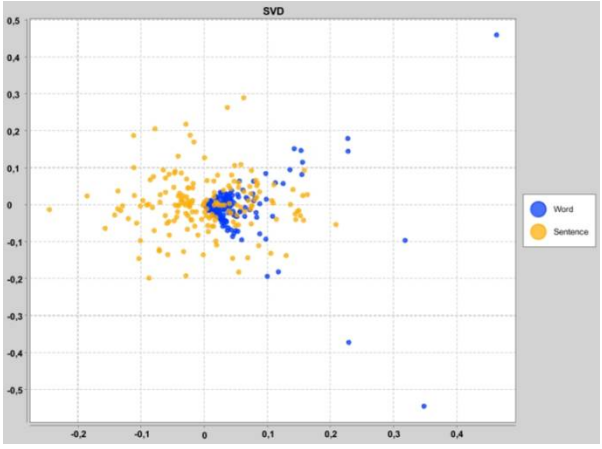

Рис. 7. Проекція сингулярного розкладання для тексту зі слабкими семантичними зв’язками: Word – координати-стеми, Sentence – координати-речення.

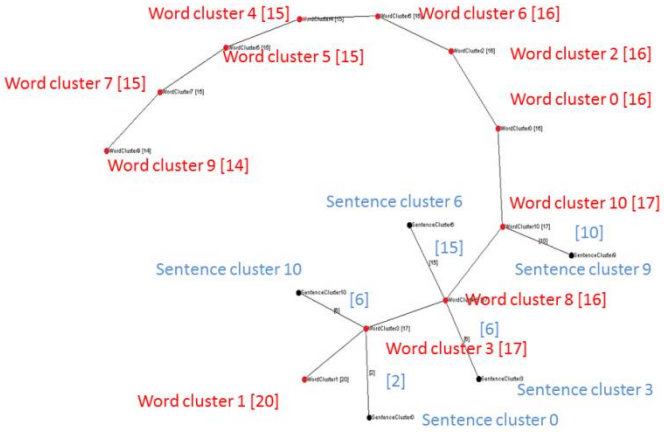

Рис. 8. Результат роботи системи для тексту зі слабкими семантичними зв’язками – семантична мережа документа. WordCluster відповідають кластерам-стемам, SentenceCluste відповідаюсть кластерам-реченням.

Незважаючи на те, що обсяг автоматично згенерованого тексту співпадав з прикладом, наведеним раніше, його семантична мережа має яскраво виражені відмінності. Це спостереження дозволяє зробити припущення не тільки про адекватність семантичної моделі, а так само і про можливості її застосування за межами завдання побудови систем запит-відповідь. Такі дані, як кількість кластерів-речень та кластерів-стем, кількість зв’язків та їх вага, ваги кластерів-слів, що містить в собі семантична мережа, можна використовувати для навчання моделей, які входять, наприклад, в системи автоматичного визначення плагіату або зв'язності тексту.

Література References

[1] Н.Н. Леонтьева. Автоматичне розуміння текстів. Системи, моделі, ресурси. [Текст]/Н.Н. Леонтьева//Москва – 2006

[2] М.В. Мозговой. Машиний семантичний аналіз російської мови и його застосування. [Текст]/Мозговой М. В.//СПбГУ, Санкт-Петербург –116с. – 2006г.

[3] N. Chomsky (2002) Syntactic Structures [Text]/Chomsky N//Berlin, New York: Mouton de Gruyter, 2002.

[4] Н.І. Гурін Семантична мережа электронного підручника для діалогу с віртуальним викладачем [Текст] / Н.І. Гурін, Я.А. Жук //Материали міжнародної научно-техничної інтернет конференції "Інформаційні технології в науці та виробництве"//Білоруський державний технологічний університет, Мінск, 2015 г..

[5] О.Н. Швецов. Система синтеза навчальних тестів заснована на формальних граматиках [Текст]/Швецов О.Н.//журнал «Программні продукти і системи», №2(102), 2013, с 181-185.

[6] I.A. Bolshakov. (2000) The Meaning ↔ Text Model: Thirty Years After. [Текст]/I.A. Bolshakov, A.F. Gelbukh//J. International Forum on Information and Documentation, N 1, 2000.

[7] D.S. Tarasov. (2015) - Natural language generation, paraphrasing and summarization of user reviews with recurrent neural networks [Text]/Tarasov D. S./ "Computer linguistics and Intellectual Technologies", No14(vol. 1), 2015, p.607-614//Materials of international conference "Dialog", 2015.

[8] Y.D. Apresyan. (1989) Linguistic support of the ETAP-2 system [Text]/Y. D. Apresyan, I. M. Bulavskiy, L.L. Iomdin//M.: Nauka, 1989.

[9] P. Bartlett, J. Shawe-Taylor. (1998) Generalization performance of support vector machines and other pattern classifiers//Advances in Kernel Methods. — MIT Press, Cambridge, USA, 1998.