

ВЫКАРЫСТАННЕ ІНФАРМАЦЫЙНЫХ ТЭХНАЛОГІЙ У ДАСЛЕДАВАННІ КІТАБАЎ

Кітабістыка — навука, якая займаецца даследаваннем рукапіснай спадчыны татараў Вялікага княства Літоўскага. Слова “кітаб” ў вузкім сэнсе абазначае толькі адну катэгорыю татарскіх рукапісных кніг. Але часта гэтакія слова, дзякуючы яго літаральнаму значэнню “кніга”, распаўсюджваецца на ўсе рукапісы татараў Вялікага княства Літоўскага.

Сёння праблема вывучэння рукапісных тэкстаў досыць актуальная. Распазнанне і запіс рукапіснага тэксту з’яўляецца адной з самых складаных задач для сучасных інфармацыйных тэхналогій. Шматлікія спецыялісты працуюць над стварэннем полілінгвістычнай праграмы для перанясення тэксту, напісанага рукою чалавека, у кампутар і наступнай яго апрацоўкі.

Вялікім дасягненнем у рэчышчы даследавання кітабаў з дапамогай інфармацыйных тэхналогій з’яўляецца стварэнне базы дадзеных “Рукапісы татар (мусульман) Беларусі канца XVII — пачатку XX стагоддзя”. Гэтая база дадзеных змяшчае навуковае апісанне і кадыкалагічны аналіз (да тыроўка, лакалізацыя, выяўленне моўнай і жанравай прыналежнасці дакументаў і іх фрагментаў) 50 дакументаў, якія захоўваюцца ў ЦНБ імя Якуба Коласа НАН Беларусі, Нацыянальнай бібліятэцы Беларусі, Гродзенскім дзяржаўным музеі гісторыі рэлігіі, бібліятэцы Беларускага дзяржаўнага ўніверсітэта культуры і мастацтваў, Музеі гісторыі Клетчыны. На аснове гэтай базы дадзеных падрыхтаваны каталог “Рукапісы татараў Беларусі канца XVII — пачатку XX стагоддзя з дзяржаўных кнігазбораў краіны”.

У даследаванні кітабаў трэба знайсці аптымальны спосаб перанясення тэкставай інфармацыі ў кампутар, аднак можа ўзнікнуць шмат цяжкасцей, Галоўная з іх — адсутнасць якой-небудзь кампутарнай праграмы, якая палегчыць працу па пераносу рукапісных арабскаграфічных тэкстаў у электронны выгляд.

На сённяшні дзень няма ідэальнага рашэння праблемы распазнання арабскаграфічнага рукапіснага тэксту, а значыць, ёсць толькі перспектывыя распрацоўкі выкарыстання інфармацыйных тэхналогій у даследаванні кітабаў.

Для распазнання тэксту існуюць спецыяльныя OCR-праграмы. Пра іх, а таксама аб праграмах, якія выкарыстоўваюць сродкі ўводу, ніжэй. Найбольш якаснай, па меркаванні шматлікіх экспертаў, з’яўляецца ABBYY FineReader. FineReader на сёння лічыцца лідэрам распазнання тэксту з графічных дакументаў у тэкставыя. Праграма здольная самастойна правяраць арфаграфію і захоўваць распазнаную інфармацыю ў файлы фарматаў doc, pdf, html і інш. Увесь працэс (ад сканавання і да распазнання) можна праводзіць самастойна, з дапамогай майстра распазнання і перакладу ў аўтаматычным рэжыме. Аднак з’яўляецца шмат

праблем, звязаных з даследаваннем рукапіснага тэксту наогул. Па-першае, дадзеная праграма не распазнае рукапісны тэкст. Існуе разнавіднасць ABBYY FineReader — ABBYY FormReader — сістэма ўводу формаў, запоўненых ад рукі, з ужываннем тэхналогіі ICR (Intelligent Character Recognition). На дадзены момант існуе магчымасць распазнаваць толькі машыначытальныя формы. Пры гэтым для карэктнага распазнання такія формы павінны адпавядаць шэрагу ўмоў. Гаворкі аб распазнанні звычайнага тэксту, нават калі ён напісаны друкаванымі літарамі, тут не ідзе.

Варта таксама дадаць, што FormReader прызначаны для апрацоўкі вялікіх аб'ёмаў аднатыпных дакументаў з наступным захаваннем гэтай інфармацыі ў базе дадзеных. Гэта праграма прызначана для апрацоўкі інфармацыі, атрыманай ад кліентаў, карыстальнікаў, супрацоўнікаў у выглядзе анкет, запоўненых бланкаў і далейшага цэнтралізаванага захоўвання падобнай інфармацыі ў электроннай базе дадзеных такім чынам, каб заўсёды хутка можна было знайсці патрэбныя дадзеныя. Аднак дадзеная праграма непрыдатная для распазнання рукапіснага тэксту.

Існуюць таксама падобныя OCR-праграмы (SimpleOCR, CuneiForm, Recognita, Microsoft Office Document Imaging), але і ў іх нельга распазнаваць рукапісны тэкст.

Аднак, не варта думаць, што рукапісны тэкст на сённяшні дзень нельга распазнаць наогул. На самой справе праграмы распазнавання ёсць для прылад уводу дадзеных (планшэт, мыш, лічбавая ручка), але ёсць свае асаблівасці: яны працуюць толькі з гэтымі прыладамі, і ў асноўным у рэжыме рэальнага часу. Прыкладам такой праграмы можа быць CellWriter.

CellWriter працуе з першапачатковым аналізам почырку кожнага індывідуальнага карыстальніка, таму трэніроўка абавязковая. Пры запуску CellWriter запуская рэжым навучання, каб вывучыць почырк карыстальніка. Калі ўзнікаюць цяжкасці з пэўным знакам (літарай), можна, націснуўшы Train (Навучанне), задаць больш узораў почырку — праграма працягвае навучацца па меры назапашвання вопыту. Рэжым навучання падпраўляе табліцу сімвалаў, у якой кожны знак займае ўласную клетку. Вы малюеце адпаведны знак у яго клетцы, а CellWriter запамінае яго. "Навучаныя" клеткі адрозніваюцца па колеры ад "ненавучаных", таму лёгка сачыць за працэсам навучання. Калі вы неакуратна намалявалі літару, яе можна выдаліць. Калі вы пішаце ў рэжыме ўводу тэксту, CellWriter ператворыць штрыхі ў сімвалы па меры прасоўвання ручкі з адной клеткі ў другую. Калі сімвал інтэрпрэтаваны некарэктна, з дапамогай мышы можна выклікаць спіс візуальна падобных знакаў і выбраць правільны. Калі ж CellWriter не ўпэўнены ў выбары літары, ён вылучыць адпаведную клетку колерам, каб папярэдзіць, але не прыпыніць працу. Такім чынам, кітабы можна літаральна "перапісваць" і ў выніку атрымліваць электронныя варыянты для далейшай працы.

Для арабскага пісьма характэрна даволі строгае напісанне графічных элементаў, што нашмат спрашчае іх распазнаванне.

Распазнаваць арабскаграфічны рукапіс — працаёмкая задача, бо складана разабраць почырк без аналізу сэнсу. У сувязі з гэтым мадэлюецца пэўны алгарытм пры даследаванні рукапіснага тэксту:

1) падзяліць рукапісны тэкст на радкі;

2) вызначыць шырыню лініі пярэ і вышыню радка;

3) усе элементы рукапіснага тэксту злучаюцца паміж сабой адным з 3-4 спосабаў, характэрным менавіта для гэтага элемента. Трэба вылучыць гэтыя стандартныя элементы і атрымаць мадэль дадзенага элемента ў выглядзе набору дуг з улікам вышыні радка і шырыні лініі, а потым вылучыць у працэсе распазнавання;

4) усе асаблівасці напісання літар залежаць не толькі ад почырку, але і ад канкрэтнага слова. Тым не менш, ёсць агульныя прыкметы, па якіх можна вылучыць канкрэтныя літары або іх элементы ў тэксце. Потым да іх прыбудоўваецца астатняе.

Распазнаць тэкст, не маючы магчымасці «выдраць» з яго літары або лігатуры, — практычна немагчыма. Сутнасць у стварэнні алгарытму, які дазваляе вылучыць наборы рысак, што змяшчаюць літары або лігатуры з некаторым перабарам варыянтаў (меншым, чым у выпадку перабору за ўсё слова). Па задуме — будзе задзейнічаны частотны слоўнік пэўнай мовы. Калі алгарытм не можа вызначыць, якое слова напісана ў дадзеным фрагменце тэксту, то ён павінен выдаць набор магчымых варыянтаў для карэкціроўкі, у якім павінна ўтрымлівацца патрэбнае спалучэнне, пры гэтым колькасць варыянтаў павінна быць мінімальнай і ўтрымліваць у сабе правільны варыянт.

Аднак у рукапісах татараў сустракаюцца тэксты на розных мовах: арабскай, цюркскіх (чагатайскай, крымскататарскай і інш.), беларускай, польскай.

З улікам таго, што аналізуецца набор дуг кожнага элемента і прапарцыі адносна шырыні лініі і вышыні радка, фарміруецца група напісанняў элемента, потым фарміруецца група элементаў для дадзенага сімвала, а лепш для дадзенага слова, гэта значыць, фарміруецца група напісанняў дадзенага слова.

Цяжкасці могуць узнікнуць пры пэўных дэфектах рукапіснага тэксту — наезд ліній адна на адну, зліпанне суседніх літар і іх перасячэнне.

Такі спосаб выдатна падыходзіць для арабскаграфічнага тэксту на арабскай мове з выкарыстаннем частотнага слоўніка гэтай мовы. Па сутнасці гэты спосаб прыдатны для кітабаў на цюркскіх, беларускай і польскай мовах. Толькі тут трэба звярнуць увагу на фанетычнае значэнне літар і некаторыя складанасці пры выкарыстанні частотных слоўнікаў гэтых моў.

Выкарыстанне інфармацыйных тэхналогій будзе мець вялікі ўплыў на далейшыя даследаванні кітабаў, бо гэта дазволіць хутчэй пераходзіць ад фазы ўводу тэксту ў базу да непасрэднага яго вывучэння.

ЛІТАРАТУРА

1. Антонович, А.К. Белорусские тексты, писанные арабским письмом, и их графико-орфографическая система. — Вильнюс, 1968.
2. Рукапісы татарэў Беларусі канца XVII — пачатку XX ст. з дзяржаўных кнігазбораў краіны: каталог / склад.: М.У. Тарэлка, А.І. Цітавец. — Мінск, 2011.