Issue 9, 2021

https://doi.org/10.23939/sisn2021.09.075

УДК 510.635:004.891(045)

ОСОБЛИВОСТІ БАЗИ ЗНАНЬ СИСТЕМИ АВТОМАТИЗОВАНОЇ ПОБУДОВИ ЛОГІКО-ЛІНГВІСТИЧНИХ МОДЕЛЕЙ ТЕКСТОВИХ ДОКУМЕНТІВ

Анастасія Вавіленкова

Національний авіаційний університет
vavilenkovaa@gmail.com, ORDIC 0000-0002-9630-4951

Окреслено проблему пошуку змістовних одиниць у електронних текстових документах та проаналізовано основні недоліки відомих підходів до видобування знань із текстової інформації. Досліджено особливості побудови логіко-лінгвістичних моделей електронних текстових документів, зокрема описано та досліджено особливості баз знань системи автоматизованої побудови логіко-лінгвістичних моделей україномовних текстових документів. Запропоновано схему формалізації текстової інформації на основі побудови логіко-лінгвістичної моделі електронного текстового документа. У ній першим етапом є формування логіко-лінгвістичних моделей речень природної мови. Для цього використано спеціально розроблений метод автоматизованого формування логіко-лінгвістичних моделей, що ґрунтується на здійсненні синтаксичного аналізу речень природної мови, використанні бази даних у вигляді тезаурусу слів природної мови та бази правил для виявлення логічних зв'язків. Це уможливилось завдяки базі знань 1, яку розробила автор. Ця база використовується для визначення ролі кожного зі слів електронного текстового документа та є продукційною моделлю із формалізованими правилами української мови для формування словосполучень, які можуть утворювати між собою члени речення природної мови. Базу знань 2 створено для пошуку зв'язків між реченнями, що входять до складу електронного текстового документа, вона є сукупністю продукцій, які відображають принципи синтезу логіко-лінгвістичних моделей речень природної мови, тобто правила об'єднання та заміни структурних компонентів логіко-лінгвістичних моделей – речень природної мови. База знань 3, використана для побудови лінгвістичної складової логіко-лінгвістичної моделі текстового документа, є множиною продукцій, що містить правила формування мереж переходів для інтерпретації тематичної прогресії тексту. На конкретних текстових фрагментах продемонстровано застосування розроблених формалізованих правил. Механізм використання запропонованих баз знань дає змогу простежити процес формування логіко-лінгвістичних моделей електронних текстових документів.

Ключові слова: змістовні одиниці; природна мова; електронний текстовий документ; логіко-лінгвістична модель; база знань; продукційна модель.

Постановка проблеми

Вирішення проблеми видобування знань з електронних текстових документів досі залишається одним із пріоритетних завдань у сфері штучного інтелекту, інформаційних технологій, комп'ютерної лінгвістики та інформаційного пошуку, незважаючи на величезну кількість теоретичних припущень

та масу створених програмних продуктів, що так чи інакше здійснюють пошук ключових слів та концептів у текстовій інформації.

Проте пошук змістовних одиниць у електронних текстових документах неможливо реалізувати лише на основі виявлення певних статистичних закономірностей, зокрема, аналізуванням сусідніх слів речення, ігноруванням знаків пунктуації та невизначенням частин мови, якими є слова речень природної мови. Зміст тексту ретельно формується за правилами об'єднання текстових фрагментів за допомогою сюжетної лінії, для цього використовують семантичні зв'язки та різноманітні лінгвістичні засоби (тематичні та рематичні засоби), за правилами поєднання окремих речень в абзаці (засоби когезії) засобами зв'язку простих речень у структурі складного (сурядний та підрядний зв'язок), а також за правилами поєднання слів речень природної мови на основі їхніх граматичних характеристик [1].

Всі ці особливості побудови електронних текстових документів дає змогу врахувати формальна модель подання знань електронного текстового документа – логіко-лінгвістична модель [2]. Особливості побудови цієї моделі розглянуто у статті.

Аналіз останніх досліджень та публікацій

Сьогодні для пошуку ключових фраз у текстових фрагментах користуються готовими програмними рішеннями, що містять лінгвістичну компоненту роботи з природномовними текстами та ґрунтуються на синтаксичному парсингу [3–5], в основу якого покладено теорію контекстних граматик Хомського. Однак зазвичай не зазначають критерії оцінювання якості виявлених зв'язків, через що можуть виникати похибки під час подальшого глибинного аналізу текстової інформації.

Останніми роками все більше уваги приділяють тематиці Natural Language Processing. Безліч конференцій відбуваються і в Україні, і за кордоном, тематика їхніх секцій розширяється на все більшу кількість предметних областей [6, 7], що свідчить про актуальність використання результатів опрацювання текстів, написаних природною мовою. Наприклад, у роботах Д. В. Ланде розглянуто застосування методів порівняльного аналізу текстової інформації у правовій сфері [8], Ф. Самем та Д. Ландегем описують у своїх працях розпізнавання текстової інформації у фінансових електронних документах [9], застосовування NLP дуже поширене сьогодні й у медицині [10].

Одним із ефективних способів формалізації текстової інформації є логіко-лінгвістичне моделювання, яке застосовують для інтерпретації зв'язків між концептами речень природної мови за допомогою логіки предикатів першого порядку [11, 12].

Формулювання цілі статті

Розроблення та використання окремих форм логіко-лінгвістичних моделей [13] дало змогу формалізувати процес видобування інформації з речень природної мови. Для цього використовують спеціально розроблений метод автоматизованого формування логіко-лінгвістичних моделей, що ґрунтується на здійсненні синтаксичного аналізу речень природної мови, використанні баз даних у вигляді тезаурусу слів природної мови та бази правил для виявлення логічних зв'язків. Тоді масив логіко-лінгвістичних моделей речень природної мови певного тексту може виступати семантико-синтаксичною складовою логіко-лінгвістичної моделі електронного текстового документа, формування якого також потребує дотримання специфічних правил та виявлення логічних зв'язків між реченнями тексту.

Сучасні програмні продукти опрацювання текстової інформації здебільшого орієнтовані на англійську мову, оскільки для тестування результатів використовують напрацьовані бази, чого не можна сказати про українську мову. Тому метою цієї статті є описання та дослідження особливостей баз знань системи автоматизованої побудови логіко-лінгвістичних моделей україномовних текстових документів.

Виклад основного матеріалу

Процес формалізованого представлення текстової інформації, наданої у вигляді електронних текстових документів або їх фрагментів, розподіляють на ітерації, які проходять речення природної мови, для того щоб основні їх концепти зайняли правильні позиції у логіко-лінгвістичній моделі [2].

Аналізування текстової інформації неможливе без застосування спеціальних баз знань, які дають змогу виявити логічні зв'язки між словами у реченнях природної мови. Для функціонування системи автоматизованої побудови логіко-лінгвістичних моделей текстових документів [13] використано три бази знань, які побудовані у вигляді продукційних моделей представлення знань та містять формальні правила української мови, інтерпретовані у вигляді продукцій.

База знань 1 використовується для визначення ролі кожного зі слів електронного текстового документа та формування логіко-лінгвістичних моделей речень природної мови. В українській, як і у будь-якій іншій флективній мові, існують правила узгодження граматичних форм між словами речення, характерний порядок слів та взаємозв'язки між конкретними частинами мови, завдяки чому в реченні виражається думка. Автор виокремила правила [2], асоціювавши їх із математичними змінними, щоб узагальнити процес знаходження логічно зв'язаних конструкцій (словосполучень).

Групу слів, пов'язаних між собою логічними зв'язками, позначатимемо sp_j, $j=\overline{1,m}$, де m – кількість словосполучень у реченні. За правилами української мови словосполучення можуть утворювати такі члени речення [13]:

– "означення – підмет" – $sp_j = g \cup x$ – суб'єкт логіко-лінгвістичної моделі та його характеристика;

– "присудок – додаток" – $sp_j = p \cup y$ – предикат та об'єкт логіко-лінгвістичної моделі;

– "означення – додаток" – $sp_j = q \cup y$ – об'єкт логіко-лінгвістичної моделі та його характеристика;

– "додаток – додаток" – $sp_j = y \cup z$ – об'єкт та предмет відношення логіко-лінгвістичної моделі;

– "означення – додаток" – $sp_j = r \cup z$ – предмет логіко-лінгвістичної моделі та його характеристика;

– "обставина – присудок" – $sp_j = h \cup p$ – предикат та його характеристика.

Враховуючи можливі типи зв'язків у словосполученнях української мови, кожне правило бази знань 1 можна подати у вигляді правила "modus ponens":

– антецедент або посилання 1 – безпосереднє формалізоване правило, елементи якого об'єднані логічними операціями;

– антецедент або посилання 2 – виявлений тип зв'язку у поточному реченні природної мови;

– консеквент або висновок – формула для формальної інтерпретації виявленого виду зв'язку між словами речення природної мови.

Наприклад, щоб виявити логічні зв'язки у реченні природної мови "*Кожен з однорідних членів входить у зв'язок з головним словом словосполучення*", використаємо такі правила поєднання граматичних форм слів:

1) "дієприкметник – прийменник – іменник" – "*кожен зі членів*",

$$if\left((cm(S_i)=6)\,and\,(cm(S_{i+1})=9)\,and\,(cm(S_{i+2})=1)\right)and$$
$$\left(g(S_{i+2})\neq 1\right)then\left(S_j = S_i \cup S_{i+1} \cup S_{i+2}\right);$$

2) "прикметник – іменник" – *"однорідних членів"*,

$$if\left((cm(S_i)=2)\,and\,(cm(S_{i+1})=1)\right)and\left(g(S_i)=g(S_{i+1})\right)$$
$$and\left(n(S_i)=n(S_{i+1})\right)and\left(k2(S_i)=k2(S_{i+1})\right);$$
$$then\left(S_j = S_i \cup S_{i+1}\right)$$

3) "дієслово – прийменник – іменник" – "*входить у зв'язок*",

$$if\left((cm(S_i)=5)\,and\,(cm(S_{i+1})=9)\;and\,(cm(S_{i+2})=1)\right)and$$
$$\left(g(S_{i+2})\neq 1\right)then\left(S_j = S_i \cup S_{i+1} \cup S_{i+2}\right);$$

4) "іменник – прийменник – іменник" – *зв'язок зі словом*",

$$if\left((cm(S_i)=1)\,and\,(cm(S_{i+1})=9)\,and\,(cm(S_{i+2})=1)\right)and$$
$$\left((g(S_{i+2})=2)\vee(g(S_{i+2})=4)\right)\vee\left(g(S_{i+2})\neq 1\right) \quad ;$$
$$then\left(S_j=S_i\cup S_{i+1}\cup S_{i+2}\right)$$

5) "прикметник – іменник" – *"головним словом"*,

$$if\left((cm(S_i)=2)\,and\,(cm(S_{i+1})=1)\right)and\left(g(S_i)=g(S_{i+1})\right)$$
$$and\left(n(S_i)=n(S_{i+1})\right)and\left(k2(S_i)=k2(S_{i+1})\right) \quad ;$$
$$then\left(S_j=S_i\cup S_{i+1}\right)$$

6) "іменник – іменник" – *"словом словосполучення"*,

$$if\left((cm(S_i)=1)\,and\,(cm(S_{i+1})=1)\right)and$$
$$\left((g(S_{i+1})=2)\vee(g(S_{i+1})=4)\right)\vee\left(g(S_{i+1})\neq 1\right).$$
$$then\left(S_j=S_i\cup S_{i+1}\right)$$

Висновок на основі правила "modus ponens" дає можливість сформувати набір словосполучень для кожного речення природної мови.

Після формування словосполучень для кожного речення природної мови створюються їх логіко-лінгвістичні моделі, компоненти яких визначено на основі підстановки в загальну форму конкретних значень зі словосполучень.

Базу знань 2 створено для пошуку зв'язків між реченнями, що входять до складу електронного текстового документа. Вона являє собою сукупність продукцій, що відображають принципи синтезу [14], тобто об'єднання та заміни структурних компонентів логіко-лінгвістичних моделей речень природної мови. База знань 2 містить одинадцять правил синтезу та їх різні інтерпретації, внаслідок чого формується масив логіко-лінгвістичних моделей речень електронного текстового документа, а кожній логіко-лінгвістичній моделі $L^{S_\delta(\gamma)}$ приписується масив характеристик l_δ, кожна з яких відповідає певній компоненті логіко-лінгвістичної моделі речення S_δ, пов'язаній з ним за змістом:

$$t'^{(\gamma)}=\begin{cases} L^{S_1(\gamma)}=\bigwedge\limits_{p\in P^{S_1}}\bigwedge\limits_{h\in H_p^{S_1}}L_p^{S_1(\gamma)}(h), \\ L^{S_2(\gamma)}=\bigwedge\limits_{p\in P^{S_2}}\bigwedge\limits_{h\in H_p^{S_2}}L_p^{S_2(\gamma)}(h), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L^{S_\delta(\gamma)}=\bigwedge\limits_{p\in P^{S_\delta}}\bigwedge\limits_{h\in H_p^{S_\delta}}L_p^{S_\delta(\gamma)}(h), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ L^{S_{N(t)}(\gamma)}=\bigwedge\limits_{p\in P^{S_{N(t)}}}\bigwedge\limits_{h\in H_p^{S_{N(t)}}}L_p^{S_{N(t)}(\gamma)}(h), \end{cases}$$

$$\begin{cases} G_1(l_1):U\to u_k(S_e), e\neq 1,\ e=\overline{1,N(t)}, \\ G_2(l_2):U\to u_k(S_e), e\neq 2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ G_\delta(l_\delta):U\to u_k(S_e), \delta\neq e, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ G_{N(t)}(l_{N(t)}):U\to u_k(S_e), e\neq N(t). \end{cases}$$

Наприклад, для текстового фрагмента "*Якщо вирішення проблеми неоднозначне, то процес прийняття рішень потребує структуризації, яка дозволить визначити етапи та процедури, спрямовані на її вирішення. Підготовка та прийняття рішень у процесі управління є набором процедур, що об'єднуються в окремі етапи. Завдяки цьому можна побудувати загальну схему розроблення науково обґрунтованих рішень*" буде сформовано такий масив логіко-лінгвістичних моделей:

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

L^{S_1} = вирішення (0,0,проблеми,0,0,0,неоднозначне) →
потребує (процес,прийняття_рішень,структуризації,0,0,0,0)&
[дозволить_визначити(структуризація,0,етапи,спрямовані,вирішення,проблеми,0)&
дозволить_визначити (структуризація,0,процедури,спрямовані,вирішення,проблеми,0)].

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

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

L^{S_3} = можна_побудувати(0,0,схему,розробки,рішень,наукових,0)&
можна_побудувати(0,0,схему,розробки,рішень,обгрунтованих,0).

$L^{S_2} = p_{31} \_ p_{32}(0,0,y_3,q_3,z_3,r_{31},0)$ & $p_{31} \_ p_{32}(0,0,y_3,q_3,z_3,r_{32},0)$.

У вектор характеристик першого речення S_1 будуть внесені суб'єкт та його характеристика *"прийняття рішень"*, а також об'єкт та його характеристика *"набір процедур"* із другого речення: $l_1 = \{x_{22}, q_{22}, y_2', q_2'\}$. Відповідно до вектора характеристик другого речення текстового фрагмента увійдуть слова *"прийняття рішень"*, *"процедури"* з першого речення та слово *"рішень"* з третього речення: $l_2 = \{q_{11}, q_{12}, z_3\}$. До вектора характеристик третього речення потрапили $l_3 = \{q_{12}', g_{22}\}$, тобто слово *"рішень"*.

Отже, в результаті застосування принципів синтезу логіко-лінгвістичних моделей утворився масив логіко-лінгвістичних моделей речень природної мови, що є семантико-синтаксичною складовою логіко-лінгвістичної моделі електронного текстового документа, та масивом характеристик, що слугує основою для формування її лінгвістичної складової.

База знань 3, що використовується для побудови лінгвістичної складової логіко-лінгвістичної моделі текстового документа, є множиною продукцій, що містить правила формування мереж переходів для інтерпретації тематичної прогресії тексту.

Визначення типу тематичної прогресії із множини $Y = \{1,2,3,4,5\}$, що вжита у абзаці $a_k \in A$ (1 – проста лінійна прогресія; 2 – прогресія із наскрізною темою; 3 – прогресія з похідними темами; 4 – прогресія із розщепленою темою; 5 – прогресія із тематичним стрибком) відбувається завдяки використанню правил побудови абстрактних моделей розгортання інформації, які і формують базу знань 3.

Ще однією характеристикою абзацу тексту є множина рематичних домінант $R = \{1,2,3,4,5,6\}$ (1 – предметна; 2 – статальна; 3 – динамічна; 4 – якісна; 5 – імпресивна; 6 – комбінована). Рематичні домінанти визначаються узагальненням результатів синтезу (отриманої множини продукцій та масивів характеристик кожного речення), а також на основі отриманих із бази даних характеристик ключових слів [13].

Для наведеного вище текстового фрагмента характерна третя абстрактна модель, що відповідає прогресії із похідними темами, тобто коли суб'єкт першого речення частково виступає суб'єктом або предметом для всіх наступних речень, як у цьому випадку "*прийняття рішень*".

Внаслідок встановлення властивостей абзаців вдається виявити тематичну спрямованість та засоби побудови логічних зв'язків у них, завдяки чому формалізується процес визначення міжфразових зв'язків.

Побудова логіко-лінгвістичних моделей текстових документів можлива завдяки використанню описаних баз знань, а наведені вище приклади детально показують, який процес аналізу, починаючи

із найнижчого рівня пошуку словосполучень речень природної мови до виявлення логічних зв'язків між частинами текстів, проходить електронний текстовий документ.

Отже, на вхід системи автоматизованої побудови логіко-лінгвістичних моделей текстових документів подається текст або текстовий фрагмент. Його розділяють на речення, для кожного з яких будується логіко-лінгвістична модель за допомогою використання правил бази знань 1, що програмно являють собою окремі класи (рис. 1), у яких слово із заданого тексту та його характеристики порівнюються із зазначеними умовами у класі.

```
protected void rule01phrase(int val) {
    if(this.sentence[val].getPrioritet() == 0){
      for (int i = val + 1; i < this.sentence.length; i++) {
        //=========================================
        if (((this.sentence[val].selectedProperties("прикметник", 0, 0, 0, 0, 0, 0))
                && (this.sentence[i].selectedProperties("іменник", 0, 0, 0, 0, 0, 0))
                )
                && ((this.sentence[val].equalls(this.sentence[i], this.sentence[val].getC
                    this.sentence[val].getCurrent().getN(),0, 0, 0, 0)))

                ) {
            setText("\nП-1. Словосочетание   -", this.sentence[val], this.sentence[i]);

if(i+1 < this.sentence.length){
    if((this.sentence[i+1].selectedProperties("дієслово",0,0,0,0,0,0))){
    return;
}}
     if(!testSpoluchnikAndImennik(val,i)){
    return;
    }
```

Рис. 1. Приклад класу для програмної реалізації правила бази знань 1

Правила бази знань 2 являють собою класи із умовами заміни компонент логіко-лінгвістичних моделей логічно пов'язаних між собою речень природної мови на зразок:

```
public boolean _selectedProperties(String valNameLanguage, int valG, int valN, int valK2, int valT,
int valH, int valL) {
        boolean equally = false;
        Iterator it = this.properties.iterator();
        int[] array;
        boolean G = true,N = true,K2 = true,T = true,H = true,L = true;
        PropertyWord tempPropertyWord;
        while (it.hasNext()) {
           tempPropertyWord = (PropertyWord) it.next();
           array = tempPropertyWord.getProperties();
        if (tempPropertyWord.getNameLangPart().equals(valNameLanguage))
       { equally = true;
      for (int j : array) {
       tempPropertyWord.setColumn(j);
       if (valG > 0) {
       G = (tempPropertyWord.getG() !=valG); }
        if (valN > 0) {
        N = (tempPropertyWord.getN() !=valN);//equally = (tempPropertyWord.getN() == valN)}
        if (valK2 > 0) {
       K2 = (tempPropertyWord.getK2() !=valK2);//equally = (tempPropertyWord.getK2() ==
valK2);}
```

```
if (valT > 0) {
T = (tempPropertyWord.getT() !=valT);//equally = (tempPropertyWord.getT() == valT);}
if (valH > 0) {
H = (tempPropertyWord.getH() !=valH);//equally = (tempPropertyWord.getH() == valH);}
if (valL > 0) {
 L = (tempPropertyWord.getL() !=valL);//equally = (tempPropertyWord.getL() == valL);}
```

На основі правил бази знань 2 формується синтаксична складова логіко-лінгвістичної моделі електронного текстового документа.

Після виявлення зв'язків між реченнями заданого тексту та створення для кожного із них вектора характеристик застосовується база знань 3, тобто фактично формальні моделі речень природної мови, що входять до складу електронного текстового документа, послідовності їх вживання та їх характеристики зіставляються із можливими комбінаціями, сформованими у шаблони типу:

$$
\begin{aligned}
&t = \{L^{S_1}, L^{S_2}, ..., L^{S_\delta}, ..., L^{S_{N(t)}}\} = \\
&\{p_1(x_1, g_1, y_1, q_1, z_1, r_1, h_1)\ \& \\
&p_2(z_1, g_2, y_2, q_2, z_2, r_2, h_2)\ \& ...\ \& \\
&..... \\
&p_\partial(x_\partial, g_\partial, y_{\partial-1}, q_\partial, z_\partial, r_\partial, h_\partial)\ \& ...\ \& \\
&..... \\
&p_{N(t)}(z_{N(t)-1}, g_{N(t)}, y_{N(t)}, q_{N(t)}, z_{N(t)}, r_{N(t)}, h_{N(t)})\}.
\end{aligned}
$$

У результаті застосування методу автоматизованого створення логіко-лінгвістичних моделей електронних текстових документів [13], основні етапи якого описано вище, формуються синтаксична та лінгвістична складові логіко-лінгвістичної моделі текстового документа.

Для оцінювання формалізованих правил розроблених баз знань проаналізовано тези доповідей, дипломи, статті та дисертації.

Висновки

Застосування формалізованих правил трьох описаних у статті баз, що являють собою продукційні моделі представлення знань, дає змогу простежити послідовність кроків для семантико-синтаксичного аналізу електронних текстових документів з метою виявлення у них змісту. Із наведених прикладів зрозуміло, що формування тематичної спрямованості тексту неможливе без правильної структуризації текстової інформації на нижчих рівнях, тобто під час формування логічних зв'язків між та в середині речень природної мови. Саме тому, щоб забезпечити видобування знань з електронних текстових документів, розроблено три різні за функціональним призначенням бази даних: перша – щоб встановити зв'язки між словами, об'єднуючи їх у словосполучення та формуючи логіко-лінгвістичні моделі простих речень природної мови; друга – для виявлення зв'язків між реченнями, а третя – для встановлення міжфразових зв'язків. Оскільки продукції бази даних обробляють продукції з різними текстовими компонентами та містять антецеденти різних граматичних характеристик текстової інформації, об'єднати бази знань неможливо.

Бази знань, які розробила автор, можуть слугувати потужним інструментом у системах екстракції знань із електронних текстових документів не лише для української, але і для будь-якої флективної мови з відповідним урахуванням специфіки кожної конкретної природної мови.

Список літератури

1. Филиппов, К. А. (2008). Лингвистика текста: курс лекций. Спб.: Изд-во С.-Петерб. ун-та, 336.

2. Vavilenkova, A. (2020). Modelling of the context links between the natural language sentences. *Proceedings of the 9th International Scientific and Practical Conference "Information Control Systems & Technologies" (ICST2020)*, 282–293.

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

4. Bengfort, B. Syntax Parsing With CoreNLP and NLTK. Retrieved from: https://www.districtdatalabs.com/syntax-parsing-with-corenlp-and-nltk. (Дата звернення: 05.03.2021).

5. Gupta, M. Syntactic. Constituency Parsing usiong the CYK algorithm in NLP. Retrieved from: https://medium.com/data-science-in-your-pocket/syntactic-constituency-parsing-using-the-cyk-algorithm-in-nlp-eff9c2912b09. (Дата звернення: 04.05.2020).

6. NLPIR 2020: Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, Association for Computing Machinery, New York, United States, Seoul Republic of Korea. Retrieved from: https://dl.acm.org/doi/proceedings/10.1145/3443279. (Дата звернення: 05.03.2021).

7. NLPAI 2021: 2nd International Conference on Natural Language Processing and Artificial Intelligence. China. Retrieved from: http://www.nlpai.org/. (Дата звернення: 05.03.2021).

8. Ланде, Д. В. (2014). Елементи комп'ютерної лінгвістики в правовій інформатиці. Київ, НДІІП НАПрН, 168.

9. Sumam, F., Landeghem, J. V., Moens, M.-F. (2019). Transfer learning for named entity recognition in financial and biomedical documents. *Information* 2019, *10*(8), 248. Retrieved from: https://doi.org/10.3390/info10080248.

10. Chen, X., Xie. H., Cheng, G., Poon, L., Leng, M., and Wang, F. (2020). Trends and features of the applications of natural language processing techniques for clinical trials text analysis. *Applied Sciences,* 10, 2157. doi:10.3390/app10062157.

11. Khairova, N., Mamyrbayev, O., Mukhsina, K. and Kolesnyk, A. (2020). Logical-linguistic model for multilingual Open Information Extraction. *Cogent Engineering.* doi: 10.1080/23311916.2020.1714829.

12. Khairova, N., Petrasova, S. and Gautam A. P. S. (2016). The logical-linguistic model of fact extraction from english texts. *Communications in Computer and Information Science,* Vol. 639. Springer, Cham. Retrieved from: https://doi.org/10.1007/978-3-319-46254-7_51.

13. Вавіленкова, А. І. (2017), Аналіз і синтез логіко-лінгвістичних моделей речень природної мови: монографія. К.: ТОВ "СІК ГРУП УКРАЇНА", 152.

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

References

1. Fillipov, K. A. (2008). Text Linguistics. SpB Publisher, 336.

2. Vavilenkova, A. (2020). Modelling of the context links between the natural language sentences. *Proceedings of the 9th International Scientific and Practical Conference "Information Control Systems & Technologies" (ICST2020)*, 282–293.

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

4. Bengfort, B. Syntax Parsing With CoreNLP and NLTK. Available at: https://www.districtdatalabs.com/syntax-parsing-with-corenlp-and-nltk. (Accessed: 5 March 2021).

5. Gupta, M. Syntactic/ Constituency Parsing usiong the CYK algorithm in NLP. Available at: tps://medium.com/data-science-in-your-pocket/syntactic-constituency-parsing-using-the-cyk-algorithm-in-nlp-eff9c2912b09. (Accessed: 4 May 2020).

6. NLPIR 2020: Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval, Association for Computing Machinery, New York, United States, Seoul Republic of Korea. Available at: https://dl.acm.org/doi/proceedings/10.1145/3443279. (Accessed: 5 March 2021).

7. NLPAI 2021: 2nd International Conference on Natural Language Processing and Artificial Intelligence. China. Available at: http://www.nlpai.org/. (Accessed: 5 March 2021).

8. Lande, D. V. (2014). The Elements of Computer Linguistics in Legal Informatics. Kyiv, NDIIP NAPrH, 168.

9. Sumam, F., Landeghem, J. V., Moens, M.-F. (2019). Transfer Learning for Named Entity Recognition in Financial and Biomedical Documents. *Information*, *10*(8), 248. Available at: https://doi.org/10.3390/info10080248.

10. Chen, X., Xie. H., Cheng, G., Poon, L., Leng, M., and Wang, F. (2020). Trends and features of the applications of natural language processing techniques for clinical trials text analysis. *Applied Sciences*, 10, 2157. doi:10.3390/app10062157.

11. Khairova, N., Mamyrbayev, O., Mukhsina, K. and Kolesnyk, A. (2020). Logical-linguistic model for multilingual Open Information Extraction. *Cogent Engineering*, doi: 10.1080/23311916.2020.1714829.

12. Khairova, N., Petrasova, S. and Gautam A. P. S. (2016). The logical-linguistic model of fact extraction from english texts. *Communications in Computer and Information Science*, Vol. 639. Springer, Cham. Available at: https://doi.org/10.1007/978-3-319-46254-7_51.

13. Vavilenkova, A.I. (2017), Analysis and Synthesis of logic and linguistic models for natural language sentences, TOV "SIK GROUP UKRAINE", Kyiv, 152.

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

FEATURES OF THE KNOWLEDGE BASE OF THE SYSTEM OF AUTOMATED CONSTRUCTION OF LOGIC AND LINGUISTIC MODELS OF TEXT DOCUMENTS

Anastasiia Vavilenkova
National Aviation University,
vavilenkovaa@gmail.com, ORDIC 0000-0002-9630-4951

The article outlines the problem of finding meaningful units in electronic text documents and analyzes the main shortcomings of existing approaches of extracting knowledge from textual information. The article is devoted to the study of the peculiarities of the process of construction of logic and linguistic models of electronic text documents, in particular the description and research of the peculiarities of knowledge bases of the system of automated construction of logic and linguistic models of Ukrainian-language text documents. The author proposes a scheme of formalization of textual information based on the construction of a logic and linguistic model of an electronic text document. The first stage of construction is the formation of logical and linguistic models of natural language sentences, which uses a specially developed method of automated formation of logical and linguistic models. This method is based on parsing sentences of natural language, using words of natural language as a thesaurus database and using a database of rules to identify logical connections. This in turn is made possible by the author's developed knowledge base 1, which is used to determine the role of each word in an electronic text document and serves as a production model with formalized rules of the Ukrainian language for forming phrases that can form members of sentence of natural language. The knowledge base 2 was created by the author to find connections between sentences that are part of an electronic text document and is a set of productions that reflect the principles of synthesis of logic and linguistic models of sentences of natural language, ie the rules of combining and replacing structural components of logic and linguistic models of sentences of natural language. The knowledge base 3, used to build the linguistic component of the logic and linguistic model of a text document, is a set of productions that contains the rules of forming of transition networks to interpret the thematic progression of the text. The application of the developed formalized rules was demonstrated on specific text fragments. Applying the developed knowledge bases allows to trace the process of formation of logic and linguistic models of electronic text documents.

Key words: meaningful units, natural language, electronic text document, logic and linguistic model, knowledge base, production model.