БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФИЛОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра прикладной лингвистики

ПРИКЛАДНАЯ ЛИНГВИСТИКА: НАСЛЕДИЕ И СОВРЕМЕННОСТЬ

Материалы
II Международной научно-практической конференции, посвященной 85-летию филологического факультета Белорусского государственного университета

Минск, 22–23 марта 2024 г.

Научное электронное издание

МИНСК, БГУ, 2024

ISBN 978-985-881-656-8

УДК 81'33(06)
ББК 81.1я431

Прикладная лингвистика: наследие и современность [Электронный ресурс] : материалы II Междунар. науч.-практ. конф., посвящ. 85-летию филол. фак. Белорус. гос. ун-та, Минск, 22–23 марта 2024 г. / Белорус. гос. ун-т ; редкол.: О. М. Дорогокупец-Новицкая (гл. ред.) [и др.]. – Минск : БГУ, 2024. – 1 электрон. опт. диск (CD-ROM). – ISBN 978-985-881-656-8.

Рассматриваются актуальные вопросы и проблемы компьютерной и корпусной лингвистики, дискурс-анализа, судебной лингвистической экспертизы, лексикографии, переводоведения, лингводидактики и др.

Минимальные системные требования:
PC, Pentium 4 или выше; RAM 1 Гб; Windows XP/7/10;
Adobe Acrobat

Оригинал-макет подготовлен в программе Microsoft Word

В авторской редакции

Ответственный за выпуск *О. М. Дорогокупец-Новицкая*
Компьютерная верстка *А. А. Клименковой*

Подписано к использованию 25.09.2024. Объем 0,98 МБ

Белорусский государственный университет.
Управление редакционно-издательской работы.
Пр. Независимости, 4, 220030, Минск.
Телефон: (017) 259-70-70.
e-mail: urir@bsu.by
http://elib.bsu.by

СЕКЦИЯ II. КОМПЬЮТЕРНАЯ И КОРПУСНАЯ ЛИНГВИСТИКА

СЕКЦИЯ III. ПРОБЛЕМЫ ДИСКУРС-АНАЛИЗА И ЛИНГВИСТИЧЕСКОЙ ЭКСПЕРТИЗЫ ТЕКСТА

УДК 004.932.75:811.161.3’322.2

АЎТАМАТЫЧНАЯ АПРАЦОЎКА НАТУРАЛЬНАЙ МОВЫ: ПЕРАДАПРАЦОЎКА ТЭКСТАВАЙ ІНФАРМАЦЫІ З CORPUS.BY

Д. А. Бяляўскі[1)] **, Г. С. Кухарэвіч**[2)]**, Я. С. Зяноўка**[3)]**, А. А. Бакуновіч**[4)]**, А. Я. Драгун**[5)]**, М. А. Хацькова**[6)]**, Ю. С. Гецэвіч**[7)]

[1–7)] *Аб’яднаны інстытут праблем інфарматыкі НАН Беларусі, вул. Сурганава 6, 220012, г. Мінск, Беларусь,*
dzianis.bialiauski@gmail.com, empeiricos@gmail.com, evgeniakacan@gmail.com, bakunovich.andrei.work@gmail.com, ndrahun@gmail.com, margaryta.kazlova@gmail.com, yuras.hetsevich@gmail.com

Артыкул прысвечаны сучасным тэхналогіям апрацоўкі натуральнай мовы як асобнай падзадачы штучнага інтэлекту. На прыкладзе платформы corpus.by прыведзены асобныя інструменты, што ажыццяўляюць перадапрацоўку тэксту на беларускай мове, у прыватнасці сэрвісы па ўдакладненні інфармацыі аб сімвалах і часцінах мовы, падліку частотнасці слоў, такенізацыі і лематызацыі тэкставай інфармацыі.

Ключавыя словы: аўтаматычная апрацоўка тэксту; камп’ютарныя тэхналогіі; перадапрацоўка тэксту; лематызацыя; прыкладная лінгвістыка.

NATURAL LANGUAGE PROCESSING: TEXT PREPROCESSING BY CORPUS.BY

Dz. Bialiauski[1)]**, G. Kukharevich**[2)]**, Ya. Zianouka**[3)]**, A. Bakunovich**[4)]**, A. Drahun**[5)]**, M. Khatskova**[6)]**, Yu. Hetsevich**[7)]

[1–7)] *United Institute of Informatics Problems of National Academy of Sciences of Belarus, 6 Surganova St., 220012, Minsk, Belarus,*
dzianis.bialiauski@gmail.com, empeiricos@gmail.com, evgeniakacan@gmail.com, bakunovich.andrei.work@gmail.com, ndrahun@gmail.com, margaryta.kazlova@gmail.com, yuras.hetsevich@gmail.com

The article is devoted to modern technologies of natural language processing which is a separate subtask of artificial intelligence. Some tools for text preprocessing in the Belarusian language in corpus.by platform are described. They are the services for clarifying information about symbols and parts of speech, calculating the frequency of words, tokenization and lematization of text information.

Keywords: automatic text processing; computer technologies; text preprocessing; lemmatization; applied linguistics.

Аўтаматычная апрацоўка натуральнай мовы і маўлення (Natural Language Processing, NLP) – гэта вобласць даследавання і тэхналогій, звязаных з аналізам і інтэрпрэтацыяй чалавечай мовы з дапамогай камп'ютараў. Асноўнай задачай NLP з'яўляецца распрацоўка метадаў, пры дапамозе якіх камп'ютары могуць разумець, аналізаваць і генераваць чалавечую мову. Гэта тэхналогія размешчана на стыку камп'ютарных навук, метадаў штучнага інтэлекту, машыннага навучання і лінгвістыкі.

Натуральная апрацоўка тэксту даследуе пытанні ўзаемадзеяння паміж камп'ютарамі і людзьмі, спрыяючы аўтаматызацыі і паляпшэнню працэсаў апрацоўкі і аналізу тэкставай і гукавой інфармацыі. З развіццём штучнага інтэлекту і машыннага навучання, метады NLP становяцца ўсё больш дакладнымі і эфектыўнымі, што спрашчае працу камп'ютарных лінгвістаў, філолагаў і ўсіх зацікаўленых. Камп'ютарная перадапрацоўка тэксту як адзін з важных крокаў NLP адказвае за працэс апрацоўкі і падрыхтоўкі тэкставых даных перад іх аналізам або выкарыстаннем у розных інфармацыйных тэхналогіях. Гэты этап адказвае за забеспячэнне якаснай працы алгарытмаў апрацоўкі тэксту, менавіта таму, што дапамагае ліквідаваць шумы, лішнія сімвалы, структураваць тэкст і прывесці яго да зручнага фармату, прыдатнага да камп'ютарнага аналізу і мадэлявання ў розных сферах прымянення. Лабараторыя распазнавання і сінтэзу маўлення АІПІ НАН Беларусі [1] распрацавала платформу для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў corpus.by [2], якая прадстаўляе карыстальніку набор інструментальных сродкаў (сэрвісаў) для рэалізацыі аўтаматычнай апрацоўкі тэксту, маўлення і іншых даных.

Распрацаваныя сэрвісы платформы групуюцца ў тэматычныя дамены для больш зручнага выкарыстання ў канкрэтных практычных сферах. Яны забяспечваюць просты і ўстойлівы доступ да сродкаў і інструментаў апрацоўкі электроннага тэксту для аналізу, выяўлення, даследавання або аб'яднання набораў даных на беларускай, рускай і англійскай мовах [3]. У дадзеным артыкуле разгледзім набор інструментаў, якія будуць карысныя для перадапрацоўкі тэкставай інфармацыі. Гэта *«Генератар інфармацыі аб сімвалах», «Падлік частотнасці слоў», «Вызначэнне часцін мовы», «Такенізатар», «Лематызатар».*

Сэрвіс *«Генератар інфармацыі аб сімвалах»* дае магчымасць атрымаць назвы сімвалаў кадыроўкі Windows-1251 (стандартнай 8-бітнай кадыроўкі для Microsoft Windows). Мэтай дадзенага сэрвіса з'яўляецца вырашэнне праблемы агучвання тэксту, у якім сустракаюцца незнаёмыя сінтэзатару маўлення сімвалы [4]. Сэрвіс таксама дапамагае разабрацца з парадкам ужывання сімвалаў у тэксце.

На ўваход сэрвісу падаецца электронны тэкст ці любая адвольная паслядоўнасць электронных сімвалаў. Пасля апрацоўкі ўведзеных даных карыстальнік атрымлівае спіс назваў сімвалаў па парадку іх знаходжання ва ўваходным тэксце. Напрыклад, можна высветліць, на якім сімвалам у слоўніку ці транскрыпцыі пастаўлены націск, калі ён адлюстроўваецца незразумела ў нейкім рэдактары. У будучым функцыянал вырашэння гэтай задачы будзе ўбудаваны ў сэрвіс «*Падлік частотнасці сімвалаў*», які на цяперашні час адлюстроўвае статыстыку і кантэкст ужывання сімвала ў тэксце.

Сэрвіс *«Вызначэнне часцін мовы»* дазваляе карыстальніку даведацца, да якой часціны мовы належыць пэўнае слова ў рэжыме анлайн. Сэрвіс апрацоўвае тэксты на беларускай ці рускай мове, пасля чаго выдае карыстальніку спіс слоў, у якім пазначана, да якой часціны мовы адносіцца кожнае слова. Згодна з канцэпцыяй сэрвісу, часціна мовы можа быць вылучана толькі на падставе сукупнасці пэўных крытэрыяў. Увага надаецца наступным фактарам пэўнай адзінкі: што яна звычайна абазначае (прадмет, дзеянне, якасць і г. д.); у якіх граматычных формах яна можа ўзнікаць; якія характэрныя словаўтваральныя сродкі яна мае; якія функцыі яна выконвае ў сказе. «Вызначэнне часцін мовы» выкарыстоўвае шэраг слоўнікаў, якія карыстальнік абірае самастойна ці адразу ўвесь набор па змоўчванні.

Аўтаматычнае вызначэнне часцін мовы з’яўляецца важным інструментам у камп’ютарнай лінгвістыцы для дакладнага разумення значэнняў слоў у сказе. Гэта дае магчымасць больш эфектыўна аналізаваць тэкст, што палягчае розныя задачы апрацоўкі натуральнай мовы, такія як машынны пераклад, вылучэнне ключавой інфармацыі і семантычны аналіз. Акрамя таго, ведаючы часціны мовы ў сказе, можна даследаваць яго структуру і сінтаксічныя адносіны паміж словамі. Гэта карысна для пабудовы дрэў сінтаксічнага аналізу і разумення ўзроўню граматычнай складанасці сказа. Таксама правільнае вызначэнне часцін мовы дапамагае пошукавым сістэмам больш дакладна апрацоўваць запыты карыстальнікаў і прадастаўляць больш рэлевантныя вынікі пошуку.

Сэрвіс *«Падлік частотнасці слоў»* вырашае задачу па атрыманні статыстыкі ўжывання адвольных сімвальных паслядоўнасцей у электронным тэксце. Ён дазваляе апрацоўваць тэксты на многіх натуральных і штучных сімвальных мовах (калі сімвалы алфавіта мовы ўведзены ў адмысловае поле). У выпадку працы з беларускай мовай сэрвіс вызначае літары *У* і *Ў* як звычайныя асобныя сімвалы, таму, напрыклад, слова *«ўзвышша»* ў кантэкстах *«пад узвышшам»* і *«на ўзвышшы»* будзе вызначана як два асобныя словаўжыванні. На дадзены момант праграма вызначае толькі абсалютную колькасць ужыванняў слоў.

Сэрвіс будзе карысны:

– для лексікографаў, якія складаюць частотныя слоўнікі;

– для лінгвістаў, якія займаюцца вызначэннем славеснага каркаса пэўнай мовы і даследаваннямі ў галіне лінгвістычнай тыпалогіі на падставе аналізу пісьмовых крыніц;

– для спецыялістаў, хто вычытвае тэксты (знаходжанне памылковых ужыванняў слоў паводле шаблона памылковага ўжывання);

– для патрэб стылістычнай карэктуры тэкстаў – вызначэння частага ўжывання пэўнага слова (і наступнай ручной замены яго сінанімічным) і знаходжання «слоў-паразітаў».

Сэрвіс *«Такенізатар»* прызначаны для вылучэння ў тэксце токенаў. Такенізацыя (*англ. tokenizing, лексічны аналіз*) – працэс аналітычнага разбору ўваходнай паслядоўнасці знакаў на распазнаныя групы – лексемы, з мэтай атрымання на выхадзе ідэнтыфікаваных паслядоўнасцяў, так званых «токенаў». *Токен* (лексічны аналіз) – паслядоўнасць знакаў у лексічным аналізе ў інфарматыцы, адпаведны лексеме. Аб'ект, які ствараецца з лексемы ў працэсе лексічнага аналізу (такенізацыі). Лексічны аналіз выкарыстоўваецца ў кампілятарах і інтэрпрэтатарах зыходнага коду моў праграмавання, і ў розных парсерах слоў натуральных моў.

Аўтаматычная такенізацыя тэксту садзейнічае выкананню шэрагу задач NLP:

1. Падзел тэксту на асобныя элементы: такенізацыя дазваляе разбіць тэкст на асобныя токены, такія як словы, лічбы, знакі прыпынку і іншыя элементы. Гэта дапамагае камп'ютару лепш разумець структуру тэксту і працаваць з ім больш эфектыўна.

2. Падрыхтоўка даных для аналізу: такенізацыя рыхтуе тэкст для далейшага аналізу, менавіта вылучэнне ключавых слоў, вызначэнне тональнасці, сінтаксічны аналіз і многія іншыя задачы ў галіне апрацоўкі натуральнай мовы.

3. Забеспячэнне дакладнасці: правільная такенізацыя дапамагае пазбегнуць памылак у інтэрпрэтацыі тэксту камп'ютарнымі праграмамі. Няправільна падзелены тэкст можа прывесці да няправільнага разумення сэнсу сказаў і скажэння вынікаў аналізу.

4. Лематызацыя і стэмінг: для многіх метадаў аналізу натуральнай мовы, такіх як лематызацыя або стэмінг, неабходныя правільна такенізаваныя тэкставыя даныя.

5. Паляпшэнне якасці машыннага навучання: такенізаваны тэкст дае камп'ютарным мадэлям больш структураваную інфармацыю, што спрыяе працы алгарытмаў машыннага навучання і павышэнню іх эфектыўнасці.

Сэрвіс «*Лематызатар*» прызначаны для вызначэння пачатковых форм слоў [5]. На ўваход сэрвісу падаецца адвольны тэкст на беларускай або рускай мове. Вынікам працы сэрвіса з'яўляецца спіс слоў уваходнага тэксту з іх пачатковымі формамі, а таксама спіс слоў, пачатковую форму якіх не ўдалося вызначыць. Агульны выгляд, у якім будзе прадстаўлены вынік, можа быць наладжаны згодна з патрэбамі карыстальніка.

Метад лематызацыі прымяняецца ў пошукавых алгарытмах у працэсе схематызацыі вэб-дакументаў, а таксама пры іх індэксіраванні. Нягледзячы на высокі тэхналагічны ўзровень сучасных пошукавых сістэм, падобная апрацоўка не заўсёды бывае дакладнай, паколькі пошукавы робат часта ўлічвае толькі адну з магчымых лем словаформы, прыведзенай у тэксце дакумента. Таму далейшае развіццё метадаў лематызацыі, чаму сэрвіс прызваны паспрыяць, з'яўляецца прыярытэтнай тэхналагічнай задачай.

Выкарыстанне лематызацыі значна паляпшае якасць аналізу сайтаў і дакументаў. Калі ў камерцыйнай сферы (напрыклад, для палягчэння знаходжання тавараў і паслуг у інтэрнэце, а таксама для іх прасоўвання) якасная лематызацыя будзе адной з «радавых» пераваг, то пры апрацоўцы дакументаў медыцынскага, юрыдычнага, розных тэхналагічных даменаў дадзеная акалічнасць крытычна важная. Сістэмы лематызацыі рускіх тэкстаў распрацаваныя на сённяшні дзень дастаткова добра, у той час як для беларускай мовы сітуацыя выглядае інакш. Многія працы беларускіх вучоных даступныя чытачам на беларускай мове. Правільная лематызацыя як асноўных тэкстаў, так і дапаможных даных (назваў артыкулаў, звестак пра аўтараў, спісаў літаратуры) можа быць паспяхова прыменена ў дзейнасці бібліятэк.

Прыватнымі выпадкамі прымянення лематызацыі могуць быць крыміналістычная лінгвістычная экспертыза, аналіз тэкстаў на прадмет плагіяту, аналіз мовы тэкстаў пісьменніка, аналіз электронных вучэбных тэкстаў і электронных тэкстаў, падрыхтаваных навучэнцамі, у сістэмах адаптыўнага навучання.

У камп'ютарнай лінгвістыцы лематызацыя часта вызначаецца як метад марфалагічнага аналізу, у працэсе якога ад лексемы павінны быць адкінутыя ўсе флектыўныя элементы, якія не адпавядаюць пачатковай форме слова. Для атрымання дапаможных даных, у прыватнасці для вызначэння стандартнай структуры пачатковай формы слоў пэўнай часціны мовы, сістэма лематызацыі можа выкарыстоўваць пошук па слоўніку.

Такім чынам, прымяненне камп'ютарных сродкаў апрацоўкі натуранай мовы вельмі шырокае: ад аўтаматычнага перакладу і рэдагавання тэксту да аналізу сацыяльных медыя і распрацоўкі чат-ботаў. Правільна

праведзеная перадапрацоўка тэкставай і гукавой інфармацыі дазваляе палепшыць якасць і вынікі іх аналізу машыннымі алгарытмамі. Разгледжаныя сэрвісы інтэрнэт-платформы corpus.by з‘яўляюцца карыснымі і эфектыўнымі інструментамі аўтаматычнай перадапрацоўкі тэкстаў на беларускай, рускай і англійскай мовах. Паасобку кожны сэрвіс дае магчымасць вырашыць пэўную камп’ютарна-лінгвістычную задачу, а ў сукупнасці яны дазваляюць атрымаць якасны вынік комплекснага аналізу электроннага тэксту. Укараненне беларускай мовы ў інфармацыйныя тэхналогіі, стварэнне электронных слоўнікаў і новых праграм для апрацоўкі менавіта беларускай мовы на сённяшні дзень з’яўляецца актуальнай задачай і не страціць сваёй актуальнасці дзякуючы пастаяннаму пашырэнню ролі камп’ютарных тэхналогій у сучасным асяроддзі.

Бібліяграфічныя спасылкі

1. Лабараторыя распазнавання і сінтэзу маўлення [Электронны рэсурс]. URL: http://ssrlab.by/ (дата зваротку: 11.12.2023).

2. Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў Corpus.by [Электронны рэсурс]. URL: http: //www.Corpus.by/ (дата зваротку: 02.02.2024).

3. *Дзенісюк Д. А., Зяноўка Я. С., Драгун А. Я*. Платформа для апрацоўкі тэкставай і гукавой інфармацыі для розных тэматычных даменаў беларускай мовы // Языковая личность и эффективная коммуникация в современном поликультурном мире : материалы VI Междунар. науч.-практ. конф., посвящ. 100-летию Белорус. гос. ун-та (29–30 окт. 2020 г.) / редкол.: С. В. Воробьева (гл. ред.) [и др.]. Минск : БГУ, 2020. С. 69–74.

4. *Гецэвіч Ю.С., Дыдо В. В., Бяляўскі Д. А*. Тэхналогіі аўтаматычнай апрацоўкі і аналізу маўлення з прымяненнем штучнага інтэлекту // II Форум IT-Академграда «Искусственный интеллект в Беларуси» : доклады (12–13 октября 2023 г.). Минск : ОИПИ НАН Беларуси, 2023. С. 71–78.

5. *Дзенісюк Д. А., Маеўскі С. С., Зяноўка Я. С., Гецэвіч Ю. С.* Аўтаматызаваная лематызацыя тэкставай інфармацыі беларускай і рускай моў // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2020) : доклады XIX Международной конференции (19 ноября 2020 г.) / под науч. ред. А. В. Тузикова, Р. Б. Григянца, В. Н. Венгерова. Минск : ОИПИ НАН Беларуси, 2020. С. 246–252.