



Arantza Diaz de Ilarraza
EHUko Informatika Fakultateko irakaslea

Hizkuntza prozesamenduaren teknikaz balia gaitezke

osasuna / geltokiak / teknologia

3 BEGIRADA:

Interneten aurki ditzakegun dokumentuen kantitatea ikaragarria da, eta, gainera, esponentzialki handitzen da urtero. Hori dela eta, gure eskura testu-informazio bolumen itzela dugu, eta hazten doa. Hala ere, testu-informazio hori erabiltzea ez da erraza; laguntza behar dugu, eta konputagailuek lagun diezagukete zeregin horretan. Gaur egun sistema informatikoez datuak erraz prozesatzen dituzte, baina testua prozesatzea ez da lan erraza. Testua prozesatzea esaten dugunean, funtsean, testua ulertzea adierazi nahi dugu.

Hizkuntza prozesamenduaren arloaren (*natural language processing*) azken helburua hizkuntza automatikoki ulertzea da. Horretarako, hizkuntzaren mailak automatikoki tratatu behar dira: morfologikoa, sintaktikoa, semantikoa eta pragmatikoa. Arlo horretako ikerketa 1930ean hasi zen. Garai hartako ikertzaileen lehen helburuak itzulpen automatikoarekin lotuta zeuden, baina urte gutxiren buletan konturatu ziren ezarritako helburu horiek ez zirela hain lorgarriak, eta 1960ko hamarkadan beste norabide bat hartu zuten. Gaur egun hizkuntzaren prozesamendua pil-pilean dagoen arloa da, konputagailuaren eta gizakiaren arteko edo gizakien arteko komunikazioaren garrantzia gero eta handiagoa baita.

Historikoki, hiru lan-paradigma identifikatu dira hizkuntzaren prozesamenduan: ezaguera linguistikoan edo erregeletan oinarritua, datuetan oinarritua eta hibridoa (bai ezagueran eta bai datuetan oinarritua). Azkeneko urteetan, datuetan oinarritutako paradigma da gehien erabiltzen dena, gero eta datu gehiago ditugulako, eta hori da paradigma horren oinarria: datuetatik ikastea; zenbat eta datu gehiago, orduan eta hobeto ikasiko du sistemak.

Jende askok ez daki teknologia hori zertan datzan, baina gure bizitzaren alderdi funtsezkoa bilakatu dela esan behar dugu. Egunero ditugu teknologia hori erabiltzen deneko adibideak, edo haren erabilerak gure zeregin askoren garapena erraztuko ligukeen kasuak.

Aipatutako adibide bat akats ortografikoak eta gramatikalak saihesteko erabil-

itzen ditugun zuzentzaileak ditugu. Gaur egun zuzentzaile ortografikoak, estilo zuzentzaileak edo gramatikalak erabiltzea oso arrunta da, edozein testu gure hizkuntzan edo ondo menperatzen ez dugun hizkuntza batean idatzi behar dugunean.

Itzulpen automatikoa, berriz, baliagarri gertatzen ari da beste hizkuntza batean idatzita dagoen testu baten ulermen orokorra lortzeko, eta zenbait hizkuntzaren artean lortzen diren emaitzak nahiko onak dira, batez ere domeinu espezifikotetan (meteorologia, arlo juridikoa, osasungintza...); ez dira, noski, perfektuak, baina lagundu laguntzen dute itzulpenaren, postedizioa erraztuz eta produktibitatea hobetuz. Sistemok historikoki erregeletan oinarritutako paradigma jarraitu dute, baina badira urte batzuk datuetan oinarritutako teknikak gailendu direna, batez ere itzulgai diren hizkuntzetan idatzitako testu paralelo kopuru handiak eskuragarri direnetik.

Bestalde, hizkuntzaren prozesamenduko teknikak gero eta gehiago erabiltzen dira bilatzaileetan, galderen erantzun egokiagoak lortzeko. Gure galderari erantzuteko, bilatzaileek galderaren eta testuen arteko adostasuna aztertzen du-

te (zati komunak), eta gehien hurbiltzen diren dokumentuak itzultzen dituzte emaitza gisa. Antzekotasun hori detektatzeko, oso baliagarriak dira hizkuntzaren prozesamenduko teknika aurreratuak (semantikan oinarritutakoak).

Oraindik, baina, ez dira lortzen guk nahiko/beharko genituzkeen emaitzak. Galdera baten erantzunean, galderako hitzak agertzen diren dokumentuak ez ezik, hitz horiekin semantikoki erlazionatutakoak dauzkatena ere eskuratu nahiko genituzke; adibidez, “neurtzeko tresnei” buruz jakin nahi baldin badugu, erantzun gisa “barometroak”, “termometroak”, “higrometroak” eta abar dauzkatena web orriak ere eskuratu nahiko genituzke.

Datozen urteetan teknologia honen baliagarritasuna gero eta nabarmenagoa izango da eguneroko beharretan. Konputagailuak gero eta hobeto dabilta itzulpen automatikoan, dokumentuen sailkapenean, bilaketetan eta baita ahozaren ezagutzan eta hizketaren sintesian ere, baina oraindik ez dute hizkuntzaren esanahia ulertzen. Bitartean, baina, balia gaitezke lagungarriak izan daitezkeen aplikazioez. •

Xuxen euskarazko zuzentzaile ortografiko eta gramatikalarren webgunea.

Xuxen.eus Xuxen 5 Informazio orokorra Bertsioak FAQ Prentsa-aretoa Kontakua Surtu

Idazten ari garela, **gorriz** azpimarratzen ditu gure akats ortografikoak: **zihur, xagu, akanpaleku, iharduera, konbidatu...** Eta horiek nola zuzendu proposatzen diren erabiltzeko aukerak ematen ditu. **Astebururako hainbat iharduera antolatu dituzte.**

berdez edo urdinez azpimarratzen ditu hitzak, zuzenak izan arren Euskaltzaindiak beste forma bat hobesten duen kasuetan, hala nola **bordatu, fondo, naitu...** **Ainitz irakurtzen duzu?**

zenbait akats gramatikak ere berdez edo urdinez azpimarratu eta zuzentzeko proposamena egiten du: **gosea daukat, ikusiko gara...** **Bihar ikusiko gara.**

Office berrietan (Office 2013 arte) erabiltzeko prest dago Xuxen 5. Beraz, gure ordenagailura deskargatzen badugu, aipatutako hizkuntza-zuzenketa eta -prozesatze bat idazten ari garenean, adibidez. Eta erabiltzailearen hiztegia osatzeko aukera ere izango dugu, Xuxen 5-ekin batera erabiltzeko. Xuxen 5-ekin batera erabiltzeko, sistema hauetan guztietan txertatzeko **pluginak** berriak ditugu, zuzenketa ortografiko eguneratuak egin ditzaten: [Office 2010/2013](#)

elhuyar komunikazio eta teknologia **ixa** ikuntza eta teknologia

Zuzentzaile ortografikoak eta estilo zuzentzaileak erabiltzea arrunta da gaur egun.