ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ И РЕЧИ

УДК 004.912

Ю.С. Гецэвіч, І.В. Рэентовіч

ЛІНГВІСТЫЧНЫ АНАЛІЗ ДЛЯ БЕЛАРУСКАГА КОРПУСА ТЭКСТАЎ З ПРЫМЯНЕННЕМ МЕТАДАЎ АПРАЦОЎКІ НАТУРАЛЬНАЙ МОВЫ І МАШЫННАГА НАВУЧАННЯ

Аналізуюцца праблемы лакалізацыі розных марфалагічных, лексічных і сінтаксічных элементаў з дапамогай беларускага модуля праграмы NooJ. У тым ліку выпраўляюцца памылкі, якія сустракаюцца ў беларускіх тэкстах, будуюцца мадэлі мовы і тэгіравання часцін мовы. Праводзіцца апрацоўка беларускага корпуса тэкстаў на натуральнай мове з дапамогай распрацаванага алгарытму з выкарыстаннем машыннага навучання.

Уводзіны

Стварэнне нацыянальных корпусаў тэкстаў, а таксама корпусаў, прызначаных для вырашэння вузкаспецыялізаваных задач для пэўных сфер дзейнасці, з'яўляецца на сённяшні дзень актуальнай задачай. З 1960-х гг. па цяперашні час была сфармавана значная колькасць корпусаў тэкстаў у розных краінах свету (Вялікабрытаніі, ЗША, Францыі, Расіі і інш.). Згодна агульнапрызнанай класіфікацыі існуюць нацыянальныя корпусы тэкстаў, корпусы замежных моў, паралельныя корпусы. Вывучэнне корпусаў тэкстаў дазваляе атрымліваць дакладныя даныя аб лексічным складзе моў, адносных частотах ужывання тых ці іншых слоў, спалучальнасці граматычных з'яў паміж сабой і г. д. [1].

У Беларусі таксама інтэнсіўна вядзецца праца па распрацоўцы агульнанацыянальнага корпуса мовы [2]. Навуковыя і вышэйшыя адукацыйныя ўстановы нашай краіны непасрэдна займаюцца вырашэннем дадзенай задачы [3]. Першым беларускім корпусам тэкстаў, даступным у Інтэрнэце, стаў навукова-тэхнічны Corpus Albaruthenicum. Даследчай групай з лабараторыі распазнавання і сінтэзу маўлення Аб'яднанага інстытута праблем інфарматыкі Нацыянальнай акадэміі навук Беларусі быў сабраны першы мільённы корпус тэкстаў для беларускага модуля праграмы NooJ [4]. Дадзены корпус складаецца з тэкстаў розных тэматычных напрамкаў (мастацкіх твораў, навуковых, даведчых, вучэбных і іншых выданняў). Асноўнае яго прызначэнне – вырашэнне вузканакіраваных задач, што датычацца аптымізацыі і пашырэння распрацовак высакаякасных лінгвістычных алгарытмаў для папярэдняй апрацоўкі розных электронных тэкстаў.

У наш час напрамкам, які актыўна развіваецца, з'яўляюцца распазнаванне і сінтэз маўлення па тэксце: камп'ютар агучвае праграмы для інвалідаў па зроку, чытае электронныя кнігі, спявае па тэкстах і нотах. Непазбежна ўзнікаюць памылкі распазнання, якія выпраўляюцца з дапамогай аўтаматычных метадаў на аснове слоўнікаў і марфалагічных мадэляў з ужываннем машыннага навучання. Аднак, не гледзячы на ўсе дасягненні ў гэтай галіне, застаюцца праблемы, якія не ўдалося вырашыць цалкам: саманавучанне сістэм, іх здольнасць да самастойнага папаўнення слоўнікаў, прымянення да розных прадметных галін і інш.

Асноўныя праблемы сінтэзу маўлення па тэксце:

– калі слоўнікі націскаў няпоўныя, сінтэзатар маўлення няправільна робіць акцэнтуацыю, а калі поўныя, то невядома, як іх хутка папаўняць новымі словамі для ўсёй парадыгмы;

– няпоўны слоўнік часцін мовы, з-за чаго сінтэзатар маўлення няправільна робіць сінтагмы.

Мэтай працы з'яўляецца апісанне этапаў працэсу лінгвістычнай апрацоўкі тэкстаў беларускага корпуса з прымяненнем метадаў машыннага навучання для далейшага выкарыстання абноўленага корпуса ў працэсе сінтэзу беларускага маўлення. Пасля апрацоўкі корпуса з вялі-

кай верагоднасцю можна будзе даведацца, як хутка рашаць праблемы сінтэзу ў прынцыпе. Таму становяцца важнымі і аналітыка па корпусе, і лакальнае вырашэнне праблем.

Рашэнне практычна кожнай задачы аўтаматычнай апрацоўкі тэкстаў так ці інакш уключае ў сябе аналіз тэксту на некалькіх узроўнях прадстаўлення [5], а менавіта:

графематычны – выдзяленне з масіву даных сказаў і слоў (токенаў);

марфалагічны – выдзяленне граматычнай асновы слова, вызначэнне часцін мовы, прывядзенне слова да слоўнікавай формы;

сінтаксічны – выяўленне сінтаксічных сувязей паміж словамі ў сказах, пабудова сінтаксічнай структуры сказа;

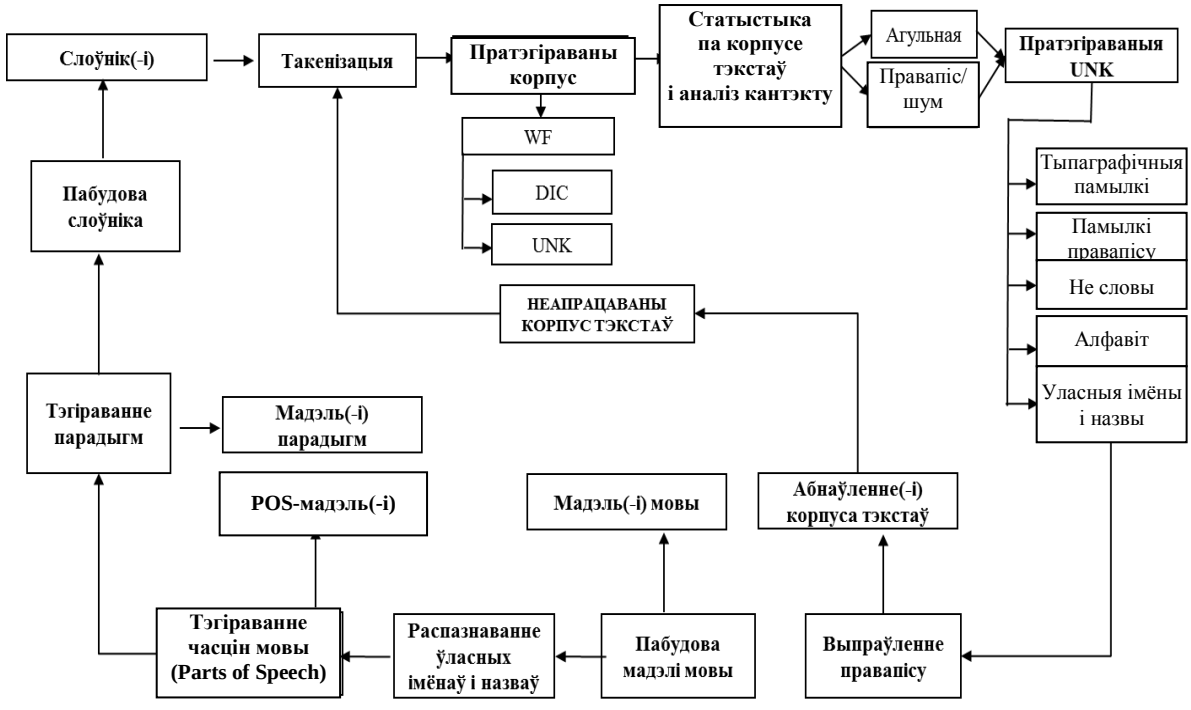
семантычны – выяўленне семантычных сувязей паміж словамі і сінтаксічнымі групамі, устанаўленне семантычных адносін.

Кожны такі аналіз – самастойная задача, якая не мае ўласнага практычнага прымянення, але актыўна выкарыстоўваецца для вырашэння большасці агульных пытанняў. Многія даследчыя сістэмы прымяняюцца або для апрабацыі метадаў і правядзення вылічальных эксперыментаў, або ў якасці састаўных частак (ці бібліятэк) для сістэм, што рашаюць ту ці іншую прыкладную задачу. Прыкладам такіх сістэм могуць служыць сродкі NLTK для графематычнага аналізу і такенізацыі, марфалагічны аналізатар mysystem, сінтаксічны парсер «ЭТАП3» і інш. [5, 6].

1. Агульная схема працэсу апрацоўкі корпуса тэкстаў

Беларускі корпус быў створаны з дапамогай праграмнага забеспячэння – інтэграванага лінгвістычнага асяроддзя распрацовак NooJ [7] – спецыяльна для беларускага модуля, які з'яўляецца лінгвістычным рэсурсам і дадаткова ўсталёўваецца ў дадзеную праграму. Корпус складаецца з 338 тэкстаў, праанатаваных ў вышэйназванай праграме ў ходзе агульнага лінгвістычнага аналізу з выкарыстаннем спецыяльнага слоўніка general_be.nod [8]. Слоўнік general_be.nod утрымлівае 2 153 082 словаформы, 138 200 слоў, 2 280 828 словаформаў з улікам аманіміі, 111 759 словаформаў з іншай аналітычнай інфармацыяй.

Апрацоўка тэкстаў праводзілася ў адпаведнасці са схемай, прадстаўленай на мал. 1.



Мал. 1. Схема працэсу апрацоўкі корпуса тэкстаў

2. Лексічны аналіз корпуса тэкстаў

На *першым этапе* апрацоўкі корпус тэкстаў быў разбіты на сказы і словаформы, праведзена параўнанне з існуючым слоўнікам і выведзена статыстыка.

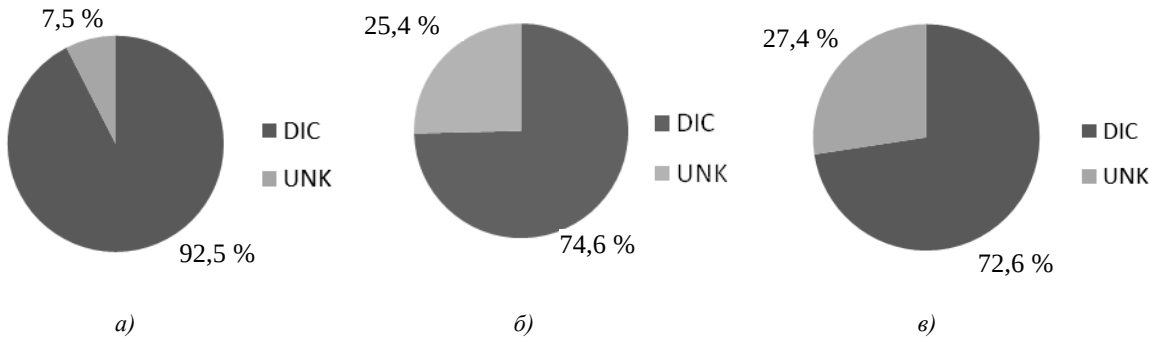
Пасля лексічнага аналізу корпуса былі атрыманы яго колькасныя вынікі па ўсіх словаформах увогуле (WF – wordforms), усіх вядомых (DIC – базавы слоўнікавы састаў дадзенага корпуса) і ўсіх невядомых словаформах (UNK_{total} – unknown wordforms), усіх унікальных вядомых (DIC_{unique}) і ўсіх унікальных невядомых словаформах (UNK_{unique}), усіх унікальных вядомых ($DIC_{unique\ lowered}$), усіх унікальных невядомых ($UNK_{unique\ lowered}$) словаформах, прыведзеных да ніжняга рэгістру.

Ніжэй прыводзіцца адпаведная статыстычная табліца і дыяграмы для параўнання гэтых даных (мал. 2).

Табліца 1

Вынікі сінтаксічнага і граматычнага разбораў корпуса тэкстаў (парсінга)

Запыт	Усяго словаформаў	Унікальных словаформаў	Унікальных словаформаў, прыведзеных да ніжняга рэгістру
DIC	1 747 723	148 402	127 554
UNK	142 234	50 513	48 146
DIC+UNK	1 889 957	198 915	175 700



а) *б)* *в)*

Мал. 2. Дыяграмы адлюстравання словаформаў: *а)* усіх у корпусе; *б)* усіх унікальных; *в)* усіх унікальных, прыведзеных да ніжняга рэгістру

Аналіз дыяграм, атрыманых на аснове даных папярэдняй апрацоўкі тэкстаў з дапамогай праграмы NooJ, паказаў, што 1/14 частку ад агульнага аб'ёму ўсіх словаформаў, а таксама 1/4 аб'ёму ўнікальных словаформаў дадзенага корпуса беларускамоўных тэкстаў складаюць невядомыя (непраанатаваныя) словы.

Такім чынам, маюць месца наступныя сцверджанні:

1. Паколькі суадносіны ўсіх невядомых словаформаў у корпусе (UNK_{total} = 142 234) да колькаснага размеркавання ўсіх унікальных невядомых словаформаў (UNK_{unique} = 50 513) у выніку даюць

$$\frac{UNK_{total}}{UNK_{unique}} = \frac{142234}{50513} \approx 2{,}82,$$

то ў сярэднім у корпусе тэкстаў сустракаецца тры супадзення кожнага ўнікальнага невядомага слова.

2. Паколькі суадносіны ўсіх невядомых словаформаў у корпусе (UNK_{total} = 142 234) да колькаснага размеркавання ўсіх унікальных невядомых словаформаў, прыведзеных да ніжняга рэгістру ($UNK_{unique\ lowered}$ = 48 146), у выніку даюць

$$\frac{UNK_{total}}{UNK_{uniquelowered}} = \frac{142234}{48146} \approx 2{,}952,$$

то ў сярэднім у корпусе тэкстаў сустракаецца тры супадзення кожнага ўнікальнага невядомага слова, прыведзенага да ніжняга рэгістру.

Атрыманыя статыстычныя даныя па корпусе тэкстаў паставілі наступныя першарадныя падзадачы:

– высветліць, колькі невядомых словаформаў могуць сапраўды папоўніць агульны слоўнік general_be.nod (маецца на ўвазе новая лексіка-граматычная інфармацыя), даць статыстыку па такіх словах, поўнасцю праанатаваць іх, правесці працэдуру па выключэнні мнагазначнасці, назначыць парадыгмы і саставіць на іх базе новы слоўнік, дадатковы да асноўнага;

– параўнаць невядомыя словаформы дадзенага корпуса з тымі, што ёсць у асноўным слоўніку, і вызначыць, колькі з іх супадаюць;

– высветліць, колькі словаформаў па той ці іншай прычыне (з-за памылак правапісу, няправільнай сінтаксічнай апрацоўкі тэкстаў і г. д.) былі не праанатаваны праграмай NooJ у працэсе лінгвістычнага аналізу;

– даць статыстыку па непраанатаваных словах;

– апрацаваць усе тэксты корпуса з мэтай выпраўлення праблем, што ўзніклі, і затым яшчэ раз правесці лінгвістычны аналіз корпуса (гл. мал. 1).

3. Выпраўленне крытычных памылак для зніжэння «зашумлення» корпуса тэкстаў

На *другім этапе* апрацоўкі корпуса тэкстаў праведзены аналіз статыстыкі і выяўлены асноўныя тыпы памылак, якія сустракаюцца ў беларускіх тэкстах і не дазваляюць праанатаваць словы з памылкамі (табл. 2): змешанае напісанне літар кірылічнага і лацінскага алфавітаў (напрыклад, лацінская літара «і» сустракаецца ў беларускіх словах у 16,72 % выпадкаў, лацінская «у» – у 3,99 % выпадкаў).

У выніку аналізу статыстычных дадзеных абагульнены выпадкі, калі такенайзер можа працаваць некарэктна:

– няправільнае (у прыватнасці, злітнае) напісанне складаных слоў, якое не адпавядае лексіка-граматычным правілам беларускай мовы, напрыклад: *агітацыйнапрапагандысцкая, адміністрацыйнадырэктыўныя*;

– адсутнасць прабелу паміж словамі (*абрываефразу, абутварэнні, неўдалося*);

– знак пераносу ў слове (*асаблі-выя, дысцып-ліне, сялен-няў*);

– словы, лексічна напісаныя праз дэфіс (*вя-алі-кай, ст-о-о-й, ш-ш-ырк*);

– словы з няправільным (некарэктным) правапісам (*у бўдзе, бсларускія, каіпулю, калектьгеізацыі, цбпла*);

– залішнія прабелы ў словах (*адра␣␣зу, выкарыс␣␣тоўваліся*);

– словы з апострафам (*аб'яднаны, з 'едзены, з' явілася, раз 'юшыцца*).

Лексіка-граматычная ідэнтыфікацыя невядомых словаформаў:

– невядомыя ці мала вядомыя беларускія словы, напрыклад: *драганты, зѐлкі, кіпецень, чапуля, дапяцца, нажыліцца, зіхотны, ціхманы*;

– словы беларускага класічнага правапісу (*абласьцях, аналёгічная, у вадно імгненьне, лятарэі, намэнклятуры*);

– замежныя словы (*acquisition, Akademie, aussi, армии, существующий*);

– беларускія або рускія словы ў замежнай транслітэрацыі (*pif-paf, belarusugady, knihi, korrespondent*);

– абрэвіятуры і скарачэнні (*ААН, АКЗП-6, стар., грэч., Інбелкульт, італ*);

– уласныя імёны і назвы (*Абрыцкі, Агеевы, Біман, Вялікабрытаніі, Дзятлавічы, Паоло, Севярын*).

Табліца 2

Статыстыка беларускага корпуса для пабудовы мадэлі каналу яго «зашумленасці» (Noisy Channel Model) пры параўнанні асобных літар беларускага і лацінскага алфавітаў

Letter BLR	Single BLR	Total BLR (TB)	Letter Lat	Mixed Cyrillic (MC)	Single Lat	Total Lat (TL)	Probability (TL/(TB+TL)),%	MC Fixed	Single Fixed	Total Fixed
а	4	99 351	a	64	0	64	0,06	64	0	64
б	0	14 813	b	1	8	9	0,06	1	8	9
е	415	28 000	e	11	0	11	0,04	11	0	11
і	0	36 765	i	813	6 569	7 382	16,72	797	6 559	7 356
к	0	37 076	k	0	10	10	0,03	0	10	10
м	623	24 053	m	4	55	59	0,24	4	55	59
н	236	45 474	h	3	4	7	0,02	3	4	7
о	0	28 731	o	27	11	38	0,13	27	9	36
п	185	19 934	n	7	0	7	0,04	7	0	7
р	148	41 010	p	25	6	31	0,08	25	2	27
с	302	37 126	c	28	29	57	0,15	28	3	31
т	410	25 071	t	4	10	14	0,06	4	5	9
у	0	26 114	y	19	1 066	1 085	3,99	19	1 065	1 084
х	231	10 841	x	30	49	79	0,72	30	0	30
Total	2554	474 359		1 036	7 817	8 853	22,34	1 020	7 720	8 740

Ніжэй прыведзены расшыфроўкі пазначэнняў табл. 2.

Letter BLR – літара беларускага алфавіту.

Single BLR – колькасць асобна ўжытых літар беларускага алфавіту ў корпусе тэкстаў.

Total BLR (TB) – агульная колькасць ужывання пэўнай беларускамоўнай літары ў корпусе тэкстаў.

Letter Lat – літара лацінскага алфавіту.

Mixed Cyrillic – колькасць выпадкаў ужывання ў беларускамоўных словах лацінскіх літар замест кірылічных.

Single Lat – колькасць асобна ўжытых літар лацінскага алфавіту ў корпусе тэкстаў.

Total Lat (TL) – агульная колькасць ужывання пэўнай лацінскай літары ў корпусе.

Probability (TL/(TB+TL)) – формула для вылічэння верагоднасці «зашумлення» корпуса тэкстаў літарамі лацінскага алфавіту.

MC Fixed – колькасць выпраўленых выпадкаў недакладнага ўжывання ў беларускамоўных словах лацінскіх літар замест кірылічных.

Single Fixed – колькасць выпраўленых выпадкаў асобнага недакладнага ўжывання літар (беларускіх, лацінскіх) у корпусе тэкстаў.

Total Fixed – агульная колькасць выпраўленых выпадкаў недакладнага ўжывання літар у корпусе тэкстаў.

Total – выніковая сума.

На *трэцім этапе* апрацоўкі корпуса тэкстаў для выпраўлення крытычных памылак, пералічаных вышэй, выкарыстоўваліся наступныя спосабы карэктыроўкі для зніжэння «зашумлення» корпуса тэкстаў:

– замена лацінскіх літар на кірылічныя і кірылічных на лацінскія;

– выключэнне спецыяльных сімвалаў;

– выпраўленне рымскіх лічбаў;

– даданне дэфісу ў слове;

– даданне прабелу паміж словамі;

– выдаленне знака пераносу ў адпаведных словах і аб’яднанне частак такіх слоў у адно цэлае;

– выдаленне залішніх прабелаў паміж словамі;

– распазнаванне і ідэнтыфікацыя слоў з апострафам як адно цэлае;

– выкананне працэсу распазнавання ўласных імёнаў і назваў, а таксама слоў беларускага класічнага правапісу.

4. Распрацоўка мадэлі мовы

Для эфектыўнай працы сістэм сінтэзу беларускага маўлення неабходна распрацаваць мадэль мовы. У цяперашні час асноўным падыходам да пабудовы моўных мадэляў для сістэм распазнання маўлення з'яўляецца выкарыстанне апарата статыстычных метадаў. Аўтарамі быў прыменены клас мадэляў, які выкарыстоўвае дрэвы рашэнняў для ацэнкі размеркавання верагоднасцяў чарговага слова.

У выніку тэгіравання часцін мовы вызначана, што значную частку словаформаў у корпусе пакрываюць назоўнікі (33,7 %), дзеясловы (26,2 %), прыметнікі (25,8 %). Так як гэтая частка застанецца нязменнай для невядомых словаформаў, то вышэйназваныя часціны мовы былі абраны для машыннага навучання. У дадзеным выпадку для машыннага навучання былі выкарыстаны алгарытмы нейроннай сеткі [9], дрэва прыняцця рашэнняў [10, 11] і кластэрызацыі. Для таго каб натрэніраваць і праверыць усе магчымыя парадыгмы слова з дапамогай вышэйназваных алгарытмаў, было ўзята 70 % (143 808) усіх вядомых словаформаў. Затым 30 % (27 397) усіх вядомых словаформаў былі рэалізаваны выведзенай мадэллю алгарытмаў машыннага навучання, каб праверыць дакладнасць дадзенай мадэлі. Дакладнасць мадэлі склала 80–90 %. Вынікі часцінамоўнага тэгіравання адлюстраваны ў табл. 3.

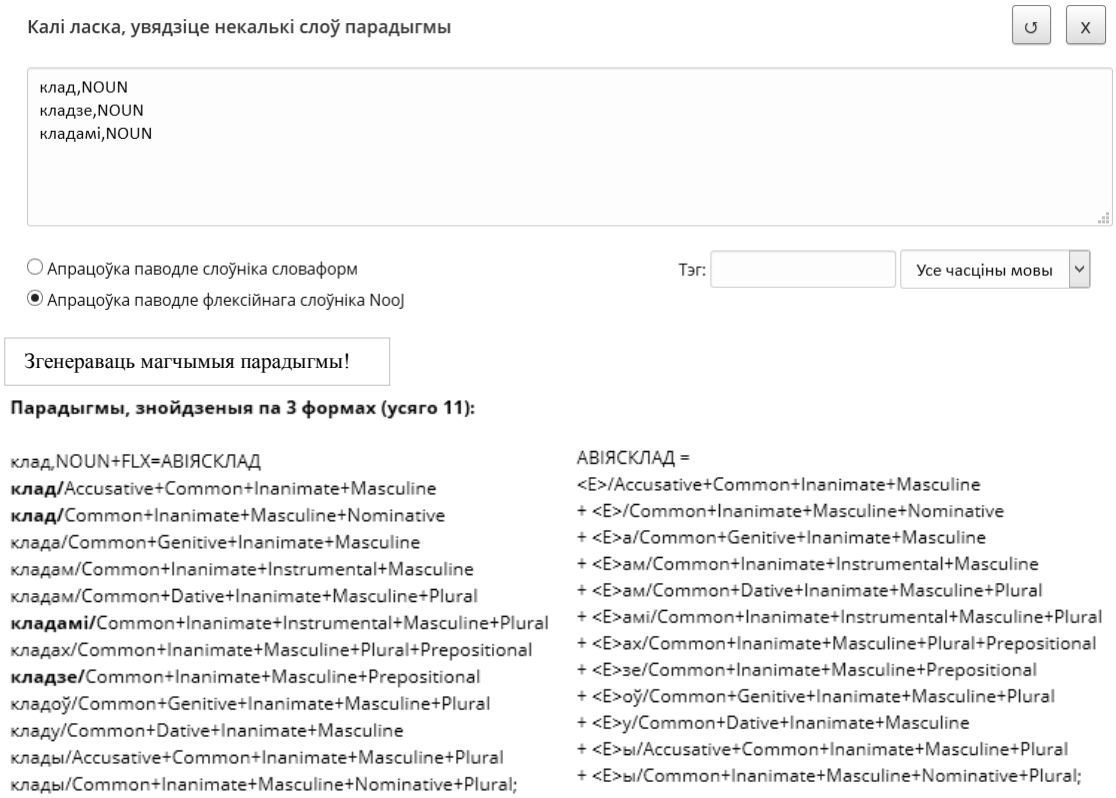
Табліца 3

Статыстычныя даныя пасля аўтаматычнага тэгіравання часцін мовы

Часціна мовы		Прадказаны клас, адз.		Прадказаны іншы клас, адз.		Карэктнасць мадэлі, %	Адчувальнасць мадэлі, %
Дзеяслоў, дзеепрыслоўе	Фактычны клас	Праўдзіва дакладны	4656	Памылкова адмоўны	1913	91,5	70,9
	Іншы фактычны клас	Памылкова дакладны	420	Праўдзіва адмоўны	20 408		
Назоўнік	Фактычны клас	Праўдзіва дакладны	10 882	Памылкова адмоўны	658	79,2	94,3
	Іншы фактычны клас	Памылкова дакладны	5043	Праўдзіва адмоўны	10 814		
Прыметнік, дзеепрыметнік	Фактычны клас	Праўдзіва дакладны	5233	Памылкова адмоўны	1238	91,4	80,9
	Іншы фактычны клас	Памылкова дакладны	1122	Праўдзіва адмоўны	19 804		

Каб праанатаваць знойдзеныя новыя невядомыя словы быў прыменены анлайн-сэрвіс «Генератар парадыгмы слова» [12], які з'яўляецца эфектыўным сродкам для тэгіравання парадыгмы слова. Механізм яго работы прадстаўлены на мал. 3.

Пасля апрацоўкі невядомага слова з дапамогай генератара парадыгмы слова дабаўляецца ў слоўнік, узбагачаючы наяўны агульны слоўнік новымі словамі. Інфармацыя, дададзеная ў слоўнік такім чынам, можа прадстаўляць вялікі інтарэс як для лексікаграфаў, так і для іншых моўных тэхналогій [13–15].



Мал. 3. Механізм работы генератара парадыгмы слова

Вышэйпрыведзеныя этапы працэсу апрацоўкі беларускамоўнага корпуса тэкстаў дадуць магчымасць даследчыкам больш дакладна аналізаваць з пункту гледжання лінгвістыкі розныя беларускія тэксты.

Заключэнне

У выніку апрацоўкі мільённага беларускага корпуса тэкстаў на натуральнай мове з дапамогай распрацаванага алгарытму і спецыяльнага слоўніка general_be.nod выяўлены і выпраўлены крытычныя памылкі для зніжэння «зашумлення» корпуса тэкстаў. Распрацавана мадэль тэгіравання часцін мовы з прымяненнем машыннага навучання, дакладнасць выкарыстання якой пры лінгвістычнай апрацоўцы беларускіх тэкстаў складае 80–90 %, што дазволіла папоўніць слоўнік новымі словамі. Рэалізаваны працэс папярэдняй апрацоўкі тэксту, дзе мелася магчымасць уводзіць невядомыя словы, тэгіраваць для іх часціны мовы ці прапускаць лексічную інфармацыю або да разбору, або падчас ліквідацыі неадназначнасці.

Распрацаваныя мадэлі і алгарытм апрацоўкі корпуса тэкстаў зрабілі магчымым значна палепшыць апрацоўку (выпраўленне правапісу, пабудову мадэлі мовы, распазнаванне ўласных імёнаў і назваў, тэгіраванне часцін мовы, тэгіраванне парадыгм слоў, пабудову слоўнікаў) і прадвызначылі яго прымяненне ў сінтэзе маўлення. Рэалізацыя праведзенага даследавання палепшыць лінгвістычны аналіз беларускамоўных корпусаў тэкстаў і паскорыць іх бесперапынную інтэграцыю ў розныя сістэмы сінтэзу беларускага маўлення.

Спіс літаратуры

1. Kennedy, G. An Introduction to Corpus Linguistics / G. Kennedy. – London : Longman, 1998. – 315 p.

2. Belarusian N-corpus [Electronic resource]. – 2015. – Mode of access : http://bnkorpus.info/. – Date of access : 22.06.2017.

3. Барковіч, А.А. Беларускі корпус тэкстаў : інтэрнэт-дыскурс / А.А. Барковіч // Веснік Беларус. дзярж. ун-та. Сер. 4. Філалогія. Журналістыка. Педагогіка. – 2013. – № 2. – С. 26–29.

4. The First One-Million Corpus for the Belarusian NooJ Module / I. Reentovich [et al.] // Automatic Processing of Natural-Language Electronic Texts with NooJ : 9th Intern. Conf. «NooJ 2015». – Springer International Publishing, 2016. – P. 3–15.

5. Холоденко, А.Б. Использование лексических и синтаксических анализаторов в задачах распознавания для естественных языков / А.Б. Холоденко // Интеллектуальные системы. – 1999. – № 1–2. – С. 185–193.

6. Автоматическая обработка текстов на естественном языке и анализ данных / Е.И. Большакова [и др.]. – М. : Изд-во НИУ ВШЭ, 2017. – 269 с.

7. Silberztein, M. NooJ Manual / M. Silberztein [Electronic resource]. – 2003. – Mode of access : www.nooj4nlp.net. – Date of access : 22.06.2017.

8. Hetsevich, Yu. Overview of Belarusian and Russian Dictionaries and Their Adaptation for NooJ / Yu. Hetsevich, S. Hetsevich // Automatic Processing of Various Levels of Linguistic Phenomena: Selected Papers from the NooJ 2011 Intern. Conf. – Newcastle : Cambridge Scholars Publishing, 2012. – P. 29–40.

9. Kriesel, D. A Brief Introduction to Neural Networks / D. Kriesel [Electronic resource]. – 2005. – Mode of access *:* http://www.dkriesel.com. – Date of access : 22.06.2017.

10. Quinlan, J.R. Simplifying Decision Trees / J.R. Quinlan // Intern. J. of Man-Machine Studies. – 1987. – Vol. 27, no. 3. – P. 221–234.

11. Cha, S.-H. A Genetic Algorithm for Constructing Compact Binary Decision Trees / S.-H. Cha, C.C. Tappert // J. of Pattern Recognition Research. – 2009. – Vol. 4, no. 1. – P. 1–13.

12. Генератар парадыгмы слова // Лабараторыя распазнавання і сінтэзу маўлення [Электронны рэсурс]. – 2017. – Рэжым доступу : http://ssrlab.by/5047. – Дата доступу : 13.05.2017.

13. Oliveira, H.G. Towards the Automatic Enrichment of a Thesaurus with Information in Dictionaries / H.G. Oliveira, P. Gomes // Expert Systems. – 2013. – Vol. 30, no. 4. – P. 320–332.

14. The Enrichment of Lexical Resources Through Incremental Parsebanking / V. Rosén [et al.] // Language Resources and Evaluation. – 2016. – Vol. 50, no. 2. – P. 291–319.

15. Computer Treatment of Slavic and East European Languages / ed. R. Garabik // Third Intern. Seminar, Bratislava, Slovakia, 10–12 Nov. 2005. – Bratislave : VEDA, 2005. – 246 p.

Паступіла 28.09.2017

Аб`яднаны інстытут праблем
інфарматыкі НАН Беларусі,
Мінск, Сурганава, 6
e-mail: Yury.Hetsevich@gmail.com,
ivan.reentovich@gmail.com

Yu.S. Hetsevich, I.V. Reentovich

LINGUISTIC ANALYSIS FOR THE BELARUSIAN CORPUS WITH THE APPLICATION OF NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING TECHNIQUES

The article focuses on the problems existing in text-to-speech synthesis. Different morphological, lexical and syntactical elements were localized with the help of the Belarusian unit of NooJ program. Those types of errors, which occur in Belarusian texts, were analyzed and corrected. Language model and part of speech tagging model were built. The natural language processing of Belarusian corpus with the help of developed algorithm using machine learning was carried out. The precision of developed models of machine learning has been 80–90 %. The dictionary was enriched with new words for the further using it in the systems of Belarusian speech synthesis.