УДК 004.912

Амонс Олександ Анатолієвич
кандидат технічних наук, доцент кафедри АУТС,
Національного технічного університету України «КПІ»,
Плехова Ірина Михайлівна
студентка кафедри АУТС,
Національного технічного університету України «КПІ»

Амонс Александр Анатольевич
кандидат технических наук, доцент кафедры АУТС,
Национального технического университета Украины «КПИ»,
Плехова Ирина Михайловна
студентка кафедры АУТС,
Национального технического университета Украины «КПИ»

Amons Olexandr
Candidate of Engineering Sciences, Associate Professor,
National Technical University of Ukraine «KPI»
Plekhova Iryna
student,
National Technical University of Ukraine «KPI»

АБСТРАКТНЕ РЕФЕРУВАННЯ НА ОСНОВІ ВИБОРУ ФРАЗ ТА ЇХ ЗЛИТТЯ

АБСТРАКТНОЕ РЕФЕРИРОВАНИЕ НА ОСНОВЕ ВЫБОРА ФРАЗ И ИХ СЛИЯНИЕ

ABSTRACTIVE DOCUMENT SUMMARIZATION VIA PHRASE SELECTION AND MERGING

Анотація. *Дана стаття присвячена розробці методу генерації реферату по текстовому документу. До уваги взяті такий підхід як абстрактне реферування.*

Ключові слова: *реферування, генерування реферату, абстрактний реферат.*

Аннотация. *Данная статья посвящена разработке метода генерации реферата по текстовому документу. Вниманию взяты такой подход как абстрактное реферирования.*

Ключевые слова: *реферирование, генерация реферата, абстрактный реферат.*

Summary. *This article is dedicated to the development of the method for creating summarization of the text document. Approaches were taking into account such as the generation of summarization.*

Key words: *summarization, abstract-based summarization, abstract summary.*

Постановка проблеми

Мистецтво реферування — витяг найважливіших і найхарактерніших фрагментів одного чи декількох джерел інформації — невід'ємна частина повсякденного життя. Результатом реферування документів є вторинні документи — реферати.

Ознайомлення з рефератами дає змогу оперативно одержати коротку інформацію про зміст первинних документів і завдяки цьому максимально правильно вирішити питання про необхідність використання їх. Тому робота в напрямі збільшення ефективності інформації є на сьогодні дуже важливою й актуальною. Отже, проведення досліджень у напрямі автоматизованого реферування тексту є перспективним і необхідним для сучасного суспільства.

Виклад основного матеріалу

Протягом останніх років з'явилося багато публікацій, в яких розглядаються проблеми автоматичного реферування. Візьмемо до уваги традиційну задачу

побудови реферату. В існуючих методах реферування можна виділити три напрямки: екстракція інформативних частин, стисненні вихідного документу та генерування реферату [3].

Більшість систем реферування використовують методи екстракції речень. Цей напрямок є найбільш вивченим. Ранні дослідження в основному відносяться до «жадібної» стратегії у виборі речень [4]. Спочатку кожному реченню в документі присвоюється оцінка ваги. Потім вибираються речення, які мають найбільшу вагу серед інших. Надмірність контролюється під час відбору в залежності від схожості з вже вибраними реченнями.

Компресійні підходи були засновані, щоб вирішити зазначені вище обмеження. Як природне розширення методу екстракції речень, ранні роботи пропонують використовувати двоетапний підхід [7] [8] [9]. На перший етапі проходить вибір речень, а на другому — видаляються несуттєві або надлишкові блоки в реченнях.

З іншого боку, підходи, що засновані на абстракції можуть генерувати нові речення, використовуючи факти з різних частин вихідного документу. Сумарний перегляд був також досліджений для підвищення якості автоматичного реферування шляхом заміни іменних фраз або посилань на власні назви, імена в кінцевому рефераті [9].

Більшість існуючих методів обробки та аналізу документів зосереджені на витягненні фактів з тексту. В той же час вид, в якому представлені дані, є не менш важливим, адже реферати створюються для людей. Також зв'язність тексту та границі переходу в рефераті є актуальною проблемою і до сьогодні. Тому реферат має бути зручним для швидкого сприйняття людиною. Метою даної статті є розробка алгоритму аналізу та обробки заданого документа для побудови реферату.

Пропонований підхід має декілька етапів. Спочатку необхідно зробити синтаксичний та граматичний аналізи тексту та побудувати діаграми речень. Для цього може бути використаний Stanford parser [10]. Результат аналізу та обробки тексту стороннім аналізатором зображений на рис. 1.

Виділяємо іменні та дієслівні фрази (ІД та ФД) і відповідно обчислюємо їх вагу та складаємо матриці сумісності. Система спочатку розділяє речення в документі на набір іменних фраз (ІФ-и), отриманих від предметних частин дерева речення, та набір дієслівних фраз (ДФ-и), що представляють потенційні ключові концепції і ключові факти, відповідно.

Після цього ми вибираємо ІФ-и та ДФ-и з дерева наступним чином: ІФ-и та ДФ-и, що є прямими нащадками вузла речення (представлена вузлом S). Для прикладу розглянемо дерево, зображене на рисунку 1, відповідне речення розбивається на фрази «An armed man», «walked into an Amish school, sent the boys outside and tied up and shot the girls, killing three of them», «walked into an Amish school», «sent the boys outside», and «tied up and shot the girls, killing three of them». Через рекурсивну операція, що вибирає фрази можем мати перекриття інформації.

Вага розраховується для кожної фрази і вказує на її важливість. В нашій системі використовуємо метод на основі концепта [11]. Ключовою характеристикою є те, що базовою одиницею являється фраза, а не речення.

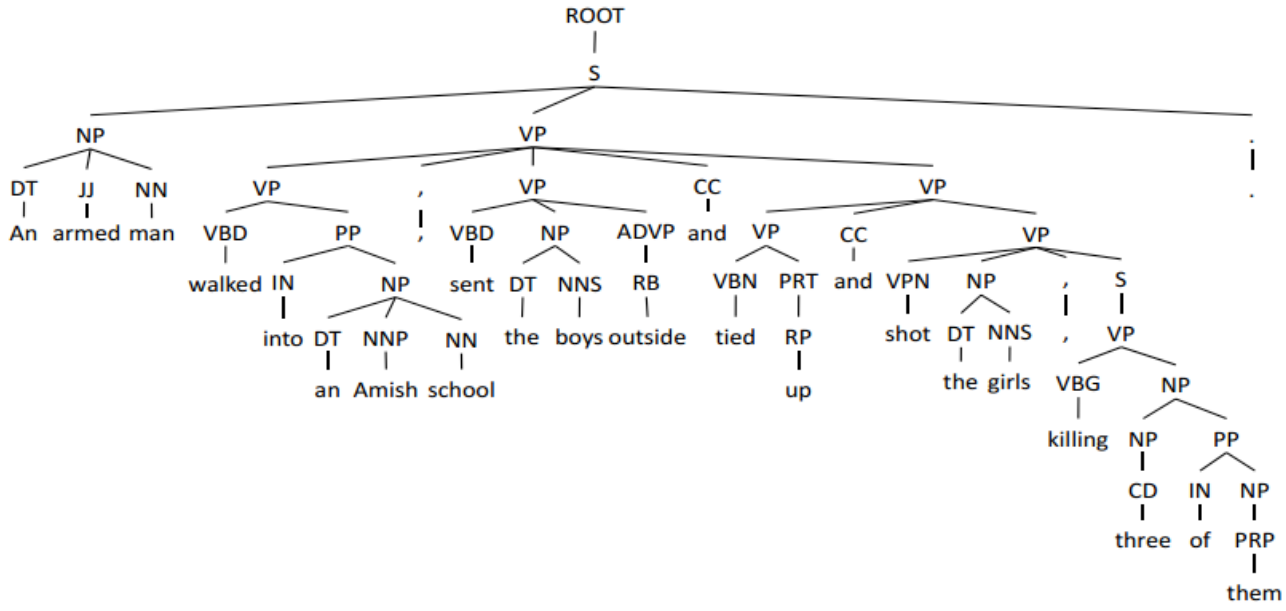

Рис. 1. Граф розбору речення з документу новин

Для знаходження тотожних ІФ (різні назви одного й того ж об'єкту) використовуємо Stanford coreference resolution package [6]. Для того щоб знайти тотожні ДФ-и, Jaccard Index використовується як міра схожості. Зокрема, кожен ДФ представляється у вигляді набору його понять і значення індексу розраховується для кожної пари ДФ-и. Якщо значення більше, ніж порогове значення, два ДФ-и визначаються в якості альтернативи один для одного.

Потім ми визначимо індикаторну матрицю $\Gamma_{|N||V|}$, в якій $\Gamma[i, j] = 1$, якщо ІФ *Ni* та ДФ *Vj* приходять з того ж вузла ***S*** в дереві вибраного речення, в іншому випадку, $\Gamma[i, j] = 0$. Нехай $\tilde{N}_i$ та $\tilde{V}_i$ представляють альтернативні фрази для *Ni* та *Vi*, як описано вище. Матриця сумісності $\tilde{\Gamma}_{|N||V|}$ визначається наступним чином:

$$\tilde{\Gamma}[p, q] = \begin{cases} 1 \text{ if } N_p \in \tilde{\mathbf{N}}_i \wedge \Gamma[i,q]=1 \\ 1 \text{ if } V_q \in \tilde{\mathbf{V}}_j \wedge \Gamma[p,j]=1 \\ 1 \text{ if } \Gamma[p, q] = 1 \\ 0 \text{ otherwise} \end{cases} \quad (1)$$

де $\tilde{\Gamma}[p, q]= 1$ означає, що *Np* та *Vq* сумісні / дозволені для побудови нової фрази. $\tilde{\Gamma}$ — матриця остаточної сумісності, яку ми використовуємо в оптимізації. У першому випадку, якщо *Np* і *N*i є тотожні, *Np* може замінити *Ni* і служити в якості іменника для його дієслівної фрази. Другий випадок має на увазі, що *Vq* дуже схожий до *Vj, Vq* може бути приєднаний до *Np*.

Загальна цільова функція оптимізації нашого формулювання для вибору ІФ та ДФ визначається наступним чином:

$$\max\left\{\sum_i \alpha_i S_i^N - \sum_{i<j}\alpha_{ij}\left(S_i^N + S_j^N\right)R_{ij}^N + \sum_i \beta_i S_i^V - \sum_{i<j}\beta_{ij}\left(S_i^V + S_j^V\right)R_{ij}^V\right\} \quad (2)$$

де α_i та β_i є індикаторами вибору для NP N_i та VP V_i відповідно. S_i^N та S_i^V є характеристичними оцінками для N_i та V_i. α_{ij} та β_{ij} є показниками суміжності пар (N_i, N_j) та (V_i, V_j). R_{ij}^N та R_{ij}^V є показники подібності пар (N_i, N_j) та (V_i, V_j). Якщо N_i та N_j суміжні, то $R_{ij}^V = 1$. В іншому випадку, схожість обчислюється за описаним вище способом на основі Jaccard Index методі. Вказані обмеження просумовані в таблиці 1.

Позначення	Опис
N_i, V_i	Іменна фраза *i* та дієслівна фраза *j*
α_i, β_i	Індикатори вибору N_i та V_i
α_{ij}, β_{ij}	Індикатори суміжності пар та
S_i^N, S_i^V	Характеристична оцінка та
R_{ij}^N, R_{ij}^V	Подібність пар та пар
$\Gamma_{\|N\|\|V\|}$	та з одного і того ж речення
$\tilde{\mathrm{N}}_i$, $\tilde{\mathrm{V}}_j$	Альтернатива фразам N_i та V_i
$\tilde{\Gamma}_{\|N\|\|V\|}$	$\tilde{\Gamma}[i,j]$ означає, що та сумісні для створення нового речення

Зокрема, ми максимізували характеристичну оцінку вибраних ІФ і ДФ, як зазначено на першій і третій складовій рівняння 2, і штрафуємо вибір подібних пар NP і подібних пар VP як зазначено в другому та четвертому членах рівняння. У той же час, вибір фрази регулюється набором обмежень таким чином, щоб обрані фрази могли бути використані для генерації правильних речень.

Однією з характерних рис нашої цільової функції є те, що ІФ і ДФ трактуються по-різному, тобто є різні виборчі/штрафні терміни для ІФ та ДФ. Така конструкція дозволяє уникнути помилкового штрафу між ІФ і ДФ.

Так, наприклад, в результаті алгоритму було створено дві пропозиції: перше речення є «*the gunman shot...*» з NP «*the gunman*», а інша пропозиція має VP «*confirmed the gunman died*». Очевидно, що ми не повинні вважати це надмірністю між ними, тому що згадувати того, хто стріляв необхідно в обох реченнях.

Результати порівняння з іншими системами приведені у таблиці 1. До уваги візьмемо System 22 [5].

Таблиця 1

Порівняльна характеристика оцінок систем

System	Q1	Q2	Q3	Q4	Q5	AVG
Наша	4.12	3.90	3.90	3.30	2.83	3.61
System 22	4.13	3.50	3.97	2.97	2.87	3.49

По параметрам зазначених у [2] наша система має перевагу по трьом позиціям, що є досить гарними результатами. Результати наведені в таблиці 4. У середньому, дві системи близькі одна до одної по результатам. System 22 використовує метод вилучення на основі, який вибирає оригінальні пропозиції, тому бал в Q1 граматичності є майже однаковим. Для Q4 фокуса, наш показник вище, ніж System 22, що вказує на фокусування на основних моментах в документі за рахунок вибору їх та вставлення їх в існуючі речення. Рахунок Q2 показує, що реферат має менше повторень порівняно з рефератом згенерованим System 22. Зокрема, середній бал нашої системи і System 22 є 3,61 і 3,33 відповідно.

Також важливо зазначити, що обробка тексту системами майже однакова, але наша система генерує реферат на 5% від загального часу швидше за System 22, за рахунок не використання сторонніх лінгвістичних ресурсів таких як Wikipedia (в System 22).

Висновки і пропозиції

В даній роботі представлена система автоматичної обробки текстового документу для створення реферату по ньому. Запропонований алгоритм, що використовується в системі, комбінує в собі відомі вже методи обробки тексту, але з викладеними вище

модифікаціями. За основу алгоритму було взято концепції абстрактного реферування. Матриця сумісності дозволяє нам підібрати правильні фрази для створення нового речення. Такий підхід дозволяє вибрати найбільш вагомі факти та твердження з тексту і представити їх у логічно правильній формі, щодо недоліків, то додання семантичної мережі допоможе вирішити проблеми граматики новоутворених речень.

Список літератури

1. Celikyilmaz A., Hakkani T. Discovery of topically coherent sentences for extractive summarization. // HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. — Volume 1. — p. 491–499.

2. Filatova, E., Hatzivassiloglou, V. Event-based Extractive summarization. // In: Proceedings of ACL 2004 Workshop on Summarization — 2004 — p. 104–111.

3. Електронний ресурс — http://ua-referat.com — Останні етапи складання тексту реферату, його оформлення та редагування.

4. Lin, H., Bilmes J. Multi-document summarization via budgeted maximization.

5. Pierre-Etienne Genest, Lapalme G. Framework for abstractive summarization using text-to-text generation. // In MTTG — 2011. — p. 64–73.

6. Pierre-Etienne Genest and Guy Lapalme. Fully abstractive approach to guided summarization. // In ACL — 2012 — p. 354–358.

7. Nenkova A. Entity-driven rewrite for multidocument summarization. // Third International Joint Conference on Natural Language Processing, IJCNLP, —2008 — p. 118–125.

8. Ganesan K., Zhai C., Han J. A graph-based approach to abstractive summarization of highly redundant opinions. // COLING-2010 — p. 340–348.

9. Lidong Bing, Piji Li, Yi Liao, Wai Lam, Weiwei Guo, and Rebecca Passonneau. Abstractive Multi-Document Summarization via Phrase Selection and Merging. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL'15). Beijing, China. July 26–31, 2015.

10. Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. // ACL — 2003 — p. 423–430.

11. Huiying Li, Yue Hu, Zeyuan Li, Xiaojun Wan, and Jianguo Xiao. Pkutm participation in tac 2011. // Proceedings of TAC — 2011.