УДК 510.635:004.891 (045)

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

ПРАВИЛА СИНТЕЗУ ЛОГІКО-ЛІНГВІСТИЧНИХ МОДЕЛЕЙ РЕЧЕНЬ ПРИРОДНОЇ МОВИ

Постановка проблеми та її зв'язок з важливими науковими завданнями

Боротьба з плагіатом наукових робіт та порушення авторських прав вимагають сьогодні наявності алгоритмів змістовного аналізу ефектронних текстових документів. Це, в свою чергу, неможливо без глибинного лінгвістичного аналізу текстів, написаних природною мовою. Необхідним є пошук нових принципів, у рамках яких було б можливим проектування якісно нових систем обробки великих та динамічних масивів даних [1, с. 10].

Задачі змістовної автоматичної обробки текстової інформації, пов'язані з проблемами реєстрації інформації, захистом інтелектуальної власності, інноваційною діяльністю, відповідністю документації юридичних та інформаційно-аналітичних організацій певним встановленим нормам.

Основною проблемою на шляху змістовної автоматичної обробки текстової інформації є вирішення протиріччя між існуванням підходів та алгоритмів виявлення текстових збігів на основі логічних моделей логіки предикатів та відсутністю механізмів автоматичної побудови таких моделей.

Такий формальний апарат дозволив би аналізувати текстову інформацію за єдиним принципом, систематизувати процес змістовного пошуку та уникнути неоднозначності при аналізі текстових документів.

Описана проблема підпадає під пріоритетний тематичний напрям фундаментальних досліджень і науково-технічних розробок на період до 2020 року «Технології та засоби розробки програмних продуктів і систем» постанови Кабінету Міністрів України № 556 від 23.08.2016 р.

Автором статті пропонується вирішувати цю проблему шляхом застосування логіко-лінгвістичного моделювання.

Аналіз останніх досліджень і публікацій

Проблематикою змістовного аналізу текстів займаються у наш час, в першу чергу, комп'ютерні лінгвісти. Так, у праці [2, с. 31] В. А. Широков описує концепцію станів мовних одиниць, що лежить в основі створення та функціонування лексикографічних систем. Визначенню лексичних компонент та спробам моделювання змісту речень природної мови присвячено праця Вів'єна Еванса та Мелані Грін «Когнітивна лінгвістика» [3, с. 106]. Хорошу теоретичну базу для досліджень структури тексту та логічних зв'язків у ньому дає Р. І. Гальперін у своїй монографії «Текст як об'єкт лінгвістичного дослідження» [4, с. 73].

Автори книги «Інтелектуальні інформаційні технології» розглядають моделювання знань як основу функціонування інтелектуальних автоматизованих систем [5, с. 135].

Проблемам аналізу тексту та виявленню у ньому змістовних зв'язків присвячено праці Дж. Лайонза [6, с. 147], М. В. Нікітіна [7, с. 345], Ю. Д. Апресяна [8, с. 316].

Незважаючи на таку велику теоретичну базу розробок у сфері аналізу тексту, автоматичного засобу для здійснення коректного лінгвістичного аналізу,а також для вилучення змісту електронних текстових документів досі не існує.

Мета статті

Будь-який електронний текстовий документ не залежно від своєї предметної області повинен мати організаційну структуру, притаманну довільному тексту. Тобто для формування змісту тексту та його безпосереднього існування необхідна наявність певного набору якісних характеристик та ознак: когезія, когерентність, адресованість, інформативність, ситуативність, типологічна інтертекстуальність [9, с. 20]. Сучасні системи обробки даних не використовують в своїй роботі механізмів автоматичного лінгвістичного

аналізу і, відповідно, не виявляються перераховані вище ознаки текстових документів.

Тому метою статті є створення формальних правил, за якими можна буде знайти логічні зв'язки між реченнями природної мови, з яких складається електронний текстовий документ.

Виклад основного матеріалу досліджень

Текст — це завершена з точки зору його автора, проте відкрита для множини інтерпретацій в змістовному та інтенсіональному плані, лінійна послідовність мовних знаків, виражених графічно, семантико-змістовна взаємодія яких створює композиційну єдність, підтримуючи лексико-граматичні відношення між окремими елементами створеної таким чином структури [9, с. 19].

У граматичному аспекті зв'язність тексту визначається законами узгодження, правилами побудови висловлювань з використанням морфологічних та синтаксичних засобів мови. Тому у даній статті електронний текстовий документ розглядається як множина взаємопов'язаних логіко-лінгвістичних моделей речень природної мови, що входять до складу тексту, тобто його семантико-синтаксична складова [10, с. 103]. Таки чином, об'єднання та заміна структурних компонентів логіко-лінгвістичних моделей на основі виявлення способів логічного зв'язку буде представляти собою синтез логіко-лінгвістичних моделей речень природної мови [11, с. 177].

Правила синтезу логіко-лінгвістичних моделей являють собою умови, при виконанні яких у логіко-лінгвістичних моделях здійснюються заміни тотожних за змістом компонент, а також утворюються одновимірні масиви, елементи яких є словами чи словосполученнями, що пов'язують речення природної мови одне з одним у частинах тексту.

Правило 1. Якщо у тексті є послідовність речень природної мови $S_1,...,S_i,...,S_k$, кожному з яких поставлена у відповідність логіко-лінгвістична модель $L^{S_1},...,L^{S_i},...,L^{S_k}$, у яких суб'єкти $x_1,...,x_i,...,x_k$ тотожні, виражені такими частинами мови, як іменники, займенники та чисельники або є синонімами, то необхідно здійснити заміну суб'єктів: $x_1,...,x_i \equiv x_1,...,x_k \equiv x_1$ та сформувати одновимірний масив зв'язків: $l_i = \{x_1,...,x_{i-1},x_{i+1},...,x_k\}$.

Наприклад, нехай задано такий фрагмент тексту: «*Комп'ютерне моделювання входить до переліку основних дисциплін. Воно вивчає основні методи та алгоритми моделювання за допомогою комп'ютерних засобів. Комп'ютерне моделювання також передбачає створення імітаційних моделей для подальшого тестування*».

Логіко-лінгвістичні моделі для речень цього фрагменту матимуть вигляд:

L^{S_1} = *входить(моделювання, комп'ютерне, переліку, 0, дисциплін, основних, 0).*

L^{S_2} = *вивчає(моделювання, комп'ютерне, методи, основні, моделювання, 0, комп'ютерних&засобів)&вивчає (моделювання, комп'ютерне, алгоритми, основні, моделювання, 0, комп'ютерних&засобів).*

L^{S_3} = *передбачає (моделювання, комп'ютерне, створення, 0, моделей, імітаційних, подальшого&тестування).*

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

$L^{S_2} = p_2(x_1, g_1, y_{21}, q_2, x_1, 0, h_2)$

$\& p_2(x_1, g_1, y_{22}, q_2, x_1, 0, h_2)$.

$L^{S_3} = p_3(x_1, g_1, y_3, 0, x_1, r_3, h_3)$.

Таким чином, одновимірний масив характеристик для заданого фрагменту буде складатися з суб'єктів речень: $l_i = \{x_1, x_2, x_3\}$.

Правило 2. Якщо у тексті є послідовність речень природної мови $S_1,...,S_i,...,S_k$, кожному з яких поставлена у відповідність логіко-лінгвістична модель $L^{S_1},...,L^{S_i},...,L^{S_k}$, у яких об'єкт або предмет відношення кожного попереднього речення тотожний, виражений займенником або є синонімом по відношенню до суб'єкта наступного речення, то необхідно здійснити заміну:

$y_1,...,x_i \equiv y_{i-1},...,x_k \equiv y_{k-1}$ або

$y_1,...,y_i \equiv z_{i-1},...,y_k \equiv z_{k-1}$

та сформувати одновимірний масив зв'язків:

$l_i = \{y_1,...,x_i,x_{i+1},...,x_k\}$ або

$l_i = \{y_1,...,z_i,z_{i+1},...,z_k\}$ відповідно.

Взаємозв'язки між об'єктами та предметами відношень попередніх речень з суб'єктами наступних можуть чергуватися.

Нехай задано текст: «*На цукрових заводах України для кристалізації утфелів використовують вакуум-апарати періодичної дії. Робота групи вакуумних апаратів організується таким чином, щоб забезпечити безперервну переробку сиропів і ефективне використання пари, яка є для вакуум-апаратів основним тепловим носієм*» [12, с. 42].

Логіко-лінгвістичні моделі для речень цього фрагменту матимуть вигляд:

L^{S_1} = *використовують (0, 0, вакуум-апарати, 0, дії, періодичної, цукрових& заводах)&*

використовують (0, 0, вакуум-апарати, 0, дії, періодичної, заводах&України)&

використовують (0, 0, вакуум-апарати, 0, дії, періодичної, кристалізації&утфелів).

L^{S_2} *= організується (робота, 0, групи, 0, апаратів, вакуумних,0)→*

(забезпечити (робота, 0, переробку, безперервну, сиропів, 0, 0)&

забезпечити (робота, 0, використання, ефективне, пари, 0, 0)&

є(пара, 0, носієм, основним,вакуум-апаратів, 0, 0) &

є(пара, 0, носієм, тепловим, вакуум-апаратів, 0, 0))).

$$L^{S_1} = p_1(0,0,\ y_1,0,z_1,r_1,h_1)\ \&$$
$$\&\, p_1(0,0,\ y_1,0,z_1,r_1,h_2)\ \&$$
$$\&\, p_1(0,0,\ y_1,0,z_1,r_1,h_3);$$
$$L^{S_2} = p_2(x_2,0,\ y_2,0,y_1,0,h_2) \rightarrow$$
$$\rightarrow (p_2'(x_2,0,\ y_{21}',q_{21}',y_1,0,0)\ \&$$
$$\&\, p_2'(x_2,0,\ y_{22}',q_{22}',z_{22}',0,0)\&;$$
$$(p_2''(x_2'',0,y_2'',q_{21}'',y_1,0,0)$$
$$\&\, p_2''(x_2'',0,y_2'',q_{22}'',y_1,0,0))).$$

У цьому випадку масив характеристик для заданого фрагменту буде містити предмети відношень другого речення фрагменту і матиме вигляд: $l_i = \{y_1, z_2, z_{21}', z_2''\}$.

Правило 3. Якщо у тексті є послідовність речень природної мови $S_1,...,S_i,...,S_k$, кожному з яких поставлена у відповідність логіко-лінгвістична модель $L^{S_1},...,L^{S_i},...,L^{S_k}$, у яких суб'єкт або об'єкт першого речення поступово виступає характеристиками суб'єктів, об'єктів та предметів відношень всіх наступних речень, то необхідно здійснити заміну: $x_1,...,g_i \equiv x_1,...,r_k \equiv x_1$ або $y_1,...,g_i \equiv y_1,...,r_k \equiv y_1$ та сформувати одновимірний масив зв'язків: $l_i = \{x_1,...,g_i,q_{i+1},...,r_k\}$ або $l_i = \{y_1,...,g_i,q_{i+1},...,r_k\}$ відповідно. Характеристики суб'єктів, об'єктів та предметів відношень із наступних речень можуть чергуватися.

Нехай задано такий фрагмент: «*Ситуаційне управління дає можливість повної, якісної формалізації моделі об'єкта управління із застосуванням мови ситуаційного управління, або інших методів, що використовуються для цих цілей. Можливість і потреба застосування принципу ситуаційного управління виходять із ряду особливостей ситуаційного управління*» [12, с. 43].

Логіко-лінгвістичні моделі для речень цього фрагменту матимуть вигляд:

L^{S_1} *= дає&можливість(управління, ситуаційне, формалізації, повної, моделі, 0, 0)& дає&можливість (управління, ситуаційне, формалізації, якісної, моделі, 0, 0)&*

(модель (об'єкта, управління, застосуванням, 0, мови, ситуаційного&управління,0) $\vee$

модель (об'єкта, управління, застосуванням, 0, методів, інших,0)&

використовуються (методи, інші, цілей, цих, 0,0,0)).

L^{S_2} *= виходять (можливість, 0, застосування, 0, принципу, ситуаційного&управління,0)&*

виходять (потреба, 0, застосування, 0, принципу, ситуаційного&управління,0) &

виходять (можливість, 0, ряду,0, особливостей, ситуаційного&управління,0)&

виходять (потреба, 0, ряду, 0, особливостей, ситуаційного&управління,0).

$$L^{S_1} = p_1(x_1,g_1,\ y_1,q_{11},z_1,0,0)\ \&$$
$$\&\, p_1(x_1,g_1,\ y_1,q_{12},z_1,0,0)\&;$$
$$(z_1(x_1',x_1,\ y_1',0,z_{11}',x_1,0) \vee$$
$$\vee z_1(x_1',x_1,\ y_1',0,z_{12}',r_{12}',0)\ \&$$
$$\&\, p_1''(z_{12}',r_{12}',y_1'',q_1'',0,0,0)).$$
$$L^{S_2} = p_2(x_{21},0,\ y_{21},0,z_{21},x_1,0)\ \&$$
$$\&\, p_2(x_{22},0,\ y_{21},0,z_{21},x_1,0)\&;$$
$$p_2(x_{21},0,\ y_{22},0,z_{22},x_1,0)\ \&$$
$$\&\, p_2(x_{22},0,\ y_{22},0,z_{22},x_1,0).$$

Масив характеристик для заданого фрагменту буде містити характеристики об'єктів та предметів відношень другого речення фрагменту і матиме вигляд: $l_i = \{x_1, r_{11}', r_{21}\}$.

Отже, з прикладів видно, що правила синтезу логіко-лінгвістичних моделей речень природної мови допомагають простежити спосіб формування логічних зв'язків у текстових фрагментах.

Висновки

Правила синтезу логіко-лінгвістичних моделей речень природної мови можна застосовувати як для речень, розташованих у тексті послідовно, так і для речень з декількох абзаців. Розмірність одномірного масиву характеристик, утвореного внаслідок синтезу, відповідає кількості речень у тексті, пов'язаних між собою за змістом. Порівняння масивів характеристик кожного з речень одного з текстів з аналогічними масивами характеристик іншого тексту дає змогу знайти змістовний збіг. При порівнянні векторів характеристик важливим аспектом є правила, за якими їх було сформовано, тобто принципи утворення масивів.

Таким чином, якщо елементами одновимірних масивів характеристик будуть синоніми або елементи інваріантних форм, але правила здійснення синтезу логіко-лінгвістичних моделей речень однакові, то відсоток рерайту при цьому буде значно вищим, ніж якщо правила не будуть співпадати. Якщо дана стаття буде перевірена сервісами пошуку плагіату, то відсоток унікальності буде коливатися в межах 90–95 % і головним джерелом запозичення буде [12, с. 43], проте зміст статті абсолютно не пов'язаний з темою ситуаційного управління, а лише використовує матеріали джерела для прикладів. При синтезі логіко-лінгвістичних моделей речень матеріалів статті запозичений текст не буде синтезуватися з іншими реченнями, що входять до аналізу джерел, актуальності, цілі статті та ін. Таким чином, відсутність логічного зв'язку за правилами синтезу буде свідчити про різний зміст текстів. У цьому випадку низький відсоток рерайту буде вказувати на коректну роботу сервісу визначення унікальності, так як у статті є посилання на джерело [12, с. 43], що уже свідчить про відсутність плагіату.

Розширення та удосконалення правил синтезу для різноманітних методів логічного зв'язку між частинами тексту забезпечить підвищення якості пошуку змісту з використанням формальних логіко-лінгвістичних моделей.

ЛІТЕРАТУРА

1. **Ландэ Д. В.** Интернетика: навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Снарский, И. В. Безсуднов. — М. : Либроком, 2009. — 264 с.

2. **Широков В. А.** Лінгвістичні та технологічні основи тлумачної лексикографії / В. А. Широков, В. М. Білоноженко, О. В. Бугаков та ін. — К. : Довіра, 2010. — 295 с.

3. **Evans V., Green M.** Cognitive Linguistics. — Edinburg: Edinburg university press Publ., 2006, — 830 p.

4. **Гальперин И. Р.** Текст как объект лингвистического исследования. Изд. 5-тое, стереотипное / И. Р. Гальперин. — М. : КомКнига, 2007. — 144 с.

5. **Башмаков А. И.** Интеллектуальные информационные технологии: учеб. пособие / А. И. Башмаков, И. А. Башмаков. — М. : Изд-во МГТУ им. Баумана, 2005. — 304 с.

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

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

8. **Апресян Ю. Д.** Лексическая семантика: в 2-х т: Т. 1. / Ю. Д. Апресян. — М. : «Восточная литература», 1995. — 422 с.

9. **Чернявская З. Е.** Лингвистика текста: поликодовость, интертекстуальность, интердискурсивность / З. Е. Чернявская. — М. : Книжный дом «Либроком», 2009. — 248 с.

10. **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.

11. **Вавиленкова А. И.** Основные принципы синтеза логико-лингвистических моделей / А. И. Вавиленкова // Кибернетика и системный анализ. — 2015. — Т. 51, № 5. — С. 176–185.

12. **Прокопенко Ю. В.** Застосування ситуаційного підходу для формування алгоритмів управління вакуум-апаратом періодичної дії / Ю. В. Прокопенко, А. П. Ладанюк // Східно-європейський журнал передових технологій. — 2015. — № 3/2 (75). — С. 42–47. doi: 10.15587/1729-4061.2015.43758.

Вавіленкова Анастасія Ігорівна

ПРАВИЛА СИНТЕЗУ ЛОГІКО-ЛІНГВІСТИЧНИХ МОДЕЛЕЙ РЕЧЕНЬ ПРИРОДНОЇ МОВИ

У статті сформульовано основну проблему змістовної автоматичної обробки текстової інформації. Описано формальні правила, за якими можна буде знайти логічні зв'язки між реченнями природної мови, з яких складається електронний текстовий документ. Правила синтезу логіко-лінгвістичних моделей являють собою умови, при виконанні яких у логіко-лінгвістичних моделях здійснюються заміни тотожних за змістом компонент, а також утворюються одновимірні масиви, елементи яких є словами чи словосполученнями, що пов'язують речення природної мови одне з одним у частинах тексту. Правила синтезу логіко-лінгвістичних моделей речень природної мови можна застосовувати як для речень, розташованих у тексті послідовно, так і для речень з декількох абзаців. Розширення та удосконалення правил синтезу для різноманітних методів логічного зв'язку між частинами тексту забезпечить підвищення якості пошуку змісту з використанням формальних логіко-лінгвістичних моделей.

***Ключові слова*:** логіко-лінгвістична модель; синтез; природна мова; логічні зв'язки; зміст.

Vavilenkova Anastasia Igorivna

PRINCIPLES OF SYNTHESIS OF LOGIC AND LINGUISTIC MODELS OF NATURAL LANGUAGE SENTENCES

The paper formulates the basic problem of the content automatic processing of textual information. An author describe formal rules by which we can found logical connections between natural language sentences that compose the electronic text document. Principles of synthesis of logic and linguistic models (LLM) are the conditions under which, within LLM, we can replace components identical by content and form one-dimensional arrays, elements of which are words or phrases that connect natural language sentences to each other in text's parts. The above principles can be applied to text's sentences, which are arranged sequential, as well as for sentences from several paragraphs. Expansion and improvement of principles of synthesis for various methods of logical connection between the parts of the text will improve the search content quality using formal logic and linguistic models.

Key words: logical and linguistic model; synthesis; natural language; logical connections;content.

Вавиленкова Анастасия Игоревна

ПРАВИЛА СИНТЕЗА ЛОГИКО-ЛИНГВИСТИЧЕСКИХ МОДЕЛЕЙ ПРЕДЛОЖЕНИЙ ЕСТЕСТВЕННОГО ЯЗЫКА

В статье сформулирована основная проблема содержательной автоматической обработки текстовой информации. Описаны формальные правила, по которым можно будет найти логическую связь между предложениями естественного языка, из которых состоит электронный текстовый документ. Правила синтеза логико-лингвистических моделей представляют собой условия, при выполнении которых в логико-лингвистических моделях осуществляется замена тождественных по содержанию компонент, а также создаются одномерные массивы, элементы которых являются словами или словосочетаниями, которые, в свою очередь, связывают предложения естественного языка в частях текста. Правила синтеза логико-лингвистических моделей предложений естественного языка можно применять как для предложений, стоящих последовательно один за другим, так и для предложений нескольких абзацев. Расширение и усовершенствование правил синтеза для разнообразных методов логической связи между частями текста обеспечим повышение качества поиска содержания с использование логико-лингвистических моделей.

Ключевые слова: логико-лингвистическая модель; синтез; естественный язык; логические связи; содержание.

Стаття надійшла до редакції 22.02.2017 р.
Прийнято до друку 24.02.2017 р.
Рецензент – д-р техн. наук, проф. Ю. К. Зіатдінов