https://doi.org/10.31649/1997-9266-2021-157-4-70-77

УДК 510.635:004.891(045)4

А. І. Вавіленкова[1]

ШЛЯХИ ВІДНОВЛЕННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ, ПРЕДСТАВЛЕНОЇ У ВИГЛЯДІ ЛОГІКО-ЛІНГВІСТИЧНОЇ МОДЕЛІ

[1]Національний авіаційний університет, Київ

Обґрунтовано актуальність вирішення проблеми пошуку змістовних зв'язків в електронних текстових документах з метою подальшого їх порівняння за змістом та удосконалення роботи систем виявлення плагіату. При цьому важливим етапом є оцінювання достовірності сформованих формальних моделей. Тому метою цієї статті є дослідження алгоритму автоматичного аналізу логіко-лінгвістичних моделей електронних текстових документів для відтворення текстової інформації, що об'єднує в собі основні властивості тексту та його складових частин, відображає основні взаємозв'язки між структурними компонентами. Логіко-лінгвістична модель текстового документу представляє собою впорядковану четвірку та масив логіко-лінгвістичних моделей речень природної мови, що входять до тексту. Автором запропоновано декілька шляхів відновлення текстової інформації, що відштовхуються від структури логіко-лінгвістичної моделі електронного текстового документу, яка містить лінгвістичну та семантико-синтаксичну складову. Описано схеми здійснення відновлення текстової інформації, вибрано комбінований спосіб, що передбачає аналіз семантико-синтаксичної складової паралельно з аналізом текстової бази, зокрема, її компоненти – множини пропозицій, що містить зв'язки між логіко-лінгвістичними моделями речень тексту електронного текстового документу. Розроблено алгоритм відновлення текстової інформації, представленої у вигляді формальної логіко-лінгвістичної моделі електронного текстового документу, описано його етапи. Всі кроки алгоритму продемонстровано на прикладі аналізу конкретної заданої логіко-лінгвістичної моделі фрагменту електронного текстового документу. Проведено експерименти щодо відновлення текстової інформації для текстів наукового стилю. Виявлено, що до основних факторів, що впливають на відновлення текстової інформації, належить зняття омонімії, а також різна інтерпретація синонімічних конструкцій та інваріантних форм логіко-лінгвістичних моделей речень природної мови.

Ключові слова: текстова інформація, природна мова, логіко-лінгвістична модель, алгоритм відновлення.

Вступ

Незважаючи на велику кількість різноманітних систем та он-лайн сервісів здійснення порівняльного аналізу електронних текстових документів, сьогодні в мережі Інтернет та серед академічних робіт продовжує зростати кількість неунікального контенту.

Це говорить про недосконаліть алгоритмічної бази для здійснення порівняльного аналізу. Тому *основною проблемою* для змістовного аналізу електронних текстових документів є відсутність алгоритмів автоматичного здійснення лінгвістичного аналізу текстів на основі побудови формальних моделей та алгоритмів відновлення окремих фрагментів тексту і взаємозв'язків між ними для здійснення якісної перевірки на плагіат. Ця тематика підпадає під один з пріоритетних тематичних напрямків наукових досліджень і науково-технічних розробок у категорії «Інформаційні та комунікаційні технології», а саме інтелектуальні інформаційні та інформаційно-аналітичні технології згідно з постановою Кабінету Міністрів України №556 від 23.08.2016 р. та поправок до неї у постанові № 380 від 21.04.2021р. [1]. Також в Україні діє проєкт сприяння академічній доброчесності SAIUP за підтримки Міністерства освіти і науки України [2], що свідчить про актуальність розроблення систем виявлення плагіату у сфері освіти.

Аналіз досліджень і публікацій у сфері запобігання академічному плагіату [3]—[6] доводить, що процес прийняття рішень щодо унікальності тих чи інших електронних текстових документів великого розміру є автоматизованим і у більшості випадків покладається на користувача системи або експертні комісії. Для здійснення автоматичного порівняльного аналізу електронних тексто-

вих документів за змістом створюються формальні моделі представлення знань кожного тексту [7]—[10], які надалі порівнюються. Зокрема, у своїх роботах Д. В. Ланде [11] дає обґрунтування використанню різних моделей та алгоритмів пошуку текстової інформації. Вирішення проблеми побудови формальної моделі тексту з лінгвістичної точки зору можна знайти у роботах І. Р. Гальперіна [12]. В результаті багаторічних досліджень і узагальнення здобутих знань у сфері комп'ютерної лінгвістики та інформаційних технологій, розроблено формальну логіко-лінгвістичну модель представлення електронного текстового документу [13], що базується на виявленні логічних зв'язків між складовими тесту.

Невирішеною частиною проблеми пошуку змістовних зв'язків в електронних текстових документах є оцінювання достовірності отриманих формальних моделей. Її можна вирішити, застосувавши зворотний до синтезу складових моделі [14] процес — аналіз, результатом якого буде відновлений текст.

Тому *метою статті* є дослідження алгоритму автоматичного аналізу логіко-лінгвістичних моделей електронних текстових документів для відтворення текстової інформації, що дасть змогу оцінювати точність побудови формальних логічних моделей текстів для їх порівняння за змістом.

Результати дослідження

Аналіз логіко-лінгвістичних моделей речень природної мови є процесом дослідження окремих компонент логіко-лінгвістичних моделей речень природної мови та їх характеристик з метою виявлення та встановлення логічних зв'язків між словами, які інтерпретуються цими компонентами [13].

Логіко-лінгвістична модель текстового документу об'єднує в собі основні властивості тексту та його складових частин, відображає основні взаємозв'язки між структурними компонентами [15], представляє собою впорядковану четвірку та масив логіко-лінгвістичних моделей речень природної мови, що входять до тексту.

Існує декілька шляхів відновлення текстової інформації:

– аналіз параметрів лінгвістичної складової з подальшим відновленням тексту за масивом логіко-лінгвістичних складових, на основі чого спочатку відновлюється структура складних синтаксичних частин, кількість абзаців та логічні зв'язки між реченнями, а після цього в уже відновлену структуру тексту підставляється текст речень з семантико-синтаксичної складової (рис. 1а);

– аналіз семантико-синтаксичної складової, тобто відновлення структури речень за компонентами їх логіко-лінгвістичних моделей, на базі чого буде отримано текст, а за подальшого аналізу він буде розбиватися на абзаци та буде здійснено заміни між деякими складовими речень згідно з правилами синтезу логіко-лінгвістичних моделей (рис. 1б);

– комбінований спосіб, що передбачає аналіз семантико-синтаксичної складової паралельно з аналізом текстової бази, зокрема, її компоненти — множини пропозицій, що містить зв'язки між логіко-лінгвістичними моделями речень тексту електронного текстового документу (рис. 1в).

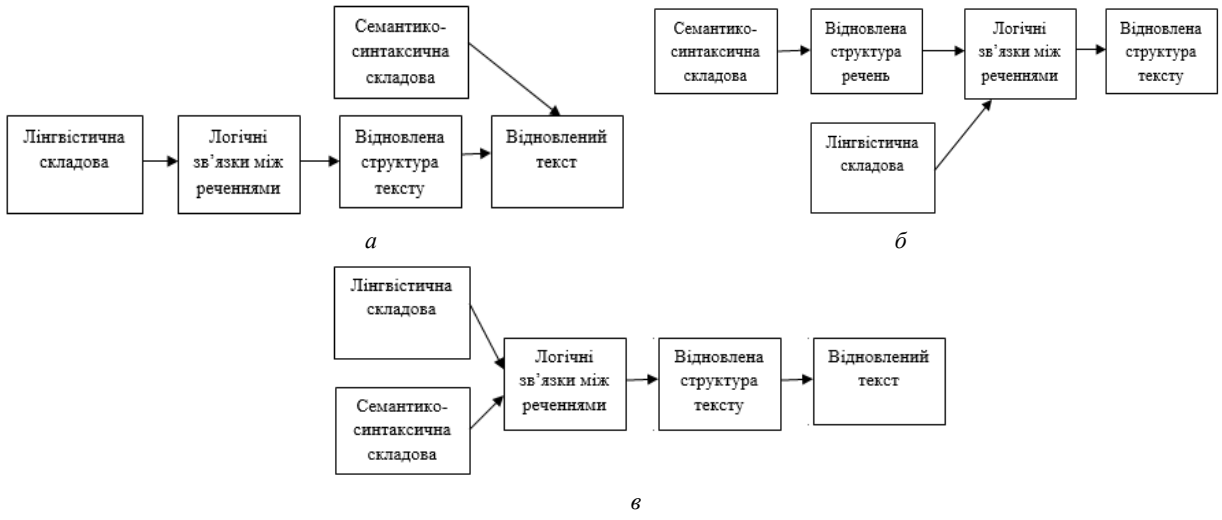

Рис. 1. Шляхи відновлення текстової інформації

Лінгвістична складова — це абстрактна модель, що формується на основі правил побудови зв'язного тексту та його змістовних категорій, які виражаються цими правилами [15]—[16].

Нехай для певного тексту t_1 задано таку лінгвістичну складову:

$$t_1 = < CQ_1, F_1, B_1, A_1 >;$$

$t_1 = <\{1\}, \{f_1\}, <\{$вирішення, проблема, прийняття, рішення, процедура, структуризація$\}$,

$\{$вирішення проблеми, прийняття рішень$\}, \{l_1 = \{x_{22}, q_{22}, y_2', q_2'\}, l_2 = \{q_{11}, q_{12}, z_3\}, l_3 = \{q_{12}', g_{22}\}\} >, <1,1,1>>$,

де $CQ_1 \in C$ — стиль тексту, що може приймати одне із значень елементів множини слів текстів $C = \{1, 2, 3, 4\}$, що розглядаються (1 — науковий, 2 — публіцистичний, 3 — художній, 4 — офіційно-діловий); $F_1 = \{f_1, ..., f_j, ..., f_m\}$ — множина складних синтаксичних частин тексту, $j = \overline{1, m}$; m — кількість складних синтаксичних частин; B_1 — текстова база, що складається з набору ключових слів тексту та взаємопов'язаних пропозицій, і яку можна представити у вигляді трійки: $B_1 = \langle K_1, SJ_1, D_1 \rangle$, K_1 — множина ключових слів тексту; SJ_1 — множина ключових словосполучень тексту; D_1 — множина пропозицій; $A_1 = \{a_1, ..., a_k, ..., a_q\}$ — множина абзаців тексту, $k = \overline{1, q}$, q — кількість абзаців.

Кожний абзац у свою чергу описується трійкою: $a_k = \langle H, Y, R \rangle$, $H = \{1, 2\}$ — множина типів зв'язків між реченнями (ланцюговий чи паралельний); $Y = \{1, 2, 3, 4, 5\}$ — множина типів тематичних прогресій, що вжиті у абзаці $a_k \in A_1$; $R = \{1, 2, 3, 4, 5, 6\}$ — множина рематичних домінант у абзаці.

Синтаксис у роботах автора передбачає формування логіко-лінгвістичних моделей речень природної мови на основі формальних правил побудови словосполучень природної мови та принципах виявлення засобів вираження змістовних відношень у цих словосполученнях [13]. Для цього тексту t_1 задано таку семантико-синтаксичну складову, отриману при побудові логіко-лінгвістичних моделей текстового фрагменту [17]:

$$t_1' = \begin{cases} L_{p^{(\lambda)}}^{S_1}\left(x_1^{(\lambda)}, g_1^{(\lambda)}, y_1^{(\lambda)}, q_1^{(\lambda)}, z_1^{(\lambda)}, r_1^{(\lambda)}, h_1^{(\lambda)}\right), \\ L_{p^{(\lambda)}}^{S_2}\left(x_2^{(\lambda)}, g_2^{(\lambda)}, y_2^{(\lambda)}, q_2^{(\lambda)}, z_2^{(\lambda)}, r_2^{(\lambda)}, h_2^{(\lambda)}\right), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ L_{p^{(\lambda)}}^{S_\delta}\left(x_\delta^{(\lambda)}, g_\delta^{(\lambda)}, y_\delta^{(\lambda)}, q_\delta^{(\lambda)}, z_\delta^{(\lambda)}, r_\delta^{(\lambda)}, h_\delta^{(\lambda)}\right), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ L_{p^{(\lambda)}}^{S_{N(t_1)}}\left(x_{N(t_1)}^{(\lambda)}, g_{N(t_1)}^{(\lambda)}, y_{N(t_1)}^{(\lambda)}, q_{N(t_1)}^{(\lambda)}, z_{N(t_1)}^{(\lambda)}, r_{N(t_1)}^{(\lambda)}, h_{N(t_1)}^{(\lambda)}\right), \end{cases}$$

де $L^{S_{\delta_1}}$ — логіко-лінгвістична модель речення природної мови текстового фрагменту, $\delta = \overline{1, N(t_1)}$ — номер речення у тексті, $N(t_1)$ — загальна кількість речень у тексті t_1.

$$L^{S_1} = p_1(0,0,y_1,0,0,0,h_1) \to p_1'(x_1', q_{11}'\_q_{12}', y_1', 0,0,0,0) \,\&$$
$$\left[p_{11}''\_p_{12}''(y_1', 0, y_1'', q_1'', z_{11}'', r_1'', 0) \,\&\, p_{11}''\_p_{12}''(y_1', 0, y_1'', q_1'', z_{12}'', r_1'', 0)\right];$$

$$L^{S_2} = p_2(x_{21}, 0, y_2, q_2, z_2, r_2, 0) \,\&\, p_2(x_{22}, g_{22}, y_2, q_2, z_2, r_2, 0) \,\&\, p_2'(y_2, q_2, y_2', q_2', 0,0,0) \to$$

$$L^{S_3} = p_{31}\_p_{32}(0,0,y_3,q_3,z_3,r_{31},0) \,\&\, p_{31}\_p_{32}(0,0,y_3,q_3,z_3,r_{32},0);$$

$$L^{S_1} = \text{вирішення}\begin{pmatrix}\text{0,0,проблеми,}\\ \text{0,0,0,неоднозначне}\end{pmatrix} \to \text{потребує}\begin{pmatrix}\text{процес, прийняття\_рішень,}\\ \text{структуризації,0,0,0,0}\end{pmatrix}\&$$

$$\begin{bmatrix}\text{дозволить\_визначити}\begin{pmatrix}\text{структуризація,0,етапи,}\\ \text{спрямовані,вирішення,проблеми,0}\end{pmatrix}\& \\ \text{дозволить\_визначити}\begin{pmatrix}\text{структуризація,0,}\\ \text{процедури,спрямовані,вирішення,проблеми,0}\end{pmatrix}\end{bmatrix}.$$

$$L^{S_2} = \text{є}\ (\text{підготовка},0,\text{набір, процедур,процесі,управління},0)\&$$
$$\text{є}\ (\text{прийняття, рішень, набір, процедур,процесі,управління},0)\&$$
$$\text{об'єднуються}\ (\text{набір, процедур, етапи,окремі},0,0,0)\rightarrow$$

$$L^{S_3} = \text{можна\_побудувати}(0,0,\text{схему,розробки,рішень,наукових},0)\&$$
$$\text{можна\_побудувати}(0,0,\text{схему,розробки,рішень,обгрунтованих},0).$$

Алгоритм відновлення текстової інформації за заданою формальною логіко-лінгвістичною моделлю базується на комбінованому способі та складається з таких етапів (рис. 2).

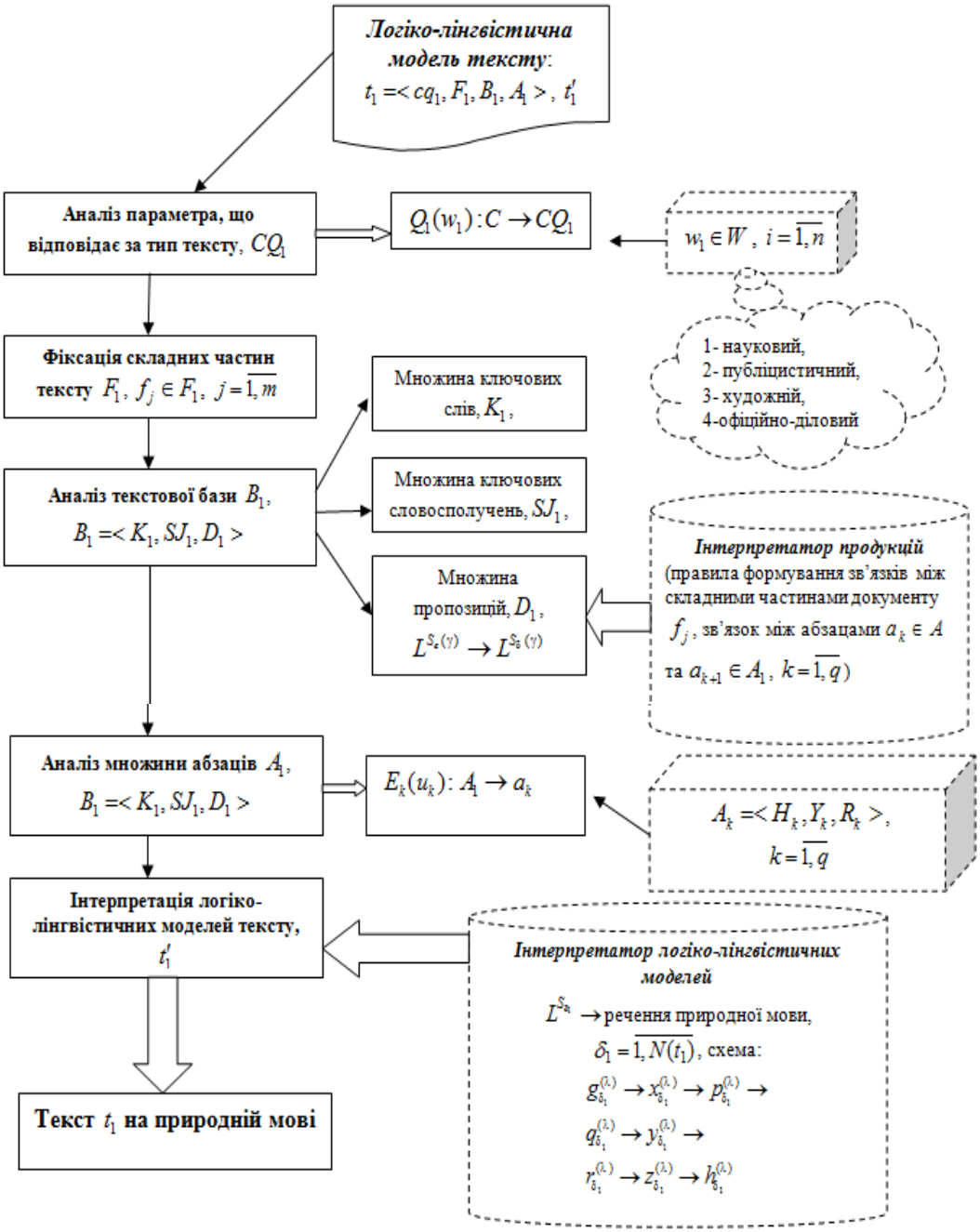

Рис. 2. Алгоритм відновлення текстової інформації із заданої формальної моделі

Проаналізувати перший параметр лінгвістичної складової моделі CQ_1=1. Це означає, що текстовий фрагмент наукового стилю, для якого характерні точність, однозначність слова, часта повторюваність ключових слів, стилістично нейтральні слова, загальнонаукові терміни; переважають

іменники, використовуються дієслова узагальненого значення, вживаються займенники третьої особи; часте використання дієприкметникових та дієприслівникових зворотів.

1. Зафіксувати кількість складних частин електронного текстового документу $f_j \in F_1$, $j = \overline{1,m}$, m — кількість складних синтаксичних частин. У цьому випадку $m = 1$, тобто присутня одна складна синтаксична частина, розглядається текстовий фрагмент.

2. Проаналізувати третій параметр лінгвістичної складової — текстову базу B_1, що є однією із значущих змінних, містить множину ключових слів, множину ключових словосполучень, а також множину пропозицій, які формують основу для вилучення знань з електронного текстового документа.

Таким чином, до множини ключових слів текстової інформації, що інтерпретована заданою формальною моделлю, належать слова $K_1 = \{$вирішення, проблема, прийняття, рішення, процедура, структуризація$\}$. Множина ключових словосполучень містить $SJ_1 = \{$вирішення проблеми, прийняття рішень$\}$. Також задані вектори характеристик: $l_1 = \{x_{22}, q_{22}, y'_2, q'_2\}$, $l_1 = \{$прийняття, рішень, набір, процедур$\}$; $l_2 = \{q_{11}, q_{12}, z_3\}$, $l_2 = \{$прийняття, рішень, процедури$\}$; $l_3 = \{q'_{12}, g_{22}\}$ $l_3 = \{$рішень, рішень$\}$. Враховуючи задані вектори характеристик, в результаті роботи інтерпретатора продукцій формується множина пропозицій, тобто відбувається пошук таких моделей $L^{S_e(\gamma)}$, які за змістом передують або з яких випливають моделі $L^{S_\delta(\gamma)}$.

$$L^{S_e(\gamma)} \to L^{S_\delta(\gamma)};$$

$$D_1 = \left\{ L^{S_1} \to L^{S_2}, L^{S_2} \to L^{S_3} \right\}.$$

1. Відновити словосполучення для кожного речення природної мови на основі логіко-лінгвістичних моделей із семантико-синтаксичної складової. Отримані словосполучення:

– для речення S_1 — «вирішення проблеми, вирішення неоднозначне, процес прийняття, прийняття рішень, потребує структуризації, дозволить визначити етапи, дозволить визначити процедури, етапи вирішення, процедури вирішення, спрямовані етапи, спрямовані процедури, вирішення проблеми»;

– для речення S_2 — «є набір, набір процедур, прийняття рішень, процесі управління, об'єднуються етапи, окремі етапи»;

– для речення S_3 — «можна побудувати схему, схему розробки, схему рішень, наукових рішень, обґрунтованих рішень».

2. Відновити прямий порядок слів у реченні природної мови шляхом об'єднання отриманих словосполучень, а також враховуючи елементи текстової бази з лінгвістичної складової (п. 3 алгоритму). Інтерпретатор перетворює модель $L^{S_{\delta_1}}, \delta_1 = \overline{1, N(t_1)}$ у речення природної мови шляхом синтезу простих предикатів за такою схемою:

$$g_{\delta_1}^{(\lambda)} \to x_{\delta_1}^{(\lambda)} \to p_{\delta_1}^{(\lambda)} \to q_{\delta_1}^{(\lambda)} \to y_{\delta_1}^{(\lambda)} \to r_{\delta_1}^{(\lambda)} \to z_{\delta_1}^{(\lambda)} \to h_{\delta_1}^{(\lambda)};$$

S_1 = «Якщо вирішення проблеми неоднозначне, то процес прийняття рішень потребує структуризації, що дозволить визначити етапи і процедури, спрямовані на вирішення проблеми».

S_2 = «Підготовка і прийняття рішень у процесі управління є набором процедур, вони об'єднуються в окремі етапи».

S_3 = «Тому можна побудувати загальну схему розробки наукових, обґрунтованих рішень».

1. Проаналізувати множину абзаців $A_1 = \{1\}$ та перевірити відновлені у п. 5 логічні зв'язки між реченнями, зокрема, тип зв'язку між реченнями $H_k = 1$, що означає ланцюжковий зв'язок; $Y_k = \{1\}$, тобто в абзаці використано просту лінійну прогресію, що характеризується послідовним розгортанням інформації, коли рема попереднього речення стає темою наступного речення [13]; $R_k = \{1\}$, тобто використана предметна рематична домінанта.

2. Відновити текст з речень природної мови, отриманих у п. 5 алгоритму з урахуванням параметрів лінгвістичної складової.

Після застосування описаного алгоритму проведено аналіз відтвореного тексту та здійснено його порівняння з еталонним, що показало повне відновлення текстового фрагменту за змістом і 88 % відновлення слів.

Часову складність виконання алгоритму можна оцінити, як $O(8\nu N(t))$, де ν — кількість атомарних предикатів у логіко-лінгвістичній моделі, $N(t)$ — загальна кількість речень у текстовому фрагменті. Таким чином, часова складність алгоритму відновлення текстової інформації із заданої логіко-лінгвістичної моделі прямо пропорційна кількості компонент логіко-лінгвістичної моделі речення природної мови, кількості простих речень у складному та кількості речень у тексті.

Висновки

Запропонований алгоритм дозволяє автоматично відновлювати текстову інформацію на основі комбінованого способу аналізу лінгвістичної та семантико-синтаксичної складових логіко-лінгвістичних моделей електронних текстових документів. Формальний апарат трансформації тексту у логіко-лінгвістичну модель і навпаки виступає єдиним засобом автоматизації процесу обробки текстової інформації [13].

Проведено дослідження шляхом застосування алгоритму аналізу логіко-лінгвістичних моделей електронних текстових документів до уже побудованих моделей. На вхід системи подавалися масиви логіко-лінгвістичних моделей типу речень природної мови, в результаті чого отримувався текст. Результати проведених досліджень наведено у таблиці.

Результати застосування алгоритму аналізу логіко-лінгвістичних моделей для відновлення електронних текстових документів

Параметр /Тип тексту	Кількість слів	Відсоток відновлення, %	Особливості граматичних форм	Тип абстрактної моделі
Тези 1	868	95	Прийменники, сполучники	1 та 2
Тези 2	807	75	Синоніми, синтаксична деривація	1, 2 та 3
Диплом бакалавра	10809	70	Службові слова, конверсиви	1—5
Стаття 1	1917	89	Омоніми	1—5
Стаття 2	5095	96	Прийменники	1, 3, 4
Диплом магістра	20655	74	Службові слова, конверсиви	1—5
Дисертаційна робота	60081	70	Синоніми, омоніми, конверсиви	1—5
Стаття 3	2531	92	Прийменники	1 та 2
Стаття 4	1854	86	Прийменники	1—3
Тези 3	1022	95	Прийменники	1
Тези 4	853	91	Прийменники	1
Диплом бакалавра 1	11862	90	Прийменники, сполучники	1 та 2
Диплом магістра 1	21754	85	Службові слова, конверсиви	1—5

На основі проведених досліджень зроблено висновки про те, що 63 % текстів відновлено з точністю 90—100 %, 22 % текстів — з точністю 80—89 %, що пов’язано зі зняттям омонімії, вживанням у природних мовах різних видів службових слів у тих самих граматичних формах за їх різного змістовного навантаження, та 15 % текстів відновлено з точністю 70—79 %, що пов’язане з відновленням синонімічних конструкцій та інваріантних форм логіко-лінгвістичних моделей.

Алгоритм автоматичного аналізу логіко-лінгвістичних моделей електронних текстових документів призначено для відтворення текстової інформації, що дасть змогу оцінювати точність побудови формальних логічних моделей текстів для їх порівняння за змістом.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Кабінет Міністрів України, *Постанова* № 556. **[**Електронний ресурс]. Режим доступу: https://zakon.rada.gov.ua/laws /show/380-2021-%D0%BF#n10 .

[2] Проєкт «Ініціатива академічної доброчесності та якості освіти». [Електронний ресурс]. Режим доступу: https://saiup.org.ua/pro-proekt/.

[3] «Академічна доброчесність: виклики сучасності,» в *Збірник наукових есе учасників дистанційного етапу наукового стажування для освітян* (Республіка Польща, Варшава, 28.01 – 08.02.2019). Варшава, 2019. 171 с.

[4] *Understanding & Preventing Plagiarism* [Online]. Available: https://www.accreditedschools online.org/resources/preventing-plagiarism/. Accessed on: April 15, 2021.

[5] Nur E. Hafsa, “Plagiarism: A Global Phenomenon,” in *Journal of Education and Practice*, vol. 12, no.3, pp. 53-59, 2021.

[6] A. Vavilenkova “Regularity of context units identification in electronic text documents,” *CEUR Workshop Proceedings*, 2845, pp. 1-10, 2021.

[7] J. Kahre, *The mathematical theory of information*, New York: Springer Science, 2002, 502 p.

[8] М. М. Глибовець, А. М. Глибовець, і М. В. Поляков, *Інтелектуальні мережі*, Дніпропетровськ: Нова ідеологія, 2014, 462 с.

[8] V. Shyrokov, "System semantics of explanatory dictionaries," *Cognitive Studies*, no12, pp. 95-106, 2015.

[9] N. Khairova, S. Petrasova, and A. P. S. Gautam, "The Logical-Linguistic Model of Fact Extraction from English Texts," *Communications in Computer and Information Science*, vol. 639. Springer, Cham. 2016. https://doi.org/10.1007/978-3-319-46254-7_51 .

[10] О. Г. Оксіюк, «Моделі подання знань в інтелектуальних системах навчання,» *Збірник наукових праць Військового Інституту Київського національного університету імені Тараса Шевченка*, т. 28, с. 98-101, 2010.

[11] Д. В. Ланде, І. Ю. Субач, і Ю. Є. Бояринова, *Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки*: навч. посіб., К.: ІСЗЗІ КПІ ім. Ігоря Сікорського, 2018, 300 с.

[12] И. Р. Гальперин, *Текст как объект лингвистического исследования*. 5-е изд. Москва: КомКнига, 2007, 144 с.

[13] А. І. Вавіленкова, *Аналіз і синтез логіко-лінгвістичних моделей речень природної мови*, монографія, К.: ТОВ «СІК ГРУП УКРАЇНА», 2017, 152 с.

[14] A. Vavilenkova, "Basic principles of the synthesis of logical-linguistic models," *Cybernetics and systems analysis*, vol. 51(5), pp. 826-834, 2015. http:// doi.org/10.1007/s10559-015-9776-z .

[15] К. А. Филлипов, *Лингвистика текста.* Курс лекций, Спб.: Издательство С.-Петербургского университета, 2008, 336 с.

[16] O. V. Bisikalo, W. Wojcik, O. V. Yahimovich, and S. Smailova, "Method of determining of keywords in English texts based on the DKPro Core," *Technology Audit and Production Reserves*, 1/2(21), pp. 26-30, 2015. https://doi.org/10.15587/2312-8372.2015.37274 .

[17] А. І. Вавіленкова, «Особливості бази знань системи автоматизованої побудови логіко-лінгвістичних моделей текстових документів,» *Вісник Національного університету «Львівська політехніка». Серія «Інформаційні систем та мережі»*: зб. наук. праць, № 9, с. 75-83, 2021. https://doi.org/10.23939/sisn2021.09.075 .

Рекомендована кафедрою автоматизації та інтелектуальних інформаційних технологій ВНТУ

Стаття надійшла до редакції 25.07.2021

Вавіленкова Анастасія Ігорівна — д-р техн. наук, доцент, професор кафедри комп'ютеризованих систем управління, e-mail: vavilenkovaa@gmail.com .

Національний авіаційний університет, Київ

A. I. Vavilenkova[1]

Ways of Recovery of Text Information Represented in the Form of a Logic and Linguistic Model

[1]National Aviation University, Kyiv

The materials of the article substantiate the urgency of solving the problem of identifying meaningful links in electronic text documents in order to further compare their content and improve the operation of plagiarism detection systems. An important step is to assess the reliability of the formed formal models. Therefore, the aim of this article is to study the algorithm of automatic analysis of logic and linguistic models of electronic text documents for the reproduction of textual information, which combines the basic properties of the text and its components. The logic and linguistic model of a text document reflects the main relationships between structural components; it is an ordered quadruple and an array of logic and linguistic models of sentences of natural language, which are included in the text. The author proposes several ways of restoration of textual information, starting from the structure of the logic and linguistic model of an electronic text document, which contains a linguistic and semantic-syntactic component. The article describes the schemes of text information recovery, it chooses the combined method, which provides the analysis of semantic-syntactic component in parallel with the analysis of the text base, in particular, its components - a set of sentences containing connections between logic and linguistic models of sentences of the text within electronic text document. It has been developed an algorithm for recovery of textual information presented in the form of a formal logic and linguistic model of an electronic text document, and there have been described the stages of abovementioned algorithm. All steps of the algorithm are demonstrated on the example of analysis of a specific given logic and linguistic model of a fragment of an electronic text document. The author conducted experiments on the restoration of textual information for scientific style texts. During experiments, it was revealed, that the main factors influencing the restoration of textual information include the removal of homonymy, as well as different interpretations of synonymous constructions and invariant forms of logic and linguistic models of sentences of natural language.

Keywords: textual information, natural language, logic and linguistic model, recovery algorithm.

Vavilenkova Anastasiia I. — Dr. Sc. (Eng.), Associate Professor, Professor of the Chair of Computerized Control Systems, e-mail: vavilenkovaa@gmail.com

А. И. Вавиленкова[1]

Пути восстановления текстовой информации, представленной в виде логико-лингвистической модели

[1]Национальный авиационный университет, Киев

Обоснована актуальность решения проблемы поиска смысловых связей в электронных текстовых документах с целью дальнейшего их сравнения по смыслу и усовершенствования работы систем обнаружения плагиата. При этом важным этапом является оценивание достоверности построенных формальных моделей. Поэтому целью данной статьи есть исследование алгоритма автоматического анализа логико-лингвистических моделей электронных текстовых документов для воспроизведения текстовой информации, которая объединяет в себе основные свойства текста и его составляющих частей, отображает основные связи между структурными компонентами. Логико-лингвистическая модель текстового документа представляет собой упорядоченную четверку и массив логико-лингвистических моделей предложений естественного языка, которые входят в текст. Автором предложено несколько путей восстановления текстовой информации, которые отталкиваются от структуры логико-лингвистической модели электронного текстового документа, которая содержит лингвистическую и семантико-синтаксическую составляющие. Описаны схемы осуществления восстановления текстовой информации, выбран комбинированные способ, который предполагает анализ семантико-синтаксической составляющей параллельно с анализом текстовой базы, в частности, ее компонента — множества предположений, которое содержит связи между логико-лингвистическими моделями предложений текста электронного текстового документа. Разработан алгоритм восстановления текстовой информации, представленной в виде формальной логико-лингвистической модели электронного текстового документа, описаны его этапы. Все шаги алгоритма продемонстрированы на примере анализа конкретной заданной логико-лингвистической модели фрагмента электронного текстового документа. Проведены эксперименты по восстановлению текстовой информации для текстов научного стиля. Выявлено, что к основным факторам, которые влияют на восстановление текстовой информации, относится снятие омонимии, а также различная интерпретация синонимических конструкций и инвариантных форм логико-лингвистических моделей предложений естественного языка.

Ключевые слова: текстовая информация, естественный язык, логико-лингвистическая модель, алгоритм восстановления.

Вавиленкова Анастасия Игоревна — д-р техн. наук, доцент, профессор кафедры компьютеризированных систем управления, e-mail: vavilenkovaa@gmail.com