УДК 510.635:004.891 (045)
DOI: 10.18372/2310-5461.34.11605

А. І. Вавіленкова — канд. техн. наук, доц.
Національний авіаційний університет
ID ORCID 0000-0002-9630-4951
a_vavilenkova@mail.ru

ОСНОВИ АВТОМАТИЧНОЇ ОБРОБКИ ЕЛЕКТРОННИХ ТЕКСТІВ ПРИ БОРОТЬБІ З РЕРАЙТОМ

Постановка проблеми та її зв'язок з важливими науковими завданнями

Сьогодні, коли всі мають прямий доступ до різноманітних електронних ресурсів, все більш актуальним стає питання про плагіат робіт у освітній сфері, плагіат у сфері музики, копіювання контенту сайту, дублювання зображень, запозичення ідей та ін. Під плагіатом у будь-якій галузі згідно з законом України «Про авторське право і суміжні права» (ст. 50 «Порушення авторського права і суміжних прав» п. в) розуміють оприлюднення, повністю або частково, чужого твору під іменем особи, яка не є автором цього твору [1]. Під плагіатом у академічній сфері розуміють публікацію чужого тексту під власним іменем або запозичення його фрагменту без зазначення джерела [2]. За статистикою опитування, яку можна знайти в мережі Інтернет [3], 26 % людей, що навчаються, здійснюють дослівний плагіат без посилань на джерела, 18% — перекладають іноземні роботи без зазначення джерел, 49 % — переписують текст з джерела власними словами, не вказуючи посилань, решта відсотків припадає на фальсифікацію, відсутність посилань на прямі цитати, поєднання власного та запозиченого тексту і т. д. Ці показники не можуть претендувати на еталонні, так як:

– в Україні відсутня єдина база всіх акад.-мічних робіт;

– не існує єдиних критеріїв визначення плагіату для всіх існуючих сьогодні систем перевірки на унікальність;

– не існує єдиної шкали для оцінки відсотку збігів;

– не існує критеріїв допустимості, за якими автоматично знайдений збіг не буде вважатися плагіатом (наприклад, стандартні кліше, що обов'язково присутні на титульних сторінках дипломних та дисертаційних робіт).

Таким чином, для визначення унікальності електронних текстових документів потрібно дотримуватися певних критеріїв, проте цю процедуру можна віднести до технічних аспектів функціонування тих чи інших автоматичних систем. Натомість лінгвістичний аспект виявлення плагіату серед текстових документів вимагає наявності методів та алгоритмів, що дозволили б формально здійснювати змістовний порівняльний аналіз речень та фрагментів природної мови.

Аналіз останніх досліджень і публікацій

Насамперед виявлення плагіату передбачає використання для цього комп'ютерних методів, які повинні не лише знаходити повний збіг групи слів, але й виявляти перефразування, зміну порядку слів у реченні, вживання синонімів та інших засобів створення нового унікального тексту з уже існуючого. Такий тип плагіату (рерайтинг) найважче оцінити існуючим системам пошуку унікальності електронних текстових документів [4–6].

У наш час єдиним програмним продуктом, незважаючи на заяви розробників різних сервісів виявлення плагіату про використання ними унікальних алгоритмів [7], що дозволяє здійснювати автоматичний лінгвістичний аналіз, а відповідно, має можливість виявлення засобів перефразування, є програма ABBYY Compreno [8]. Для аналізу текстової інформації система використовує дерева розбору, отримані семантико-синтаксичним парсером, онтології (моделі предметної галузі), правила вилучення інформації та правила ідентифікації. Складністю в роботі цієї системи є застосування правил до різних фрагментів дерева розбору, тобто постійний та повний перегляд шаблонів і тверджень бази знань.

Способи перефразування та опис засобів когезії у природно мовних текстах наводять у своїх працях Д. Лайонз [9] та М. В. Нікітін [10].

Проте, незважаючи на велику кількість публікацій у сфері комп'ютерної лінгвістики та різноманітні сервіси для перевірки унікальності електронних текстових документів, відсоток плагіату академічних робіт з кожним роком зростає завдяки збільшенню кількості рерайтингових та копірайтингових агенств. А отже, потрібна розробка нових принципів та методів автоматичного лінгвістичного аналізу, що дозволили б виявляти змістовні збіги в електронних текстових документах, чому і присвячені матеріали даної статті.

Цілі статті

Як показує статистика, для перетворення авторського тексту на унікальний для машинного сприйняття текст, застосовується рерайтинг, автоматичних методів дослідження якого поки що не запропоновано. Тому метою статті є формулювання основних критеріїв, за якими здійснюється рерайтинг електронних текстових документів, та опис формальних умов їх виявлення за допомогою логіко-лінгвістичного моделювання.

Виклад основного матеріалу досліджень

Матеріали досліджень присвячені проблемі виявлення поверхневого рерайту. При цьому на предмет перефразування перевіряється кожне речення тексту, структура якого завдяки рерайтингу змінена, а унікальність наближається до 100 %.

Формальні умови, запропоновані автором виведені на базі основних способів рерайтингу, які застосовуються сьогодні [11]:

– зміна конструкції речення природної мови;

– заміна прямої мови;

– використання синонімів, гіперонімів та конверсивів;

– зміна порядку речень та абзаців у тексті;

– видалення деяких слів та словосполучень з тексту.

Нехай логіко-лінгвістичні моделі [12] першого (оригіналу) та речення-рерайту у загальному випадку позначаються відповідно:

$$L^{S_1} =$$

$$= \bigwedge_{z_1 \in Z_{p1}^{S_1}(x_1,g_1,y_1,q_1,h_1)} \bigwedge_{r_1 \in R_{p1}^{S_1}(x_1,g_1,y_1,q_1,z_1,h_1)} p_1(x_1,g_1,y_1,q_1,z_1,r_1,h_1),$$

$$L^{S_2} =$$

$$= \bigwedge_{z_2 \in Z_{p2}^{S_2}(x_2,g_2,y_2,q_2,h_2)} \bigwedge_{r_2 \in R_{p2}^{S_2}(x_2,g_2,y_2,q_2,z_2,h_2)} p_2(x_2,g_2,y_2,q_2,z_2,r_2,h_2).$$

Умова 1. Якщо одне речення природної мови написано у стверджувальній формі, а друге — у вигляді умови, то у першому випадку логіко-лінгвістична модель являє собою атомарний предикат, а у другому — два атомарних предиката, з'єднаних логічною операцією імплікації, і при цьому зберігається еквівалентність наступних компонент логіко-лінгвістичних моделей: $y_1 \equiv y_2$, $z_1 \equiv z_2$, $h_1 \equiv \tilde{h}_2$ (антоніми), $x_1 \equiv x_2'$, $g_1 \equiv g_2'$, $p_1 \equiv \vec{p}_2$ (конверсиви), $y_1 \equiv \hat{p}_2'$ (спільнокореневі), $x_2 \equiv 0$, то речення однакові за змістом.

Наприклад, є два речення: «*Флективні мови погано піддаються опису формалізмами*» та «*Якщо формалізмами не вдається правильно зробити опис, то вважають, що описуються флективні мови*».

Перевіримо за першою умовою, чи являється друге і третє речення рерайтом першого.

Для цього необхідно побудувати логіко-лінгвістичні моделі цих двох речень:

$$L^{S_1} = p_1(x_1,g_1,y_1,q_1,z_1,r_1,h_1) = \\ = p_1(x_1,g_1,y_1,0,z_1,0,h_1),$$

L^{S_1} = *піддаються (мови, флективні, опису, 0, формалізмами, 0, погано*).

$$L^{S_2} = p_2(x_2,g_2,y_2,q_2,z_2,r_2,h_2) \to \\ \to p_2'(x_2',g_2',y_2',q_2',z_2',r_2',h_2') = \\ = \neg p_{21} \,\&\, p_{22}(0,0,y_2,0,z_2,0,h_2) \to \\ \to p_2'(x_2',g_2',0,0,0,0,h_2'),$$

L^{S_2} = *¬вдається&зробити (0,0,опис,0, формалізмами,0,правильно)→описуються (мови, флективні,0,0,0,0,0)*.

Умова 2. Якщо для перефразування речення вжитий спосіб перенесення дії у пасивний стан, а в логіко-лінгвістичних моделях двох речень, що порівнюються, простежується еквівалентність таких компонент $x_2 \equiv 0$, $x_1 \equiv y_2$, $g_1 \equiv q_2$, $z_1 \equiv z_2$, $h_1 \equiv \hat{h}_2$ (синоніми), то одне з речень природної мови являється рерайтом другого.

Нехай є два речення: «*Флективні мови погано піддаються опису формалізмами*» та «*Для флективних мов дуже важко підібрати формалізми*».

Логіко-лінгвістичні моделі речень будуть мати вигляд, відповідно:

$$L^{S_1} = p_1(x_1,g_1,y_1,q_1,z_1,r_1,h_1) = \\ = p_1(x_1,g_1,y_1,0,z_1,0,h_1),$$

L^{S_1} = *піддаються (мови, флективні, опису, 0, формалізмами, 0, погано*).

$$L^{S_2} = p_2(x_2,g_2,y_2,q_2,z_2,r_2,h_2) = \\ = p_2(0,0,y_2,q_2,z_2,0,h_{21} \,\&\, h_{22}),$$

L^{S_2} = *підібрати (0,0,мов, флективних, формалізми,0,дуже&важко)*.

Умова 3. Якщо у двох реченнях природної мови було здійснено пряму заміну слів синонімами, то логіко-лінгвістичні моделі таких речень будуть еквівалентними, тобто $p_1 \equiv \hat{p}_2$ (спільнокореневі), $x_1 \equiv \hat{x}_2$, $z_1 \equiv \hat{y}_2$, $h_1 \equiv \hat{h}_2$ (синоніми), при

цьому характеристики об'єктів, суб'єктів та предметів відношень можуть бути упущені або додані, тому їх еквівалентність не являється обов'язковою, то речення однакові за змістом, і одне з них являється рерайтом іншого.

Наприклад, нехай є два речення: «*Флективні мови погано піддаються опису формалізмами*» та «*Природні мови описуються математичними формулами не дуже добре*».

Логіко-лінгвістичні моделі таких речень будуть мати вигляд:

$$L^{S_1} = p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1) = \\ = p_1(x_1, g_1, y_1, 0, z_1, 0, h_1),$$

L^{S_1} = *піддаються (мови, флективні, опису, 0, формалізмами, 0, погано*).

$$L^{S_2} = p_2(x_2, g_2, y_2, q_2, z_2, r_2, h_2) = \\ = p_2(x_2, g_2, y_2, q_2, 0, 0, \neg h_{21} \,\&\, h_{22}),$$

L^{S_2} = *описуються(мови, природні, формулами, математичними, 0,0,* $\neg$ *дуже&добре)*.

Умова 4. Якщо дія, про яку йдеться у першому реченні, замінена на дієприслівний зворот у другому, а логіко-лінгвістичні моделі таких речень матимуть однакову структуру атомарних предикатів, проте у логіко-лінгвістичній моделі другого речення буде наявний ще один предикат, при цьому $x_1 \equiv y_2'$, $g_1 \equiv q_2'$, $z_1 \equiv z_2'$, $x_2' \equiv 0$, $y_1 \equiv \hat{p}_2'$ (спільнокореневі), то одне з речень природної мови являється рерайтом другого.

Наприклад, два речення «*Флективні мови погано піддаються опису формалізмами*» та «*Погано піддаючись опису, флективні мови все-таки намагаються описати формалізмами*».

Логіко-лінгвістичні моделі речень мають вигляд:

$$L^{S_1} = p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1) = \\ = p_1(x_1, g_1, y_1, 0, z_1, 0, h_1),$$

L^{S_1} = *піддаються (мови, флективні, опису, 0, формалізмами, 0, погано*).

$$L^{S_2} = p_2(x_2, g_2, y_2, q_2, z_2, r_2, h_2) \,\& \\ \& p_2'(x_2', g_2', y_2', q_2', z_2', r_2', h_2') = \\ = p_2(x_2, g_2, y_2, 0, 0, 0, h_2) \,\& \\ \& p_{21}' \,\&\, p_{22}'(0, 0, y_2', q_2', z_2', 0, h_2'),$$

L^{S_2} = *піддаються (мови, флективні, опису, 0, 0, 0, погано*)*& намагаються& описати (0,0, мови,флективні,формалізмами,0,все-таки)*.

Умова 5. Якщо у першому реченні, що порівнюється, присутня пряма мова, а у другому відсутня, і при цьому логіко-лінгвістичні моделі речень мають такий вигляд

$$L^{S_1} = p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1) \to \\ \to p_1'(x_1', g_1', y_1', q_1', z_1', r_1', h_1'), \\ L^{S_2} = p_2(\hat{x}_1', g_2, \hat{p}_1', q_2, z_2, r_2, h_2) \to \\ \to p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1),$$

то зміст речень однаковий, а одне з них є рерайтом другого.

Наприклад, нехай є два речення «*"Флективні мови погано піддаються опису формалізмами", — констатують вчені*» та «*Вченими була дана констатація факту про те, що флективні мови погано піддаються опису формалізмами*».

$$L^{S_1} = p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1) \to \\ \to p_1'(x_1', g_1', y_1', q_1', z_1', r_1', h_1') = \\ = p_1(x_1, g_1, y_1, 0, z_1, 0, h_1) \to \\ \to p_1'(x_1', 0, 0, 0, 0, 0, 0),$$

L^{S_1} = *піддаються (мови, флективні, опису, 0, формалізмами, 0, погано*)→

констатують(вчені,0,0,0,0,0,0).

$$L^{S_2} = p_2(x_2, g_2, y_2, q_2, z_2, r_2, h_2) \to \\ \to p_2'(x_2', g_2', y_2', q_2', z_2', r_2', h_2') = \\ = p_{21} \,\&\, p_{22}(x_2, 0, y_2, 0, z_2, 0, 0) \,\& \\ \& p_2'(x_2', g_2', y_2', 0, z_2', 0, h_2'),$$

L^{S_2} = *була&дана(вченими,0,констатація, 0,факту,0,0)*→

піддаються(мови, флективні,опису,0, формалізмами,0,погано).

Умова 6. Якщо дія в першому реченні замінюється однаковим за змістом іменником, при цьому логіко-лінгвістична модель другого речення представляє собою атомарні предикати, з'єднані логічною операцією імплікації з однаковими відношеннями, а $x_1 \equiv y_{21}, g_1 \equiv q_{21}$, $y_1 \equiv z_{21}$, $z_1 \equiv r_{21}$, $h_1 \equiv g_{22}$, то одне з речень природної мови являється рерайтом другого.

Нехай є два речення: «*Флективні мови погано піддаються опису формалізмами*» та «*Піддавання флективних мов опису формалізмами — дуже складна справа*».

Логіко-лінгвістичні моделі речень будуть мати вигляд:

$$L^{S_1} = p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1) = \\ = p_1(x_1, g_1, y_1, 0, z_1, 0, h_1),$$

L^{S_1} = *піддаються (мови, флективні, опису, 0, формалізмами, 0, погано*).

$$L^{S_2} = p_2(x_{21}, g_{21}, y_{21}, q_{21}, z_{21}, r_{21}, h_{21})\,\&$$
$$\&\, p_2(x_{22}, g_{22}, y_{22}, q_{22}, z_{22}, r_{22}, h_{22}) =$$
$$= p_2(0,0, y_{21}, q_{21}, z_{21}, r_{21}, 0)\,\&$$
$$\&\, p_2(x_{22}, g_{22}, 0,0,0,0, h_{22}),$$

L^{S_2} = піддавання (0,0,мов, флективних, опису, формалізмами, 0)→

піддавання(справа,складна,0,0,0,0,дуже).

Умова 7. Якщо в реченні була здійснена зміна структури за рахунок об'єднання або роз'єднання змістовних складових, то в логіко-лінгвістичних моделях це відобразиться на кількості атомарних предикатів з однаковими компонентами, що фактично, дублюють одна одну, при цьому $x_1 \equiv y_2$, $g_1 \equiv q_2$, $z_1 \equiv z_2'$, $x_2 \equiv 0$, $g_2 \equiv 0$, $y_1 \equiv \hat{p}_2$ (спільнокореневі), $h_1 \equiv \hat{h}_2 \equiv \hat{h}_2'$ (синоніми), то над реченнями, що порівнюються, здійснено рерайтинг.

Наприклад, нехай є речення «*Флективні мови погано піддаються опису формалізмами*» і «*Описувати флективні мови дуже складно. Ще складніше описувати флективні мови формалізмами*».

Логіко-лінгвістичні моделі речень будуть мати вигляд:

$$L^{S_1} = p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1) =$$
$$= p_1(x_1, g_1, y_1, 0, z_1, 0, h_1),$$

L^{S_1} = піддаються (мови, флективні, опису, 0, формалізмами, 0, погано).

$$L^{S_2} = p_2(x_2, g_2, y_2, q_2, z_2, r_2, h_2)\,\&$$
$$\&\, p_2'(x_2', g_2', y_2', q_2', z_2', r_2', h_2') =$$
$$= p_2(0,0, y_2, q_2, 0,0, h_2)\,\&$$
$$\&\, p_2(0,0, y_2, q_2, z_2', 0, h_2'),$$

L^{S_2} = описувати(0,0,мови,флективні,0,0, дуже&складно)&

описувати(0,0,мови,флективні,формалізмами, 0,складніше).

Висновки

Для успішного проведення пошуку рерайту необхідно здійснювати перевірку всіх запропонованих у статті умов паралельно, що дасть можливість виявляти прийоми комбінованого рерайтингу.

На рівні аналізу фраз та текстів в цілому виникають нові умови виявлення рерайтингу, які пов'язані, в першу чергу, зі зміною стилю текстової інформації та зміною порядку речень у тексті. Часто рерайтинг починається не з початку тексту, а з кінця, що дає можливість зберегти зміст (ідею), проте переписати речення тексту зовсім іншими словами.

Таким чином, дослідження різних методів рерайтингу дало можливість виявити закономірності та можливості заміни елементів логіко-лінгвістичних моделей для формального встановлення їх тотожності.

ЛІТЕРАТУРА

1. Стаття 50. Порушення авторського права і суміжних прав [Електронний ресурс]. — Режим доступу: http://kodeksy.com.ua/pro_avtors_ke_pravo_i_sumizhni_prava/statja-50.htm.

2. Перевірка на плагіат [Електронний ресурс]. — Режим доступу: http://library.kubg.edu.ua/informatsiya/naukovtsiam/4-informatsiia/naukovtsiam/253-perevirka-na-plahiat.html.

3. Епідемія академічного плагіату в цифрах [Електронний ресурс]. — Режим доступу: http://studway.com.ua/plagiat-2/.

4. StrikePlagiarism [Електронний ресурс]. — Режим доступу:http://strikeplagiarism.com.

5. Проверка уникальности текста Advego Plagiatus [Электронный ресурс]. — Режим доступа: http://advego.ru/.

6. Text.ru [Електронний ресурс]. — Режим доступу: https://text.ru/about.

7. **Вавіленкова А. І.** Способи виявлення логічних зв'язків між частинами текстових документів / А. І. Вавіленкова // Вісник Національного технічного університету «Харківський політехнічний університет»: зб. наук. праць. — (Серія «Нові рішення в сучасних технологіях»). — 2016. — № 12 (1184). — С. 101 — 105. doi: 10.20998/2413-4295. 2016.12.14.

8. Алгоритм извлечения информации в ABBYY Compreno [Электронный ресурс]. — Режим доступа: https://habrahabr.ru/company/abbyy/blog/269191/.

9. **Лайонз Дж.** Лингвистическая семантика: монография / Дж. Лайонз. — М. : Языки славянской культуры, 2003. — 400 с.

10. **Никитин М. В.** Курс лингвистической семантики / М. В. Никитин. — СПб. : из-во РГПУ им. Герцена, 2007. — 819 с.

11. Рерайтинг [Электронный ресурс]. — Режим доступа: http://uniofweb.ru/wiki/rewriting/#full_version.

12. **Вавіленкова А. І.** Критерії аналізу логіко-лінгвістичних моделей речень природної мови / А. І. Вавіленкова // Вісник Національного технічного університету «Харківський політехнічний університет»: зб. наук. праць. — (Серія «Нові рішення в сучасних технологіях»). — 2017. — № 7 (1229). — С. 118–122. doi: 10.20998/2413-4295.2017.07.16.

Вавіленкова А. І.

ОСНОВИ АВТОМАТИЧНОЇ ОБРОБКИ ЕЛЕКТРОННИХ ТЕКСТІВ ПРИ БОРОТЬБІ З РЕРАЙТОМ

Матеріали статті присвячені проблемі виявлення поверхневого рерайту. При цьому на предмет перефразування перевіряється кожне речення тексту, структура якого завдяки рерайтингу змінена. Метою статті є формулювання основних критеріїв, за якими здійснюється рерайтинг електронних текстових документів, та опис формальних умов їх виявлення за допомогою логіко-лінгвістичного моделювання. Описано сім умов, за яких при порівнянні логіко-лінгвістичних моделей можна виявити рерайт і які виведені на базі основних способів рерайтингу, що використовуються сьогодні. Дослідження різних методів рерайтингу дало можливість виявити закономірності та можливості заміни елементів логіко-лінгвістичних моделей для формального встановлення їх тотожності.

Ключові слова: логіко-лінгвістична модель; рерайт; тотожність; зміст.

Вавиленкова А. И.

ОСНОВЫ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЭЛЕКТРОННЫХ ТЕКСТОВ ПРИ БОРЬБЕ С РЕРАЙТОМ

Материалы статьи посвящены проблеме выявления поверхностного рерайта. При этом на предмет перефразирования проверяется каждое предложение текста, структура которого изменена благодаря рерайтингу. Целью статьи является формулировка основных критериев, по которым осуществляется рерайтинг электронных текстовых документов, а также описание формальных условий их выявления с помощью логико-лингвистического моделирования. Описаны семь условий, с помощью которых при сравнении логико-лингвистических моделей можно выявить рерайт и которые выведены на базе основных способов рерайтинга, которые используются сегодня. Исследование различных методов рерайтинга дало возможность выявить закономерности и возможности замены элементов логико-лингвистических моделей для формального восстановления их тождества.

Ключевые слова: логико-лингвистическая модель; рерайт; тождество; содержание.

Vavilenkova A. I.

THE FUNDAMENTALS OF AUTOMATIC PROCESSING OF ELECTRONIC TEXTS IN ACTIONS AGAINST REWRITING

The article is devoted to the problem of detecting so called surface rewrite. In this case the object of paraphrasing is checked in every sentence which structure is changed due to rewriting. The aim of the article is to formulate the basic criteria of rewriting of electronic text documents and a description of the formal conditions of their detection using logic and linguistic modeling. The study describes seven conditions under which, using comparison of logic and linguistic models can detect rewriting and which are derived based on the main modern methods of rewriting. Research of different methods of rewriting make it possible to identify patterns and the possibility of replacing the logic and linguistic models for the formal establishment of their identity

Keywords: logical and linguistic model; rewriting; identity; content.

Стаття надійшла до редакції 30.05.2017 р.
Прийнято до друку 31.05.2017 р.
Рецензент — д–р техн. наук, проф. Зіадінов Ю. К.