Zer entzun, hura idatzi

Kortabitarte Egiguren, Irati
Elhuyar Zientziaren Komunikazioa

EITB

ETBko *Gaur Egun* programak erabiltzen dituzte, besteak beste, hizketaren tratamendua egiteko sistemak trebatzeko.

Idatziz jasotakoa bilatzea erraza da sarean. Horretarako, kontsultatu nahi dugun hitza bilatzailean idaztea besterik ez dugu. Bilaketa horietan, ordea, audio-fitxategietan esandakoak galtzen ditugu besteak beste, betiere, audio-fitxategi horietan esandakoaren azalpenak testu idatzian jasotzen ez badira.

AHOZKO HIZKETA EZAGUTZEA ETA HURA TESTU BIHURTZEA EZ DA LAN ERRAZA. Hitzak ez dira ongi bereizten bata bestetik, intonazioa kontuan izan behar da, eta, gainera, seinale fisikoen zarata ere oztopo da. Horren harira, merkatu handia zabaldu da ahozko hizketa prozesatzen eta ulertzen duten sistementzat. Alegia, ahozkoa testu idatzi bihurtuko diguten tresnentzat.

Sistema horiek batez ere telefono bidezko zerbitzuetan integratzen dira oraingoz: aurretiko hitzordua, produktu-eskaerak, ikuskizunetarako erreserba-eskea eta abar. Baina badaude bestelakoak ere: diktaketa automatikoa, adibidez. Azken horretan dihardute lanean, hain zuzen ere, EHUko Sistemen Ingeniaritza eta Automatika sailean, besteak beste.

Hizketaren tratamendua egiteko sistema asko eta ongi trebatu behar da. Alegia, sistemak nolabaiteko entrenamendua jaso behar du, makina-ikasketa deritzona. Horretarako, batetik, telebista nahiz irratietako fitxategiak, audioak nahiz soinuak behar dira; eta bestetik, komunikabide horietan esan denaren erreferentziazko testuak. EHUko ikertzaileek, adibidez, ETBko *Gaur Egun* eta *Teleberri* programak erabiltzen dituzte maiz, sistema trebatzeko. Ez da beharrezkoa hitzez hitz zer esan den jakitea; bai, ordea, esandakoaren laburpen bat jasotzeko gai izatea sistema. Azken finean, soinu eta hitzen arteko erlazioa ulertzen saiatzen da.

Ikasketa-prozesua amaitu ostean, edozein *Gaur Egun*-etan edo *Teleberri*-tan esandakoa ulertzeko gai behar luke izan sistemak. Ikastea prozesu motela izan arren, sistemak behin arauak edo informazioa barneratuta duenean, hau da, erreferentziazko material egokia duenean, nahiko azkar erakusten du emaitza. Kasu honetan, ahoz esandakoaren testu idatzia. Azken finean, helburua da audio edo soinu batetik testua lortzea.

Txikia handi

Egia da merkatuan aurki daitezkeen horrelako aplikazio gehienek hizkuntza

Babeslea:

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Proiektua

Proiektuaren laburpena

Ikerkuntza-talde hau hizketaren ezagutza eleaniztunen alorrean aritzen da, euskararako eta haren inguruko hizkuntzetarako. Bereziki, euskal komunikabideetako albistegien hizketatik informazioa automatikoki eskuratzeko hainbat tresna eta baliabide garatzen dituzte. Horretarako, informazio hori ahalik eta modu eraginkorrenean eskuratzeko teknikak ikertzen dituzte, eta, batez ere, baliabide urriko hizkuntzetarako metodoak garatzen dituzte —hala nola euskararako—.

Zuzendaria

Miren Karmele López de Ipiña doktorea.

Lantaldea

M.K. López de Ipiña[1], N. Barroso[1], N. Gilisagasti[1], I. Ariztimuño[1], A. Ezeiza[1], N. Ezeiza[2] eta M. Hernández[2].

Saila

Sistemen Ingeniaritza eta Automatika.

Fakultatea

[1]Donostiako Unibertsitate Eskola Politeknikoa eta [2]Informatika Fakultatea.

Taldea

EHU

Ezkerretik hasita, Ixabel Ariztimuño, Nora Barroso, Aitzol Ezeiza, Karmele Lopez de Ipiña eta Nerea Ezeiza.

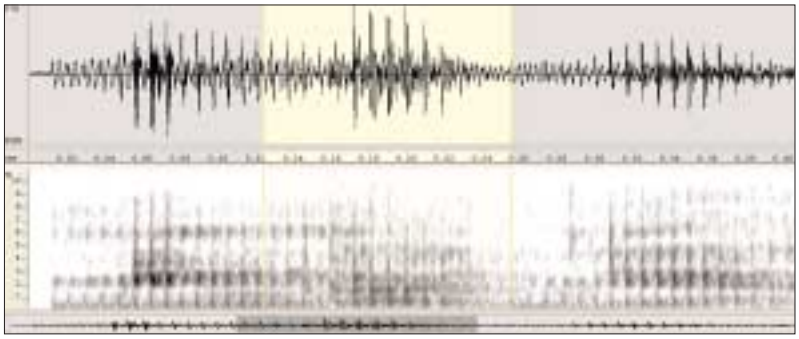
'handiak' dituztela helburu; ingelesa, batik bat. Dena den, Donostiako Unibertsitate Eskola Politeknikoko ikertzaileek, EHUko IXA, GTTS eta Adimen Konputazionala taldeekin elkarlanean, euskararekin dihardute lanean. Hizkuntza 'handi' eta 'txiki' horien arteko ageriko ezberdintasun nagusia erreferentziazko datu-kopuruan datza. Mota horretako ingelesezko tresnek ikaragarrizko datu piloa izaten dute; euskarazkoen erreferentziazko materiala, berriz, dezente txikiagoa da. Horregatik, datu gutxi horiek hobeto eta zehaztasun handiagoz aprobetxatzeko teknika berriak bilatzen ari dira ikertzaileak.

Zehaztasun-maila hori lortzeko, zenbait ekuazio matematiko erabiltzen dituzte. Datu-multzo eta audio-fitxategietatik informazio aproposa emango duten ezaugarri garrantzitsuenak aurkitzen saiatzen dira. Dena den, nahiko zaila da hautaketa hori egitea; alegia, jasoko den eta baztertuko den informazioa aukeratzea. Normalean, maiztasunarekin eta intonazioarekin lan egiten dute, une bakoitzean sistema jasotzen ari den informazio-mota bereizteko (galdera bat edo adierazpen-perpaus bat den bereizteko, adibidez).

> *"helburua da audio edo soinu batetik testua lortzea; alegia, soinuen eta hitzen arteko erlazioa lortzea"*

Sistema horiek hizkuntzaren mende daude erabat, eta hizkuntza bakoitzak bere tresna du. Baina, EHUko ikertzaileek, adibidez, euskararekin ez ezik, gaztelaniarekin eta frantsesarekin ere egiten dute lan. *Teleberri* programak edo *Infozazpi*-ko saioak aztertzen dituztenean, esaterako, bi helburu nagusi dituzte: batetik, gaztelania eta frantsesa ulertu nahi dituzte —euskararekin batera—, eta, bestetik, mota horietako sistemetan euskararen eta beste bi hizkuntza horien artean dauden antzekotasunak bilatu nahi dituzte, euskarazko tresnak hobeto trebatu ahal izateko.

Bide horretan, gaur egun, tresna berean hizkuntza bat baino gehiago erabiltzeko aukera aztertzen duten hainbat saiakuntza egiten ari dira. Horixe da, hain zuzen ere, EHUko ikertzaileen etorkizuneko erronka: euskara, gaztelania eta frantsesa ulertzeko gai izango den sistema bat garatzea.



EHU

Ahoz esandakoaren maiztasuna eta intonazioa lagungarri dira sistema jasotzen ari den informazio-mota bereizteko.