Ю.С. Гецэвіч, А.М. Скопінава (Мінск, АІПІ НАН Беларусі)

ЛІНГВІСТЫЧНЫЯ РЭСУРСЫ ДЛЯ ПЕРАЎТВАРЭННЯ КОЛЬКАСНЫХ ВЫРАЗАЎ З АДЗІНКАМІ ВЫМЯРЭННЯ ТЫПУ “ЛІЧБА–СІМВАЛ» У СЛОВАЗЛУЧЭННІ ДЛЯ БЕЛАРУСКАЙ І РУСКАЙ МОЎ

Апрацоўка натуральнай мовы з’яўляецца адным з самых актуальных навукова-даследчых накірункаў XXI стагоддзя. Ён прадугледжвае вырашэнне розных камп’ютэрна-лінгвістычных задач [1, с. 333], адной з якіх можна назваць апрацоўку складана ці спецыфічна структураванай інфармацыі ў электронных тэкстах: табліц, формул, схем, спасылак, зносак і г.д. У дадзеным дакладзе аўтары канцэнтруюцца на колькасных выразах з адзінкамі вымярэння (КВАВ) — спалучэннях колькасных паказчыкаў (перададзеных на пісьме пасродкам лічбаў) і пазначэнняў мерных адзінак (літарных сімвалаў), напрыклад: *1,72 г/см³*, *19640 км*, *55°*, *6,387 сутак*, *1212±16 км*, *1,9×10 ²¹ кг* і г.д. Будову КВАВ можна ўявіць фармальна, што адлюстравана ў табліцы 1. У ёй пад *X* маецца на ўвазе колькасны паказчык (лік), а пад *Y* — сімвалы літар; знак вертыкальнай рысы размяжоўвае прыклады.

Табліца 1

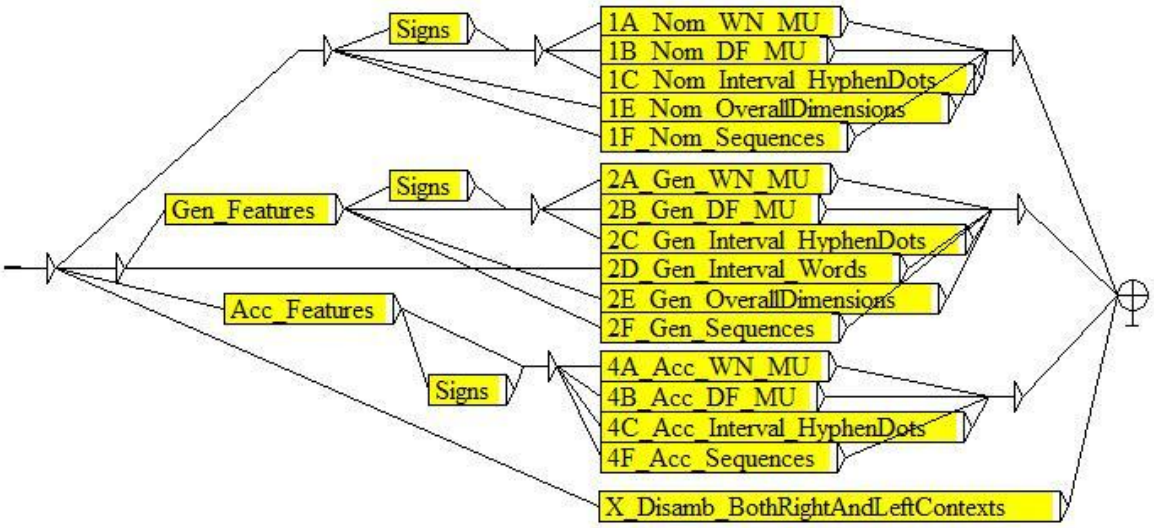
Мадэлі фармальнага ўяўлення КВАВ

№	Мадэль	Прыклад
1.	X Y	12 % | 40-50 тыс. м
2.	X Y/Y	0,5144444 м/с
3.	X-[… ..]X Y	1-1,5 года | +13... +19 °C
4.	X-[… ..]X Y/Y	0,1-5,7·10⁻² м/с
5.	~[+-±><]X Y	±0,3° | > 6 Зв
6.	~[+-±<>]X Y/Y	~107 К/с | ~9,8 м/с²
7.	X,[i] X Y	2 і 4 метры
8.	X, X, X Y	5, 6, 7 шт.
9.	X Y — X Y	0,1 Гц — 300 кГц
10.	X × X Y	1136×640 пікселяў
11.	X × X × XY	146,8 × 75,3 × 8,9 мм
…	…	…

Неабходнасць правільнай апрацоўкі падобных лічба-сімвальных канструкцый актуальная і для навуковай, і для бытавой сферы жыцця, напрыклад, для дадзеных ад штучных спадарожнікаў і касмічных зондаў; медыцынскіх аналізаў (тэмпература, крывяны ціск, пульс, цукар, халестэрын, гемаглабін…); навуковых даследаванняў; кулінарных рэцэптаў; прагнозаў надвор’я; этыкетак на спажывецкіх таварах; апісанняў тавараў у анлайн-крамах; каментарыяў да спартыўных мерапрыемстваў; турыстычных і іншых даведнікаў і г.д.

У працах [2; 3] аўтарамі былі прапанаваныя алгарытмы-рашэнні апрацоўкі КВАВ у выглядзе канчатковых аўтаматаў, якія дазваляюць

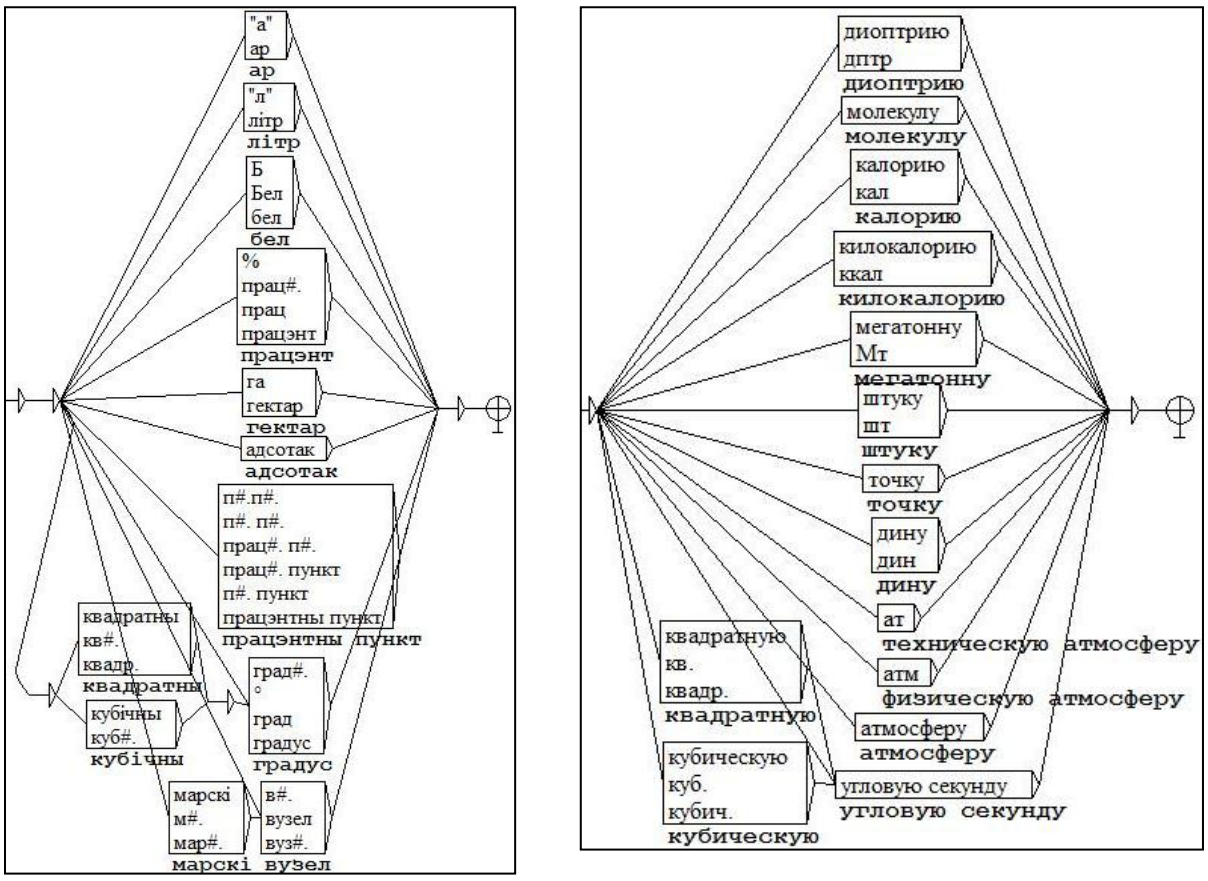
эмуляваць працу марфалагічных і сінтаксічных граматык. Для іх стварэння быў выкарыстаны наладжвальны лінгвістычны працэсар NooJ. На малюнку 1 прадстаўлена апошняя мадыфікацыя алгарытму, які зараз апрацоўвае КВАВ для беларускай і рускай моў ужо ў трох склонах: назоўным, родным і вінавальным (раней гаворка вялася толькі пра назоўны). Акрамя гэтага, будова дадзенага алгарытму заснаваная ўжо на мадэлях будовы саміх КВАВ, напрыклад, галіна з графам *1E_Nom_OverallDimensions* спрацуе для мадэляў КВАВ пад нумарамі 10 і 11 у табліцы 1.



Мал. 1. Агульны выгляд алгарытму для апрацоўкі КВАВ

Апрацоўка КВАВ прадугледжвае іх папярэднюю ідэнтыфікацыю і аналіз. Пад ідэнтыфікацыяй будзем разумець непасрэдны пошук у электронных тэкстах колькасных паказчыкаў (лічбаў), пазначэнняў адзінак (сімвалаў), а таксама спецыфічнага акружэння, якое можа ўплываць на граматычныя катэгорыі ідэнтыфікаванага КВАВ, напрыклад: *5 м — 5 метраў* (назоўны/вінавальны склон), *каля 5 м — каля пяці метраў* (родны склон), *роўны 5 м — роўны пяці метрам* (давальны склон) і г.д. Вызначэнне такога ўплыву на КВАВ з боку акружэння і з'яўляецца аналізам. Таксама важна прааналізаваць і элементы КВАВ: прыналежнасць паказчыка колькасці да цэлых ці дзесятковых лікаў, са знакамі ці без знакаў; прыналежнасць ідэнтыфікаванай адзінкі вымярэння да Сістэмы Інтэрнацыянальнай (СІ), распрацаванай Міжнародным бюро мер і вагаў [4]. Нарэшце апрацоўка КВАВ заключаецца ў іх пераўтварэнні ў словы і канчатковым "склейванні» ўсіх элементаў у граматычна правільныя словазлучэнні.

Трэба адзначыць, што нягледзячы на машыннае паходжанне алгарытмаў, іх карэктная праца немагчымая без адпаведных лінгвістычных рэсурсаў, пад якімі мы маем на ўвазе наборы разгортак колькасных паказчыкаў (лічбаў) і пазначэнняў адзінак вымярэння (літарных сімвалаў) у цэлыя словы. Іх распрацоўка і была задачай дадзенага даклада (малюнак 2).



(а) ***(б)***

Мал. 2. Лінгвістычны рэсурс у выглядзе графа для апрацоўкі пазначэнняў дадатковых адзінак вымярэння для беларускай *(а)* і рускай *(б)* моў

Падкрэслім, што лінгвістычныя рэсурсы непасрэдна ўбудаваныя ў алгарытм у выглядзе асобных графаў, а не падключаныя, напрыклад, у якасці самастойных слоўнікаў. На дадзены момант апрацоўваецца 120 адзінак вымярэння ў розных відах запісу. Дзеля зручнасці стварэння і паўнаты рэсурсаў патрэбна было перш за ўсё скласці ліст з адзінкамі вымярэння і размежаваць адзінкі ў некалькі груп. За аснову была выкарыстаная класіфікацыя Міжнароднага бюро мер і вагаў: базавыя адзінкі СІ, вытворныя ад адзінак СІ і пазасістэмныя адзінкі. Пазней, па меры таго як алгарытм увесь час тэставаўся на беларуска- і рускамоўных тэкставых масівах навукова-тэхнічнай тэматыкі, спіс адзінак значна папоўніўся. Так, усе астатнія адзінкі, не апісаныя СІ, а таксама словы, якія ўмоўна можна лічыць адзінкамі вымярэння (напрыклад, *шт.* ад *штука*), увайшлі ў алгарытм асобным дадатковым класам Extra. На малюнку 2 як раз прадстаўлены выгляд гэтага графу менавіта для беларускай і рускай моў. Вынікі выкарыстання алгарытма дэманструюцца ў табліцы 2.

Табліца 2

Фрагменты вынікаў апрацоўкі алгарытмам беларуска- *(а)* і рускамоўнай *(б)* навукова-тэхнічнай тэкставай інфармацыі

Выгляд у тэксце	Пасля апрацоўкі алгарытмам
(а)	
займае 36,4 % **1136х640 пікселяў** **на 1,3 мегапікселя** **на 1,5-7 адсоткаў** **1430 мАг** **500-600 кв. метраў**	**займае трыццаць шэсць цэлых чатыры дзясятыя працэнта** **адна тысяча сто трыццаць шэсць на шэсцьсот сорак пікселя** **на адну цэлую тры дзясятыя мегапікселя** **на адну цэлую пяць дзясятых дэфіс сем адсоткаў** **адна тысяча чатырыста трыццаць міліампер-гадзін** **пяцьсот дэфіс шэсцьсот квадратных метраў**
(б)	
0,01 МДж **~ 0,1 МДж** **более 40 га** **20-210 мм** **+25° С** **свыше 2000 м**	**ноль целых одна сотая мегаджоуля** **около нуля целых одной десятой мегаджоуля** **более сорока гектаров** **двадцать дефис двести десять миллиметров** **плюс двадцать пять градусов Цельсия** **свыше двух тысяч метров**

У заключэнне падкрэслім: значнасць лінгвістычных рэсурсаў для камп’ютэрных рашэнняў задач, звязаных з апрацоўкай тэкставай інфармацыі, немагчыма пераацаніць. Чым лепш распрацаваныя рэсурсы, тым вышэйшае значэнне паўнаты ўсяго алгарытму — найбольш важнага паказчыка якасці алгарытму, акрамя дакладнасці. Паўната вылічваецца як вынік дзялення колькасці КВАВ, якія алгарытм правільна апрацаваў, на рэальную колькасць КВАВ ва ўсім тэкставым мностве, якую падлічыў эксперт. Зараз паўната алгарытму дасягае 75 %. У бліжэйшых планах аўтараў палепшыць гэтае значэнне праз тэставанне алгарытма і наступную дапрацоўку яго лінгвістычных рэсурсаў на дадатковым тэкставым мностве. У доўгатэрміновай перспектыве мае быць укараненне алгарытму ў сістэму сінтэзу маўлення па тэксце дзеля большай сэнсавай дакладнасці і правільнасці, а таксама для правільнай інтанацыйнай і прасадычнай афарбоўкі тэкстаў для агучвання.

ЛІТАРАТУРА

1. Гецэвіч, Ю.С. Метад пабудовы кампанентаў сінтэзу маўлення па тэксце для натуральна-маўленчага інтэрфейса пры дапамозе NooJ / Ю.С. Гецэвіч, А.М. Скопінава, Т.І. Окрут // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS–2014): материалы IV Междунар. науч.-техн. конф. (Минск, 20–22 февраля 2014 года) / редкол.: В.В. Голенков (отв. ред.) [и др.]. – Минск: БГУИР, 2014 г. – С. 333–338.

2. Hetsevich, Yu.S. Transforming quantitative expressions with measurement units into orthographical words for text-to-speech synthesis to Belarusian and Russian / Yu.S. Hetsevich, A.M. Skopinava // Вестник МГЛУ. Сер. 1, Филология. – 2013. – № 3. – С. 133–144.

3. Гецэвіч, Ю.С. Мадэляванне і распрацоўка сістэм пошуку колькасных выразаў з адзінкамі вымярэння ў электронных тэкстах на беларускай і рускай мовах / Ю.С. Гецэвіч, А.М. Скопінава,

А.Ф. Есіс // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2013): доклады XII Международной конференции (Минск, 20 ноября 2013 г.). – Минск: ОИПИ НАН Беларуси, 2013. – С. 282–287.

4. Апісанне СІ на сайце Міжнароднага бюро мер і вагаў [Электронны рэсурс]. – 2006. – Рэжым доступу: http://www.bipm.org/en/si/si_brochure/general.html. – Дата доступу: 15.04.2014.