

## КАМП'ЮТЭРНЫ ПЕРАКЛАД: МАТЭМАТЫЧНЫЯ МАДЭЛІ МОЎНЫХ ПРАЦЭСАЎ

*Іванова А.В., 4 курс,*

*Канцавая Г.М., к.філ.н., дацэнт,*

*УА «Брэсцкі дзяржаўны ўніверсітэт імя А.С. Пушкіна»*

Камп'ютэрны пераклад – пераўтварэнне з дапамогай камп'ютэра тэксту з адной натуральнай мовы ў семантычна эквівалентны тэкст на іншую натуральную мову. Камп'ютэрны пераклад не толькі падае інструментарый рашэння прыкладных задач, але і адкрывае новыя гарызонты лінгвістычных даследаванняў, фарміруе новае праблемнае поле і змяняе наша разуменне мовы. «Праблема машыннага перакладу аказалася значна больш складанай, чым уяўлялася. Затое падчас яе вывучэння мы атрымалі такія звесткі аб мове, на якія не маглі нават разлічваць... Як аб'ект навуковага даследавання яна аказалася значна больш складанай і багатай, чым мы лічылі» (Лотман Ю.М.).

Да нядаўняга часу было прынята дзяліць сістэмы перакладу на класы: transfer і interlingua. Алгарытмы перакладу transfer будуцца як кампазіцыя трох працэсаў: аналіз уваходнага сказа ў тэрмінах структур уваходнай мовы, пераўтварэнне гэтай структуры ў структуру выходнай мовы, сінтэз выходнага сказа па атрыманай структуры. Алгарытмы тыпу interlingua прадугледжваюць папярэдняю наўнасць некаторай метамовы структур, на якой можна апісаць усе структуры як уваходнай, так і выходнай моў увогуле. Алгарытм перакладу transfer больш просты: аналіз уваходнага сказа ў тэрмінах метамовы і затым сінтэз з метаструктуры адпавядага сказа выходнай мовы. Складанасць у гэтым выпадку заключаецца ў распрацоўцы самой метамовы і апісанні натуральнай мовы адпаведнымі тэрмінамі.

Апошнім часам атрымліваюць развіццё тэхналогіі статыстычнага машыннага перакладу, пры якіх у сістэму загрузаюцца тэксты вельмі вялікага аб'ёму: матэрыялы на мове гатовага перакладу альбо двухмоўныя тэксты, якія ўтрымліваюць сказы на зыходнай мове і адпаведныя ім пераклады, выкананыя чалавекам уручную. Для пабудовы мадэлі перакладу выкарыстоўваюцца метады саманавучання на аснове статыстычнага аналізу.

Аднак аналіз масіваў тэкстаў з выкарыстаннем выключна ручных метадаў з'яўляецца вельмі працаёмкім. Толькі сучасныя метады матэматычнай статыстыкі ў апрацоўцы карпусоў раўналежных тэкстаў адкрылі шляхі скарачэння працаёмкасці шматлікіх работ у галіне двухмоўнай лексікаграфіі, стварэння праграмных інструментаў для перакладчыкаў і сістэм машыннага перакладу. Абстрагуючыся ад некаторых дасягненняў структурнай лінгвістыкі на карысць прастаты прапануемых мадэляў, большасць сучасных сістэм статыстычнай апрацоўкі раўналежных тэкстаў сутыкаецца з неабходнасцю велізарных трэніровачных карпусоў для апісання нават досыць трывіяльных з пункту гледжання традыцыйнай парадыгмы з'яў. Варта сказаць такога падыходу перавешваюць яго недахопы толькі на аснове выкарыстання адэкватных задачы матэматычных мадэляў моўных працэсаў.

Мадэль камп'ютэрнага статыстычнага перакладу ўяўляе сабой мадэль мовы і мадэль канала. Мадэль мовы павінна прысвойваць ацэнку верагоднасці любому сказу мэтавай мовы, а мадэль перакладу павінна прысвойваць ацэнку верагоднасці сказа арыгінала пры ўмове вызначанага сказа на мэтавай мове. Тэхналогія статыстычнага перакладу практычна рэалізуецца ў рэжыме навучання і рэжыме функцыянавання. Для навучання сістэмы выкарыстоўваецца трэніровачны корпус раўналежных тэкстаў, і з дапамогай лінейнага праграмавання шукаюцца такія значэнні табліц перакладных адпаведнасцяў, якія максімізуюць верагоднасць уваходнай часткі корпуса пры наяўнай іншай частцы паводле абранай мадэлі перакладу. На ўваходнай частцы таго ж корпуса будуюцца мадэль рускай мовы. Мадэль функцыянавання можа быць апісаная наступным чынам: на аснове атрыманых дадзеных для незнаёмага іншамоўнага сказа шукаецца зыходны сказ, які максімізуе здабытак верагоднасцяў, якія прысвойваюцца мадэллю мовы і мадэллю перакладу.

У якасці мадэлі мовы ў сістэмах статыстычнага перакладу выкарыстоўваюцца пераважна розныя мадыфікацыі n-грамнай мадэлі, якая сцвярджае, што граматычнасць выбару чарговага слова пры фарміраванні тэксту вызначаецца толькі тым, якія (n - 1) слоў ідуць перад ім. Верагоднасць кожнага n-грама вызначаецца па тым, як ён часта сустракаецца ў трэніровачным корпусе.

Вялікую ролю адыгрывае тут паўната трэніровачных карпусоў. У студзені 2008 года Еўрапейская Камісія прапанавала распрацоўшчыкам сістэм камп'ютэрнага перакладу бясплатны доступ амаль да мільёна перакладзеных сказаў на 22 мовах. Выкарыстанне іх у кантэксце адэкватных матэматычных мадэляў моўных працэсаў заклікана істотна павысіць якасць разнастайных моўных інструментаў, уключаючы праграмы праверкі граматыкі і арфаграфіі, анлайнавыя слоўнікі і аўтаматычныя перакладчыкі. Асабліва карысна гэта для распрацоўшчыкаў статыстычнага машыннага перакладу, якім трэба сабраць корпус з тысячы фраз, якія супадаюць на розных мовах, каб затым можна было знайсці найбольш верагодны пераклад любой зададзенай фразы.

Выкарыстанне статыстычных метадаў у лексікаграфіі (як двухмоўнай, так і аднамоўнай дыяхранічнай) дае ў распараджэнне лінгвіста інструменты для эфектыўнага пошуку невідавочных адпаведнасцяў у вялікіх карпусах тэксту.