HEDAPENA INFORMAZIOAREN BERRESKURAPENEAN: HITZEN ADIERA-DESANBIGUAZIOAREN ETA ANTZEKOTASUN SEMANTIKOAREN EKARPENAK

Tesiaren egilea: Arantxa Otegi Usandizaga
Unibertsitatea: Euskal Herriko Unibertsitatea
Saila: Lengoaia eta Sistema Informatikoak Saila
Tesi-zuzendaria: Eneko Agirre Bengoa eta Xabier Arregi Iparragirre
Tesiaren laburpena:

Informazioaren berreskurapena (IB) erabiltzaile baten informazio-beharra asetuko duten dokumentuak bilatzean datza. Horrela bada, IB sistemak erabiltzaileari dokumentu adierazgarriak, alegia, erabiltzaileak behar duen informazioa eduki dezaketen dokumentuak topatzen lagunduko dio, betiere erabiltzaileak egindakoan oinarrituz. Hain ezagunak eta erabiliak diren *Google* eta *Yahoo!*, esaterako, web-bilatzaileak IB sistemen adibide garbiak dira.

IB sistema perfektu batek dokumentu adierazgarriak bakarrik berreskuratu beharko lituzke, eta ez-adierazgarriak baztertu. Alabaina, sistema perfektuak ez dira existitzen. IB sistemek aurre egin beharreko arazo nagusietako bat kontsulta eta dokumentuen arteko parekatze-arazoa deiturikoa da: dokumentu bat kontsulta batentzako adierazgarria izan daiteke, nahiz eta bietan erabilitako hitzak guztiz berdinak ez izan; eta, alderantziz, dokumentu bat ez-adierazgarria izan daiteke kontsulta batentzat, nahiz eta termino batzuk komunean eduki. Lehena ideia edo gauza bera adierazteko hitz edo esamolde bat baino gehiago erabili ditzakegulako (sinonimia) gerta daiteke. Bigarrena, berriz, testuinguruaren arabera hainbat interpretazio izan ditzaketen hitzek (anbiguotasuna) eragiten dezakete. Hau kontuan izanik, IB sistema batek dokumentu bat adierazgarri edo ez-adierazgarri bezala sailkatzerakoan kontuan hartzen duen irizpide bakarra kontsultako hitzak egotea (edo ez egotea) denean, zaila suerta daiteke dokumentu egokiak topatzea, bai eta adierazgarriak ez direnak baztertzea ere. Horren aurrean, bidezkoa dirudi pentsatzeak hitz horien esanahiak kontuan hartuz gero berreskurapen arrakastatsuago bat egiteko aukera gehiago egongo direla.

IBaren hastapenetatik gaur arte parekatze-arazoaren inguruan ikerketa-lan dezente egin badira ere, oraindik guztiz ebatzi gabe jarraitzen du, eta bilatzaile askok ez dute aintzat hartzen. Tesi-lan honetan hizkuntzaren prozesamenduaren (HP) bidez arazo hau arintzerik ba ote den aztertu da.

Hitz gutxitan esanda, kontsulten eta dokumentuen hedapena egiten dugu HPko bi teknikaz baliatuz: hitzen adiera-desanbiguazioa eta ahaidetasun semantikoa. Alde batetik, teknika hauetako bakoitzerako hedapen-prozesu bat proposatzen

dugu, non kontsulta eta dokumentuetako hitzen sinonimoak eta bestelako ahaidetasuna duten hitzak lortuko ditugun. Bestetik, hedapenetik lortutako hitz horiek, kontsulta eta dokumentuetako jatorrizko hitzekin batera, IB sistemaren prozesuan txertatu eta ustiatzeko modu eraginkor bat azaltzen dugu kasu bakoitzerako. Are gehiago, erabiliko dugun hedapen-teknikak kontsulta eta dokumentuak itzultzeko balio duenez, hedapen-teknika hori erabiliz hizkuntza arteko berreskurapenean hobekuntzak lortzen direla erakutsiko dugu.

Hiru datu-multzotan egindako esperimentu eta analisiek erakusten dute tesi-lan honetan proposatutako hedapen-metodoek parekatze-arazoari aurre egiteko balio dutela eta, ondorioz, baita IB sistemaren eraginkortasuna hobetzeko ere.