Варановіч В.В. Непаўната машынных слоўнікаў: прычыны і спосабы змяншэння непаўнаты// Пятые чтения, посвященные памяти профессоров В.А. Карпова и С.М. Прохоровой (Минск, филологический факультет БГУ, 18-19 марта 2011 г.). Сборник научных статей. Мн., 2011, с.93-96.

ВАРАНОВІЧ В.В.

НЕПАЎНАТА МАШЫННЫХ СЛОЎНІКАЎ: ПРЫЧЫНЫ І СПОСАБЫ ЗМЯНШЭННЯ НЕПАЎНАТЫ

Адным з асноўных кампанентаў лінгвістычнай базы ведаў у сістэмах аўтаматычнай апрацоўкі тэксту (сістэмы машыннага перакладу, пошукавыя і экспертныя сістэмы і інш.), з'яўляецца машынны слоўнік. Напаўняльнасць і паўната слоўніка шмат у чым вызначаюць паспяховасць задач, якія вырашаюцца ў галіне аўтаматычнай апрацоўкі тэкстаў.

Важнай праблемай, якая стаіць перад распрацоўшчыкамі праграмных прадуктаў, апрацоўваючых натуральную мову, з'яўляецца аб'ём машыннага слоўніка. З аднаго боку, існуе імкненне максімальна ахапіць лексіку пэўнай мовы і размясціць яе ў слоўніку, што зніме праблему распазнавання незнаёмых слоў пры апрацоўцы тэкстаў. Аднак, як вядома, поўны ахоп лексікі практычна немагчымы, акрамя таго, існуе шэраг натуральных абмежаванняў на аб'ём слоўніка, якія накладаюцца самой моўнай сістэмай.

Адной з найбольш значных прычын непаўнаты любога машыннага слоўніка з'яўляецца наяўнасць вялікай колькасці ўласных імёнаў. Зафіксаваць абсалютна ўсе наяўныя антрапонімы і тапонімы практычна немагчыма: «Прадметаў, вартых прысваення індывідуальнага імя, так шмат, што імёны ўласныя знаходзяцца быццам бы за рамкамі асноўнага лексічнага складу моў» [1, с. 10]. Акрамя таго, існуе пэўная колькасць імёнаў, аманімічных агульным назоўнікам, што ўскладняе задачу апрацоўкі натуральна-моўнага тэксту, у прыватнасці, у сістэмах машыннага перакладу. Напрыклад, пры перакладзе з беларускай мовы на рускую назва горада *Чэрвень* павінна перакладацца як *Червень*, а назва месяца *чэрвень* – як *июнь*. Аднак некаторая (і дастаткова вялікая) колькасць найбольш частотных уласных імёнаў усё-такі павінна прысутнічаць у слоўніку, паколькі пры транслітараванні ўзнікае вялікая колькасць памылак

(напрыклад, назва горада *Несвиж* можа быць перададзена на беларускай мове і як *Несвіж*, хоць правільны варыянт – *Нясвіж*), а многія імёны перадаюцца на замежных мовах, зыходзячы з традыцыі, а не актуальных правілаў: так, горад *Вільнюс* у беларускамоўных тэкстах часцей называецца *Вільня*, хоць на сённяшні дзень нарматыўным з'яўляецца варыянт *Вільнюс*.

Яшчэ адзін значны пласт лексікі, праблемны для машынных слоўнікаў – дымінутывы і аўгментатывы. Аналіз корпусаў тэкстаў рускай і беларускай моў паказвае, што практычна любы прадметны назоўнік у гэтых мовах можа мець памяншальную ці павелічальную форму. У маўленні сустракаюцца нават такія экзатычныя ўтварэнні, як *скотчык* ці *каханнейка*. Па даным Нацыянальнага корпуса рускай мовы [2], каля 30 тыс. дакументаў з агульнага аб'ёму корпуса (каля 49 тыс. дакументаў) утрымліваюць такія словы. Аднак размяшчэнне дымінутываў і аўгментатываў у машынным слоўніку значна павялічыць яго аб'ём, а значыць, і ўскладніць апрацоўку. Паколькі ўтварэнне гэтых формаў, за рэдкім выключэннем, лёгка паддаецца апісанню, мэтазгодна апрацоўваць іх з дапамогай правілаў, а не машыннага слоўніка.

Існуюць таксама пэўныя абмежаванні на размяшчэнне ў машынных слоўніках устарэлых слоў і жарганізмаў. Паколькі выкарыстанне слоў гэтых лексічных пластоў характэрна толькі для тэкстаў мастацкага стылю, можна выкарыстоўваць асобны слоўнік устарэлай і зніжанай лексікі толькі пры апрацоўцы мастацкіх тэкстаў.

Практычна любы слоўнік будзе няпоўным з-за наяўнасці ў натуральных тэкстах неалагізмаў і аказіяналізмаў. Безумоўна, любы слоўнік – у той ці іншай ступені – дынамічны, але ўсё адно немагчыма зафіксаваць усе новыя словы, якія з’яўляюцца літаральна кожны дзень. Для пастаяннага папаўнення машыннага слоўніка неабходна праводзіць рэгулярны статыстычны аналіз новых тэкстаў з мэтай выяўлення новых слоў і далейшага папаўнення слоўніка. Прычым пытанне ўнясення ці неўнясення лексемы ў слоўнік павінна вырашацца ў кожным выпадку індывідуальна, паколькі некаторыя словы (напрыклад, *снукерыст* ці *смартфон*) безумоўна, маюць усе шанцы ўвайсці ў пласт актыўнай лексікі, а іншыя – толькі патэнцыяльную магчымасць (*свiтчар*, *зомбаскрынка*).

Яшчэ адно важнае пытанне, якое ўзнікае пры напаўненні слоўніка – размяшчэнне патэнцыяльных словаформаў, і ў першую чаргу, формаў множнага ліку для назоўнікаў Singularia tantum. Як вядома, дастаткова вялікая група лексікі беларускай і рускай мовы

(абстрактныя імёны, рэчыўныя, зборныя назоўнікі і інш.) не ўтварае формаў множнага ліку, аднак корпусны аналіз узусу паказвае, што вельмі многія такія словаформы тым не менш сустракаюцца ў натуральна-моўных тэкстах: *бензіны*, *любові*, *прадбачлівасці* і да т.п. У “Граматычным слоўніку” А.А. Зализняка [3] адлюстраваны практычна ўсе патэнцыяльныя формы для лексем рускай мовы. Такім чынам, пры стварэнні машынных слоўнікаў трэба ўлічваць і такія новаўтварэнні. У гэтым выпадку, па меркаванні І.В. Соўпеля, найлепшым выхадам з'яўляецца пабудова слоўніка па гнездавым прынцыпе, а не слоўніка словаформаў ці асноў: «Пад гнездавым слоўнікам разумеецца сукупнасць словаформаў адной асновы, прадстаўленых у памяці ЭВМ сумесна з адпаведнымі ім кодамі» [4, с. 52]. У такім слоўніку прадстаўлены не асобныя словаформы, а асновы з адпаведным кодам, які адлюстроўвае сістэму канчаткаў, утвараючых цэлае словаўтваральнае гняздо:

кадр*148
1а*
1*у
1*_
1*ам
1*ы …
2*авы
2*авага
2*аваму …
3*авік
3*авіка
3*авіку …
і г.д.

«Гнездавая структура машыннага слоўніка дазваляе скараціць патрэбны для захавання слоўніка аб'ём памяці ЭВМ прыкладна ў 3–4 разы» [4, з. 52], акрамя таго, такая структура ўлічвае магчымасць з'яўлення словаформаў, якія ў цяперашні час адсутнічаюць у актыўным выкарыстанні, але якія ўтвараюцца па прадуктыўных мадэлях словаўтварэння пэўнай мовы.

Вопыт стварэння машынных марфалагічных слоўнікаў паказвае, што напаўняльнасць слоўніка ў многім залежыць ад канкрэтнай задачы, у якой дадзены слоўнік будзе прымяняцца. Так, створаны ў Навукова-даследчай лабараторыі інтэлектуальных інфармацыйных сістэм БДУ універсальны марфалагічны слоўнік беларускай мовы даволі значна мяняўся пры выкарыстанні ў розных праграмных

прадуктах. Напрыклад, пры напаўненні двухмоўнага слоўніка для сістэмы машыннага пераводу значная частка лексем беларускай мовы не ўвайшла ў слоўнік, паколькі пры перакладзе важна выкарыстоўваць найбольш ужывальнае і нейтральнае слова з шэрагу сінонімаў. Напрыклад, з руска-беларускага слоўніка была выключана лексема *слугаваць ‘служить’*, паколькі існуе больш нейтральны варыянт перакладу – *служыць*. Таксама скарачэнне аб'ёму слоўніка дазваляе пазбегнуць многіх выпадкаў аманіміі, якая прыводзіць да памылак. Так, з двухмоўнага слоўніка быў выключаны рускі назоўнік *под* ‘гарызантальная паверхня ў печы’, паколькі значна больш ужывальным з'яўляецца прыназоўнік *под*.

Пры напаўненні слоўніка для пошукавай сістэмы, наадварот, стаяла задача максімальна поўна ахапіць лексіку мовы, прычым важна пры гэтым улічваць сістэмныя адносіны між лексемамі, у першую чаргу, сінанімічныя. Таму слоўнік для пошукавых сістэм павінен уяўляць сабой тэзаўрус у выглядзе сінанімічных радоў з указаннем некаторых іншых адносін (гіперонімы, сінгулятывы і інш.): гіпероним: *агароджа*; *паркан, плыт, тын, частакол, штыкетнік, шчыкетнік*; сінгулятыў: *штакеціна*. У гэтым выпадку ў слоўнік можна ўключаць і пасіўную лексіку (устарэлыя словы, гутарковая лексіка), паколькі асноўная задача пошукавай сістэмы – пошук інфармацыі незалежна ад яе слоўнага афармлення.

Літаратура

1. Ермолович, Д.И. Имена собственные на стыке языков и культур / Д.И. Ермолович. – М.: Р.Валент, 2001. – 133 с.
2. Национальный корпус русского языка [Электронный ресурс]. – 2003-2011. – Режим доступа: http://ruscorpora.ru/index.html. – Дата доступа: 17.02.2011.
3. Зализняк А.А. Грамматический словарь русского языка. Словоизменение / А.А. Зализняк. – М.: Русский язык, 1977. – 880 с.
4. Совпель, И.В. Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста / И.В. Совпель. – Мн.: Вышэйшая школа, 1991. – 118 с.