OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК [004.522+004.934+004.91]:004.89

АЛГАРЫТМ І ЛІНГВІСТЫЧНЫЯ РЭСУРСЫ ДЛЯ НАРМАЛІЗАЦЫІ ТЭКСТАЎ ГЕАГРАФІЧНАГА ДАМЕНА

Гецэвіч Ю.С., Качан Я.С., Лысы С.І., Маракуліна П.А., Крывальцэвіч А.В.

Аб'яднаны інстытут праблем інфарматыкі НАН Беларусі, Мінск, Рэспубліка Беларусь

yury.hetsevich@gmail.com

evgeniakacan@gmail.com

stanislau.lysy@gmail.com

marakulina.polina@gmail.com

elena.krivaltsevich@gmail.com

У дадзеным артыкуле апісваецца алгарытм лінгвістычнай апрацоўкі і нармалізацыі тэкстаў геаграфічнага дамена на прыкладзе вучэбнага дапаможніка "Геаграфія Беларусі". Прыведзена паслядоўнасць крокаў вылучэння ўсіх катэгорый сімвалаў, лікаў і іншых ужыванняў, неабходных для апрацоўкі.
Ключавыя словы: лінгвістычная апрацоўка, нармалізацыя, адзінка вымярэння, семантычная катэгорыя**.**

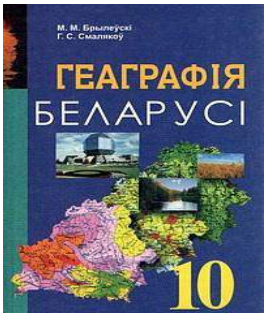
Уводзіны

Адной з задач сістэм сінтэзу і распазнавання маўлення з'яўляецца агучванне апрацаванага тэксту. На сённяшні дзень падобныя сістэмы дастаткова якасна ажыццяўляюць усе пастаўленыя перад імі мэты. Аднак, нягледзячы на яскравыя поспехі, у працэсе сінтэзу маўлення застаецца яшчэ адна нявырашаная праблема: успрыманне і агучванне незразумелых для машыны сімвалаў (лікі, скарачэнні, даты і г.д.) [Гецэвіч, 2014]. Ужо існуюць артыкулы паводле апісанай праблемы [Барадзіна, 2015]. Таму аўтары дадзенага артыкула, азнаёміўшыся з даследаваннямі папярэдніх навукоўцаў, лічаць важным распрацаваць праверэныя нармалізаваныя мноствы фраз і сказаў на прыкладзе геаграфічнага дамена для далейшага тэставання сістэм сінтэзу і распазнавання маўлення. У будучыні атрыманы спіс будзе пакладзены ў аснову аўтаматызаванага працэсу нармалізацыі ўсіх невядомых сімвалаў.

1. Збор матэрыялаў і методыка іх апрацоўкі

У якасці матэрыялаў даследавання выкарыстоўваліся вучэбныя дапаможнікі па геаграфіі з пятага па адзінаццаты клас. За аснову быў выбраны вучэбны дапаможнік "Геаграфія Беларусі" для дзясятага класа ўстаноў агульнай сярэдняй адукацыі з беларускай мовай навучання (малюнак 1). Дапаможнік складаецца з дзесяці частак: геаграфічнае становішча і даследаванні Беларусі; прыродныя ўмовы і рэсурсы Беларусі; геаграфічныя ландшафты, экалагічныя праблемы; прыроднае раяніраванне Беларусі; насельніцтва; агульная характарыстыка гаспадаркі; геаграфія вытворчай сферы; геаграфія невытворчай сферы; Беларусь у сусветнай супольнасці; вобласці Беларусі. Такім чынам, дадзеная кніга дае магчымасць сабраць матэрыял не толькі геаграфічнага дамена, але і скарачэнні і сімвалы іншых даменаў.

Увесь матэрыял перагледжаны і былі абраны па тры прыклады на кожную семантычную падгрупу.



Малюнак 1 – Вучэбны дапаможнік "Геаграфія Беларусі" для дзясятага класа ўстаноў агульнай сярэдняй адукацыі з беларускай мовай навучання

Библиотека БГУИР

2. Класіфікацыя сабраных матэрыялаў

Сабраны матэрыял быў афармлены як спіс сказаў з ненармалізаванымі дадзенымі з наступнымі семантычнымі катэгорыямі:

1) Адзінкі вымярэння:
   а) Маса (т, кг, г, мг).
   б) Тэмпература (градусаў па Цэльсію).
   в) Плошча (га, км², м², см²).
   г) Аб'ём (м³, см³).
   д) Вага (кг, г, мг).
   е) Час (г,ст,г, хв).
   ж) Даўжыня і адлегласць (км, м, см).
   з) Магутнасць (Вт, кВт, мВт).
   і) Напружанне (мм, см, м, м,км/с).
   к) Геаграфічныя каардынаты (пд/пн ш, у/з д, градусаў °, хвілін ').
   л) Шчыльнасць насельніцтва (чал/км², чал/м², тыс.чал/км²,млн. чал., тыс. чал.).
2) Працэнты, праміле.
3) Скарачэнні (інш., г.д., г., ст., стст.).
4) Лікі.
5) Матэматычныя знакі (-,+,/,=).
6) Абрэвіятуры.
7) Ініцыялы.
8) Даты.
9) Выпадковыя дадзеныя.
10) Статыстычныя дадзеныя.

Кожная з катэгорый таксама падзяляецца на падгрупы ў залежнасці ад кантэксту, у якім сустракаецца той ці іншы выраз. Напрыклад, *Больш за 20 прадпрыемстваў выкарыстоўваюць натуральныя і штучныя скуры*. Выраз *больш за* патрабуе ад залежнага слова вінавальнага склону. Таму словаспалучэнне *20 прадпрыемстваў* пішацца ў вінавальным склоне.

Такім чынам, на гэтым этапе атрымаўся наступны спіс (табліца 1):

Табліца 1 – Першапачатковы спіс сабранага матэрыялу па семантычных катэгорыях (фрагмент)

Адзінкі вымярэння	Разрад	Прыклад
Плошча	км/м²	Яна займае плошчу 207,6 тыс. км².
Аб'ём	Тыс/млн м³	Сумарны памер лесакарыстання можа скласці больш за 19 млн м³ драўніны.
Маса	Тыс/млн т	У 2010 г. у рэспубліцы было выраблена 4,5 млн цэменту (для параўнання: у 1913 г.33 тыс. т).

3. Экспертная нармалізацыя прадстаўленых семантычных класаў

Большасць ненармалізаваных дадзеных складаюць спалучэнні лікаў з адзінкамі вымярэння. Таму першапачаткова неабходна надаць увагу лікам: перавесці іх у колькасныя ці парадкавыя лічэбнікі (табліца 2). Ужыванне ўсіх лічэбнікаў залежыць ад папярэдняга прыназоўніка ці спалучэння слоў. Так, напрыклад, такія словы як *амаль, прыкладна, складае і інш.* патрабуюць пасля сябе лічэбнікі ў назоўным ці вінавальным склонах. Ніжэй прыведзена табліца прыкладаў прыназоўнікаў з улікам склонаў ужывання лічэбнікаў з іх залежнымі словамі (табліца 3).

Табліца 2 – Пераўтварэнне лікаў у колькасныя лічэбнікі

Запіс лікаў (шаблонамі)	Колькасныя лічэбнікі ў наз. скл.	Канчаткі адзінак вымярэння			
		Мужчынскага роду		Жаночага роду	
1	Адзін/адна	Нулявы канчатак	мільён, працэнт	-а	тысяча
2-4	два, тры, чатыры	-а	мільёна, працэнта	-ы	тысячы
5-20	пяць…дваццаць	-аў	мільёнаў, працэнтаў	нулявы канчатак	тысяч
[2-9]1	дваццаць адзін/адна – дзевяноста адзін/адна	Нулявы канчатак	мільён, працэнт	-а	тысяча
[2-9][2-4]	дваццаць дзве – дзевяноста чатыры	-а	мільёна, працэнта	-ы	тысячы
[2-9][5-9]	Дваццаць пяць тысяч – дзевяноста тысяч	-аў	мільёнаў, працэнтаў	нулявы канчатак	тысяч
[1-9]01	Сто адзін/адна – дзевяцьсот адзін/адна	Нулявы канчатак	мільён, працэнт	-а	тысяча
[1-9]0[2-4]	Сто два/дзве – дзевяцьсот чатыры	-а	мільёна, працэнта	-ы	тысячы
[1-9]0[5-9]	Сто пяць – дзевяцьсот дзевяноста дзевяць	-аў	мільёнаў, працэнтаў	нулявы канчатак	тысяч

Бібліятэка БГУІР

Табліца 3 – Спіс канчаткаў адзінак вымярэння ў спалучэнні з колькаснымі лічэбнікамі

Прыназоўнікі	Склон ужывання лічэбнікаў	Канчаткі адзінак вымярэння			
		Канчаткі мужчынскага роду		Канчаткі жаночага роду (тысячы)	
		мн. л.	адз. л.	мн. л.	адз. л.
Empty	N, V (назоўны, вінавальны)	нул. к.	-аў/оў	-у	нул. к.
у/ў	Р (месны)	-е	-ах	-ы	-ах
Па/праз	V (вінавальны)	-ы	-оў	-ы	нул. к.
на	Т (творны)	-е	-ах	-ы	-ах
з-за/з/каля/да/ад	R (родны)	-а	-аў/оў	-ы	нул. к.

Акрамя лічэбнікаў у працэсе нармалізацыі прыкладаў аўтары таксама сустрэліся з наступнымі пытаннямі:

1. Вызначэнне прынцыпу нармалізацыі абрэвіятур. Усе абрэвіятуры падзелены на дзве групы: тыя абрэвіятуры, што ўтвораны з пачатковых літар элементаў зыходнага словазлучэння і чытаюцца не па алфавітных назвах літар, а як звычайнае слова, называюцца акронімамі (*ЮНЭСКА, НАТА, ААН і інш.*); абрэвіятуры, якія ўтвораны часткова з назваў пачатковых літар, часткова з пачатковых гукаў слоў зыходнага словазлучэння, называюцца літарна-гукавымі (*СССР, ВУП, ВКЛ і інш.*). Складанаскарочаныя абрэвіятуры нармалізуюцца як звычайныя словы (*ЛітБел, БелАз і інш.*).

2. Нармалізацыя дробных лікаў. Такія лікі нармалізуюцца ў залежнасці ад кантэксту, а менавіта ад папярэдняга прыназоўніка (*каля 1/3 (адной трэцяй) насельніцтва, на 1/3 (адну трэцюю) частку насельніцтва)*.

3. Вызначэнне прынцыпу нармалізацыі скарачэнняў. З тымі скарачэннямі, які маюць зафіксаваны выгляд, не ўзнікла пытанняў. Усе яны маюць адну словаформу: *г. д. (гэтак далей), т. п. (таму падобнае*) і інш. Цяжэй апрацаваць скарачэнні, форма якіх залежыць ад азначаемага слова ці спалучэння слоў (*гг., ст., стст., р. і інш.*). Яшчэ больш пытанняў узнікае, калі скарачэнне можа прымаць розныя значэнні. Напрыклад, скарачэнне *г.* можа перакладацца як *горад, гара і год*. У такіх выпадках усё залежыць ад кантэксту.

4. Нармалізацыя адзінак вымярэння пасля дробных лікаў. Усе адзінкі вымярэння апрацоўваюцца ў залежнасці ад дробнай часткі лічэбніка (*55,3% - пяцьдзясят пяць цэлых тры дзясятыя працэнта*).

5. Нармалізацыя дадзеных з прамежкамі. Яшчэ адной складанасцю з'яўляюцца адзінкі вымярэння ў нейкі пэўны перыяд. Напрыклад, тыя дадзеныя, перад якімі ўжываецца прыназоўнік, расшыфроўваюцца ў неабходным склоне праз коску*: у 2-3 (два, тры) разы*. Тыя лікі, што ўжываюцца ў назоўным ці вінавальным склоне, таксама ў круглых дужках, нармалізуюцца з дапамогай спалучэння прыназоўнікаў *з…па, ад …да* (*складаюць 200-1000 чалавек – складаюць ад ста да тысячы чалавек*).

Такім чынам, атрымаўся спіс нармалізаваных выразаў геаграфічнага дамена з вызначэннем семантычных катэгорый і іх груп.

4.Распрацоўка алгарытму лінгвістычнай апрацоўкі і нармалізацыі тэкстаў на прыкладзе семантычнай групы "Насельніцтва"

У выніку праведзеных даследаванняў, аўтары артыкула распрацавалі наступны алгарытм для нармалізацыі дадзеных семантычнай катэгорыі "Шчыльнасць насельніцтва". Алгарытм складаецца з наступных крокаў:

1. Ідэнтыфікаваць адзінкі вымярэння шчыльнасці насельніцтва [Гецэвіч, 2012]. Першапачатковыя паказчыкі семантычнай групы: чалавек, людзі ў спалучэнні з такімі адзінкамі вымярэння, як *тысячы*, *мільёны*, *тысяч(ы) на квадратны метр/кіламетр*.

2. Атрымаць структурныя часткі, якія складаюцца з бліжэйшага папярэдняга кантэксту (адзін-два словы), ліка(ў) і адзінак вымярэння насельніцтва ў спалучэнні са словамі чалавек(а) ці людзі(ей).

3. Ідэнтыфікаваць вызначаны кантэкст: ці з'яўляецца гэта прыназоўнікам, дзеясловам, ці гэта круглыя дужкі:

а) калі гэта прыназоўнік (на, каля, з, у і г.д), лік пераўтвараецца ў колькасны лічэбнік у залежнасці ад склона, які патрабуе дадзены прыназоўнік (глядзіце табліцу 3);

б) калі гэта дзеяслоў (*складае, налічвае*), прыслоўе (*амаль, прыблізна*), то лік пераўтвараецца ў колькасны лічэбнік у назоўным ці вінавальным склоне (глядзіце табліцу 3);

в) калі структурная частка знаходзіцца ў круглых дужках, лік пераўтвараецца ў колькасны лічэбнік у назоўным склоне;

г) калі ў дужках знаходзяцца два лікі, напісаныя праз працяжнік без папярэдняга кантэксту, то ў працэсе нармалізацыі трэба выкарыстоўваць такія спалучэнні прыназоўнікаў, як *ад… да…* (лічэбнікі апрацоўваюцца ў родным

Библиотека БГУИР

склоне), *з… па…* (адпаведна ў родным і вінавальным склонах) [Hetsevich, 2013].

4. Згенераваць колькасныя лічэбнікі ў адпаведнасці з неабходным склонам (глядзіце табліцу 3).

5. Вызначыць, да якога разраду адносяцца колькасныя лічэбнікі: да цэлых ці дробных састаўных лічэбнікаў. Калі гэта цэлы лічэбнік, адзінкі вымярэння апрацоўваюцца ў залежнасці ад склону, у якім ужываецца лічэбнік. Калі гэта дробны лічэбнік, адзінкі вымярэння апрацоўваюцца ў залежнасці ад дробнай часткі лічэбніка ў неабходным склоне (глядзіце табліцу 2).

6. Згенераваць атрыманыя вынікі і ўнесці ў сказ нармалізаваныя выразы.

Заключэнне

У дадзеным артыкуле аўтары апісалі падрабязны алгарытм ручнога пошуку і апрацоўкі выразаў геаграфічнага дамена, якія патрабуюць нармалізацыі. На дадзены момант спіс неабходны для тэставання якасці працы сістэм распазнавання і сінтэзу маўлення, але ў далейшым будзе пакладзены ў аснову вырашэння задачы аўтаматызаванага папаўнення базы дадзеных для любога дамена.

Бібліяграфічны спіс

[Барадзіна, 2015] Барадзіна, Ю.С. Апрацоўка колькасных выразаў з адзінкамі вымярэння: ад навукова-тэхнічнага тэксту да тэлеметрыі / Ю.С. Барадзіна, Ю.С. Гецэвіч // Мова і літаратура ў XXI стагоддзі: актуальныя аспекты даследавання : матэрыялы III Рэсп. навук.-практ. канф. маладых навукоўцаў / БДУ ; пад рэд. П. І. Навойчык. — Мінск : Бел. дзярж. ун-т., 2015. — С. 7-12.

[Гецэвіч, 2014] Гецэвіч, Ю.С. Лінгвістычныя рэсурсы для пераўтварэння колькасных выразаў з адзінкамі вымярэння тыпу “лічба-сімвал” у словазлучэнні для беларускай і рускай моў / Ю.С. Гецэвіч, А.М. Скопінава // Карповские научные чтения, выпуск 8 : сб. науч. ст. : в 2 ч. / Бел. гос. ун-т ; редкол. : А.И. Головня (отв. ред.) [и др.]. – Минск : “ИВЦ Минфина”, 2014. – Ч. 1. – С. 236-240.

[Hetsevich, 2013] Hetsevich, Yu. Identification of Expressions with Units of Measurement in Scientific, Technical & Legal Texts in Belarusian and Russian // Yu. Hetsevich, A. Skopinava // Proceedings of the Workshop on Integrating IR technologies for Professional Search [Electronic resource]. – 2013. Mode of access : http://ceur-ws.org/Vol-968/irps_6.pdf. – Date of access : 24.03.2013

[Гецэвіч, 2012] Гецэвіч, Ю.С. Ідэнтыфікацыя выразаў з адзінкамі вымярэння ў навукова-тэхнічных і прававых тэкстах на беларускай і рускай мовах / Ю.С. Гецэвіч, А.М. Скопінава // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2012) : доклады XI Международной конференции (Минск, 15 ноября 2012 г.). – Минск : ОИПИ НАН Беларуси, 2012. – С. 260–265.

ALGORYTHM AND LINGUISTIC RESOURCES FOR TEXT NORMALIZATION OF GEOGRAPHIC DOMAIN

Hetsevich Y.S.*, Lysy S.I.*, Kachan E.S.*, Marakulina P.A*, Krivaltsevich A.V.*

** United Institute of Informatics Problems, National Academy of Sciences, Minsk, Belarus*

yury.hetsevich@gmail.com

evgeniakacan@gmail.com

stanislau.lysy@gmail.com

marakulina.polina@gmail.com

elena.krivaltsevich@gmail.com

Introduction

This article covers the problem of linguistic processing and text normalization of geographic domain. It introduces steps of symbols categorization, numbers and other cases for text processing.

The problem is that not all characters in text can be perceived and vocalized with automatic algorithms.

Main Part

Materials of Geography course books in the Belarusian language, in terms of “Geography of Belarus” for 10th form of secondary school were studied and analyzed. All materials were collected in a non-systematic fashion with three examples on each semantic subgroup.

Next categories were extracted: measurement units, percent, permille, acronyms, numbers, mathematical characters, abbreviations, initials, date, random data, statistical data.

Also the algorithm of categorization of characters, numbers and other cases for text processing was proposed.

Conclusion

In this article, authors explained in detail the algorithm of manual search and processing of expressions in geographical domain for the purpose of normalization.

All this materials should be used as a basis for automatic appending of database for any domain in future.

Библиотека БГУИР