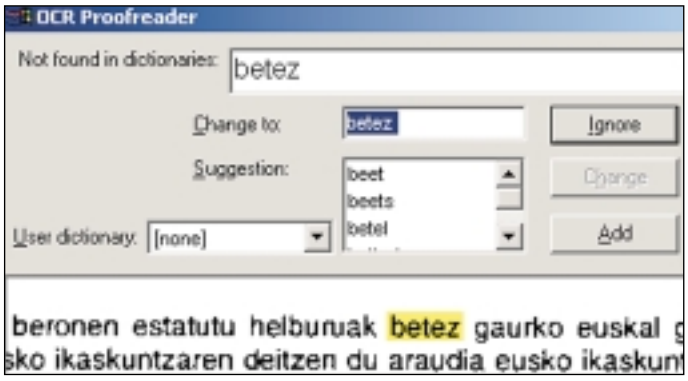
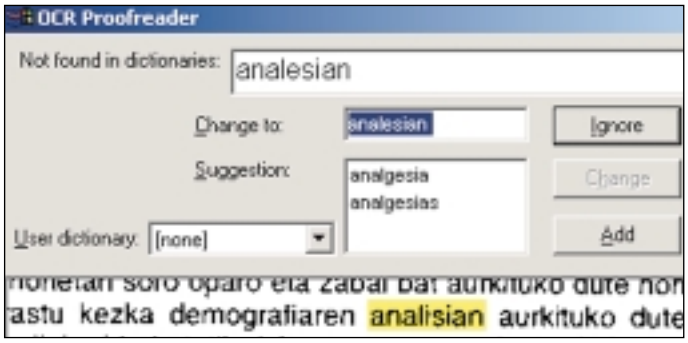
Euskarazko OCRa

Edurne Martinez Iraola
ELEKA ingeniaritza linguistikoa



ARTXIBOKOA

Euskarazko testu bat ingelesezko informazioa erabiliz zuzentzean, akatsak sortzen dira eta programak proposamen desegokiak egiten ditu.



ARTXIBOKOA

Euskarazko testu bat gaztelaniazko informazioa erabiliz zuzentzean, akatsak sortzen dira eta programak proposamen desegokiak egiten ditu.

OCR bezala ezagutzen dugun teknologia aspaldikoa bada ere, ez zegoen oraindik euskaraz lan egitea ahalbidetzen zigun horrelako aplikaziorik. ELEKAk garatu berri duen plug-inaren bitartez, merkatuan ezagunena den OCR programak euskarazko testuak eskaneatzeko aukera eskaintzen digu. Proiektu honek Eusko Jaurlaritzaren babesa izan du, eta emaitza laster izango da kalean.

INFORMAZIOA ESKURATU, AZTERTU ETA JASOTZEKO BIDEAK ALDATZEN DOAZ. Garai batean informazioa jasotzeko bide egokiena liburu inprimatua zen; gaur egun, ordea, bestelako aukerak eskatzen ditugu: informazioa bilatu, kopiatu eta mugitu, gure erara sailkatu, aldatu eta manipulatzeko aukera. Horiek guztiak orain arte ezagututako testu tradizionalek ematen ez zizkiguten aukerak dira, baina egungo gizarte digitalean gauzak oso bestelakoak dira.

Euskal merkatuan oso zabaldua dago OCRaren erabilera, nahiz eta horrek ondoren zuzenketa-lan handia eskatzen duen. Euskal Herrian egunkari, aldizkari eta argitaletxe asko ditugu, eta, kasu gehienetan, horien funts dokumentala ez dago formatu digitalean gordeta. Interneten zabalkundearekin, ordea, beharrezkoa bihurtu da funts dokumental horiek guztiak behar bezala digitalizatu eta jasota izatea katalogazio- eta bilaketa-sistema azkarragoak antolatzeko.

> *"OCRak egiten duena zera da: eskaneatutako irudia testu bihurtu"*

OCRa (Optical Character Recognition), idatzitako edo inprimatutako testu-karaktereen ordenagailu bidezko ezagutza da. Horrek esan nahi du OCR softwarea erabiltzen dugunean karaktere bakoitza eskaneatzen dugula, argazki bat balitz bezala, eta ondoren eskaneatutako irudi hori aztertu eta karaktere-kode arrunt batera itzultzen dela (ASCII, esaterako).

OCR sistemaren doitasuna hiru faktorek mugatzen dute: dokumentu originalaren kalitateak, eskanerrak sortu duen irudiaren kalitateak eta azken horren gainean OCR softwareak egiten duen interpretazioak. Azkenekoaz arituko gara hemen.

OCRak egiten duena, hitz gutxitan esanda, zera da: eskaneatutako irudia testu bihurtu. Horretarako, irudia osatzen duten puntuak aztertzen ditu, eta tartean dauden hutsuneak bereiztu. Prozesu horri segmentazioa deitzen zaio eta hiru pausotan egiten da: lehenengo lerroak bereizten dira (lerrokako segmentazioa), ondoren hitzen isolamendua egiten da (hitz-segmentazioa), eta, azkenik, karaktereak bereizten dira (karaktere-segmentazioa). Azkeneko fase hori errazagoa da karaktere guztiak zabalera berekoak badira; asko konplikatzen da, aldiz, karaktereek elkar ukitzen badute, beste puntuazio-markekin nahasten badira edo zabalera karaktererearen formaren araberakoa bada.

Karaktere mailako ezagutza egiteko, beharrezkoa da OCR sistemak eskaneatu dugun testuko hizkuntzaren karaktere guztiak ezagutzea. Karaktereekin zalantzarik sortuko balitzaio, berriz, hitza osatu arte itxarongo luke; prozesu horretan baliagarria izango da hizkuntza horretako hiztegi bat edukitzea harekin parekatu ahal izateko. Horrela, probabilitate-joko batez eta hiztegiko hitza den ala ez ebaluatuz, karaktere bat ala bestea hautatuko du sistemak.

Dirudienez, hizkuntza horretako alfabetoa eta hiztegi bat edukitzea nahikoa litzateke OCRa modu egokian aplikatzeko, baina euskararen kasuan ez da horrela suertatzen. Kasu horretan ezin da hitz posibleen zerrenda oso bat eman, hots, ezin da hiztegi bat sortu, hizkuntza deklinatua izanik, hitz-erro bakoitzetik, hitz-forma gehiegi ateratzen baitira. Tresna linguistikoek laguntza handia emango dute pauso honetan; hau da, euskararen ezaugarri nagusiak landuz hobekuntza handiak lor ditzakegu OCR sistema bat garatzerako garaian. Esaterako, euskaraz egiten diren karaktere- edo hitz-elkarketak (ts, tz, tx, edo marren erabilera) ez dira hain arruntak Europako gainerako hizkuntzetan.

Gaur egun erabiltzen diren OCR software gehienekin, euskarazko testu bat aztertu nahi dugunean, erdal hizkuntza bateko hiztegia erabili behar izaten dugu. Hala ere, horrelakoetan hobe da hiztegirik ez erabiltzea beste hizkuntza bateko hiztegia erabiltzea baino, testuan akats gehiago ez egitearren. Esaterako, ingelesezko hiztegi bat erabiltzen ari bagara, "sei" hitzaren agerraldi gehienak "set" hitzarengatik ordezkatuko ditu ia seguru. Gaztelaniazkoa erabiltzen ari bagara, berriz, "energia" hitzaren agerpenak "energía" (tildearekin) hitzarengatik ordezkatuko ditu.

> *"gaur egun erabiltzen diren OCR software gehienekin, euskarazko testu bat aztertu nahi dugunean, erdal hizkuntza bateko hiztegia erabili behar izaten dugu"*

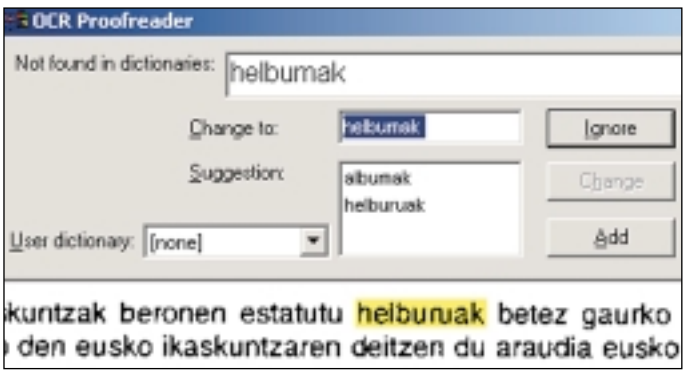
Datuak azkarrago bidaltzeko softwarea

Segundoko 8 gigabyte igortzeko gai den softwarea prestatu dute. Hau da, ohiko modema baino 153.000 bider eta ADSL arrunta baino ia 6.000 bider azkarrago. Software berriak FAST izena du eta datuak transmititzeko erabiltzen diren ohiko mekanismoez baliatzen da, baina kongestioaren seinaleen aurrean leunago erreakzionatzen du. TCP/IP protokoloak, Interneteko transmisio protokoloak, igorleak jasotako erroreen arabera erabakitzen du bideetan kongestioa dagoela eta, ondoren, abiada izugarri jaisten du. FASTek, berriz, informazioa bidali eta jaso dela adierazten dion mezuaren arteko denbora-tartearen arabera neurtzen du kongestio-maila. Denbora-tartea handituz gero, abiada apur bat jaisten du. Modu horretan, datuen transmisio-efizientzia asko hobetzea lortu dute ikertzaileek. Ordubete Kalifornia eta CERNen Genevako zentroaren artean datuak bidaltzen pasatuta, 925 gigabyteko batez besteko abiadura erdietsi zuten egoera arruntean. TCP/IP bidez baino ia lau bider azkarrago. Errekorra, berriz, 10 linea batera erabilita lortu zuten. Softwarea Kaliforniako Teknologia Institutuan garatu dute.

ARTXIBOKOA

ELEKAn garatu dugun proiektuaren emaitza zera da: egun gehien erabiltzen den OCR softwareari, Omnipage programari, euskarazko zuzenketa gehitu zaio, euskararen informazio morfologikoarekin batera. Programa hori, euskararen kasurako, eskaneatutako irudia karaktere bihurtzeko urratsa emateko prestatua dago. Orain arte, ordea, ez zegoen ondoren egin behar den hitzen egiaztapen eta zuzenketa-faserako prestatua (hizkuntza nagusien kasurako egina badago ere: ingelesa, alemana ...). Hurrengo asmoak Xuxen moduko OCR zuzentzaile bat gehitzea izango da Microsoft Word nahiz OpenOffice testu-prozesadoreentzat, Omnipage erabiltzen ez duten erabiltzaileen esku jarri ahal izateko euskarazko OCR sistema.

Beraz, euskarazko tresna linguistikoak gehituz, euskarazko testuak ahal den hobekien digitalizatzen dituen tresna garatu da. Hau da, testuak digitalizatzean euskara automatikoki behar bezala ulertu eta zuzentzen duen tresna baten garapena egin du ELEKAk. Proiektu hau garatzeko, Eusko Jaurlaritzaren Hizkuntza Politikarako Sailordetzaren laguntza izan du, eta hura arduratuko da aplikazio honen banaketaz.



ARTXIBOKOA

Euskarazko testu bat euskarazko informazioa erabiliz zuzentzean, emaitza hobeak lortzen dira eta programak proposamen egokiak egiten ditu.