МОДИФІКОВАНИЙ МЕТОД АВТОМАТИЧНОГО РЕФЕРУВАННЯ ТЕКСТІВ З ВИКОРИСТАННЯМ ТЕМАТИЧНО ЗВ'ЯЗАНОГО РАНЖУВАННЯ РЕЧЕНЬ

УДК 004.912

ЗАБОЛОТНЯ Тетяна Миколаївна
к.т.н., доцент кафедри ПЗКС ФПМ НТУУ «КПІ»,
e-mail: tatiana104@yandex.ua.

ФЕДЧЕНКО Наталія Володимирівна,
магістрант ФПМ НТУУ «КПІ», контактний
Наукові інтереси: автоматична обробка текстів.
e-mail: nataliia.fedchenko@gmail.com.

ВСТУП

В сучасному світі нас оточує величезна кількість текстової інформації, обсяг якої збільшується з кожним днем. В цій ситуації особливої значущості набуває задача автоматизації обробки великих обсягів текстових даних, зокрема отримання стиснутого подання документів – рефератів або анотацій.

Аналіз існуючих підходів до автоматичного реферування природномовних текстів [1-3], в тому числі зведеного, показав, що на сьогоднішній день практично реалізовано в основному методи квазіреферування, засновані на підході вилучення з вихідних документів речень або абзаців. Найпростіші такі методи дозволяють виділяти найбільш значущі текстові одиниці, найбільш складні – дозволяють формувати з них текст короткого реферату або анотації. Істотним недоліком даного підходу є відсутність зв'язності одержуваного тексту.

Методи складання короткого викладу тексту на основі використання результатів попереднього аналізу останнього, стиснення його семантичної структури та наступного синтезу реферату [1], незважаючи на свою перспективність, поки здебільшого залишаються в рамках науково-дослідних робіт. Більшість цих методів дозволяє отримати зв'язний реферат (анотацію), але, разом з тим, вони є орієнтованими на особливості конкретної природної мови та існуючі лінгвістичні ресурси для цієї мови. Це викликає додаткові труднощі при використанні даних методів для роботи з різними мовами при створенні багатомовних систем зведеного реферування. Крім того, можливості існуючих лінгвістичних ресурсів досить обмежені. Для їх створення та підтримки потрібні значні витрати, пов'язані з роботою лінгвістів, експертів з різних предметних галузей, спеціалістів з систем штучного інтелекту.

Таким чином, актуальною задачею є розробка методу автоматичного реферування для отримання зведених рефератів текстів (у тому числі, і складених різними

мовами) у вигляді зв'язного тексту без необхідності залучення великої кількості допоміжних лінгвістичних ресурсів.

ПОСТАНОВКА ЗАДАЧІ

Для зведеного автоматичного реферування часто застосовуються методи ранжування речень вхідного тексту. Такі методи досить ефективно виконують задачу виділення важливих частин текстів з урахуванням фактору «інформаційної новизни» (тобто речення, які потрапляють у реферат, не повторюють зміст одне одного), але жодним чином не визначають процедуру формування зв'язного тексту реферату.

Таким чином, **метою** даного дослідження є покращення зв'язності текстів зведених рефератів шляхом розробки модифікованого методу автоматичного реферування з використанням тематично зв'язаного ранжування речень.

У відповідності до поставленої мети **задачами** дослідження є:

- аналіз існуючих методів автоматичного реферування;
- обгрунтування та визначення нового способу оцінювання важливості речень для їх відбору з колекції документів;
- розроблення модифікованого методу автоматичного реферування з використанням ранжування речень;
- створення алгоритму формування зв'язного тексту підсумкового реферату в межах модифікованого методу автоматичного реферування.

АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ АВТОМАТИЧНОГО РЕФЕРУВАННЯ

Методи автоматичного реферування поділяються на дві категорії: *методи, що не передбачають опори на знання,* та *методи з опорою на знання* [4].

1. *Методи, що не передбачають опори на знання.* Більшість відомих систем автореферування сьогодні використовує варіації статистичних методів аналізу, часто ігноруючи при цьому лінгвістичну взаємопов'язаність і семантику природномовних текстів. У таких системах автоматичне реферування є по суті *квазіреферуванням*. Розвинений синтаксичний розбір і застосування баз знань або хоча б тезаурусів зустрічаються рідко.

Квазіреферування зводиться до екстрагування (витягання) з документів мінімальних релевантних фрагментів. При цьому, на відміну від побудови короткого викладу тексту, воно ґрунтується на аналізі поверхово-синтетичних відносин лексичних одиниць у тексті і не вимагає звернення до семантичного етапу аналізу. Методами квазіреферування передбачається виділення характерних фрагментів тексту за допомогою зіставлення фразових шаблонів, в результаті чого виділяються блоки з найбільшою лексичною і статистичною релевантністю. Основу аналітичного етапу квазіреферування становить процедура обчислення вагових коефіцієнтів для кожного блоку тексту відповідно до таких характеристик, як розташування цього блоку в оригіналі, частота появи в тексті, частота використання в ключових реченнях тощо.

2. *Методи з опорою на знання*. На відміну від лінійної моделі, методам формування стислого викладу змісту документа потрібні потужні обчислювальні ресурси для підключення інструментів автоматичної обробки тексту, в тому числі граматик і словників для синтаксичного розбору та генерації природномовних конструкцій. Крім того, для реалізації цих методів потрібні бази знань, в яких відображені поняття предметної галузі та взаємозв'язки між ними, що необхідно для

прийняття рішень під час аналізу та визначення найбільш важливої інформації.

Методи формування короткого викладу змісту документів передбачають реалізацію двох основних способів оцінювання важливості речень вхідного тексту.

2.1. Перший спирається на традиційний синтаксичний розбір речень. При цьому застосовується також семантична інформація для анотування дерев розбору. Процедури порівняння маніпулюють безпосередньо деревами з метою видалення і перегрупування їх частин, наприклад, шляхом скорочення гілок на підставі деяких структурних критеріїв, таких як дужки або вбудовані умовні чи підлеглі речення. Після такої процедури дерево розбору істотно спрощується, стаючи, по суті, структурною витримкою вихідного тексту.

2.2. Другий спосіб іде корінням в системи штучного інтелекту і спирається на розуміння природної мови. Синтаксичний розбір також є складовою частиною такого аналізу, але дерева розбору в цьому випадку не породжуються. Навпаки, формуються концептуальні репрезентативні структури всієї вихідної інформації, які акумулюються в текстовій базі знань. В якості структур можуть бути використані формули логіки предикатів або такі представлення як семантична мережа чи набір фреймів [1].

Для автоматичного реферування часто застосовують ранжування речень набору документів і відбір найбільш значущих з них для включення до реферату. В якості речення-запиту виступає тема документа, сформульована як речення-заголовок кластеру. У цьому випадку тема повинна бути сформульована максимально чітко і детально, тому що методи засновані на аналізі лексики і не можуть враховувати можливу семантичну близькість лексично «віддалених» речень. Для автоматичного реферування виділяється набір речень, найбільш близьких заданій темі кластера; також обов'язковим є застосування алгоритму відсікання «схожих» речень, що особливо актуально для зведеного реферування.

Автоматичне реферування набору документів з використанням алгоритму ранжування складається з двох етапів:

1. Обчислення рангу кожного речення – вирішує задачу ранжування всіх речень у відповідності до їх «близькості» заданій темі кластера.

2. Застосування алгоритму відсікання речень, найбільш схожих на ті, що вже потрапили в оглядовий реферат – вирішує задачу виключення з оглядового реферату однакових або близьких речень.

В результаті кілька речень з великим рангом вибираються для включення до результуючого реферату. Порядок речень в оглядовому рефераті в загальному випадку жодним чином не визначається.

Методи ранжування зарекомендували себе з позитивного боку як досить ефективні для вирішення задач автоматичного зведеного реферування [5]. Проведений аналіз методів виявив такі їх ключові переваги:

a) методи є обчислювально простими;

b) вони дозволяють за допомогою формальної математичної моделі описувати складні структури з множиною внутрішніх зв'язків, якими є кластери документів, складені природною мовою;

c) вони є ефективним інструментом для ранжування речень на основі формального математичного апарату без специфікації конкретних засобів з аналізу природномовних конструкцій. Як базовий набір можна використовувати

морфологічний словник і словник стоп-слів;

d) оцінка методів показала їх високу ефективність і можливість застосування для зведеного реферування кластерів документів.

До основного критичного недоліку методів варто віднести те, що вони ніяк не визначають процедуру формування зв'язного тексту реферату. Можливості методів обмежуються ранжуванням речень з урахуванням фактору «інформаційної новизни».

З огляду на всі наведені вище переваги та недоліки, авторами запропоновано провести модифікацію методу автоматичного реферування, описану нижче.

МОДИФІКОВАНИЙ МЕТОД АВТОМАТИЧНОГО РЕФЕРУВАННЯ

Запропонований нижче метод автоматичного реферування базується на використанні тематично зв'язаного ранжування речень. Суть методу полягає в ранжуванні всіх речень з колекції документів і виборі найбільш інформативних з них для включення в кінцевий реферат, а також в генерації зв'язного тексту реферату з урахуванням фактору «інформаційної новизни».

Метод автоматичного реферування з використанням ранжування речень складається з таких етапів:

1. Попередній аналіз та обробка текстів з вхідної колекції документів.
2. Розбиття колекції документів на речення та слова.
3. Обчислення рангу кожного речення.
4. Вибір теми реферату.
5. Відбір найважливіших речень з урахування ступеня подібності між ними та генерація зв'язного тексту реферату.

Нижче описаний кожен етап даного методу детальніше.

Попередній аналіз та обробка текстів. Для спрощення процесу реферування колекції документів перед початком оцінювання речень запропоновано видалити частини текстів, які не є інформативними і тому не повинні потрапити до реферату. В даній роботі визначено такі типи речень та словосполучень, що мають бути видалені перед початком реферування:

1) питальні речення;
2) окличні речення;
3) вставні слова і словосполучення;
4) слова-зв'язки;
5) уточнюючі слова.

Розбиття колекції документів на речення та слова. Для *розбиття тексту на речення* авторами запропоновано такий алгоритм. Аналізуємо символи вхідного тексту зліва направо:

1) шукаємо «!», «?» або «.» (якщо ці символи стоять в дужках або лапках, то їх пропускаємо);

2) якщо знайдено «!» або «?», то це означає кінець речення, тому ідемо на крок 1.

3) якщо знайдено «.», то, якщо далі йдуть ще 2 крапки (а разом – три крапки), то дивимось наступне слово: якщо воно написане з маленької літери, то це не кінець речення, якщо написане з великої літери – кінець речення. Йдемо на крок 1;

4) якщо знайдена одна крапка «.», то перевіряємо наступне слово, як в п. 3, – якщо воно з маленької літери – це скорочення.

Далі необхідно виконати *розбиття речень на слова*. Для цього запропоновано наступний алгоритм:

1) аналізуємо кожне речення як ланцюжок символів, зліва направо;

2) використовуючи роздільники, виділяємо перше слово. Шукаємо його канонічну форму та частину мови за допо-

могою морфологічного аналізатора aot [6];

3) якщо слово є в словнику, отримуємо для нього лексико-граматичний код. Якщо аналізатор повертає декілька варіантів, беремо перший. Приписуємо слову знайдений код. Поряд з лексико-граматичним кодом отримуємо відповідну канонічну форму слова, яка буде використовуватися для статистичної оцінки на етапі зважування;

4) у разі, коли слово не вдається ідентифікувати, використовуючи словники (наприклад, для слів, написаних іншою мовою), йому приписується код іменника, а в як канонічна форма береться саме слово.

Результатом блоку лексико-граматичного аналізу є масив речень з виділеними словами, для кожного з яких відомий його лексико-граматичний код і канонічна форма.

Ранжування речень. Розглянемо питання визначення нового способу оцінювання важливості речень для подальшого їх ранжування. Для отримання вектору ранжування речень потрібно побудувати матрицю зв'язків між ними. Для цього і проводиться оцінювання важливості всіх речень.

Спочатку всі речення подаються у вигляді вектору:

$$S_i = \left[W_i^1, W_i^2, ..., W_i^n \right], \quad (1)$$

де W_i^j – вага j-го слова в і-му реченні.

На відміну від класичних методів, в яких для обчислення ваги слова використовується стандартна метрика $TF \cdot IDF$, у даній роботі пропонується скористатися наступною формулою:

$$W_i^j = \left(TF \cdot IDF + K_k + K_q + K_t \right) \cdot K_s, \quad (2)$$

де: K_k – підвищуючий ваговий коефіцієнт для ключових слів заданої теми;

K_q – підвищуючий ваговий коефіцієнт для слів із запиту користувача;

K_t – підвищуючий ваговий коефіцієнт для слів, які є у заголовках вхідних текстів;

K_s – коефіцієнт важливості, який залежить від частини мови;

$TF \cdot IDF$ – статичний показник, що використовується для оцінювання важливості слів у контексті документу, що є частиною колекції документів чи корпусу [7].

Остаточна вага речення обчислюється як середнє арифметичне ваги всіх слів у реченні:

$$S_i = \frac{\sum_{j=1}^{WordCount(S_i)} W_j}{WordCount(S_i)}, \quad (3)$$

де W_j – вага j-го слова в реченні, $WordCount(S_i)$ – кількість слів в і-му реченні.

Запропонований спосіб оцінювання важливості речень дозволяє отримувати інформативні реферати текстів з будь-якої предметної галузі, оскільки він враховує ключові слова. Зазвичай, в текстах заголовок має велику інформативність, тому вага слів, які входять в заголовок або підзаголовок, збільшується. Беручи до уваги слова із запиту користувача, даний спосіб дозволяє отримувати реферати, які відповідають інформаційним потребам користувачів. Врахування ж важливості частин мови дозволяє мінімізувати вплив неінформативних частин мови, які часто зустрічаютьсяв текстах, та підвищити вплив інформативних (таких як іменники, дієслова).

Таким чином, всі речення подаються векторами з мірою відстані:

$$Sim\left(\overline{S_i}, \overline{S_j}\right) = \frac{\overline{S_i} \cdot \overline{S_j}}{|S_i| \cdot |S_j|}. \quad (4)$$

Наступним кроком методу реферування є формування матриці зв'язків між реченнями. Якщо і та j рівні, то m_{ij} присвоюється 0:

$$M_{ij} = Sim\left(\overline{S_i}, \overline{S_j}\right). \quad (5)$$

Далі обчислюється вектор ранжування речень $\overline{R_t}$ відносно речення-заголовка.

Вибір теми реферату. Для вибору основної теми зведеного реферату використовуються заголовки текстів з вхідної колекції документів. Всі заголовки подаються у вигляді вектору (1), а потім обчислюється вага кожного з них за формулою (3). В якості основної теми обирається заголовок, який має найбільшу вагу.

Генерація реферату. Етап генерації представляє собою вибір з вхідної колекції текстів заданої кількості речень з найбільшою вагою з урахуванням фактору «інформаційної новизни» (тобто речення, які потрапляють у реферат не повторюють зміст одне одного). Для формування зв'язного тексту кінцевого реферату авторами запропоновано наступний алгоритм:

1. Задається відсоток стиснення вхідних текстів $k_{pressing}$.

$$k_{pressing} = \frac{W - W_{sum}}{W} \cdot 100\% \quad (6)$$

де *W* – загальна кількість слів у колекції текстів, W_{sum} – кількість слів у рефераті. Відсоток стиснення рахується саме з використанням кількості слів, а не речень, як в [8], оскільки речення можуть мати різну довжину, і в залежності від того, речення з якою довжиною потрапили до реферату, може значно змінюватись об'єм кінцевого реферату.

2. Всі речення сортуються за зменшенням їх ваги *S*. Вибираються речення з найбільшою вагою для їх можливого включення в реферат. Речення вибираються, доки сумарна кількість слів у них не перевищує W_{sum}.

3. Задається максимальний коефіцієнт подібності двох речень $k_{analogy}$. Він необхідний для того, щоб відсікти схожі за змістом речення. Цим вирішується завдання врахування «інформаційної новизни».

4. Для кожної пари відібраних для включення в реферат речень $\overline{x_i}$, $\overline{x_j}$ перевіряється, чи не перевищує значення $Sim\left(\overline{x_i}, \overline{x_j}\right)$ коефіцієнта подібності $k_{analogy}$. Якщо перевищує, переходимо до п. 5, в іншому випадку – п. 7.

5. З двох речень $\overline{x_i}$, $\overline{x_j}$ вибираємо те, що має меншу вагу (S_i або S_j), та відкидаємо його. Це речення не потрапить до реферату, оскільки за змістом воно дуже подібне до більш вагомого речення і не вносить жодної інформаційної новизни.

6. З метою заміни речення, яке було видалене зі списку тих, що потрапляють в реферат (п. 4), вибираємо речення з найбільшою вагою зі списку тих, що ще не були вибрані, та виконуємо п. 4 для нього.

7. Для тих речень, які не були відібрані для включення в реферат, потрібно обнулити значення коефіцієнтів зв'язності в матриці *M* та у векторі $\overline{R_t}$. Тобто, якщо речення *i* не потрапить до реферату, то коефіцієнтам r_i та m_{ij} (для всіх *j*) потрібно присвоїти 0.

8. З вектору ранжування речень відносно речення-заголовка $\overline{R_t}$ вибирається найбільше значення r_{max}. Коефіцієнт *max* відповідає реченню, яке має найвищий коефіцієнт зв'язності з темою. Це речення ставиться на перше місце в реферат. Вводиться змінна *current*, якій присвоюється значення *max*.

В матриці зв'язків між реченнями *M* обираємо в рядку *current* максимальне значення $m_{current,max}$. Коефіцієнт *max* відповідає реченню, яке найбільш пов'язане з попереднім, що було включене в реферат, тому має бути наступним. Присвоюємо значенню *current* значення *max* та повторюємо даний крок, доки всі обрані речення не будуть розміщені рефераті.

ВИСНОВКИ

Таким чином, у даній роботі проаналізовано переваги та недоліки існуючих методів автоматичного реферування, у тому числі, зведеного, та запропоновано модифікований метод з використанням ранжування речень. Він полягає у введенні нових критеріїв для оцінювання важливості слів, а також у генерації зв'язного тексту підсумкового реферату з урахуванням фактору інформаційної новизни. Даний метод може бути використаний для реферування текстів з різних предметних галузей та складених різними природними мовами.

Напрямком для подальшого продовження роботи над тематикою даної статті є детальне вивчення ефективності розробленого методу та виконання наступних модифікацій на основі емпірично отриманих даних. Також планується здійснити додавання до методу реферування етапу часткового синтаксичного аналізу з метою більш якісного виділення з тексту змістовних понять, що дозволить більш точно визначати ваги речень і, як наслідок, покращить точність та інформативність згенерованого реферату.

ВИКОРИСТАНА ЛІТЕРАТУРА

1. Han, U. Sistemy avtomaticheskogo referirovaniya [Elektronnij resurs] / U. Han, I. Mani // Otkrytye sistemy. 2000. №12. — Rezhym dostupu : http://www.osp.ru/os/2000/12/178370/.
2. Alyguliev, R.M. Referirovanie nabora dokumentov cherez klasterizaciyu i ranjirovanie predlojenii [Elektronnij resurs] / R.M.Alyguliev // Problemy informacionnyh tehnologii. 2010. №1. — S. 26-37. — Rezhym dostupu : http://jpit.az/storage/files/article/a012e1f2dfd736b3d13238add9f72f42.pdf.
3. Soriyan, A. Trends in Multi-document Summarization System Methods [Elektronnij resurs] / A. Soriyan, T. Omodunbi // International Journal of Computer Applications. 2014. №16. — S. 46-52. — Rezhym dostupu : http://research.ijcaonline.org/volume97/number16/pxc3897804.pdf.
4. Zadbuke, A. Automatic Summarization of News Articles using TextRank [Elektronnij resurs] / A. Zadbuke, S. Pimenta, D. Padwal, V. Wangikar // International Journal of Advanced Research in Computer Science and Software Engineering. 2016. №6. — S. 124-127. — Rezhym dostupu : http://www.ijarcsse.com/docs/papers/Special_Issue/iconect2016/sfcs02.pdf.
5. Tarasov, S.D. Metod tematicheskogo svyazannogo ranjirovaniya dlya zadach avtomaticheskogo svodnogo referirovaniya nauchno-tehnicheskih informacionnyh soobschenii : dis. kand. tehn. nauk : 05.13.01. — Sankt-Peterburg, 2011. — 211 s.
6. Avtomaticheskaja obrabotka teksta [Elektronnij resurs]. — Rezhym dostupu : http://aot.ru/.
7. TF-IDF [Elektronnij resurs]. — Rezhym dostupu : https://uk.wikipedia.org/wiki/TF-IDF.
8. Ehorov, S.V. Metod semanticheskogo szhatija teksta [Elektronnij resurs] / S.V. Ehorov, I.N. Ehorova // ISSN 2222-0631. Visnyk NTU «HPI». 2013. №54 (1027). — S. 118-123. — Rezhym dostupu : http://www.irbis-nbuv.gov.ua/cgi-in/irbis_nbuv/cgiirbis_64.exe?I21DBN=LINK&P21DBN=UJRN&Z21ID=&S21REF=10&S21CNR=20&S21STN=1&S21FMT=ASP_meta&C21COM=S&2_S21P03=FILA=&2_S21STR=vcpimm_2013_54_13.

Рецензент: д.т.н., проф. Дичка І.А.
НТУУ «Київський політехнічний інститут»