EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Doktorego-tesia

Ezagutza baseen aberasketa urruneko gainbegiraketaren bidez: analisiak eta hobekuntzak

Ander Intxaurrondo Gonzalez de Langarika

2015

EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Ezagutza baseen aberasketa urruneko gainbegiraketaren bidez: analisiak eta hobekuntzak

Ander Intxaurrondo Gonzalez de Langarikak Eneko Agirre Bengoa eta **Oier Lopez de Lacalle Lekuonaren** zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2015eko maiatza.

The acquisition of knowledge is always of use to the intellect, because it may thus drive out useless things and retain the good. For nothing can be loved or hated unless it is first known.

Leonardo da Vinci.

Eskerrak

Eneko eta Oierri, tesi hau aurrera ateratzeko emandako animo eta laguntzagatik. Zuzendari handiak izan zarete, zuengandik asko ikasi dut, eta zuei esker nago ikerkuntza munduan sartuta. Oso gertuko zuzendariak izan zarete beti, nahiz eta elkarrengandik 2000 km edo 9000 kilometrotara egon, zuek ondoko bulegoan egotea bezala izan da. Izugarri estimatuko zaituztet beti.

IXA taldeko kideei eta kide ohiei. Zuek ireki zenizkidaten ateak mundu honetara, izugarri ikasi dut zuei esker, bai masterrean, baita doktoretzako urte hauetan ere. Oso ondo pasatu dut zuekin, bulegoan, kafe orduan, bazkalorduan, bazkaritan, afaritan eta beste hainbat aixialdietan. Kriston barreak egin dut zuekin. Taldekide bikainak zarete.

Al Dr. Mihai Surdeanu. Tus conocimientos y trabajos previos sobre muchos temas han ayudado a que esta tesis salga adelante. Has sido una gran colaboración todos estos a*ñ*os. Siempre agradeceré la oportunidad que me diste para trabajar contigo en Stanford, además del buenísimo trato que me diste allí. Espero que volvamos a trabajar juntos algún día en otro proyecto.

To the Stanford NLP group, for giving me the chance to visit your university and learn from you all.

No puedo olvidarme de Marta Recasens, además de mi compañera de despacho en el edificio Gates, una grán amiga.

To my housekeeper, and a great friend, Craig Wood. You made me feel at home from the beginning of my stay in Menlo Park.

Etxekoei, hasieratik eman dizkidazuen animoengatik, momentu zailenetan hor egon zarete beti. Espero dut liburuxka honek behingoz laguntzea ulertzen zertan ibili naizen lanean urte hauetan.

Koadrilakoei eta hauen bikotekideei. Azkenean bukatu da! Hau ospatu beharra dago. Ea ospakizunean denok berriz ere elkartzea lortzen dugun.

Felipe IVko tropari: Jone, Mikel, Ane eta Ganixi. Ederrak izan dira elkarrekin pasa ditugun bi urte eta erdi hauek.

Por último, a todo el clán de Jose Mardones. Siempre apreciaré los consejos y apoyo que me habeís dado durante estos años.

Esker instituzionalak

Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Sailari, ikerketa lan hau egiteko emandako ikertzaileak prestatzeko bekarengatik (BFI10.134).

Al Ministerio de Economía y Competitividad del Gobierno de España (MINECO), por financiar parcialmente este trabajo de investigación (Proyecto SKaTeR – TIN2012-38584-C06-02 y proyecto EXTRECM – TIN2013-46616-C2-1-R)

A Ignacio y Bitori

Laburpena

Informazio erauzketa testuetatik informazio egituratua eskuratzean datza. Informazio erauzketa sistemak corpusetatik informazio garrantzitsua eskuratzen saiatzen dira, informazioa gizaki eta konputagailuentzat intuitiboa den eran itzuliz. Tesi honetan honen bi azpiatazatan jartzen dugu arreta: erlazio erauzketan, entitateen arteko erlazioak antzemateko, eta gertaera erauzketan, testuetan gertaerak antzeman eta hauei buruzko informazio zehatz eta egituratua lortzeko.

Urruneko gainbegiraketaren arabera, ezagutza base batek bi entitateren artean erlazio bat dagoela zehazten badu, eta bi entitate hauek esaldi berean agertzen badira, esaldi horrek erlazio hori adieraziko du nola edo hala. Urruneko gainbegiraketan oinarritutako teknika desberdinek benetako tuplen beharra dute aipamen zaratatsuak zuzentzeko, eta geroago teknika gainbegiratu tradizionalak entrenatzeko. Tesi honetan, aipamenetako zarataren iturburuak aztertu ditugu, eta aipamen zaratatsuak filtratzeko metodo desberdinak aztertu. Emaitzek erakusten dute gure heuristikoen konbinaketak bi oinarri lerro trinko garaitzeko gai dela.

Gainera, Twitterretik gertaera konplexuak erauzten dituen gertaera erauzketa sistema bat aurkezten dugu, urruneko gainbegiraketan oinarritutakoa. Ia denbora errealeko datu iturburu honek informazio zehatzgabea eta anbiguoa dakar, ebaluazioan eta erauzketa metodoetan eragina izanik. Ebaluazio erlaxatu bat diseinatu dugu, zeinek ezagutza baseko balioekiko antzekoak diren erauzitako balioei kreditu partziala ematen dien. Printzipio hau ere

etiketatze prozesura eraman dugu, antzekoak diren balioak ere aipamen positibotzat hartuz. Gure ekarpenak positiboki ebaluatzen ditugu lurrikaren domeinu konplexuan, 20 argumentu dituzten gertaerekin. Ezagutza basea, txio garrantzitsuak eta eskuz etiketatutako txioak publikoki eskuragarri daude.

"Europako Doktoregoa" aipamena lortzeko Euskal Herriko Unibertsitatearen eskakizunei jarraituz, tesi txosten honen ingelesezko bertsio laburtua ondorengo helbide honetan aurki daiteke:

https://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1427215138/publikoak/Thesis_summary_English.pdf

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

1 Tesi lanaren aurkezpen orokorra

1.1 Informazio erauzketa

Informazio erauzketa, komertzialki "testuen analitika" izenarekin ere ezagguna, testuetatik informazio egituratua eskuratzean datza. Informazio erauzketa sistemek corpusetatik garrantzizko informazioa bilatzen saiatzen dira, eta informazio hori gizaki eta ordenagailuentzat interpretatzeko erraza den formatu batean itzuli.

Informazio erauzketaren bidez, medikuntzako ikerkuntza literaturatik botika eta geneen arteko produktu iterazioa ikas dezakegu; pertsonen bibliografietatik jaioterria eta lanbidea, besteak beste; erakundeen txostenetatik diru sarrerak, etekinak, langileak, egoitzak eta beste hainbat; edo tornadoei buruzko berrietatik hildakoak, kalteak, haizearen abiadura,...

Informazio erauzketa ondorengo atazetan banatzen da:

- **Entitate izenen ezagutza:** sistemek testuetan izenak bilatzen dituzte, eta ondoren hauek pertsonen izen, erakunde, toki, denbora adierazpen edo zenbakizko adierazpen bezala sailkatu:
  - **Cristina Cifuentesek** [*pertsona*] , **Madrilgo** [*lekua*] fiskal nagusi **Manuel Moixi** [*pertsona*] idatzi bat bidali dio eskatuz **Soziedad Alkoholika** [*erakundea*] musika taldeak larunbatean eskaini behar duen kontzertua bertan behera uzteko.[1]

[1] Testua webgune honetatik hartua eta moldatua: http://www.berria.eus/albisteak/108782/soziedad_alkoholikaren_kontzertu_bat_debekatu_du_madrilgo_udalak.htm

- **Korreferentziaren ebazpena:** korreferentzia eta lotura anaforikoen antzematea testuetako entitateen artean:
  - LDPko buruek Mori hautatu zuten apirilean **Keizo Obuchi** orduko lehen ministroa ordezkatzeko, **hark** tronbosia izan ostean.[2]
- **Terminologia erauzketa:** corpus batetik termino garrantzitsuak automatikoki erauzi:
  - **Sare neuronalak** geruzatan antolatzen dira: **sarrera geruza**, **geruza ezkutua** (nahi adina geruza ezkutu jar daitezke) eta **irteera geruza**.[3]
- **Erlazio erauzketa:** entitate eta hauen ezaugarrien arteko erlazioak topatu:
  - **Werner Heisenberg** fisikari **alemaniarra** **Würzburg**en munduratu zen **1901ean**.[4]
    Erauzitako erlazioak:
    * Werner Heisenberg - *jaioterria* - Würzburg
      (PERTSONA - *jaioterria* - LEKUA)
    * Werner Heisenberg - *jaiotze_estatua* - Alemania
      (PERTSONA - *jaiotze_estatua* - LEKUA)
    * Werner Heisenberg - *jaiotze_urtea* - 1901
      (PERTSONA - *jaiotze_urtea* - DENBORA_ADIERAZPENA)
    * Werner Heisenberg - *lanbidea* - Fisikaria
      (PERTSONA - *lanbidea* - LANBIDEA)
- **Gertaeren erauzketa:** testuetan gertaerak aurkitu eta hauei buruko informazioa eskuratu:
  - **200.000** pertsonek protestatu dute **Pakistanen**.[5]
    Gertaerari buruz erauzitako informazioa:

[2] Adibidea Soraluze *et al.* 2012 lanetik hartu da.

[3] Testua artikulu honetatik hartu da: http://zientzia.eus/artikuluak/geneak-neuronak-eta-ordenagailuak/

[4] Testua artikulu honetatik hartu da: http://zientzia.eus/artikuluak/heisenberg-werner/

[5] Adibidea ACE 2005 corpusetik hartu da eta euskaratu, *CNN_CF_20030303.1900.06-2.apf.xml* dokumentutik.

    * *Mota:* Gatazka-Manifestazioa
    * *Aingura:* protestatu
    * *Lekua:* Pakistan
    * *Entitatea:* 200.000

Informazio erauzketa lengoaia naturalaren prozesamenduko hainbat atazetarako oso erabilgarria da, hala nola testu sinplifikaziorako (Klebanov *et al.* 2004), DBpedia eta Freebase bezalako ezagutza baseak aberasteko (Mintz *et al.* 2009), galdera-erantzun sistemetarako (Ravichandran eta Hovy 2002), testuen laburtzapen automatikorako (Mihalcea eta Tarau 2004), itzulpen automatikorako (Babych 2005), iritzien metaketarako (Pronoza *et al.* 2014), eta proteinen iteraziorako (Miyanishi eta Ohkawa 2013), besteak beste.

Tesi honek **erlazio erauzketan** eta **gertaeren erauzketan** jartzen du arreta **ezagutza baseen aberasketarako**. Ataza hauek jarraian azalduko ditugu.

Erlazio erauzketa

Erlazio erauzketak entitateen arteko erlazio semantikoak testuetan bilatu eta sailkatzen ditu, hala nola *kokapena*, *lantokia*, *sortzailea*, *bikotekidea* eta beste hainbat. Adibide bezala, erlazio erauzketa sistema batek ondorengo testua jasotzen badu:

- **Akerbeltz garagardoaren** ekoizleek fabrika berria estreinatu dute **Azkainen**.[6]

Sistemak Akerbeltz enpresaren lantoki bat Azkainen dagoela erauzten du, errepresentazio egituratu bat jarraituz, honela:

- Akerbeltz garagardoa - *lantokia* - Azkaine
  (Erakundea - *lantokia* - lekua)

Goian, erlazio tripleta bat erauzi dugu. Erlazio tripletak bi entitatek eta hauen arteko erlazioak osatzen dute. Hemen erabilitako formatua *"subjektua - erlazioa - objektua"* da; erlazio zehatz bati buruz hitz egitean, formatu hau erabiliko dugu.

[6]Testua artikulu honetatik hartu da: http://info.kazeta.eus/euskalherria/xiberotik-lapurdira-deslokalizatu-ondoren-akerbeltzek-garagardogintzan-\urrats-berriak-eman-nahi-ditu

Erlazio bateko *subjektua* **entitate nagusia** bezala ezagutzen da. *Objektua* (edo **betegarria**) ez da beti entitate bat, batzuetan atributu bat delako. Atributuak lanbideak, erlijioak edo tituluak izan daitezke pertsonentzat, eta webguneak erakundeentzat. Tesi honetan azalpenak errazteko, entitateei buruz hitz egingo dugu orokorrean, eta atributuak aipatu beharrezkoa den uneetan. Erlazio erauzketa sistema gehienak erlazio bitarrak erauzteaz arduratzen dira, baina posible da maila altuagoko erlazioak topatzea ere, nahiz eta hauek oso ohikoak ez izan (Bach eta Badaskar 2007). Modu desberdinak daude erlazio erauzketa sistemak sortzeko: eskuz idatzitako patroiak erabiliz; edo ikasketa gainbegiratua, ez-gainbegiratua edo erdigainbegiratua erabiliz.

Tesi hau **urruneko gainbegiraketan** oinarritzen da, teknika erdigainbegiratu bat, 1.2 atalean azalduko duguna. Gure esperimentuetan erabilitako erlazio mota guztiak **bitarrak** dira.

Gertaera erauzketa

Gertaeren erauzketa sistemek gertaera batean parte hartzen duten elementu desberdinen eginbeharrei buruzko informazioa erauztea dute helburu. Gertaerak mota askotakoak izan daitezke, hala nola negozioei buruzkoak (bateratutako erakundeak, porrot egin dutenak,...), gatazkei buruzkoak (manifestazioak, erasotutako estatuak,...), justiziari buruzkoak (atxiloketak, epaiketak,...), edo hondamendi naturalei buruzkoak (lurrikarak, urakanak,...). Hegazkin istripu bati buruzko ondorengo esaldia aztertzen badugu:

- **20** hildako eta **15** zauritu **US Airways** konpainiaren **Boeing 747** hegazkinaren istripuan.

Gertaerari buruzko ondorengo informazioa erauzi dezakegu:

- *Gertaera mota:* Hegazkin istripua
- *Hildakoak:* 20
- *Zaurituak:* 15
- *Konpainia:* US Airways
- *Hegazkin mota:* Boeing 747

Tesi honetan, urruneko gainbegiraketa erabiltzen duen gertaera erauzketa sistema bat garatu dugu, lurrikarei buruzko informazioa erauzteko.

Ezagutza baseen aberasketa

Ezagutza base batek egituratutako informazio konplexua biltegiratzen du ordenagailuetan. Guk Freebase[7] eta DBpedia[8] bezalako ezagutza base handietan jarriko dugu arreta; ezagutza base hauek erdiautomatikoki sortuta daude, Wikipediako artikuluen infotauletan oinarrituta.

Ezagutza baseen aberasketa, osatu gabeko ezagutza baseak hartu eta falta diren elementuak corpus handietan bilatzean datza. Beste era batera esanda, ordenagailuak testua "irakurri" behar du, eta hemendik informazioa eskuratu. Ataza hau ezagutza basea hasieratik sortuz egin daiteke, edo aurretik sortutako ezagutza base baten edukia falta den informazioarekin aberastuz eta eguneratuz. Gu bigarrenean zentratuko gara: **ezagutza base baten egitura emanda dagoelakoan, eta informazioa falta delakoan, hau aberastu**.

Bada ataza bat, NISTek antolatua, *Knowledge Base Population (KBP)* deritzona eta *Text Analysis Conference*ekoa dena. Ataza honen xedea informazio erauzketa eta galdera-erantzun sistemen komunitateak batzean datza, entitateei buruzko egitateei aurkikuntzan ikerketa sustatzeko, ezagutza baseak egitate berriekin hedatuz. KBP atazak bi azpiataza desberdin ditu (Ji eta Grishman 2011):

- **Entitate izenen desanbiguazioa (*Entity Linking*):** Entitate izenak testuetatik erauzi, eta ezagutza basean dagokien sarrerarekin lotu.
- **Erlazioen betetzea (*Slot Filling*):** Entitate jakin baten erlazioei buruzko informazioa eskuratu corpus batetik eta ezagutza basea aberastu.

Tesi honetan zehar, 2011ko **Slot Filling** atazan hartu genuen parte. Ataza hau erreferentzia bezala erabili da tesi honetako esperimentu desberdinentzat.

Gure esperimentuak **ingeleserako** egin ditugu. Hemendik aurrera, adibide guztiak **ingelesez** izango dira, salbuespen batzuk izan ezik.

1.2 Urruneko gainbegiraketa

Urruneko gainbegiraketa Mintz *et al.*-ek (2009) proposatutako paradigma bat da erlazioen erauzketarako. Hurbilketa hau ezagutza base batean aur-

[7] http://www.freebase.com/
[8] http://dbpedia.org/About

1.1 irudia – Aretha Franklini buruzko informazioa Freebaseen.

ki dezakegun informazioa etiketatu gabeko testuekin batera lotzean datza. Algoritmoak corpusetako informazioa automatikoki etiketatzen du, eta ikasketa gainbegiratu eta ez-gainbegiratuen algoritmoen abantailak konbinatzen ditu. Urruneko gainbegiraketaren helburu nagusienetako bat corpusen eskuzko etiketatzea ekiditzea da. Hurbilketa hau domeinuekiko independentea da, eta pertsona, erakunde eta tokiei buruzko informazioa erauzteko erabili da bereziki. Mintz *et al.*-ek dioten bezala:

> "Urruneko gainbegiraketaren arabera, ezagutza base bateko erlazio batean parte hartzen duten bi entitate esaldi berean agertzen badira, esaldi horrek erlazio hori adieraziko du nola edo hala."

Urruneko gainbegiraketaren algoritmoa datu base batek gainbegiratzen du, eta ez ditu sistema gainbegiratuek sortzen dituzten gainkarga eta domeinu dependentzien arazoak pairatzen. Sistema ez-gainbegiratuek ez bezala, sailkatzaileen irteerek erlazioentzako izen kanonikoak erabiltzen dituzte (Mintz *et al.* 2009).

Hala eta guztiz ere, urruneko gainbegiraketarekin lan egiteko ezagutza base baten beharra dugu. 1.1 irudiak adibidez, Aretha Franklini buruzko

Esaldia	Erlazioa
Aretha Franklin was born on **March 25, 1942**, in Memphis, Tennessee.	*date_of_birth*
Aretha Franklin (born **March 25, 1942**) began as a gospel singer.	*date_of_birth*
Aretha Franklin (born March 25, 1942) began as a gospel **singer**.	*occupation*
(...) failed to show legendary **singer Aretha Franklin** any respect.	*occupation*
Aretha Franklin is a legendary American soul **singer**.	*occupation*
Aretha Franklin was born on March 25, 1942, in **Memphis**, Tennessee.	*city_of_birth*
The 'Queen of Soul' **Aretha Franklin** was born right here in **Memphis**.	*city_of_birth*

1.1 taula – Aretha Franklini buruzko esaldi desberdinak, entitate desberdinekin bikoteak sortuz erlazio ezagun bat adierazteko. Eskuineko zutabeak pertsonaren eta letra lodiz adierazitako entitatearen arteko erlazioa zehazten du.

informazioa erakusten digu Freebase ezagutza basean[9]. Freebaseen, artikuluaren izena topatzen dugu entitate nagusi bezala (Aretha Franklin), eta ezagutza basetik erauzitako erlazio asko, adibidez[10]:

- Aretha Franklin - *date_of_birth* - March 25, 1942
- Aretha Franklin - *occupation* - Singer
- Aretha Franklin - *city_of_birth* - Memphis

Corpus baten beharra ere dugu entitate nagusiaren (Aretha Franklin) agerpenak dituzten esaldiak eskuratzeko. Behin testuinguru guztiak lortuta, ezagutza basean bilduta dauden gainerako entitateak topatu behar ditugu testuinguruetan. Topatutako entitateetako bat entitate nagusiarekin erlazionatuta badago ezagutza basearen arabera, orduan dagokion erlazioarekin etiketatuko dugu esaldia. 1.1 taulak kantari famatuari buruz aipatu berri ditugun tripletei buruzko esaldi desberdinak biltzen ditu. Aipamen hauei **aipamen positibo** deitzen diegu.

Bestalde, ezagutza basean erlazio espliziturik ez duten entitateak testuinguru berean agertzen badira, hauek erlazionatu gabeko bezala (*unrelated*) hartzen dira. Aipamen hauei **aipamen negatibo** deitzen diegu. Adibide bezala, ezagutza basean ez dago inongo erlaziorik zehaztuta abeslariaren

[9] http://www.freebase.com/m/012vd6

[10] Irakurketa errazteko, Freebaseen erlazio izenentzat izen intuitiboak erabiliko ditugu, hala nola *date_of_birth* izena */people/person/date_of_birth* ordez, *city_of_birth* izena */people/person/place_of_birth* ordez, eta *occupation* izena */people/person/profession* ordez.

eta San Francisco hiriaren artean, hori dela eta, ondorengo esaldian entitate bikotea erlazionatu gabetzat hartzen da:

- **Aretha Franklin** will perform in **San Francisco** next week.

Corpusetik entitate nagusiak agertzen diren esaldi guztiak eskuratu, eta hauetan parte hartzen duten entitateen arteko erlazioak zehaztu ondoren, ikasketa gainbegiratua burutzen dugu. Entrenamendurako, esaldi bakoitzaren ezaugarri linguistikoak eskuratzen ditugu, jarraian sailkatzaileak hauek aztertu ditzan. Ikasketa prozesua ikasketa gainbegiratuan oinarritutako informazio erauzketa sistema bat bezalakoa da, non sailkatzaileek testuinguruak eta ereduak ikasten dituzten, eredu bat erlazio mota bakoitzeko.

Orain prest gaude ereduak aplikatu eta entrenamenduan azaltzen ez ziren entitateei buruzko informazioa testuetatik erauzteko. Horretarako ezagutza basean ez dauden, edo informazio gutxi duten entitateak hartzen ditugu, entitate hauek **helburu entitateak** dira. Ondorengo urratsek testeko prozesua azaltzen dute:

1. Helburu entitatea agertzen den esaldiak topatu eta corpusetik erauzten ditugu.

2. Esaldietan dauden beste entitateak antzeman eta markatzen ditugu, entitate hauek helburu entitatearekin erlazionatuta egoteko entitate potentzialak dira.

3. Esaldietan topatutako entitate eta helburu entitatearekin entitate bikoteak sortzen ditugu.

4. Entitate bikote bakoitzaren ezaugarri linguistikoak erauzten ditugu, eta erabaki ea elkarren artean erlaziorik ote dagoen.

Demagun gure testeko datu multzoan Morgan Freeman dela helburu entitatea, eta esaldietako bat ondorengoa dela[11]:

- It was on this date, June 1, 1937, American actor, film director, and narrator **Morgan Freeman** was born.

Sistemak asmatu behar du ea azpimarratutako informazioa helburu entitatearekiko erlazionatuta ote dagoen. Esaldiko entitate potentzial bakoitzaren ezaugarriak aztertu ondoren, sistemak ondorengo erlazioak erauziko ditu:

[11] Esaldia webgune honetatik eskuratu da: http://freethoughtalmanac.com/?p=6861

- Morgan Freeman - *date_of_birth* - June 1, 1937
- Morgan Freeman - *occupation* - actor
- Morgan Freeman - *occupation* - director
- Morgan Freeman - *occupation* - narrator

Morgan Freemani buruz erauzitako informazioa DBpedia[12] edo Freebase[13] ezagutza baseetan ote datorren egiaztatzen badugu, bi ezagutza baseetan falta den erlazio berri bat topatu dugula konturatzen gara:

- Morgan Freeman - *occupation* - narrator

Ezagutza base hauek ez dute inon aipatzen Morgan Freemanek filma eta dokumental batzuetan narratzaile lanak egin dituenik. Behin erlazio berri bat topatu dugula, ezagutza basera gehitu dezakegu informazio hori.

1.3 Urruneko gainbegiraketaren zailtasun nagusiak

Urruneko gainbegiraketa ikasketa gainbegiratua eta ez-gainbegiratuaren onena hartzen saiatzen da. Ikasketa gainbegiratuan bezala, testuinguruak ezaugarri aberats eta konplexuekin irudikatu ditzakegu. Aukera dugu ere erlazio izen kanonikoekin lan egiteko, ezagutza baseetako erlazio izenak normalizatuta daudelako, entrenamendu eta ebaluazio faseak asko erraztuz. Bestalde, ikasketa ez-gainbegiratuan bezala, etiketatu gabeko datu kantitate handiak erabili ditzakegu.

Nahiz eta urruneko gainbegiraketak harrera ona jaso lengoaia naturalaren prozesamenduko komunitatearen partetik, ikasketa gainbegiratu eta ez-gainbegiratuaren alternatiba bezala, hurbilketa honek arazo batzuk ditu, eta hauek konpontzea komeni zaigu. Atal honek tesi honetan zehar topatu ditugun arazo desberdinak zerrendatzen ditu. Guk egindako ekarpen gehienak hauei irtenbideak topatzea dira.

[12] http://dbpedia.org:8890/page/Morgan_Freeman
[13] http://www.freebase.com/m/055c8

Aipamen negatiboen beharra

Sailkatzaileek aipamen positibo eta negatiboen artean bereizteko beharra dute. Aipamen negatiboei esker, sailkatzaileari laguntzen diogu kontuan hartzen testuinguru batean ez dagoela zertan erlaziorik egon behar bi entitateren artean. Oso beharrezkoak dira urruneko gainbegiraketak eraginkortasun ona izatea nahi badugu. Artearen egoerako urruneko gainbegiraketa sistema guztiek erabiltzen dituzten datu multzoek aipamen negatiboak dituzte.

Aipamen positibo falta

Batzuetan erlazio batzuen maiztasuna datu multzoan oso txikia da, ez direlako oso ohikoak ezagutza basean. Erlazio horientzat aipamen gutxi baditugu entrenatzeko, beste erlazioekin konparatuta, helburu entitateentzat erlazio horiei buruzko informazioa iragartzea zaila izaten da. Arazo hau okerrera doa ez badugu corpusetik erlazio batentzako inongo aipamenik eskuratu. Espero bezala, sistemak ezingo du erlazio horrentzako inongo iragarpenik egin. Arazo hau ikasketa gainbegiratuan ere topatu dezakegu. Tesi honek ez du arazo honen konponketan arretarik jartzen.

Aipamen zaratatsuak

Urruneko gainbegiraketak aipamen zaratatsu asko sortzen ditu. Baliteke hau izatea hurbilketaren arazo nagusi eta garrantzitsuena. Urruneko gainbegiraketaren arabera, aipamen batean agertzen diren bi entitateren artean erlazio bat baldin badago, aipamenak erlazio hori adierazten du, baina hau ez da beti horrela gertatzen. Aretha Franklini buruzko hainbat adibide eman ditugu 1.2 atalean, non uste hori betetzen den, baina hau ez da beti egia izaten. Tesi honetan hiru zarata mota kategorizatu ditugu:

1. **Testuinguru okerra:** Hau da zarata mota ohikoena. Gertatzen da ezagutza base batean erlazionatuta dauden bi entitate ditugunean, baina aipameneko testuinguruak ez duenean erlazio hori adierazten. 1.2 taulak 1.1 taulan erabilitako erlazioen bi adibide ditu, baina oraingoan testuinguruak ez du entitate bikoteen erlazioarekin bat egiten. Lehenengo aipamena Celine Dioni buruzko da, emakume hau ere abeslaria da, Aretha Franklin bezala, baina erlazioa hemen Celine Dionen lanbidea da. Bigarren aipamena Franklinek Memphisen emandako kontzertu bati buruzkoa da, kasualitatez bere jaioterria.

Esaldia	Erlazioa
Celine Dion is a great **singer** and a good friend of **Aretha Franklin**.	*occupation*
Aretha Franklin gave a great concert in **Memphis** last night.	*city_of_birth*

1.2 taula – Aretha Franklini buruzko testuinguru okerra. Eskuineko zutabeak letra lodiz zehaztutako entitateen artean adierazi nahi den erlazioa zehazten du.

Esaldia	Erlazioa
Celso Amorim, the Brazilian **Foreign Minister**, said the (...)	*unrelated*
Celso Amorim served as **Minister of External Relations** of Brazil.	*unrelated*

1.3 taula – Positiboak izan beharko liratekeen aipamen negatiboak, ezagutza base osatu gabea dela eta.

2. **Ezagutza base osatu gabea:** Zarata hau testuinguruak bi entitateren artean erlazio bat adierazten duenean gertatzen da, baina erlazio hori ez dugunean ezagutza base batean topatzen, ondorioz aipamena negatibotzat hartzen da datu multzoan. Erlazio erauzketa sistemak ikasiko du aipamenaren patroi horrek ez duela erlaziorik adierazten, emaitzak okertuz. Adibide bezala, Freebase ezagutza baseak informazio gutxi dauka Celso Amorimetaz, Brasileko defentsarako eta kanpoko harremanetarako ministro ohia. Ezagutza baseak ez du inon zehazten bere lana Brasileko gobernuan. Ondorioz, Celso Amorim eta Brasil agertzen diren aipamen guztiak negatibotzat hartuko dira, lanbidea adierazten duen erlazioa jaso ordez. 1.3 taulak bi aipamen negatibo ditu, non lanbidea esplizituki adierazten den.

3. **Erlazio anitza:** Ezagutza baseak bi entitateren artean erlazio bat baino gehiago zehazten dituenean gertatzen da fenomeno hau. Urruneko gainbegiraketa ez da gai erlazioak testuinguruaren arabera desanbiguatzeko, eta ezagutza baseak zehazten dituen erlazio guztiekin etiketatzen du testuingurua. Adibide bezala, har dezagun <Rupert Murdock - News Corporation> tupla, bien arteko erlazioak *founder* eta *top_member* izanik. 1.4 taulak bi entitateen arteko bi esaldi erakusten ditu, urruneko gainbegiraketak bi esaldiei bi erlazioak jarriz (eskuineko zutabea), erlazio bakarra zuzena denean (letra lodiz dagoena).

Esaldia	Erlazioak
News Corporation was founded by **Rupert Murdock**.	**founder** / *top_member*
Rupert Murdock is the CEO of **News Corporation**	*founder* / **top_member**

1.4 taula – Erlazio anitzen adibide bat. Eskuineko zutabeak esaldietan letra lodiz markatutako entitateen arteko erlazioak adierazten ditu, ezagutza basean dagoen informazioaren arabera. Letra lodiz dauden erlazioak dira testuinguruarekiko zuzenak.

Aipamen zaratatsuen erruz, erlazio erauzketa sistema bateko moduluek okerreko erlazioak ikasiko dituzte, iragarpen okerrak emanez.

Tesi honetan, **testuinguru okerra** eta **erlazio anitzaren** arazoak konpontzen ahalegintzen gara. Erabilitako ezagutza baseak guztiz osatuta daudela uste izaten dugu.

Antzeko balioak

Badira kasuak non ezagutza basean eta corpusean topatzen dugun informazioa elkarren artean antzekoak diren, baina ez berdin-berdinak. Entrenamendurako aipamenak sortzean, eskuratutako esaldietan entitateen arteko erlazioak bilatzean (baldin badago), ezagutza baseko entitateekin bat egiten duten aipamenak bilatzen ditugu. Ezagutza basearekiko antzekoa den informazioa bertan behera uzten da.

Balio zehaztugabeak, urre patroiko (ezagutza baseko) balioekin guztiz bat egiten ez dutenak, baliagarriak izan daitezke erlazio jakin batzuen patroiak ikasi eta iragarpen onak itzultzeko. Kontrako egoeran, sailkatzaileak bi esaldi jasotzen baditu, patroi berdinarekin, baina bat positiboa eta bestea negatiboa, sailkatzailea nahastu besterik ez da egingo. Hori dela eta, urruneko gainbegiraketaren eraginkortasuna kolokan jartzen da, informazio garrantzitsu asko galtzen delako. Arazo hau ez da bakarrik urruneko gainbegiraketarekin gertatzen, baita informazio erauzketako beste hainbat atazatan ere.

Har dezagun entrenamenduko datu multzoko erakunde bat, ezagutza basearen arabera 140 langile dituena. Corpuseko dokumentu batek 144 langile dituela esan dezake, agian artikulua duela gutxikoa delako eta ezagutza basearen azken bertsioa baino berriagoa. Urruneko gainbegiraketak 144 balioa alde batera utziko luke, honi dagokion esaldiak negatibo bezala utziz.

Horrez gain, ebaluatzaile tradizionalek urre patroiarekiko antzekoak diren iragarritako balioak txartzat hartzen dituzte, kontuan hartu gabe iragarritako balioa zein antzekoa den ezagutza baseko balioarekiko. Ezagutza baseetako informazioa perfektua balitz bezala hartzen da kontuan, baina ezin dugu bermatu hau beti zuzena denik. Kontuan hartuta aurreko paragrafoan jarritako adibidea, erakunde hau helburu entitatea bada, honi buruzko erlazioak erauztean, langile kantitatea 144 dela dioten iragarpen guztiak okertzat hartuko ziren. Balio zaharkitu bat zuzentzat hartuko zen, eta balio eguneratuago bat okertzat.

Tesi honetan, ezagutza basearekiko antzeko balioak dituzten aipamenak erabiltzea proposatzen dugu baita entrenamendurako ere. Ebaluazio metrika leun bat ere proposatzen dugu, ebaluazioan antzeko balioak partzialki zuzentzat hartzeko.

1.4 Tesi txostenaren ekarpenak

Atal honetan lan honen ekarpen nagusiak zerrendatzen ditugu, aldi berean tesian dagokien kapitulua adieraziz:

- Urruneko gainbegiraketaren zailtasun nagusiak aztertzen ditugu (*Kapitulu guztiak*).
- Urruneko gainbegiraketan topatu ditzakegun zarata mota desberdinak analizatu ditugu (*4. kapitulua*).
- Aipamen positiboetako testuinguru okerrak ezabatzeko heuristikoak proposatzen ditugu (*4. kapitulua*).
- Urruneko gainbegiraketarako ebaluazio erlaxatu bat proposatzen dugu (*6. kapitulua*).
- Urruneko gainbegiraketarako hedapen bat proposatzen dugu antzeko balioak etiketatzeko (*6. kapitulua*).
- Gertaeren erauzketarako ezagutza base bat eta datu multzo bat sortzen ditugu (*5. kapitulua*).
- Urruneko gainbegiraketa gertaera konplexutan aplikatzen dugu (*5. eta 6. kapituluak*).

- Urruneko gainbegiraketa sare sozialetan aplikatzen dugu (*5. eta 6. kapituluak*).

Ekarpen hauek azkenengo atalean sakonduko ditugu.

1.5 Aurrekariak IXA taldean

Tesi hau lengoaia naturalaren prozesamenduan aritzen den IXA taldean burutu da. Euskal Herriko Unibertsitateko ikerkuntza talde hau hizkuntzaren prozesamenduan aritu da hogeita bost urte baino gehiagotan. Nahiz eta talde honen ikerketa nagusiak euskararako izan, beste hizkuntzetarako ere egiten dituzte ikerketak eta hainbat tresnaren garapenak.

Fernandezek (2012) euskarazko entitate izenekin egin du lan. Bere tesian entitate izenak antzeman, sailkatu, itzuli eta desanbiguatu ditu. Horretarako, metodo gainbegiratu, ez-gainbegiratu eta erdigainbegiratuak erabili ditu ataza bakoitzean, ikasketa teknika bakoitzaren eraginkortasuna neurtzeko gaitasuna edukiz. Entitate izenen tratamendua automatizatzeko, euskarazko ezaugarri morfosintaktikoek duten eragina ere ikertu du.

Urizarrek (2012) euskarazko hitz anitzen edo *multi-word*en unitate lexikalen antzematean egin du lan, hauen analisian laguntzen duen *HABIL* sistema diseinatu eta garatuz. Bestalde, Gurrutxagak (2014) *hitza+aditza* formatuko unitate fraseologikoen erauzketa automatikoan egin du lan.

Kyoto Proiektuan[14] ezagutza ekoizten duten robotekin egin dira lanak, robot hauei *Kybot* deitzen zaie. Kybotek testuak aberasten dituzte informazio semantiko eta linguistikoekin, testuetan patroiak definituz eta hizkuntza guztietarako patroiak erauziz. Patroiak eskuz sortu dira. Kybotak klimari buruzko informazioa erauzteko erabili dira Kyoto Proiektuan hasiera batean, geroago biomedikuntzako domeinuan informazioa erauzteko ere erabili dira (Casillas *et al.* 2011), eta gaur egun beste domeinu batzuetatik informazioa eskuratzeko erabiltzen dira NewsReader Proiektuan[15].

[14]Kyoto Project: http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index.html

[15]http://www.newsreader-project.eu/

1.6 Tesi txostenaren eskema

Tesi honen motibazioa, helburuak eta ekarpenak aurkeztu berri ditugu. Jarraian, tesiaren egitura laburki azaltzen dugu:

- **1. kapitulua** - *Tesi lanaren aurkezpen orokorra.*

- **2. kapitulua** - *Artearen egoera.*

  Kapitulu honetan ezagutza-base desberdinak deskribatzen ditugu, baita datu multzo batzuk eta parte hartu dugun ataza bat ere. Atal bat eskaintzen diogu erlazio erauzketarako ikasketa teknika bakoitzari. Geroago gertaeren erauzketa eta urruneko gainbegiraketari ere jartzen dugu arreta. Amaitzeko, tesian erabilitako ikasketa metodo eta ebaluazio metrikei eskaintzen diegu tartetxo bat.

- **3. kapitulua** - *Urruneko gainbegiraketa sistema baten garapena.*

  Kapitulu honetan, ikasketa gainbegiratuko erlazio erauzketa sistema bat nola garatu dugun azaltzen dugu. Sistema honetan erabilitako ezaugarriak eta optimizazio metodoak urruneko gainbegiraketa sistema batean aplikatzen ditugu geroago. Bukatzeko, gure sistemari errore analisi bat egiten diogu, honen ahuleziak topatzeko.

- **4. kapitulua** - *Aipamen zaratatsuen garbiketa.*

  Kapitulu honetan datu multzoetan aurki ditzakegun aipamen zaratatsu mota desberdinen azterketa bat egiten dugu. Jarraian aipamen zaratatsu horiek filtratzeko teknika desberdinak aurkezten ditugu, urruneko gainbegiraketaren eraginkortasuna hobetuz.

- **5. kapitulua** - *Gertaeren erauzketarako datu multzo baten sorkuntza.*

  Kapitulu honetan, lurrikarei buruzko ezagutza base baten sorkuntzan zentratzen gara. Jarraian, lurrikarei buruzko txio ezberdinak eskuratzen ditugu Twitterretik, hauek normalizatu, eta informazioa eskuz etiketatu, pauso hauek nola egin ditugun azalduz. Ezagutza base hau eta datu multzo hau 6. kapituluan erabiltzen ditugu urruneko gainbegiraketa eta gertaeren erauzketa konbinatzen dituzten esperimentu eta analisietarako.

- **6. kapitulua** - *Gertaeren erauzketa urruneko gainbegiraketaren bidez.*

  Kapitulu honetan, urruneko gainbegiraketa aplikatzen duen gertaera erauzketa sistema bat garatzen dugu, aurreko kapituluan sortutako ezagutza basea eta datu multzoa erabiliz. Eskuz etiketatutako datu multzoa analisirako erabiltzen dugu, urruneko gainbegiraketaren beste ahulezia batzuk topatzeko. Analisiaren ondoren, ebaluazio erlaxatu bat proposatzen dugu, non ezagutza basean topatu ditzakegun balioen antzekoak diren aipamenak ontzat hartzen ditugun. Urruneko gainbegiraketako algoritmoan ere hedapen bat proposatzen dugu, antzeko balioak ere automatikoki etiketatuz. Amaitzeko, lurrikara berdinari dagokien txioen artean ezaugarriak elkarbanatzea proposatzen dugu. Proposamen hauen konbinaketak urruneko gainbegiraketaren emaitzak nabarmen hobetzen ditu.

- **7. kapitulua** - *Ondorioak eta etorkizuneko lanak.*

  Azken atal honetan tesiaren ekarpenak eta ondorio nagusiak laburtzen ditugu. Bukatzeko, etorkizunerako irekita ditugun ikerketa lerroak deskribatzen ditugu.

1.7 Tesi honen garapenetik atera diren argitalpenak

Atal honetan, tesi honekin erlazionatuta dauden argitalpenak zerrendatzen ditugu, tesian dagokien kapituluen arabera sailkatuta.

- **3. kapitulua**:
  - Ander Intxaurrondo, Oier Lopez de Lacalle eta Eneko Agirre. **UBC at Slot Filling TAC-KBP 2011.** *Proceeding of the TAC-KBP 2011 Workshop* (TAC-11). Gaithesburg, Maryland, USA, 2011.

- **4. kapitulua**:
  - Ander Intxaurrondo, Mihai Surdeanu, Oier Lopez de Lacalle eta Eneko Agirre. **Removing Noisy Mentions for Distant Supervision.** *Proceedings of the XXIX Conference of Sociedad Es-*

pañola para el Procesamiento del Lenguaje Natural (SEPLN-13). Madrid, Spain, 2013.

- **5. eta 6. kapituluak**:
  - Ander Intxaurrondo, Eneko Agirre, Oier Lopez de Lacalle eta Mihai Surdeanu. **Diamonds in the rough: Event Extraction from Imperfect Microblog Data.** *NAACL HLT 2015*. Denver, Colorado, USA. 2015.
  - Ander Intxaurrondo, Eneko Agirre eta Oier Lopez de Lacalle. **Lurrikarei buruzko informazioa eskuratzen Twitter bidez.** *Ikergazte 2015*. Durango, Bizkaia, Euskal Herria. 2015.

2

Artearen egoera

Kapitulu honetan, ezagutza base erabilienak eta gure esperimentuetan erabilitako datu multzoak deskribatuko ditugu. Jarraian, erlazio erauzketarako erabiltzen diren metodo desberdinak azalduko ditugu, urruneko gainbegiraketa erabiliz erlazioak erauzteko dauden lan garrantzitsuenak ere deskribatuz. Ondoren gertaera erauzketan zentratuko gara, dauden ikasketa teknika desberdinak azalduz, eta gertaera erauzketari urruneko gainbegiraketa aplikatzeko egin diren urratsak komentatuz. Geroago tesian zehar erabili diren ikasketa automatikorako algoritmo desberdinei buruzko azalpen laburra egingo dugu. Kapitulua gure esperimentuetan erabili ditugun ebaluazio metrikak azalduz amaituko dugu.

2.1 Ezagutza baseak eta datu multzoak

Urruneko gainbegiraketarekin lan egiteko, sarreran komentatu dugun bezala, ezagutza baseen eta datu multzoen beharra dugu. Atal honetan ezagutza base ezagun batzuk eta tesi honetan erabilitako datu multzoak deskribatuko ditugu.

2.1.1 Ezagutza baseak

Ezagutza base bat ezagutza kudeatzeko datu base berezi bat da. Ezagutzaren bilketa, antolaketa eta berreskurapena konputazionalki egiteko baliabideak hornitzen ditu.

Mikel Laboa
Mikel Laboa - *Jaioteguna* - 1934ko ekainaren 15a
Mikel Laboa - *Jaioterria* - Donostia
Mikel Laboa - *Lanbidea* - Musikagilea
Mikel Laboa - *Bizilekua* - Donostia
Mikel Laboa - *Bizilekua* - Lekeitio
Mikel Laboa - *Bikotea* - Marisol Bastida

2.1 taula – Ezagutza base batean Mikel Laboari buruz agertzen den informazioa, bai pertsonek bai ordenagailuek erraz irakurtzeko formatu batean.

Azken urteetan informazio erauzketan eta lengoaia naturalaren prozesamenduan geroz eta gehiago erabiltzen dira. Ataza hauetarako erabiltzen diren ezagutza baseak entitate desberdinek osatzen dituzte, eta entitate hauei buruzko informazioa era eskematikoan eta ulergarrian irudikatzen dute. Entitate bakoitzak erlazio batzuk ditu, erlazio bakoitzak izen bat jasotzen du eta beste entitate batekin erlazionatuta dago. Ezagutza baseetan aurki ditzakegun entitateak pertsona, erakunde, leku, denbora adierazpen eta beste hainbat motatakoak izan daitezke, bai entitate nagusia, baita erlazioan parte hartzen duen bigarren entitatea (edo betegarria). Entitateez gain, atributu desberdinak ere erlazionatzen dira, pertsona baten lanbidea edo erakunde baten webgunea adibidez.

Ezagutza baseei buruzko ideia bat egiteko, 2.1 taulan Mikel Laboari buruzko informazioa agertzen da, ezagutza base batean agertuko litzatekeen moduan. Hasteko entitate nagusia dugu (Mikel Laboa, interesatzen zaigun entitatea informazioa lortzeko) dugu, ondoren erlazioa (adibidez, jaioterria), eta amaitzeko, erlazioan parte hartzen duen entitatea (Donostia). Bi entitateren artean erlazio bat baino gehiago egon daiteke, Laboa eta Donostiaren artean jaioterria eta bizilekua adibidez.

Jarraian informazio erauzketarako erabiltzen diren ezagutza base batzuk deskribatuko ditugu.

Wikipediako infotaulak

Wikipediak badauka informazioa era egituratuan emateko modu bat, eskematizatuta, ordenagailu bidez nahi dugun formatura pasa, eta ordenagailuek ere erraz irakurri ahal izateko: infotaulak. Infotaulak Wikipediako artikulu

(a) Mikel Laboa Wikipedian. **(b)** Mikel Laboa Freebasen.

dbpedia-owl:birthPlace	▪ dbpedia:Donostia ▪ dbpedia:Gipuzkoa
dbpedia-owl:citizenship	▪ dbpedia:Euskal_Herria
dbpedia-owl:deathPlace	▪ dbpedia:Donostia ▪ dbpedia:Euskal_Herria ▪ dbpedia:Gipuzkoa
dbpedia-owl:nationality	▪ dbpedia:Euskal_Herria

(c) Mikel Laboa DBpedian.

baten eskuinaldean aurki ditzakegu, artikuluko informazioaren laburpen bat emanez. 2.1a irudian euskarazko Wikipediako Mikel Laboaren[1] infotaula dugu. Infotaulek metadatu egituratuak igortzen dituzte, beste ezagutza base batzuk sortzeko, erabilienak DBpedia eta Freebase dira.

DBpedia

DBpedia[2] ezagutza basea Wikipediatik erauzitako datuez osatuta dago. Proiektu hau Leipzigko Unibertsitateak, Berlineko Unibertsitate Libreak eta OpenLink Software enpresak kudeatzen dute. DBpediaren bitartez kontsulta konplexuak egin daitezke, Wikipedian dagoen informazio kopuru handia hobeto aztertzeko.

[1] http://eu.wikipedia.org/wiki/Mikel_Laboa
[2] http://dbpedia.org/About

Ingeleseko 2014ko iraileko bertsioan 4.58 milioi entitateei buruzko informazioa dago, horietatik 1.44 milioi pertsonak dira, eta 735 mila lekuak. Une honetan 111 hizkuntzatan dago ezagutza basea, tartean euskarazko bertsio[3] bat ere.

Mikel Laboarentzat ere badago tartea euskarazko[4] eta ingelesezko[5] DBpedian. Bi orrialdeak konparatzen baditugu, bietan informazio falta dagoela nabari da: euskarazko bertsioan jaio eta hil zen tokiei[6] buruzko informazioa agertzen da, ingelesekoan berriz informazio hau falta da, baina badauka jaio eta hil zen egunei[7] buruzko informazioa. Tesi honetan ezagutza baseen aberasketan lan egiten dugunez, Mikel Laboa bezalako entitateak egokiak dira hauei buruzko informazioa testuetan bilatu eta ezagutza base hauek betetzeko. 2.1c irudian Mikel Laboari buruzko euskarazko DBpediako artikuluaren pasarte bat dago, jaioterria, hiritartasuna eta beste informazio gehiago emanez. Nahiz eta informazioa euskaraz ageri, erlazio izenak ingelesez daude.

Freebase

Asko erabiltzen den beste ezagutza base bat Freebase[8] da, komunitate batek metadatu bidez sortutako ezagutza base librea. 2014ko urtarrilerako, 44 milioi entitate eta 2.4 mila milioi erlaziori buruzko informazioa zeukan Freebaseek. Freebaseeko edukiaren zati bat Wikipediako infotauletan oinarritzen da, baina erabiltzaileek ere gehitu dezakete informazio berria ezagutza basean. Freebase ingeleserako bakarrik dago eskuragarri une honetan.

Mikel Laboarentzat ere badago tokia Freebaseen[9]. Jaioteguna eta jaioterriari buruzko informazioa[10] aurki dezakegu, baina bere heriotzari buruz ez dago ezer. Ezagutza base honetan ere Mikel Laboari buruzko informazio ugari falta da, eta DBpedian bezala, informazio asko dugu dokumentuetatik erauzi eta bi ezagutza baseak aberasteko. 2.1b irudian Mikel Laboari buruzko Freebaseeko informazioaren zati bat dago, jaioteguna, jaioterria eta lanbideari buruzko informazioa emanez.

[3] http://eu.dbpedia.org
[4] http://eu.dbpedia.org/page/Mikel_Laboa
[5] http://dbpedia.org/page/Mikel_Laboa
[6] *dbpedia-owl:birthPlace* eta *dbpedia-owl:deathPlace*
[7] *dbpedia-owl:birthDate* eta *dbpedia-owl:deathDate*
[8] http://www.freebase.com/
[9] http://www.freebase.com/m/03nvxd
[10] */people/person/date_of_birth* eta */people/person/place_of_birth*

Wikidata

Wikidata[11] ezagutza base berri bat da. Helburua Wikipediako infotauletako informazio guztia bateratzea da, informazio guztiak koherentzia eduki dezan. Wikipediako infotaulek askotan izen desberdinak hartzen dituzte, nahiz eta esanahi berdina eduki; adibidez, erakundeen sorrera eguna adierazteko, 4 izen desberdin aurki ditzakegu: *Founded, Established, Formed* eta *Founded-date*. Wikidatari esker, izen bakarra erabiliko litzateke, eta honi esker, Wikipedian oinarritutako ezagutza base berriak sortzean koherentzia ziurtatuta edukiko genuke. Aurreko adibiderako aukeratutako terminoa *Inception* da.

Wikidataren beste helburu bat Wikipediako infotaulak eguneratzea da, aukeratutako terminoak erabiliz. DBpediak eta Freebasek terminologia batua erabiltzeko apustua egin zuten hasieratik, infotaulak betetzeko asmoz, baina Wikidata proiektua dela eta, hauek egindako lan guztia eta etorkizuna kolokan jarri da. Freebase 2015ean zehar itxiko dute, baina egindako lana alferrik ez galtzeko, Googlek eta Freebaseeko erabiltzeek Freebaseeko informazio guztia Wikidatara esportatuko dute[12].

Wikidata proiektuan oraindik lan asko dago egiteko, hala ere, Mikel Laboari buruzko zenbait informazio[13] aurki dezakegu.

Knowledge Vault

Beste ezagutza base berri bat Googlek 2014an sortutako Knowledge Vault da (Dong *et al.* 2014). Ezagutza base hau, Wikipedian oinarritu ordez, internetetik informazioa automatikoki eskuratuz sortu da. Interneteko edukia analizatzen dute, atal honetan aipatu ditugun ezagutza base desberdinen laguntzaz. Ezagutza base hau automatikoki sortutako beste ezagutza base batzuk baino 38 aldiz handiagoa da, dagoeneko 1.6 mila milioi erlazio desberdin eskuratuta.

Ezagutza base hau ikerketa proiektu bat da oraindik, eta ez dago erabiltzaileentzat ofizialki eskuragarri.

[11] http://www.wikidata.org

[12] Informazio gehiago Google+ sare sozialeko Freebasen orrialde ofizialean: https://plus.google.com/109936836907132434202/posts/bu3z2wVqcQc

[13] http://www.wikidata.org/wiki/Q379992

Beste batzuk

Amaitzeko, badira beste ezagutza base batzuk, hain erabiliak ez direnak, baina hauen existentzia gutxienez aipatu beharko genukeena:

- **YAGO**[14]: Saarbrückeneko (Alemania) Max-Planck institutuak sortua. Wikipedia, WordNet eta GeoNames datu base geografikoaren bidez sortu da. Ezagutza base honen edukia DBpedian integratu da.

- **Microsoft Satori**: Bing bilatzailean entitate izenei buruzko informazio hobea emateko helburuarekin sortua.

- **Google Knowledge Graph**[15]: Google bilatzailean entitate izenei buruzko informazio hobea emateko helburuarekin sortua. Edukia Wikidatan integratu da.

2.1.2 Datu multzoak

Erlazio eta gertaera erauzketan, gehien erabiltzen diren hiru datu multzo aipatuko ditugu. Hirurak tesi honetan erabili ditugu.

Automatic Content Extraction (ACE)

ACE ataza NIST (*National Institute of Standards and Technology*) erakundeak kudeatu zuen 2005 urtean. Atazaren helburua etiketatutako datu multzo batetik ikasketa gainbegiratuan oinarritutako erauzketa sistemak sortzea zen. 3 ikerketa helburu desberdin ditu atazak: entitateen antzematea eta jarraitzea, eta erlazio aipamen eta gertaeren erauzketa.

Datu multzo honetan, 6 entitate mota desberdinek hartzen dute parte: pertsonak, erakundeak, tokiak, armak, ibilgailuak eta zerbitzuak. Entitate mota bakoitzak bere azpimotak ere ditu. 2.2 taulan ikus ditzakegu datu multzoko entitate mota guztiak eta hauen azpimotak. 2.3 taulan berriz, datu multzoan aurki ditzakegun erlazio motak eta azpimotak daude zerrendatuta. 6 erlazio guztira, 17 azpi erlazio. Entrenamendu osoan 16771 erlazio aipamen daude etiketatuta.

[14] http://www.mpi-inf.mpg.de/yago-naga/yago/
[15] http://www.google.com/insidesearch/features/search/knowledge.html

Entitate mota	**Azpimota**
PER (Person)	Group, Intermediate, Individual
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
LOC (Location)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
GPE (Geo-political)	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
WEA (Weapon)	Biological, Blunt, Chemical, Exploding, Nuclear, Projectile, Sharp, Shooting, Unspecified
FAC (Facility)	Airport, Building-Grounds, Paths, Plant, Subarea-Facility
VEH (Vehicle)	Air, Land, Subarea-Vehicle, Unspecified, Water

2.2 taula – ACE corpuseko entitate motak eta azpimotak.

Erlazio mota	**Azpimota**
ART (Artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (Gen-Affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY	*none*
ORG-AFF (Org-Affiliation)	Employment, Founder, Investor-Shareholder, Membership, Sports-Affiliation, Student-Alum
PART-WHOLE	Artifact, Geographical, Subsidiary
PER-SOC (Person-Social)	Business, Family, Lasting-Personal
PHYS (Physical)	Located, Near

2.3 taula – ACE corpuseko erlazio motak eta azpimotak.

Datu multzo hau berri agentzien dokumentuez, telebista emanaldiez, elkarrizketetaz eta blogetako artikuluetaz osatuta dago. Guztira 1000 dokumentu baino gehiago dira eta 250 mila hitz inguru. Dokumentu bakoitzak bere XML fitxategia dauka, bertan dokumentuko entitateak, erlazioak eta gertaerak daude, eskuz etiketatuta.

Jarri dezagun adibide bezala *"Khoser River is in northern Iraq"* aipamena, non *Khoser River* eta *Iraq* entitateak erlazionaturik dauden, bien arteko erlazio mota *PART-WHOLE* da eta azpimota *Geographical*. A eranskinean, *Khoser River* eta *Iraq* entitateen arteko erlazioa dago XML kodean etiketatuta. Erlazioaz gain, bi entitateen XML kodea ere ageri da, dokumentu berdinean dituzten aipamen desberdinak zerrendatuta.

Ataza honetan, corpusak eta etiketatutako XML fitxategiak, ingeleserako gain, txinera eta arabierarako ere eskuragarri daude. Tesi honetan ingelesezko corpusa eta datu multzoarekin egiten dugu lan 3. kapituluan.

TAC-KBP lehiaketa eta *Slot Filling* ataza

TAC (*Text Analysis Conference*) biltzarra NIST erakundeak kudeatzen du, ACE bezala. KBP (*Knowledge Base Population*) atazako azpiataza eta lehiaketa bat da *Slot Filling* (*"erlazioen betetzea"* euskaraz), entitate izenei buruzko informazioa bilatu eta informazio hau ezagutza baseetan jartzeko helburuz. 1.5 milioi dokumentuz osatutako corpus erraldoi bat dago eskuragarri parte hartzaileentzat.

Lehiaketa honetan, parte hartzaileek hainbat entitateri buruzko informazioa erauzi behar dute corpusetik, eta ondoren Wikipediako infotauletan oinarritutako ezagutza base batean informazio hori gehitu. Ezagutza base hau erabili daiteke urruneko gainbegiraketarako. Esperimentuan parte hartzen duten entitateak pertsona eta erakunde motakoak dira. 2009ko edizioan, lekuei buruzko entitateek ere parte hartzen zuten.

Guk 2011 urtean hartu genuen parte, esperimentuak 3.3 atalean daude azalduta. 2.4 taulan agertzen dira urte horretan entitate hauentzako ikasketa prozesuan erabiltzen ziren eta iragarri behar ziren erlazio mota desberdinak. Pertsonentzako 26 erlazio desberdinei buruzko informazioa bilatu behar genuen, eta erakundeentzako 16.

Irakurleak 2014ko edizioari buruzko informazioa jaso nahi badu, jo dezala atazaren webgunera eta ikus dezala atazaren definizioa[16], parte hartzen

[16] http://surdeanu.info/kbp2014/KBP2014_TaskDefinition_

Pertsonak	**Erakundeak**
per:age	org:alternate_names
per:alternate_names	org:city_of_headquarters
per:cause_of_death	org:country_of_headquarters
per:charges	org:dissolved
per:children	org:founded
per:cities_of_residence	org:founded_by
per:city_of_birth	org:member_of
per:city_of_death	org:members
per:countries_of_residence	org:number_of_employees/members
per:country_of_birth	org:parents
per:country_of_death	org:political/religious_affiliation
per:date_of_birth	org:shareholders
per:date_of_death	org:stateorprovince_of_headquarters
per:employee_of	org:subsidiaries
per:member_of	org:top_members/employees
per:origin	org:website
per:other_family	
per:parents	
per:religion	
per:schools_attended	
per:siblings	
per:spouse	
per:stateorprovince_of_birth	
per:stateorprovince_of_death	
per:stateorprovinces_of_residence	
per:title	

2.4 taula – TAC-KBP Slot Filling atazako erlazioen zerrenda, 2011n.

duten erlazioen deskribapena[17] eta ebaluazio gida[18]. 2011 eta 2014ko atazen arteko aldaketak oso txikiak dira. 2011ko dokumentazioa ez dago eskuragarri NISTeko webgune ofizialean. Ezagutza basea XML bidez kodetuta dago. C eranskinak kodeketa hori nolakoa den azaltzen du adibide baten bidez.

Ataza, ingeleserako gain, gaztelerarako eta txinerako ere badago. Tesi honetako esperimentuak ingelesezko dokumentuak eta ezagutza basea erabiliz egin dira.

EnglishSlotFilling_1.1.pdf

[17] http://surdeanu.info/kbp2014/TAC_KBP_2014_Slot_Descriptions_V1.1.pdf

[18] http://surdeanu.info/kbp2014/TAC_KBP_2013_Assessment_Guidelines_V1.3.pdf

Freebasen azpimultzoa (*Riedel* et al. *2010*)

Datu multzo hau publikoki eskuragarri dago ikerketarako, eta sortzaileak Riedel *et al.* dira. Sortzeko erabili den ezagutza basea Freebase da, 2009ko abenduko kopia bat erabiliz. Aipamenak New York Times corpusetik erauzi zituzten, 2005 eta 2006 urteen arteko artikuluak erabiliz.

Corpusak bi partizio ditu: entrenamendu partizioa, 2005 eta 2006 urteetako zati baten erlazio aipamenekin, 4700 aipamen guztira, eta testeatzeko partizioa, 2007ko zati baten 1950 erlazio aipamenekin. Bi partizioetan, aipamen negatiboak automatikoki sortu ziren esaldi batean bi entitate bikote agertu eta Freebasen erlaziorik ez zegoen bakoitzean.

Entitate aipamenak corpuseko testuetan topatzeko, Stanfordeko entitate izen ezagutarazlea (Finkel *et al.* 2005) erabili zuten. Ondoren, erlazio aipamenak sortu zituzten esaldi berean agertzen ziren entitate aipamenen artean. Esaldi hauetan agertzen diren entitate aipamenen artean erlazio bat espezifikatuta badago Freebaseen, orduan esaldi horri erlazio horren marka jartzen zaio, urruneko gainbegiraketa jarraituz. Aipamen berean dauden bi entitateren artean ez badago erlaziorik, orduan *NA* marka jasotzen dute esaldiek.

Datu multzo honetan 7 entitate motek hartzen dute parte: pertsonak, erakundeak, lekuak, denbora, filmak, kirolak eta telebista saioak. Guztira 52 erlazio desberdin aurki ditzakegu, horietako batzuk 2.5 taulan ditugu. E eranskinean datu multzo hau osatzen duten erlazio guztien zerrenda aurki dezake irakurleak.

Datu multzo honetan, ezaugarriak dagoeneko erauzita daude, hala ere badago ere aukera aipamen bakoitzarekin norberak gustukoen dituen ezaugarriak sortzeko. Ezaugarri hauek aurrerago azalduko dugun Mintz *et al.*-en lanekoaren antzekoak dira: ezaugarri lexikoak, kategoria gramatikala, entitate izenak eta ezaugarri sintaktikoak erabiltzen dituzte. Kategoria gramatikalak lortzeko, OpenNLP[19] tresnak dakarren etiketatzailea erabili zuten, eta dependentziak zuhaitzak lortzeko berriz, Maltparser[20] tresna (Nivre *et al.* 2004). 4. kapituluko esperimentuetan datu multzo hau erabiltzen dugu.

[19] http://opennlp.sourceforge.net
[20] http://www.maltparser.org/

/broadcast/content/location
/broadcast/producer/location
/business/company/locations
/business/person/company
/film/film_festival/location
/location/country/capital
/location/jp_prefecture/capital
/location/us_state/capital
/people/person/profession
/people/person/place_lived
/sports/sports_team/location
/time/event/locations
(...)

2.5 taula – Freebase ezagutza basea erabilita, datu multzoan dauden erlazio izen batzuen zerrenda.

2.2 Erlazio erauzketako ikasketa motak

Sarreran aipatu bezala, erlazio erauzketarako hiru ikasketa mota desberdin daude. Bakoitzak bere alde onak eta txarrak ditu. Atal honetan ikasketa hauek azaldu eta mota bakoitzeko hainbat ikertzailek egindako lanak laburtuko ditugu.

2.2.1 Ikasketa gainbegiratua

Ikasketa gainbegiratuan, corpus bateko esaldietan entitate guztiak eskuz bilatu, hauek eskuz etiketatu, eta entitate guztien arteko erlazioak eskuz zehazten dira. Aurreko atalean aipatu dugun ACE corpusa da corpus etiketatu horietako bat.

Ikasketa gainbegiratua erabiltzen duten sistemetan, dokumentu batean dauden erlazio aipamen desberdinak aurkitu behar ditugu, sailkatzaileak erabaki beharko du ea dokumentuko aipamen bakoitzean entitateen artean erlaziorik ote dagoen edo ez. Egotekotan, erlazio hori zein den esango digu sailkatzaileak.

Ikasketa gainbegiratuan, ikasketarako, eredu bayesiar sinpleak, entropia maximoa eta SBM (ikus 2.5.1 atala) bezalako sailkatzaileak egokiak dira.

Normalean erlazio mota bakoitzarentzat eredu bat ikasten da, "denak besteen aurka"[21] edo "klase anitzak".

Ikasketa gainbegiratuaren paradigma erabiltzen duten lan nagusien artean, Zhou *et al.*-ek (2005) sintaxi eta semantikaren erabilgarritasuna frogatzen dute. Sintaxiari dagokionez, sintagma banaketa erabiliz erlazioak erauztea eraginkorra dela frogatzen dute. Semantikari dagokionez, WordNet bezalako baliabide semantikoak erabiltzen dituzte ezaugarrietan hobekuntza gehiago lortzeko. Egileek erabilitako ezaugarri multzoak gaur egun erlazio erauzketarako erreferentzia bezala erabiltzen dira, guk ere erabili ditugu 3.2 atalean egin ditugun esperimentuetan. 2007ko lanean berriz, zuhaitz kernelak[22] erabiliz egiten dituzte esperimentuak, kernel linealak erabili ordez, beraien aurreko lanaren eraginkortasuna hobetuz.

Giuliano *et al.*-ek (2006) biomedikuntzarako sistema gainbegiratu bat garatzen dute, bi datu multzo desberdinetatik geneen eta proteinen arteko erlazioak lortuz. Horretarako, hiru kernel funtzio[23] desberdinen arteko arteko konbinaketak egiten dituzte: testuinguru globaleko kernelak, testuinguru lokaleko kernelak, eta azaleko hizkuntza kernelak.

Surdeanu eta Ciaramitak (2007) pertzeptroiak erabiltzen dituzte, emaitza onak lortuz baita ere, ezaugarri sintaktikoetan bakarrik oinarrituz. Gure esperimentu batzuetan lan honen ezaugarriak erabili ditugu, baita emaitzak ere oinarri lerro bezala.

Aipatu berri ditugun lan hauetatik hiruk ACE corpusa erabiltzen dute esperimentuak egiteko. Ikasketa hau aplikatzen duten sistemek zehaztasun handiko emaitzak ematen dituzte. Adibidez, Zhou *et al.*en 2005eko lanean, ACE 2005 corpusa erabiliz, %55ko F1 neurria eskuratu zuten; Surdeanu eta Ciaramitaren lanean berriz, ACE 2005 corpusarekin %54.5eko F1 neurria.

Gaur egun erlazio erauzketari dagokionez, sistema onenak dira. Hala ere, eskuz etiketatutako entrenamendurako datu multzoak sortzea oso garestia da, denbora asko behar da kalitatezko datu multzo etiketatu bat sortzeko. Horrez gain, erabiltzen diren corpus motarekiko dependentzia handia dute erabiltzen diren erlazioek. Gainera, erlazio berriak beharko balira, edo domeinuan aldaketak egingo balira, dena berriz etiketatzeko beharra legoke.

[21] Ingelesez *One vs All*

[22] Tree-Kernel: http://en.wikipedia.org/wiki/Tree_kernel

[23] Kernel methods: http://en.wikipedia.org/wiki/Kernel_method

Ikasketa ez-gainbegiratua eta erdigainbegiratua dira ikasketa gainbegiratuaren alternatibak, erlazio erauzketan etiketatzeko beharrezkoa den eskuzko kostu guztia gutxitzeko asmoz sortuak.

Ikasketa gainbegiratuko bi erlazio erauzketa sistema

Aipatu berri dugun bezala, Zhou *et al.* (2005) eta Surdeanu eta Ciaramitaren lanen ezaugarriak erabili ditugu gure esperimentu batzuetan. Sistema hauetako ezaugarriak azaltzeaz gain, hauen arkitekturak azalduko ditugu hemen.

Zhou *et al.*-ek, ikasketarako erabilitako sailkatzailea sostengu bektoreen makina (ikus 2.5.1 atala) da, ikasketa algoritmo honen berezitasunak aurrerago azalduko ditugu kapitulu honetan. Erabilitako corpusa 2005eko ACE da.

Lan honetan, ezaugarri mota asko erabiltzen dituzte: lexikalak, sintaktikoak, eta semantikoak. Ezaugarri guztiak konbinatuz, %55eko F1 neurria lortzen dute. Halaber, hitz mailako ezaugarriak bakarrik erabiliz, %35eko F1 neurria lortzen dute. Gure esperimentuetan hitz mailarako proposatzen dituzten ezaugarriak erabili ditugu. Ezaugarri hauek ondorengo etiketa aurredefinituetaz osatuta daude, bakoitza dagokion balioarekin (E1 lehenengo entitatea, E2 bigarrena):

- WM1: E1 aipamenaren hitz multzoa.
- HM1: E1ko hitzak banatuta.
- WM2: E2 aipamenaren hitz multzoa.
- HM2: E2ko hitzak banatuta.
- HM12: HM1 eta HM2ren arteko konbinaketa.
- WBNULL: bi entitateen artean hitzik ez dagoenean.
- WBFL: bi entitateen arteko hitza, hitz bakarra dagoenean.
- WBF: bi entitateen arteko lehen hitza, hitz bat baino gehiago dagoenean.
- WBL: bi entitateen arteko azken hitza, hitz bat baino gehiago dagoenean.
- WBO: bi entitateen arteko beste hitzak, lehenengoa eta azkenekoa ezik.
- BM1F: E1en aurreko hitza.
- BM1L: E1en aurreko bigarren hitza.
- AM2F: E2ren ondorengo hitza.
- AM2L: E2ren ondorengo bigarren hitza.

Surdeanu eta Ciaramitak perzeptroi algoritmoaren[24] aldaera bat erabiltzen dute datu desorekatuetako ikasketa eraginkorragoa egiteko; aldaera honetan, bi marjina handi erabiltzen dituzte, marjina bat aipamen positibo eta negatiboen arteko desoreka modelatzeko, eta bestea ereduaren orokortzearen eraginkortasunaren arabera doitzeko.

Erabilitako ezaugarriak Zhou *et al.* (2005) eta Giuliano *et al.*-en lanetan daude oinarrituta, baina ezaugarri lexikoak, morfologikoak eta partzialki sintaktikoak bakarrik erabiltzen dituzte, informazio semantikoa eta analisi sintaktiko osorik egin gabe. Ezaugarri hauek gure esperimentu batzuetan ere erabili ditugu. Irakurketa errazteko, jarraian zerrendatu ditugun ezaugarriei aurrekoen antzeko izenak jarri diegu, eta Zhou *et al.*-en lanean agertzen ez diren ezaugarriak letra lodiz jarri:

- WM1: E1 aipameneko hitz multzoa.
- HM1: E1ko hitzak banatuta.
- **PM1:** E1en kategoria gramatikala.
- **EM1:** E1en entitate izen mota.
- WM2: E2 aipameneko hitz multzoa.
- HM2: E2ko hitzak banatuta.
- **PM2:** E2ren kategoria gramatikala.
- **EM2:** E2ren entitate izen mota.
- WBO: bi entitateen arteko hitzak.
- **DIST:** bi entitateen arteko hitz kopurua.
- **DIR:** erlazioaren norabidea, entitate nagusia aipamenean lehena edo bigarrena den zehaztuz.
- BM1F: E1en aurreko hitza.
- BM1P: E1en aurreko hitzaren kategoria gramatikala.
- AM2F: E2ren ondorengo hitza.
- AM2P: E2ren ondorengo hitzaren kategoria gramatikala .

Gure esperimentuetan bi ezaugarri motak nahasten ditugunez, etiketa aurredefinitu guztiak ez ditugu erabiliko, errepikatzen direlako. Surdeanu eta Ciaramitaren lanekoak informatiboagoak dira, hitz bakoitzaren entitate izena eta kategoria gramatikala kontuan hartzen dituztelako, baita bi entitateen arteko distantzia eta norabidea ere. 3. kapituluan ikasketa gainbegiratuen arkitekturan gehiago sakonduko dugu.

[24]Perceptron: http://en.wikipedia.org/wiki/Perceptron

2.2.2 Ikasketa ez-gainbegiratua

Ikasketa gainbegiratuaren alternatiba bezala, ikasketa ez-gainbegiratua dugu. Algoritmo hauen helburua corpus etiketaturik erabili gabe datuetako egitura edo patroiak antzematea da. Gure kasuan, erlazio mota adierazten duten egiturak dira antzeman beharrekoak.

Ikasketa mota honen abantaila ondorengoa da: etiketatu gabeko corpusen beharrik ez edukitzean, datu multzo handiekin lan egitea errazagoa da; erlazio erauzketaren kasuan, aurrez definitutako erlazioez gain erlazio berriak antzeman ditzakegu. Baina ikasketa honek ere bere desabantailak ditu: antzemandako erlazioei izen kanoniko bat ematea (adibidez, ezagutza basera lotzea) oso zaila da; ondorioz, sistema hauen ebaluazioa ez da lan erraza izaten.

Ikasketa honetan, testu sorta handietatik bi entitateren arteko hitzak erauzten dira, hitz sorta hauek multzotan (*cluster*) banatu, eta hitz sortak sinplifikatzen ditu erlazio hitzak sortzeko. Beste era batera esanda, oraingoan ez daukagu aurretik definitutako erlazio zerrendarik.

Ikasketa honen lanak bi multzotan antolatu ditzakegu: (1) *clustering*ean oinarritzen direnak, eta (2) OpenIE motakoak. Hasteko, *clustering* erabiltzen dutenak zerrendatuko ditugu.

Lin eta Pantelren lanean (2001), *X is author of Y, X wrote Y* eta beste hainbat inferentzia arau erauzten dituzte testuetatik, eta gero dependentzia bideen bidez entrenatu. Beraien hipotesiaren arabera, bi bide oso antzekoak diren testuinguruetan gertatzen badira, bideen esanahia antzekoa da. Sortu duten tresna honi *DIRT* izena jarri diote. Bada beste lan bat, Yao *et al.*-ena (2012), non *DIRT* hobetzen duten, *topic model*ak[25] erabiliz, dependentzia bideek sortu ditzaketen anbiguotasunak konponduz. *X addresses Y* kasu anbiguo baten adibidea da, ondorengo bi esaldiek erakusten duten bezala, lehena gutun baten helbideratzeari buruzkoa da, eta bigarrena hitzaldi formal bati buruzkoa:

- I **addressed** my letter to him personally.
- Will Congress finally **address** the immigration issue?

Lopez de Lacalle eta Lapatak (2013) ikasketa prozesua *topic model*ak eta lehen ordenako logika[26] konbinatuz egiten dute ikasketa.

[25]Topic-models: http://en.wikipedia.org/wiki/Topic_model
[26]First-Order Logic: http://en.wikipedia.org/wiki/First-order_logic

TextRunner, *WOE* eta *ReVerb* sistemak **Informazio erauzketa irekia (OpenIE)** bezala ezagutzen dira, webgune askotatik edozein domeinuri buruzko erlazio kantitate handiak eskuratzeko gai direlako.

Banko *et al.*-ek *TextRunner* tresna aurkezten digute (2007). Tupla bakoitzeko lortu diren aipamenak hartu eta maiztasunaren arabera probabilitate bat jasotzen du tupla bakoitzak. Eredu bayesiar sinpleak[27] erabiltzen dituzte sistema entrenatzeko, kategoria gramatikalak eta izen sintagma kateetan[28] oinarritutako ezaugarriak erabiliz. Entrenamendurako erabilitako aipamenak heuristiko batzuen bidez sortu zituzten Penn Treebanketik. 2008an, Banko eta Etzionik baldintzazko ausazko eremuak (ikus 2.5.2 atala) erabiltzen dituzte hobekuntzak lortzeko. Urtebete geroago, Zhu *et al.*-ek are gehiago hobetzen dute sistema Markov sare logikoen[29] bidez.

Wu eta Weldek *WOE* tresna aurkezten dute 2010ean. Tresna honek Wikipediako infotaulak erabiltzen ditu informazio iturri gisa, heuristiko baten bidez infotaulen informazioa eta corpusean dagokion testuak lotuz. Sistema honentzat bi aldaera aurkezten dituzte: lehenengoak kategoria gramatikalei ematen die garrantzi handia ezaugarrietan, *TextRunner* sistemaren antzeko emaitzak lortuz; bigarrenak berriz, dependentzia egituren patroiei ematen die garrantzia, emaitzak asko hobetuz, baina sistema askoz motelagoa eginez.

Fader *et al.*-ek (2011) ikasketa ez-gainbegiratua maila lexikoan aztertzen dute, erauzitako informazio inkoherenteak konpontzeko. Horretarako, muga lexiko eta sintaktiko sinple batzuk ezartzen dizkiete aditzei. Sortu duten sistema *ReVerb* deitzen da.

Xu *et al.*-ek (2013) aurreko paragrafoan aipatutako sistemen mugak aztertzen dituzte, eta hauek hobetzeko proposamenak egin, automatikoki konfirmatuz ea ezarritako erlazioak entitate-tuplentzat zentzurik ote duen ala ez.

2.2.3 Erdigainbegiratua

Erlazio erauzketarako erabiltzen den hirugarren ikasketa mota da erdigainbegiratua. Ikasketa honetan, adibide gutxi batzuk erabiltzen dira hazi bezala ikasketa prozesua abiarazteko.

Hazi hauen bidez, patroi sortak erauzten dira corpus erraldoietatik; jarraian, patroi hauen agerpen gehiago corpusean bilatu eta hazi berriak erauz-

[27] Naive Bayes: http://en.wikipedia.org/wiki/Naive_Bayes_classifier
[28] Ingelesez *chunk*.
[29] Markov logic network: http://en.wikipedia.org/wiki/Markov_logic_network

ten dira; hazi berri hauekin, patroi gehiago, guzti hau iteratiboki. Lortzen diren patroi hauek orokorrean doitasun baxua izaten dute, eta desbiderazio semantikoak dituzte.

Brinek (1999) liburu desberdinen eta hauen idazleekin egiten du lan, "(egile, izenburu)" motako tupla gutxi batzuk hazi bezala erabiliz. Interneteko testuetan hauei buruzko patroiak bilatu eta horrela liburu gehiagori buruzko informazioa lortzen du.

Riloff eta Jonesek (1999) hainbat domeinutarako egiten dituzte esperimentuak, ez bakarrik liburuentzat. Urtebete geroago, Agichtein eta Gravanok ere bide berdina jarraitzen dute, baina sortu duten *Snowball* sistemak automatikoki aztertzen ditu erauzitako patroiak, iterazio bakoitzean esanguratsuenekin geldituz.

Ravichandran eta Hovyk (2002) patroien erauzketa galdera erantzun sistema bat sortzeko aprobetxatzen dute, erlazio erauzketaren bidez sistema hauek aberastu daitezkeela erakutsiz. Horretarako, corpus batetik patroi desberdinak erauzten dituzte automatikoki, eta patroi bakoitzaren zehaztasuna neurtu galdera mota bakoitzeko.

Etzioni *et al.*-ek *KnowItAll* tresna aurkezten digute (2005). Tresna honek tupla bakoitzaren elkarrekiko informazio puntualaren[30] pixua eskuratzen du, internetetik eskuratutako corpus baten arabera kalkulatuta, ahalik eta doitasun onena lortzeko. 2008an Rozenfeld eta Feldmanek *SRES* tresna aurkezten dute, *KnowItAll* sistemaren berrinplementazio bat, domeinuekiko independentea.

Urruneko gainbegiraketa sistema erdigainbegiratutzat kontsidera daitekeen arren, hurrengo atalean azalduko dugu sakonago.

2.3 Erlazio erauzketa eta urruneko gainbegiraketa

Atal hau azpiatal desberdinetan banatu dugu. Lehenengo urruneko gainbegiraketaren hastapenak komentatuko ditugu eta ondoren paradigma honen hedapen batzuk komentatu. Lan aitzindaria Mintz *et al.*-ena da, hori dela eta sistema horren arkitektura eta ezaugarriak azalduko ditugu. Ondoren gure esperimentuetan erabilitako Mintzen aldaera bat aurkeztuko dugu. Egindako

[30] Pointwise mutual information: http://en.wikipedia.org/wiki/Pointwise_mutual_information

lan asko TAC-KBPko *Slot Filling* atazarako egin direnez, lan hauei azpiatal bat eskainiko diegu. Amaitzeko, gure esperimentuak zarataren garbiketan ere oinarritzen direnez, tokitxo bat egin diegu hau egiten duten lanei ere.

Urruneko gainbegiraketaren proposamen originala Craven eta Kumlien-en lanean (1999) aurki dezakegu. Lan hau medikuntzako domeinurako pentsatuta dago, proteina, zelula, gaixotasun eta beste hainbaten arteko erlazio bitarrak erauziz. Mintzen lanean paradigma honen hobekuntza dator, hainbat entitate desberdinetarako erabilgarria izanik, hala nola *pertsonak, lekuak, erakundeak,...*, eta domeinu askotara hedatuz. Geroztik ospea lortu du urruneko gainbegiraketak.

Yao *et al.*-ek paradigmaren hobekuntzarako lehen proposamenak aurkezten dituzte (2010), ikasketan *SampleRank* bezalako faktore grafoak (Wick *et al.* 2009) erabiliz eta inferentzia Gibbs laginketaren[31] bidez eginez.

Surdeanu *et al.*-ek (2012) Mintzen laneko sistemako algoritmoari hobekuntza asko egiten dizkiote. Atal honetan aurrerago azalduko ditugu hobekuntza horiek. Angeli *et al.*-en lanak (2014), sistema honi hobekuntzak egiten dizkiote: aurreratuagoa den bilaketa motore berri bat erabiliz, eta eskuz etiketatutako aipamen asko gehituz.

Riedel *et al.*-ek 2013an matrizeen faktorizazio[32] bidezko patroiak erabiltzen dituzte ezkutuko ezaugarri bektoreen ikasketarako. Erabiltzen dituzten patroiek ikasketa prozesurako denbora gutxiago behar dute, aurretik garatutako patroiekin konparatzen baditugu.

Weston *et al.*-ek (2013), entitate, erlazio eta *word embedding*ak[33] erabiltzen dituzte, testua eta erlazioak elkarren artean lotuz.

Li *et al.*-en lanean (2014), esperimentu berezi bat aurki dezakegu. Bertan txiolari desberdinei buruzko informazioa erauzten dute, urruneko gainbegiraketa erabiliz, txiokatu dituzten txioetatik. Helburua, bikote, ikasketa eta lanari buruzko informazioa eskuratzea, txiolari bakoitzeko. Guk ere Twitterrekin egin ditugu esperimentu batzuk, baina aurrerago azalduko dugun bezala, lurrikarei buruzko informazio kantitate handiak eskuratzeko.

Koch *et al.*-ek (2014), urruneko gainbegiraketaren emaitzak hobetzeko, entitate izenen desanbiguazioa eta korreferentziaren ebazpena argumentuak identifikatzeko konbinatzen dituzte.

[31] Gibbs sampling: http://en.wikipedia.org/wiki/Gibbs_sampling
[32] Matrix Factorization: http://en.wikipedia.org/wiki/Matrix_decomposition
[33] Word embedding: http://en.wikipedia.org/wiki/Word_embedding

2.3.1 Sistema aitzindariaren arkitektura eta ezaugarriak (*Mintz* et al. *2009*)

Sarreran aipatu dugun bezala, urruneko gainbegiraketan, bi entitatek erlazio batean parte hartzen badute, bi entitateak dituen edozein esaldik erlazio hori zehaztuko du. Sistema honen arkitekturak ondorengo urratsak jarraitzen ditu:

1. Ikasketa corpusaren sorrera. Erlazio bakoitzeko adibideak lortu:
   (a) Entitate nagusiak dituzten esaldi guztiak erauzten dira corpusetik.
   (b) Aipameneko entitate guztiak antzematen dira.
   (c) Entitate nagusia eta aurkitutako beste entitate baten artean, erlazio bat zehaztuta baldin badago Freebaseen, erlazio hori jartzen zaio aipamenari.
2. Ereduak sortu:
   (a) Ezaugarriak erauzten dira.
   (b) Sistema entrenatzen dute.
3. Helburu entitateen aberasketa:
   (a) Testerako, helburu entitatea duten esaldi guztiak erauzten dira corpusetik.
   (b) Entitate bikoteen aipamen desberdinak bildu eta hauen ezaugarriak erauzten dira, sailkatzaileak aztertu ditzan.
   (c) Sailkatzaileak entitate bikote bakoitzaren erlazioa iragarriko du.

Entrenamendua klase anitzeko erregresio logistikoko[34] sailkatzaile batekin egiten dute. Hala ere, beraien sistemak tupla bakoitzeko erlazio bakarra baino ez du onartzen.

Erabiltzen dituzten ezaugarriek entitateen arteko erlazioa deskribatzen saiatzen dira, ezaugarri lexiko eta sintaktiko desberdinak erabiliz. Gure esperimentuetan beraiek proposatutako ezaugarri lexikoak erabili ditugu. Ezaugarri hauek bi entitateen inguruko hitzak deskribatzen dituzte:

[34]Multiclass Logistic Regression: http://en.wikipedia.org/wiki/Multinomial_logistic_regression

- Bi entitateen arteko hitzak.
- Hitz hauen kategoria gramatikala.
- Etiketa bat adierazteko bi entitateetatik zein den testuan agertzen den lehena.
- Lehenengo entitatearen aurretik datozen k hitz eta hauen kategoria gramatikala.
- Bigarren entitatearen ondoren datozen k hitz eta hauen kategoria gramatikala.

Ezaugarri konjuntiboak sortzen dituzte $k \epsilon \{0, 1, 2\}$ bakoitzarentzat. Aipamenean parte hartzen duten bi entitateen izen motak ere kontuan hartzen dute ezaugarriek. Guk ere kontuan hartu ditugu gure esperimentuetan, 3. kapituluan. Ezaugarri sintaktikoak, entitateen arteko dependentzia zuhaitzetan oinarritzen dira.

Aldaerak eta berrinplementazioak

Mintzen laneko algoritmoak ondorengo arazoak ditu:

- **Entitate tupla bakoitzaren aipamen guztiak datu bakar batean kolapsatzen dira.** Beste era batera esanda, algoritmo originalak tupla berdineko aipamen guztien ezaugarriak nahasten ditu, sailkapen datu bakarra sortuz. Honek zaratarekin lan egiteko malgutasuna kentzen dio sistemari.
- **Entitate tupla bakoitzeko, erlazio-kategoria bakarra baino ez da onartzen.** Erlazio bat baino gehiago dutenentzat, pixu altuena duen erlazioarekin gelditzen dira.

Surdeanu *et al.*-en laneko sistema[35] (2012) algoritmo originalaren berrinplementazio bat da. 4. kapituluan egiten ditugun esperimentuak sistema hau erabiliz egin ditugu, baita aldaera batekin ere. Algoritmoaren berrinplementazioarentzat **Mintz*** izena erabili dugu. Berrinplementazio honek Mintzen algoritmo originalaren ahulguneak konpontzen ditu:

[35]Eskuragarri dohainik: http://nlp.stanford.edu/software/mimlre.shtml

- **Aipamen bakoitza independenteki modelatzen da,** beraz esaldi bakoitzeko sailkapen datu (edo *datum*) bat sortzen da.
- **Tuplak etiketa anitzak izan daitezke**, hau da, kategoria bat baino gehiago eduki dezakete.

Hala ere, **Mintz*** algoritmoak arazo bat dauka: datuen gehiegizko egokitzea (ingelesez *overfitting*). Arazo hau konpontzeko, aldaera bat ere dute garatuta: **Mintz++** sistema. **Mintz++** algoritmoaren helburua sistemaren gehiegizko egokitzea saihestea da. Horretarako, tupla bakoitzaren aipamen guztiak hartu, eta k multzotan banatzen dira, tupla bakoitzeko k sailkatzaile desberdin entrenatuz, kasu honetan, $k = 5$. Bukaerako emaitza bost ereduek emandako pixuen batezbestekoa da. Aldaera honek entrenamendurako denbora gehiago behar du, baina emaitzak hobetzeko gai da.

Aurretik aipatu dugu Mintzen sistema originalak klase anitzeko erregresio logistikoko sailkatzaile batekin egiten dutela entrenamendua. Berrinplementazio hauek berriz, entrenamendua itxaropenaren maximizatzearen[36] bidez egiten dute. Inferentzia metodoak tupla bakoitzeko etiketa bat baino gehiago itzuli dezake.

Surdeanuren sistemak erabiltzen dituen ezaugarriak Riedel *et al.*en datu multzoan (ikus 2.1.2 atala) erabilitako berdinak dira. Ondorengo propietateak dituzte ezaugarri hauek:

- Bi entitateen inguruko hitzak eta hitz bakoitzaren kategoria gramatikala.
- Bi entitateen entitate izen motak.
- Lehenengo entitatearen aurretik datozen bi tokenen hitz formak.
- Bigarren entitatearen ondoren datozen bi tokenen hitz formak.
- Bi entitateetatik, zein datorren lehenengo zehaztuta.

4. kapituluan urruneko gainbegiraketa sistema honekin egiten dugu lan.

Etiketa anitzen estrategia Hoffmann *et al.*-en lanean (2011) ere erabili izan da. Estrategia hau oso beharrezkoa da entitate pare batek erlazio bat baino gehiago adierazi dezakeelako, adibidez <*Balzac* - *France*> tuplak bi erlazio ditu: *place_of_death* eta *born_in*, hauek ezin dira Mintzen algoritmoarekin tratatu.

[36] Expectation Maximization: http://en.wikipedia.org/wiki/Expectation-maximization_algorithm

2.3.2 *Slot Filling* atazan aurkeztutako sistema onenak

Urruneko gainbegiraketa NIST erakundeak antolatutako *Slot Filling* atazako sistema askotan erabili da. Parte hartzaile gehienek beraien sistemetarako urruneko gainbegiraketa erabiltzen dute erlazioak erauzteko. Azpiatal honetan azken urteetan irabazi duten sistemak azaltzen ditugu.

New Yorkeko unibertsitateko Sun *et al.*-ek (2011) *Slot Filling* atazarako[37] sistema sortzeko, urruneko gainbegiraketa eta erregeletan oinarritutako galdera-erantzun sistemen teknikak konbinatzen dituzte. Urtebete geroago, Min *et al.*-en lanean (2012a), egile berdinek sistema hobetzen dute, estatistiken bidez erroreak filtratuz eta aipamen okerrei erlazio zuzenak jarriz.

Stanford Unibertsitateko Surdeanu *et al.*-ek 2010ean Mintzen laneko sistemaren hedapen bat egiten dute, antolatzaileek emandako corpusa eta ezagutza basea erabiliz, Wikipediako infotaulak eta atazan erabilitako erlazio izenak bateratuz, erlazio erauzketa informazio berreskurapen sistema batekin parekatuz. Urtebete geroago, Surdeanu *et al.*-en lanean oraindik hobekuntza gehiago aurki ditzakegu, erantzun bakoitzarentzat erlazio bat baino gehiago baimentzen duen inferentzia metodoak erabiliz, eta entitateen arteko korreferentzia aplikatuz. Bi lan hauek dira azaldu berri dugun Surdeanuren 2012ko laneko sistemaren aitzindariak. 2013 urtean, Angeli *et al.*-ek emaitzak are gehiago hobetzen dituzte, informazio berreskurapenerako bilaketa motore berri bat erabiliz, eta Surdeanuren laneko *MIML-RE* sistemarekin batera konbinatuz. Angeli *et al.*, atal honetan aurretik azaldu dugun 2014ko lanean, sistema hau are gehiago hobetzen dute.

Roth *et al.*-en lanean (2013), beren sistemaren bertsio desberdinak probatzen dituzte: patroi sintaktikoak erabiltzen dituena, Wikipedian oinarritutako egiaztatzaile bat erabiltzen duena, eta ondoren aipatzen ditugun beste hainbat. Sistema onena urruneko gainbegiraketako patroi batzuk, eskuzko idazketako patroi batzuk, analisi sintaktikorik ez, izen alternatiboen hedapenerako modulu bat, eta sistemak azkar funtzionatzeko SBM (ikus 2.5.1 atala) sailkatzaile batek osatzen dute. Sistema hau 2013ko edizioko onena izan zen.

Urruneko gainbegiraketak hainbat abantaila dituen arren, sistema orok pairatzen duen arazoa datu multzoetan aurki dezakegun aipamen zaratatsuen kantitate handia da.

[37] 2014ko atazari buruz gehiago jakiteko: http://surdeanu.info/kbp2014/KBP2014_TaskDefinition_EnglishSlotFilling_1.1.pdf

2.3.3 Zarataren garbiketa

Tesi honetan zehar ikusiko dugu nola urruneko gainbegiraketak zarata asko sortzen duen. Sortutako zarata motak, eta hauei aurre egiteko heuristiko batzuk 4. kapituluan azalduko ditugu. Badira hurbilketa asko sailkatzaileak zaratarekin batera entrenatzen laguntzeko.

Riedel *et al.*en lanean (2010), argi uzten dute bi entitate batera agertzen diren aipamen guztiek ez dutela zertan bien arteko erlazio bat adierazi behar, baina bai dauden aipamen guztietatik gutxienez batek erlazio hori adierazten duela. Eredu probabilistikoen bitartez, erlazio konkretu hori zein aipamenek duten iragartzen saiatzen dira, ondoren aipamen egokiena zein den erabaki eta sailkatzailea entrenatuz.

Urtebete geroago, Hoffmann *et al.*-ek, erlazio bat baino gehiago duten entitate tuplekin egiten dute lan, ordurako bi entitateren artean erlazio bakarra zutela uste izaten zuten sistemek. Hau konpontzeko, perpaus maila eta corpus mailako erauzketa ereduak konbinatzen dituzte.

Surdeanu *et al.* (2012) ere etiketa anitzeko ikasketan oinarritzen dira baita. Aldagai latenteak dituzten eredu grafikoak erabiltzen dituzte ikasketarako. Lan honetarako erabili zuten sistemarekin esperimentu batzuk egin ditugu, beraien emaitzak gure emaitzekin konparatu ahal izateko.

Urte berean, Takamatsu *et al.*-ek, eredu generatiboak erabiltzen dituzte aztertzeko ea erlazio mota bat zuzena den edo ez. Eredu horrek iragartzen badu aipamen batek ez duela erlazioarekin bat egiten, aipamena ezabatzen dute.

Min *et al.*-en (2013) lanean, ezagutza base osatu gabeak eta negatibo faltsuen arazoarekin egiten dute lan. Pixkat lehenago aipatu dugun Surdeanuren laneko sistemari egiten dizkiote moldaketak beraien esperimentuetan, itxaropen maximizazio algoritmoa erabiliz ikasketarako.

Azkenik, 2014an, Fan *et al.*-ek zaratarekiko tolerantzia duten bi optimizazio eredu aurkeztu dituzte, ezaugarrien bidez matrizeak sortuz, eta matrizeen faktorizazioaren bidez, zaratatsuak diren ezaugarriei tratamendu bat emanez. Emaitzak izugarri hobetzen dira, baina hala ere, eredu hauek ez dira gai testeko elementu berriak prozesatzeko, ereduak entrenamenduko datu multzoarekiko dependentzia handia duelako. Testeko elementu berriak prozesatu ahal izateko, entrenamenduko matrizeak birsortu behar dituzte. Lan honetan ere Riedel eta bere kideen datu multzoa erabili dute.

2.4 Gertaera erauzketako ikasketa motak

Gertaera erauzketan, erlazio erauzketan bezala, hiru ikasketa mota berdinak daude, bakoitza bere abantaila eta desabantailekin. Atal honetan ikasketa hauek azaldu eta mota bakoitzeko hainbat ikertzailek egindako lanak laburtuko ditugu.

Irakurlea konturatuko da proposatzen diren metodologiak erlazio erauzketan erabiltzen direnen oso antzekoak direla. Berez, gertaera erauzketa sistema gehienak erlazio erauzketa sistemetan inspiratuta daude. Gertaeren erauzketan aditzek garrantzi handia dute, hauei esker lortu daitezke gertaera motari buruzko pista batzuk. Rol semantikoek ere garrantzi handia dute, hauei esker jakin dezakegu zeintzuk diren gertaeran parte hartzen duten entitate desberdinak.

2.4.1 Ikasketa gainbegiratua

ACE atazak, entitate aipamenak detektatzeaz gain, gertaerak detektatzeko atazak ditu: gertaeren antzematea eta karakterizazioa (*Event Detection and Characterization*). Antzeman beharreko gertaerak bizitza, mugimenduak, transferentziak, negozioak eta beste hainbati buruzkoak dira. Ataza honetan erabiltzen diren dokumentu eta etiketatutako fitxategiak erlazio aipamenak eskuratzeko erabiltzen diren berdinak dira. Hala ere, gertaerak antzemateko ataza honi buruz ez dugu gehiago sakonduko, ez baitugu esperimentuak egiteko oinarri bezala erabili[38].

Grishman *et al.*-ek (2005) argumentuen testuinguruak sekuentzialki detektatzen dituzte, baita rol motak ere. Ikasketarako erabiltzen duten sailkatzailea entropia maximokoa da[39]. 2010ean, Liao eta Grishman-ek gertaeren antzematea dokumentu mailan egiten dute, dokumentu mailan dagoen informazioa oinarritzat hartuta. Eraginkortasun maila altua lortzen dute, erakutsiz dokumentu mailako informazioarekin lan egitea hobea dela esaldi mailan lan egitea baino.

Hong *et al.*-en lanean (2011), entitate moten arteko konsistentzia neurtzen dute gertaerak antzemateko. Gainera, argumentuen entitate motak rolaren

[38]Irakurleak ataza honi buruzko informazio gehiago nahi badu, ikus ondorengo dokumentua: https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

[39]Maximun Entropy, MaxEnt: http://en.wikipedia.org/wiki/Principle_of_maximum_entropy

ikasketan asko laguntzen duela frogatzen dute, rol semantikoak detektatuz, ondoren konplexuagoak direnak antzemateko.

Li *et al.*-ek 2013an pertzeptroi egituratuak erabiltzen dituzte ikasketarako. Pertzeptroien eta ezaugarri globalen bidez, artearen egoerako sistema askoren emaitzak gainditzen dituzte.

2.4.2 Ikasketa ez-gainbegiratua

Ikasketa ez-gainbegiratuetarako, ondorengo lanak topatu ditzakegu:

2005ean, Chun *et al.*-ek *GENIA* sistema aurkezten dute. Sistema honetan, medikuntzari buruzko gertaerei buruzko informazioa erauzten dute, gertaera potentzialak duen patroia ebaluatuz.

Ji eta Grishman-ren lanean (2008), gertaerak erauzteko, dokumentuak multzo desberdinetan biltzen dituzte, inongo daturik etiketatu gabe. Konfiantza handiko gertaerak entrenamendu gisa erabili daitezke beste gertaera batzuk erauzteko.

Lu eta Roth-en 2012ko lanean, gertaeren txantiloi batzuk erabiliz erauzten dituzte argumentu eta rolak. Ikasketarako baldintzazko ausazko eremuak (ikus 2.5.2 atala) erabiltzen dituzte.

2014an, Rusu *et al.*-ek hainbat teknika ez-gainbegiratu konbinatzen dituzte gertaera konplexuak multzo desberdinetan sailkatzeko. Hasteko, aditzak eta argumentuak identifikatzen dituzte, ondoren gertaerak desanbiguatzen dituzte, eta amaitzeko multzoak sortu.

2.4.3 Ikasketa erdigainbegiratua

Ikasketa erdigainbegiratuetarako, ondorengo lanak topatu ditzakegu:

Liao eta Grishman 2010a lanean, patroiak ebaluatu eta garrantzirik gabekoak alde batera uzten dituzte, anbiguotasun lexikoa alde batera utziz.

Huang eta Riloff-ek, 2012an, erlazio erauzketako ikasketa erdigainbegiratuan erabiltzen den hazien kontzeptua aplikatzen dute gertaeren erauzketan. Emaitzak sistema gainbegiratuko artearen egoeraren sistema batzuetara hurbiltzen dira.

Urte berean, Wang *et al.*-ek biomedikuntzarekin erlazionatuta dauden gertaerei buruzko informazioa eskuratzen dute, etiketatu gabeko datuak etiketatuekin batera entrenatuz. Lan hau medikuntzarako ikasketa erdigainbegiratua erabiltzen duen lehena da, egileen arabera. Ikasketa SBM (ikus 2.5.1 atala) sailkatzaile baten bidez egiten dute.

2.4.4 Gertaera erauzketa eta urruneko gainbegiraketa, erlazio zuzena duten lanen deskribapena

Urruneko gainbegiraketa gertaera erauzketan erabiltzeko oso lan gutxi dago eginda. 2011ko Benson *et al.*-en lana da gertaera erauzketa eta urruneko gainbegiraketa batu dituen lehen lana. Gure lanak egiten duen bezala, Twitterreko txioetan oinarritzen dira gertaerak erauzteko. Esperimentu hauetan, artista desberdinek New York hirian egindako emanaldiei buruzko informazioa lortzen saiatzen dira, <*artista, emanaldi_lekua*> tuplak eskuratuz. Informazioa artista eta emanaldiaren lekura mugatzen da, eta gainontzeko informazioa (adibidez, ordua edota ikusle kopurua) alde batera utziz. Gure lanaren ildo beretik doan arren, diferentzia esanguratsuena erauzitako informazio kopuruan datza: gurea 2 argumentutara mugatu beharrean 20 argumentutara hedatzen dugu erauzketa (ikus 6. kapitulua).

Duela gutxiko lan bat, 2014koa, urruneko gainbegiraketa erabiltzen duena gertaera erauzketa egiteko, Reschke *et al.*-ena da. Lan honetan hegazkin istripuei buruzko hainbat informazio erauzten dute: eguna, istripu mota, istripua gertatu den lekua, hegaldiaren zenbakia, hegazkin mota, hildakoak, zaurituak, hegazkineko langile kopurua, eta abar. Gure lana eta hau oso parekoak dira, baina beraiek berri agentzien dokumentuak aztertuz eskuratzen dute informazio hori, guk ordea, Twitter erabiltzen dugu.

Twitter bidezko gertaeren erauzketa

Azken urteetan interes handia egon da Twitter gertaera erauzketan erabiltzeko. Ondorengo lanek ez dute urruneko gainbegiraketa erabiltzen, baina argi uzten dute sare sozialak ere egokiak direla informazioa erauzteko. Aurretik aipatutakoaz gain, Petrović *et al.*-en lana oinarrizkoena litzateke (2010), bertan hainbat txiotatik gertaerak antzematen saiatzen dira. Becker *et al.*-ek, 2011n baita ere, gertaerak dituzten txioak gertaerarik ez dituztenetatik bereizten dituzte multzo edo *clusterren* bidez.

Urte berean, Sakaki *et al.*-ren lanak denbora errealean Japoniako lurrikarei buruzko txioak detektatzen ditu eta hauei buruzko informazioa eskuratu. Lan honek gurearekin antzekotasuna duen arren, desberdintasun hauek dituzte:

- **Guk mundu osoko lurrikarei buruzko txioak erabiltzen ditugu**: biok erabiltzen dugu Twitter lurrikarei buruzko txioak eskuratzeko,

ez herrialde konkretu batekoak bakarrik (beraiek Japoniako lurrikarak bakarrik antzematen saiatzen dira).

- **Gure helburua lurrikarei buruzko ezagutza base bat osatzea da**, txioetako informazioaren bidez, ez txioa herrialde bateko lurrikarei buruzko behatoki bati bidaltzea, ondoren hauek informazioa prozesatu eta horrekin lan egiteko.

- **Guk urruneko gainbegiraketa sare sozialetan egokia dela erakutsi nahi dugu, hainbat argumenturi buruzko informazioa lortuz.**

Popescu *et al.*-en lanak (2011), entitate konkretu batzuei buruzko gertaerak bilatzen ditu, txiolariaren erreakzioarekin batera.

Urtebete geroagoko Ritter *et al.*-en lana aurrerago doa, domeinu irekikoa da eta txioak kategorizatzen ditu. Erauzitako txioek egun bat espezifikatuta dute, eta egun horiekin egutegi automatiko bat[40] sortzen dute. Sistema ikasketa ez-gainbegiratuan azaldutako *OpenIE* tekniketan oinarritzen da.

2.5 Tesian erabilitako ikasketa metodoak

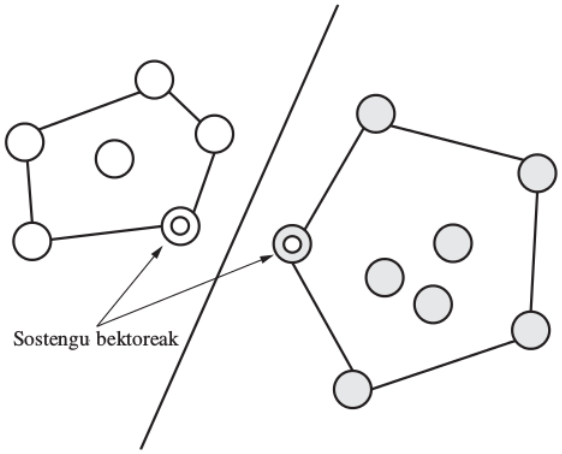
Atal honetan tesiko hainbat esperimentutan erabili ditugun bi ikasketa metodo desberdin azalduko ditugu.

2.5.1 Sostengu bektoreen makinak

Algoritmo hau eredu linealetan oinarritzen da. Oinarrian marjina handieneko hiperplanoa deitzen zaion eredu lineal berezi bat kalkulatzen du. Dena den, eredu linealek dituzten arazoak konpontzeko abilezia du. Izan ere, linealak ez diren datu multzoetarako konponbide bat ematen du. Hainbat atazatan erabiltzen da eredu lineal hau.

Har dezagun, adibidez, bi klaseko datu multzo bat, zeina linealki banagarria den; beste era batera esanda, instantzien espazioan bada hiperplano bat -zuzen bat, alegia-, zeinak instantzia guztiak sailkatzen dituen zuzenaren alde batera eta bestera. Witten eta Frank liburuko (2000) adibide batean oinarritutako 2.1 irudian, hobeto uler dezakegu azaltzen ari garena. Kontuan

[40] StatusCalendar: http://ec2-54-75-211-200.eu-west-1.compute.amazonaws.com:8000/



2.1 irudia – Sostengu bektoreen makinen interpretazio geometrikoa.

izan behar da marraz beteak dauden zirkuluak klase batekoak direla; hutsak, beste klasekoak.

Marjina handieneko hiperplanoa da bi klaseen artean banaketa handiena ematen diguna; hots, hiperplanoko alde bateko eta besteko ikasketa adibideak elkarrengandik urrutien jartzen dituena (Witten eta Frank 2000). Hiperplanotik gertuen dauden ikasketa adibideei sostengu bektore deitzen zaie. Gutxienez, sostengu bektore bana dago klase bakoitzeko. Marjina handieneko hiperplanoa eskuratu ondoren eta bi klaseetako sostengu bektoreak lorturik, gainerako ikasketa adibideak guztiak baztergarriak lirateke.

Horrela bada, sailkatzaile linealeko hiperplanoa bi elementu nagusiez eratua dago: (1) w ezaugarri bakoitzari garrantzi jakin bat esleitzen dien bektorea, eta (2) b alborapen koefizientea, jatorriko puntuaren eta hiperplanoaren arteko distantzia zehazten duena. Bi osagaiak batuz definituko dugu sailkatzaile bitarra:

$$h(\boldsymbol{x}) = \begin{cases} +1 & \text{baldin eta } \langle \boldsymbol{w} \cdot \boldsymbol{x} \rangle + b \geq 0 \\ -1 & \text{bestela} \end{cases} \tag{2.1}$$

Esan dugun moduan, ikasketarako datuak linealki banagarriak ez diren kasuetarako ere orokortu daiteke eskema hau, kernel funtzioen bitartez. Sarrerako ezaugarrien espazioa dimentsio handiagoko espazio bilaka daiteke funtzio ez linealen bat ezartzen badugu.

Praktikan, goi muga antzeko bat markatzen duen parametro bat kalkulatzea da gakoa, eta horretarako, esperimentuak egitea beste aukerarik ez dago.

Gure esperimentuetan, sailkatzaile bat sortzen dugu erlazio mota bakoitzarentzat, non aipamen positiboak sailkatzailearen mota berdinekoak diren, eta beste guztiak negatiboak. Erabilitako paradigma "bat besteen aurka" da, non k erlazioa besteen aurka ebaluatzen den (Rifkin eta Klautau 2004). Gure esperimentuetan SVMperf[41] liburutegia erabili dugu. Liburutegi honen bidez, datu multzo handiekin lan egiteko aukera dugu, eta oso azkarra da.

2.5.2 Baldintzazko ausazko eremuak

Baldintzazko ausazko eremuak (BAE, ingelesez *Conditional Random Field, CRF*), datuak sekuentzialki etiketatu eta erauzten duten eredu estokastikoak dira. 2001ean izan ziren Lafferty *et al.* ikasketa metodo hau aurkeztu zutenak.

Hitz bat etiketatzean, uneko hitzaren propietateak edo ezaugarriak jakiteaz gain, aurretik zetozen hitz kopuru baten ezaugarriak eta hauei jarritako kategoria jakitea beharrezkoa da. Aurretik dauden elementuak aztertzeaz gain, algoritmoa gai da ondoren datozen elementuen ezaugarriak aztertzeko, ziurtasun gehiago emanez ezarriko zaion etiketaren kalitateari. Adibide bezala, ingelesezko esaldi bateko hitzei dagokien kategoria gramatikalak jartzeko unean, hitz bat adjektiboa dela ondorioztatu genezake aurreko hitza *very* izan bada. Elementuen ezaugarriek garrantzi handia dute BAEetan.

BAEn abantaila nagusia ezaugarri multzo handiekin lan egiteko prestatuak daudela da, baina ondorioz ikasketa garestiagoa da.

BAE sinpleenak kate linealak (ingelesez *linear-chain CRF*) dira, non etiketatu beharreko elementu bakoitzak aurreko elementuekiko dependentzia duen. BAEak hedatu daitezke, nahi dugun elementuekiko dependentzia izateko (*general CRF*), hala ere, zenbat eta dependentzia gehiago ezarri, orduan eta konplexuagoa bihurtzen da entrenamendua. 2.2 irudiak bi BAE motak erakusten dizkigu.

Lengoaia naturalaren prozesamenduan, baldintzazko ausazko eremuak asko erabiltzen dira arazoak izaera sekuentziala duenean. Kategoria gramatikalak lortzeko, entitate izenak ezagutzeko, izen sintagmen kateak sortzeko

[41] http://svmlight.joachims.org/svm_perf.html

2.2 irudia – Baldintzazko ausazko eremu moten errepresentazio grafikoa.

eta rol semantikoak etiketatzeko ohikoak dira. Informazio erauzketarako ere oso baliagarriak dira BAE hedatuak, gertaera erauzketarako batez ere.

Tesi honen 6. kapituluko esperimentuetan, BAE sailkatzaile batekin egiten dugu lan.

2.6 Ebaluazio metrikak

Tesi honetan erabiliko ditugun ebaluazio metrikak doitasuna, estaldura, eta F1 neurria dira. Hiru neurri hauek estandarrak dira lengoaia naturalaren prozesamenduko hainbat atazetan, baita ikasketa automatikoa aplikatzen den beste ikerkuntza arlo desberdinetan ere. Zehaztasuna ez da hainbestetan erabiltzen, baina oso erabilgarria da aldi berean.

Doitasunak sistemak itzulitako emaitza zuzenen kopurua itzulitako guztiekin konparatzen du:

$$Doitasuna = \frac{\#(Emaitza\_zuzenak)}{\#(Itzulitako\_emaitzak)} \tag{2.2}$$

Estaldurak sistemak itzulitako emaitza zuzenen kopurua urre patroiak osatzen duten emaitza guztiekin konparatzen du:

$$Estaldura = \frac{\#(Emaitza\_zuzenak)}{\#(Induzitu\_behar\_direnak)} \tag{2.3}$$

F1 neurria doitasuna eta estalduraren arteko batezbesteko harmonikoa da:

$$F1Neurria = 2 * \frac{doitasuna * estaldura}{doitasuna + estaldura} \tag{2.4}$$

Zehaztasunak sistemak itzulitako emaitza zuzenen kopurua corpuseko hitz kopuruarekin konparatzen du:

$$Zehaztasuna = \frac{\#(Emaitza\_zuzenak)}{\#(Corpuseko\_hitz\_kopurua)} \tag{2.5}$$

2.6.1 ACE 2005 atazako kostu balioa

ACE corpusarekin egindako esperimentuetan atazako ebaluatzaile ofiziala erabiltzen dugu. Ebaluatzaileak bi neurri erabiltzen ditu: kostu balioa (*cost-value*) eta betiko F1 neurria.

Kostu balioa sistemaren erantzunetan lor daitekeen balio maximoaren balioa da. Ondorengo formulak adierazten du nola kalkulatu balio hau:

$$KostuBalioa = 100\% - FA - Miss - Err \tag{2.6}$$

Hauek dira parametro bakoitzaren balioak:

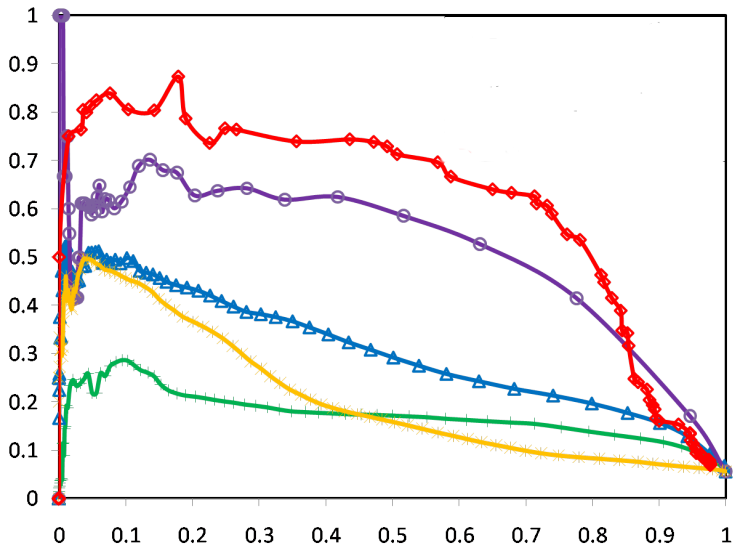
- **FA:** Sistemak bi entitateren artean erlazio bat dagoela adierazi duenean, baina urre patroiaren arabera erlaziorik ez badago.
- **Miss:** Sistemak topatu ez dituen erlazio kopurua.
- **Err:** Sistemak bi entitateren artean erlazio bat dagoela adierazi duenean, baina urre patroiaren arabera erlazioa beste bat bada.

Ebaluazio neurri honi buruz gehiago jakiteko, ikus B eranskina; eranskin honetako dokumentuak kostu-balioa sakonago azaltzen du, dokumentu hori atazako dokumentu ofiziala da.

2.6.2 Doitasun/estaldura kurbak

Sistemen eraginkortasuna ebaluatzeko, betiko neurriak erabiltzeaz gain (doitasuna, estaldura, F1), doitasun/estaldura kurbak erabiltzen ditugu. Kurba hauek sistemen eraginkortasuna hobeto ulertzen laguntzen digute, induzitutako betegarrien probabilitateen iragarpen kalitatea erakutsiz. Doitasun/estaldura kurbak asko erabiltzen dira informazio erauzketan, informazio berreskurapenean eta beste hainbat atazetan (Manning *et al.* 2008).

Kurba hauek egiteko, beharrezkoa da emaitza bakoitzak konfiantza pixu bat edukitzea. Horretarako, atalase bat erabiltzen da emaitzen konfiantza pixuentzat. Sistemak eskuratutako betegarrien probabilitatea aztertuko du, eta probabilitatea atalase horretatik gora duten betegarriak itzuliko ditu,



2.3 irudia – Doitasun/estaldura kurben adibide bat.

ondoren lortutako betegarriekin ebaluatzeko, honen doitasuna, estaldura eta F1 neurria lortuz. Atalasea maximotik minimora jaitsiz, sistemak emaitza bakarra ematetik denak ematera pasatzen da, atalasearen balio bakoitzerako doitasuna eta estaldura emanez, grafikoki irudikatzean kurbak azaleratzen ditu.

Hainbat sistema alderatu nahi direnean, atalase bakoitzaren doitasun/estaldura balioak eskuratu ondoren, bi sistemak grafiko berdinean erakutsi behar ditugu. Grafiko hauetan, guk lortzen ditugun ohiko doitasun, estaldura eta F1 balioak kurbaren azken muturrekoak dira. Kurba bat zenbat eta gorago egon planoan, orduan eta eraginkortasun hobea du, eta A sistemaren kurba B sistemarena baino gorago baldin badago, orduan kurbek argi uzten digute A sistema B baino hobea dela, emaitza errelebanteagoak itzuli dituelako.

Kurbak nolakoak diren argiago uzteko, kurba desberdinak dituen irudi bat jarriko dugu adibide bezala: 2.3 irudia[42]. Grafiko honek 5 sistema konparatzen ditu. Kurba hauek aztertuta, argi dago lerro gorriari (erronboak) dagokion sistema dela eraginkorrena, orokorrean atalase desberdinetan duen asmatze tasa oso altua delako, beste lau sistemekin konparatuta.

[42] http://www.hwkang.com/files/ICOP/ICOP/PixelWisePrecisionRecallCurve_6x8.png webgunetik eskuratua

Gure esperimentu batzuetan doitasun/estaldura kurba hauek erabiliko ditugu, gure sistemen eraginkortasuna beste oinarri lerro batzuekin konparatzeko.

3

Urruneko gainbegiraketa sistema baten garapena

Kapitulu honetan, urruneko gainbegiraketa sistema bat garatuko dugu. Lehenbizi, artearen egoeraren mailan dagoen ikasketa gainbegiratuan oinarritzen den sistema azalduko dugu. Sistema gainbegiratuan aplikatutako ezaugarriak, optimizazio prozesuak eta inferentziak urruneko gainbegiraketa sistemari aplikatuko dizkiogu ondoren. Bukatzeko, errore analisi bat egingo diogu gure sistemari, jakiteko zertan huts egin duen, aurrerago akats hauei konponbide bat bilatzeko.

3.1 Sarrera

Tesiaren planifikazioan, lehenengo urteko helburua urruneko gainbegiraketan oinarritutako erlazio erauzketa sistema bat garatzea zen. Ez zen espero sistema hau urte horietan artearen egoeran zeuden sistemen maila berdinean egotea, baina bai ahalik eta gertuen egotea eraginkortasunaren aldetik. Urruneko gainbegiraketa sistema on bat garatzea zen hurrengo asmoa, eta hau lortu ondoren sistema hauek hobera egiteko teknika desberdinak aztertzea. Sistema gainbegiratuetan lortzen diren emaitzak onak izan ohi direnez, artearen egoerara iristen den sistema baten propietateak urruneko gainbegiraketan aplikatzea zen gure asmoa, sistema honetan ere emaitza egokiak lortzeko. Bide batez, tesigilea erlazio erauzketa eta honen ikasketa metodo desberdinekin trebatzea zen beste helburuetako bat.

Ikasketa gainbegiratuko sisteman aplikatutako modulu desberdinak urruneko gainbegiraketa sisteman ere aplikatzen dira. Modulu horiek aurreprozesaketa, ezaugarrien erauzketa, sailkatzailearen optimizazioa eta inferentziaren optimizazioa dira. Bi sistemak oso antzekoak dira.

Corpusei buruz, gainbegiratutakoan eskuz etiketatutako corpusak erabiltzen dira, beraz aipamenak erraz erauzten dira corpusetik; urruneko gainbegiraketarako ez da eskuz etiketatutako corpusik erabiltzen, automatikoki etiketatutakoak baizik, horretarako bilaketa motore bat behar dugu entitate konkretu bati buruzko aipamenak eskuratzeko.

Sistema gainbegiratuan eta urruneko gainbegiraketan erabili ditugun ebaluatzeko irizpideak desberdinak dira: gainbegiratuan erlazio aipamen bakoitza bere kabuz ebaluatu dugu; urruneko gainbegiraketan entitate konkretu bati buruzko ahalik eta informazio gehien eskuratzen saiatu gara, eta informazio hori urre patroi batean zehaztutakoarekin konparatu. Dena den, erabilitako ebaluazio irizpideak bi ikasketa motetan aplikatu daitezke.

Ikasketa gainbegiratuan oinarritutako sistemaren garapena gidatzeko ebaluazioa ACE corpusean egiten dugu. Ataza honetan, dokumentu batean aurki ditzakegun ahalik eta erlazio aipamen gehien detektatzea da helburua. Ikasketa gainbegiratuan emaitza onak lortu ondoren, urruneko gainbegiraketara ematen dugu saltoa. Horretarako 2011ko TAC-KBPko *"Slot Filling"* atazan parte hartu genuen. Ataza honetan hainbat entitateri buruzko informazioa topatu behar dugu corpus erraldoi batean, topatutako informazio horrekin ezagutza base bat aberasteko.

Urruneko gainbegiraketan lortzen ditugun emaitzak ez dira onak. Sistemak zertan huts egin zuen jakiteko errore analisi bat egiten dugu. Analisi honek paradigma ulertzen laguntzeaz gain, tesiaren hurrengo urratsa definituko du: urruneko gainbegiraketaren ahulguneak detektatu eta hauek konpontzeko esperimentuak egin.

Kapitulu hau ondorengo ataletan dago banatuta: lehenengo ikasketa gainbegiratu bidez sortutako erlazio erauzketa sistemaren garapenean zentratuko gara, ondoren urruneko gainbegiraketa sistemari buruz arituko gara, eta amaitzeko ondorioak eztabaidatuko ditugu.

3.2 Erlazio erauzketa gainbegiratuaren garapena

Atal honetan, sistemaren arkitektura azalduko dugu, entrenamendurako eta testerako jarraitutako pausoekin. Jarraian sailkatzailearentzat sortu ditugun ezaugarriak ere azalduko ditugu, adibide baten bidez. Ondoren optimizazio prozesua azalduko dugu, urratsez urrats. Eta amaitzeko, lortutako emaitzak eztabaidatu.

2.1.2 atalean, ataza honetan aurki ditzakegun entitate eta erlazio motei buruzko informazioa azaldu dugu. 2.2 taulan corpuseko entitate motak eta azpimotak daude zerrendatuta, eta 2.3 taulan corpuseko erlazio motak eta azpimotak. Gogoratu ACE corpusean entitate eta erlazio aipamenak XML dokumentu batean daudela etiketatuta.

3.2.1 Sistemaren arkitektura

Azpiatal hau bi zatitan banatu dugu. Alde batetik entrenamenduko urratsak azalduko ditugu, eta bestetik testeko urratsak. Sistema hau Surdeanu eta Ciaramitaren (2007) laneko sisteman inspiratuta dago.

3.1 irudiak sistema gainbegiratu honen arkitektura erakusten du modu grafikoan, ezkerraldean entrenamenduko urratsekin, eta eskuinaldean testeko urratsekin.

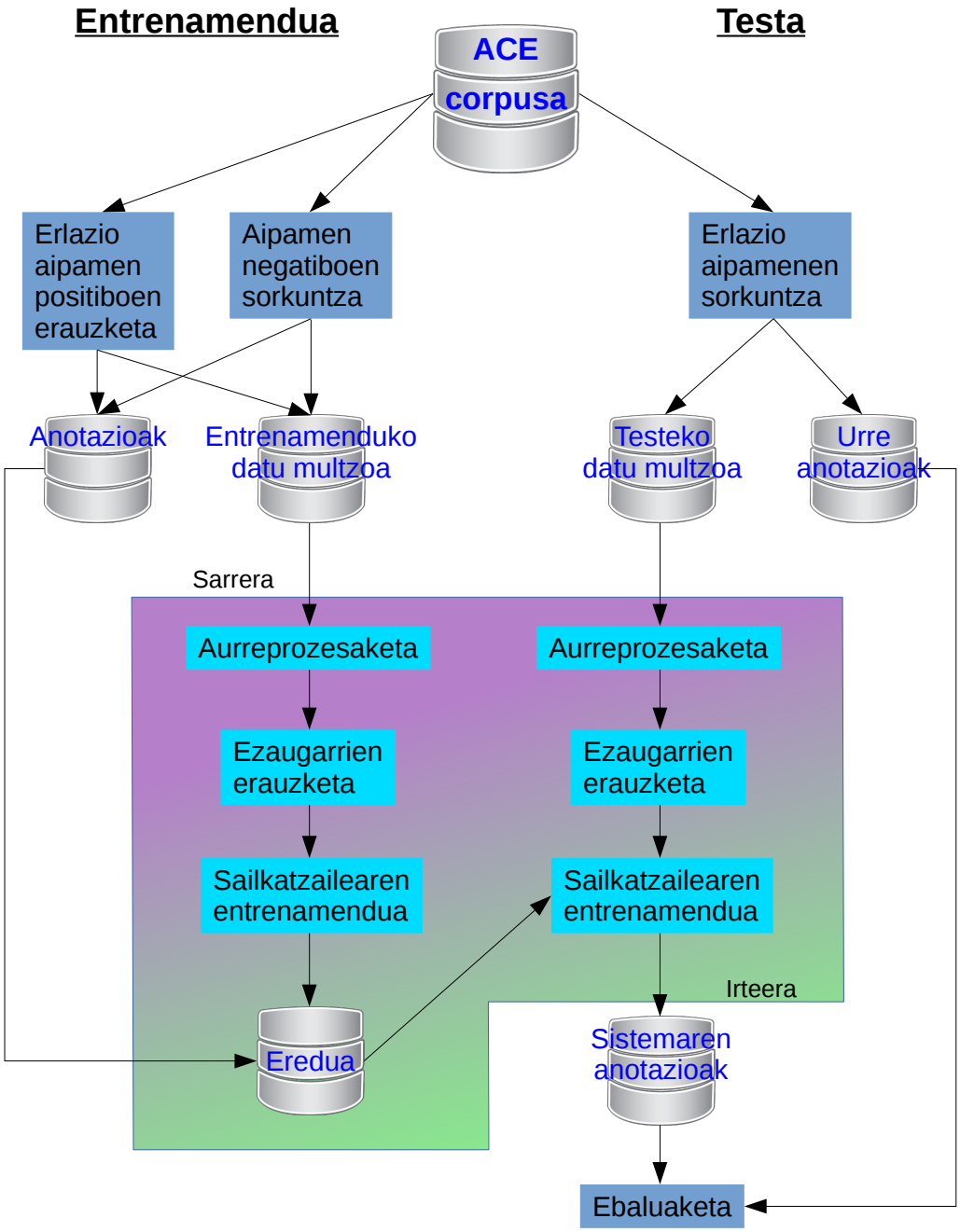
Gure sistema gainbegiratuak ondorengo pausoak jarraitzen ditu ikasketa prozesuan:

- **Erlazioa duten aipamenen erauzketa:** XML fitxategietan zehaztutako erlazio aipamenak hartzen ditugu, dagokien aipamenak testuan bilatu, eta biak agertzen diren testuinguruak erauziz. Adibide bezala, har dezagun esaldi hau:

  Jane Arraf, CNN, on the bridge over the **Khoser River** in northern **Iraq**.[1]

  Entitate nagusia *Khoser River* da, eta bigarren entitatea *Iraq*, bien arteko erlazioa *Part-Whole.Geographical* da. A eranskinean aurki dezakegu adibide hau corpusean etiketatuta.

[1] Adibidea *NN_IP_20030404.1600.00-2.apf.xml* dokumentuan topatua.



3.1 irudia – Sistema gainbegiratuaren arkitektura.

Gure esperimentuetan, testuinguruak bi entitateen arteko inguruko 5 hitzekin sortzen ditugu. Aipatu berri dugun adibidean, esaldia bigarren datorren entitatearen ondoren bukatzen da, beraz testuinguruak hurrengo esaldiaren hitz batzuk edukiko ditu, 5 hitzeko leihoa bete arte.

- **Erlaziorik gabeko aipamenen sorkuntza:** XML fitxategian erlaziorik markatuta ez dagoen entitateen arteko aipamenak sortzen ditugu, baldin eta bi entitateen artean 11 hitz baino gutxiago badaude. Aurreko adibidean, *Jane Arraf* eta *Khoser River* artean ez dago erlaziorik etiketatuta, beraz bi entitateentzat aipamen bat sortzen dugu *NIL* etiketarekin, erlazio eza zehazteko.

  Gauza bera egiten dugu erlazionatuta dauden entitateekin, baina kontrako norabidean. Aurreko adibidean, *Iraq* entitate nagusitzat hartzen badugu, eta *Khoser River* betegarri bezala, erlaziorik etiketatuta ez dagoenez, testuinguru berri bat sortu dugu, hau ere *NIL* etiketarekin.

- **Aipamenen aurreprozesaketa:** Aipamenetako hitz bakoitzarentzat, lema eta kategoria gramatikala eskuratzen ditugu, Stanfordeko CoreNLP[2] tresna erabiliz. Erlazio aipamenean parte hartzen duten entitateentzat, entitate nagusiari *entity* izena jartzen diogu, eta bigarren entitateari, edo betegarriari, *filler*.

- **Ezaugarrien erauzketa:** Aipamen bakoitzarentzat, ikasketa prozesurako ezaugarriak erauzten ditugu. Ezaugarriak 3.2.2 atalean azalduko ditugu.

- **Sailkatzailearen doitzea:** Sostengu bektoreen makina lineal bat erabiltzen dugu entrenamendurako eta optimizazio prozesurako. Azalpen gehiago 3.2.3 atalean.

Testeko prozesuan berriz, ondorengo pausoak jarraitzen dira:

- **Aipamenen sorkuntza:** Aipamenak sortzeko, entrenamenduko antzeko urratsak jarraitzen ditugu. Oraingoan, dokumentuetako esaldi guztiak hartu, etiketatuta dauden entitateak bildu, eta beraien artean erlazio aipamenak sortzen hasten gara, norabide guztietan. Aipamen hauek ez dute inongo erlaziorik zehaztuta, sistemak asmatu behar duelako erlaziorik duten ala ez, eta edukitzekotan, erlazio hori zein den.

[2] http://nlp.stanford.edu/software/corenlp.shtml

- **Aipamenen aurreprozesaketa:** Entrenamenduko pauso berdinak jarraitzen ditugu.
- **Ezaugarrien erauzketa:** Entrenamenduko pauso berdinak baita ere.
- **Sailkatzaileari aipamenak bidali:** Sailkatzailea entrenatu eta optimizatu ondoren, aipamen bakoitza bidaltzen diogu, honek bakoitzaren erlazioa asma dezan.
- **Emaitzak ebaluatu:** Urre patroia eskuz etiketatutako XML dokumentu sorta bat da, entrenamendukoaren itxura berdinarekin. Erlazio aipamena antzematean, XML fitxategi bat sortzen dugu, eta ebaluatzaile ofizialak fitxategi berria originalarekin konparatzen du. Garapen fasean erabilitako pixu atalasearen balio optimoa erabili dugu prozesu honetan.

3.2.2 Ezaugarrien erauzketa

Testuinguruak lortu ondoren, token bakoitzaren lema eta kategoria gramatikala Stanfordeko CoreNLP tresnaren bidez eskuratzen ditugu. Hauek testuinguruen ezaugarriak sortzeko erabiltzen dira.

Artearen egoeran azaldu ditugu bi sistema gainbegiratuk erabiltzen dituzten ezaugarriak (2.2.1 atala). Sistema horietako ezaugarriak erabiltzen dira gure sisteman. Adibide bezala, ausazko aipamen bat eta honen ezaugarriak jarri ditugu 3.1 taulan.

Irakurketa errazteko eta sistema bakoitzaren ezaugarriak bereizteko, lerro batez banatu ditugu. Lehenengo 9 lerroak Mintzen lanekoak dira, ondorengo 7ak Zhou *et al.*-en (2005) lanekoak, eta beste guztiak Surdeanu eta Ciaramitaren lanekoak (2007).

Lehenengo zatiko ezaugarrien egitura konplexua da. *Ezkerra* zutabean lehenengo entitatearen aurretik dauden hitzen ezaugarriak agertzen dira: hitz forma eta honen kategoria gramatikala. Ezaugarri berdinak ditugu *Erdialdea* zutaberako (bi entitateen arteko hitzak ditugu hemen) eta *Eskuina* zutaberako (bigarren iristen den entitatearen ondoren dauden hitzak). *NE1* zutabean lehenengo entitatearen ezaugarriak daude: entitate izen mota eta entitateak aipamenean jokatzen duen tokia (entitatea nagusia edo betegarria den), gauza bera *NE2* zutaberako, bigarren iritsi den entitatearentzat. *ENTITY* ezaugarriak entitate nagusia adierazten du ikasketa gainbegiratuan, urruneko gainbegiraketan entitate nagusia hau da; *FILLER* ezaugarriak berriz,

ENTITATE NAGUSIA - *erlazioa* - BIGARREN ENTITATEA : Khoser River - `PART-WHOLE_Geographical` - Iraq
AIPAMENA: *... over the* **Khoser River** *in northern* **Iraq** *. (...*

EZKERRA	NE1	ERDIALDEA	NE2	ESKUINA
[]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[]
[the/DT]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[]
[over/IN the/DT]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[]
[over/IN the/DT]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[./.]
[over/IN the/DT]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[./. -LRB-/-L]
[the/DT]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[./. -LRB-/-L]
[the/DT]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[./.]
[]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[./.]
[]	LOCATION/ENTITY	in/IN northern/JJ	LOCATION/FILLER	[./. -LRB-/-L]

EZAUGARRI MOTA	BALIOA
WBF	in
WBL	northern
BM1F	over
BM1L	the
AM2F	.
AM2L	-LRB-
DIR	ENTFILL
DIST	2
WM1	Khoser_River
HM1	Khoser
HM1	River
PM1	NN
EM1	LOCATION
WM2	Iraq
HM2	Iraq
PM2	NN
EM2	LOCATION

3.1 taula – Ikasketa gainbegiratuan erabilitako ezaugarriak. Lehenengo 9 lerroak Mintzen lanekoak dira, ondorengo 7ak Zhouren lanekoak, eta beste guztiak Surdeanuren lanekoak.

bigarren entitatea adierazten du ikasketa gainbegiratuan. *ENTITY* eta *FILLER* terminoak urruneko gainbegiraketa sistemarako pentsatuta jarri dira, hurrengo atalean azalduko dugu sistema hau.

Ezaugarri hauek 9 konbinaketa, edo 9 ezaugarri konjuntibo desberdin, egiten dituzte *Ezkerra* eta *Eskuina* zutabeetako hitzekin. Lehenengo lerroan ez dira hauek agertzen, bigarrenean *NE1* aurretik datorren hitza bakarrik dago, hirugarrenean bi hitzak agertzen dira, laugarrenean hitz horiek mantendu eta *NE2* ondoren datorren lehen hitza sartzen da, eta horrela jarraitzen dute ondorengo lerroetan.

Beste bi ezaugarri egiturak sinpleagoak dira, baina egitura aldetik beren desberdintasunak dituzte. Hauek etiketa batez eta honen balioaz egituratuta daude. Bigarren egiturako ezaugarrien etiketak ondorengoak dira, ezaugarri hauek inguruko hitzak biltzen dituzte:

- *WBF*: lehenengo entitatearen ondoren datorren hitza.
- *WBL*: bigarren entitatearen aurretik datorren hitza.
- *WBO*: lehenengo eta bigarren entitatearen artean datozen beste hitzak.
- *BM1F*: lehenengo entitatea baino bi postu aurretik dagoen hitza.
- *BM1L*: lehenengo entitatearen aurretik datorren hitza.
- *AM2F*: bigarren entitatearen ondoren datorren hitza.
- *AM2L*: bigarren entitatearen baino bi postu atzerago dagoen hitza.

Eta hirugarren egiturako ezaugarrien etiketak berriz beste hauek dira, hauek erlazioan parte hartzen duten erlazioei buruzko informazioa jasotzen dute:

- *DIR*: entitateen ordena, *ENTFILL* entitate nagusia aipamenean lehenago agertzen bada, eta *FILLENT* bigarren entitatea aipamenean lehenago agertzen bada.
- *DIST*: bi entitateen artean dagoen hitz kopurua.
- *WM1*: entitate nagusia. Hitz askoz osatutako entitate (*multiword*) bat bada, token guztiak azpimarra (_) batez elkartuta daude.
- *HW1*: entitate nagusiaren hitzak separatuta.
- *PM1*: entitate nagusiaren kategoria gramatikala.
- *EM1*: entitate nagusiaren entitate izen mota.
- *WM2*: bigarren entitatea.
- *HW2*: bigarren entitatearen burua.

- *PM2*: bigarren entitatearen kategoria gramatikala.
- *EM2*: bigarren entitatearen entitate izen mota.

Etiketetako batentzat ez badago informaziorik aipamenean, bi entitateen artean hitzik ez dagoelako adibidez, orduan etiketa hori ez da erabiltzen.

3.2.3 Sailkatzailea

Ikasketa prozesurako SBM (ikus 2.5.1 atala) sailkatzaile bat erabiltzen dugu, konkretuki SVM^{perf} sailkatzailea. Sistemaren emaitzak hobetzeko, hainbat entrenamendu desberdin egiten dira, aipamen negatiboen kantitate desberdinak erabiliz, eta sailkatzailearen parametro desberdinak erabiliz. Emaitza optimoak eman dituzten parametroak dira aukeratuak testeko datu multzoa ebaluatzeko. Optimizazioa F1 neurriarentzat egiten da, baita kostu balioarentzat (ikus 2.6.1 atala) ere, bakoitza bere aldetik.

Garapen fasea 5 partiziotako egiaztapen gurutzatua metodoan oinarritzen da. 5 partizioen bereizketa Surdeanu eta Ciaramitaren lanean egindako berdina da, egileek jarritako partizio bakoitzaren dokumentu zerrenda berdina erabiliz.

Optimizazio prozesua ondorengo urratsetan definitzen dugu:

1. Aipamen negatiboen kopuru egokiena aukeratu.
2. Kostu parametro egokiena aukeratu.
3. Pixu atalase egokiena aukeratu.

Aipamen negatiboen optimizazioa

Ikasketa prozesuan, sistemak aipamen negatiboak ere entrenatzen ditu, ondoren testean dauden aipamenak positiboetatik bereizteko. Aipamen negatiboak entrenatuko ez balira, testean jasotzen diren aipamen guztiak positibotzat hartuko ziren, jakinda testean ebaluatzeko dauden aipamen guztiak ez dutela zertan positibo izan.

Aipamen positibo eta negatibo kopurua konparatzen baditugu, datu multzoan 7 aldiz aipamen negatibo gehiago daude positiboak baino. Kantitate hau oso handia da, halako desorekaren aurrean sailkatzaileek sufritu egiten dute (Li *et al.* 2002). Joera hori saihesteko, ausaz aipamen negatibo kantitate desberdinak hartzen ditugu eta gero sistema hauekin entrenatu, azpilaginketa bat eginez (ingelesez *subsampling*). Aukeratutako negatibo kantitatea

ondorengoak dira: aipamen positibo eta negatibo kopuru berdina, negatibo kopuru bikoitza positiboekiko, hiru aldiz gehiago, 4, 5 eta 6 aldiz gehiago, eta negatibo guztiak (7 aldiz gehiago). Emaitza onenak aipamen positiboak baino 6 aldiz negatibo gehiago erabiliz lortzen ditugu kostu balioarentzat, eta 5 aldiz gehiago F1 neurriarentzat.

Sailkatzailearen optimizazioa

Aipamen negatiboen kopuru optimoena aukeratzeaz gain, sostengu bektoreen makinaren kostu parametroa (C) ere optimizatzen dugu algoritmo jale baten bidez. Bilaketak parametroaren hainbat balio probatzen ditu iteratiboki. Iterazio bakoitzean parametro balioen leihoa (balio maximo eta minimoak) eta balioen arteko tartea definitzen dugu, aurreko iterazioko balio optimoaren arabera. Iterazio bakoitzean leihoa eta tartea estutzen direnez, balio optimo lokal baterantz joatea bermatzen dugu. Gure kasuan, 3 iterazioetara murriztu dugu bilaketa.

Parametro hau optimizatzeko, 0.01 eta 20 arteko hainbat balio probatu ditugu. Optimizazioa 3 urratsetan egiten da:

1. Probatu balio guztiak, 2 unitateko gehikuntzakin, eta ebaluazio neurri onena duenarekin gelditu.

   Demagun $C = 10$ izan dela onena. Ziur gaude C horren inguruko balio txikiagoekin frogatuz, emaitzak gehiago hobetu ditzakegula.

2. Jarritako adibidea jarraituz, probatu kostu parametroa 8 eta 12 artean, 0.4 unitateko gehikuntzarekin.

   Demagun $C = 9.2$ izan dela onena. Ziur gaude inguruko are txikiagoak diren balioekin emaitzak ere hobetuko ditugula.

3. Jarritako adibidea jarraituz, eta azken saiakera eginez, probatu kostu parametroa 8.8 eta 9.6 artean, 0.08ko gehikuntzarekin. $C = 9.44$ izan bada onena, balio horrekin geldituko gara. Ez dugu beste saiakerarik egingo balio txikiagoak aztertuz.

Emaitza onenak kostu parametroari 0.66 emanez lortzen ditugu bi ebaluazio neurrientzat.

	Optimizatzeko neurria	
Optimizatutako parametroak	**F1 neurria**	**Kostu balioa**
Aipamen negatibo kopurua	*pos * 5*	*pos * 6*
Kostu parametroa	0.66	0.66
Pixu atalasea	-0.25	-0.25

3.2 taula – F1 neurria eta kostu balio optimoenak lortzeko optimizatu diren parametroen balioak. *pos* parametroak datu multzoko positibo kopurua adierazten du.

Inferentzia

SBM sailkatzaileak ebaluatu beharreko aipamen bakoitzari pixu bat ematen dio erlazio etiketa bakoitzeko. Pixu hauek positiboak eta negatiboak izan daitezke.

Erlazio etiketa egokienak hautatzeko, pixu atalase desberdinak ezarri ditugu, -1 eta 2 artekoak, eta pixua atalase hori baino altuagoa duten etiketa guztiekin gelditu. Iragarpen bakoitzeko, pixu altuena daukan erlazio etiketa da hautatua. Hainbat atalaserekin probatu ondoren, 0.25eko tarteak konprobatuz emaitza onena ematen duen atalasearekin gelditzen gara. Emaitza onenak pixu atalaseari (-0.25) emanez lortzen ditugu, bi ebaluazio neurrientzat. Oraingoan ez dugu optimizaziorik egiten inguruko balio txikiagoentzat.

Aurreko urratsetan egindako optimizazioak 0ko inferentzia atalasea erabiliz egin dira. Urrats hau aurreko parametroen balio optimoenekin egiten dugu. 3.2 taulak ebaluazio neurri bakoitzerako lortutako parametro optimoen balioak laburtzen ditu.

3.2.4 Emaitzak

F1 neurriak doitasuna eta estaldura konbinatzen ditu, neurri hau oso erabilia da sistemak ebaluatzeko, baina ACE atazan garrantzi gehiago dauka kostu balioak (2.6.1 atala). Hala ere, optimizazio prozesua bi neurrien balio onenak lortuz egiten dugu.

3.3 taulan sistemaren garapen faseko emaitzak daude. Oinarri lerro bezala Surdeanu eta Ciaramitaren laneko emaitzak erabili dira. Gure optimizazioa kostu balioan eta F1 neurrian oinarritu da, eta zoritxarrez, bietatik ezta batek ere ez du lortu oinarri lerroaren kostu balioa hobetzea. F1 neurria

Sistema	Kostu balioa	Doit.	Est.	F1 neurria
Oinarri lerroa	**34.8**	74.0	46.9	**57.4**
Kostu balioa	29.7	68.3	45.4	54.5
F1 neurria	30.3	65.3	51.2	**57.4**

3.3 taula – Sistema gainbegiratuaren garapen faseko emaitzak.

Sistema	Kostu balioa	Doit.	Est.	F1 neurria
Oinarri lerroa	**33.1**	76.1	42.5	**54.5**
Kostu balioa	22.0	65.8	35.7	46.2
F1 neurria	23.3	64.1	40.7	49.8

3.4 taula – Sistema gainbegiratuaren testeko emaitzak.

optimizatuz gero, balio hau oinarri lerroarekin berdintzea lortu dugu. Deigarria da F1 neurria optimizatzean lortutako kostu balioa altuagoa izatea, balio bera optimizatzen saiatzean baino.

3.4 taulan sistema gainbegiratuaren testeko emaitzak daude. Hemen ere garapen faseko antzeko egoera batean gaude: (1) ez dugu lortu kostu balioa hobetzea, eta honen balio hobea lortu dugu F1 neurria optimizatzean, baina (2) oraingoan ez dugu lortu F1 neurria berdintzea ez hobetzea, eta urruti gaude oinarri lerrotik bi ebaluazio neurrietan.

Bukaerako emaitzetan ikusten den bezala, ez dugu lortu gure sistemarekin artearen egoera berdintzea, baina gertu gaude, F1 neurriarekin behintzat.

3.3 Urruneko gainbegiraketan oinarritutako erlazio erauzketa sistema

Atal honetan urruneko gainbegiraketan oinarritutako erlazio erauzketa sistema baten garapenaz arituko gara. Sistema honetan erabiliko ditugun ezaugarriak eta optimizazio prozesuak aurreko atalean deskribatu ditugun berdinak dira. Sistema hau Mintz *et al.*-en laneko (2009) sisteman oinarrituta dago.

Gogora dezagun urruneko gainbegiraketaren arabera, ezagutza base batean bi entitateren artean erlazio bat zehaztuta badago, bi entitateak dauden edozein esaldik erlazio hori aditzera emango duela.

Urruneko gainbegiraketa sistema sortu eta TAC-KBP 2011 atazeko *Slot Filling* (ikus 2.1.2 atala) azpiatazan parte hartzeko aprobetxatu genuen. Ataza honen helburua, corpus batean entitate bati buruzko bilaketak eginez honek beste entitate batzuekin dituen erlazioak erauzi eta ezagutza baseak aberastea da. Ataza honetan pertsona eta erakundeei buruzko informazioa erauztea da helburua, artearen egoerari buruzko kapituluaren 2.4 taulan daude erlazio guztiak zerrendatuta.

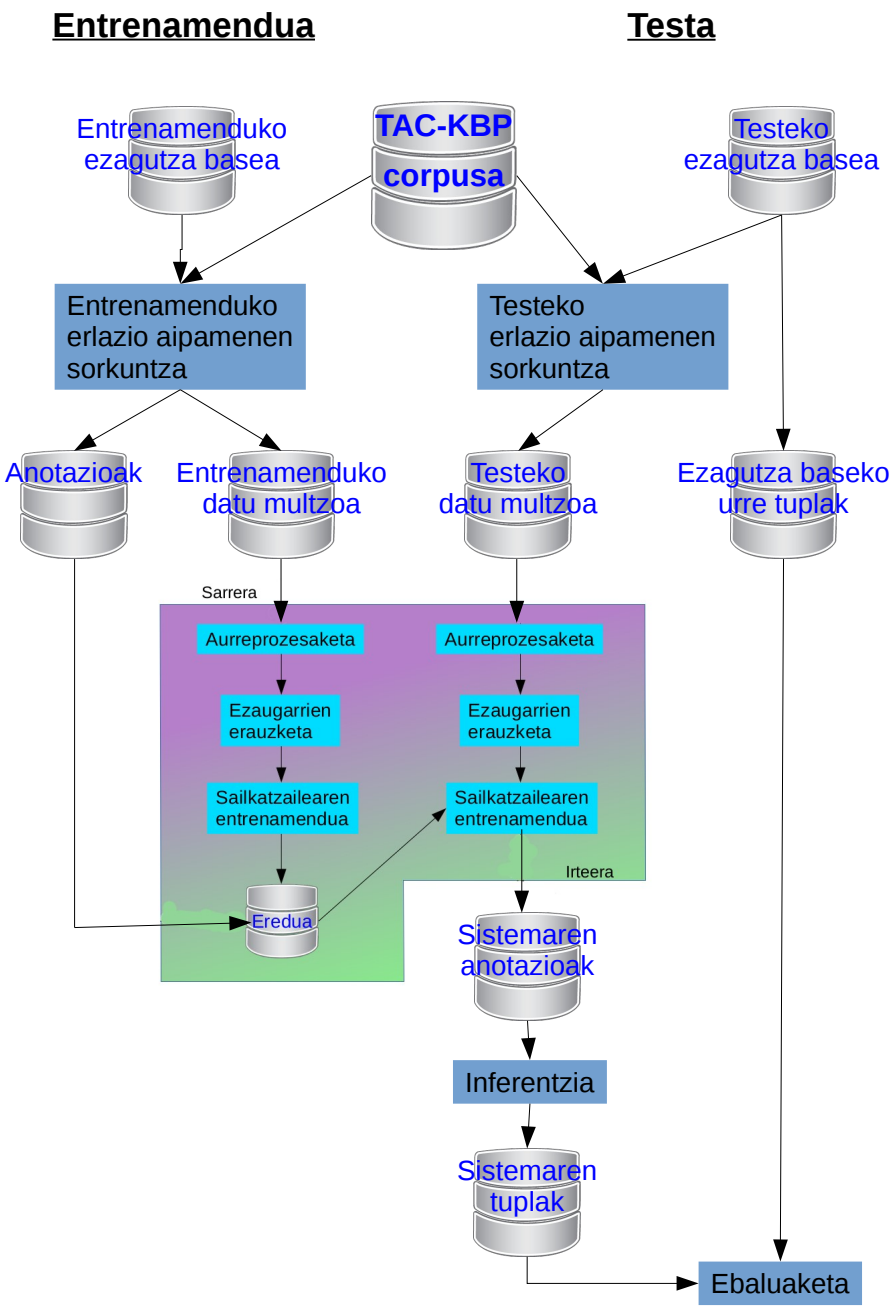
Ataza honetarako 3 exekuzio desberdin bidali genituen. Exekuzio hauen desberdintasuna aipamenen berreskurapenean eta inferentzian datza. Exekuzio hauei izen hauek jarri genien: *UBC1*, *UBC2* eta *UBC3*. Desberdintasunak aurrerago aipatuko ditugu.

3.3.1 Aipamenen berreskurapena

3.2 irudiak urruneko gainbegiraketa sistema honen arkitektura erakusten du modu grafikoan, ezkerraldean entrenamenduko urratsekin, eta eskuinaldean testeko urratsekin. Ikusten den bezala, arkitektura ikasketa gainbegiratuaren oso antzekoa da. Aipamenen berreskurapenean, etiketatzean eta ebaluaketan daude aldaketak.

Ataza honetarako erabiltzen den ezagutza basea Wikipedian dago oinarrituta. Artearen egoeran komentatu dugun bezala, ezagutza baseko infotauletan erabiltzen diren terminoak ez daude bateratuta. Hori dela eta ondo mapatu beharra dago, bertatik entitate tuplak eta hauen arteko erlazioak ondo ordenatzeko, eta atazan erabiltzen diren erlazio etiketekin bat egiteko. Gogoratu Wikidata ezagutza basearen helburua infotauletan erabiltzen diren terminoak bateratzea dela, beraz, behin proiektua nahiko aurreratuta dagoela, ezagutza base hori erabiliz gero ez genuke mapatzean denbora galdu beharko. Azken edizioetan Freebase ezagutza basea erabili dute, lana asko errazteko eta fase hau saihesteko.

3.5 taulan, mapaketaren beharraren adibide bat dugu, erakunde desberdinei buruz Wikipediako infotauletan jartzen duen informazioarekin, *org:founded* erlaziorako. 3 termino desberdin erabiltzen dira erakunde baten sorrera adierazteko, eta termino hauek antolatzaileek sortutako ezagutza basean agertzen dira baita ere, bateratu gabe. Mapaketaren bidez, termino hauek bateratzen dira, baina eskuzko lan handia suposatzen du, gauza berdina adierazten duten kasuak topatu behar direlako.



3.2 irudia – Urruneko gainbegiraketa sistemaren arkitektura.

Erakundea	Terminoa	Balioa
Cornell University	Established	1865
FDA	Formed	1906
Oracle Corporation	Founded	June 16, 1977

3.5 taula – Ingeleseko Wikipediako infotauletan erakundeen sorrerari buruz aurki ditzakegun terminoak.

Lan honetarako erabilitako mapaketa bertsioa Stanford Unibertsitateko Lengoaia Naturalaren Prozesamenduko ikerkuntza taldeak[3] jarri zigun eskuragarri.

Atazan parte hartzeagatik eskuragarri dugun corpusetik aipamenak eskuratzeko Lucene[4] bilaketa motorea erabili dugu. Sarrera bezala ezagutza baseko tuplak pasatzen dizkiogu motoreari, batetik entitate nagusia eta bestetik, bigarren entitatea (edo betegarria). Motoreak bi entitateak dauzkaten esaldi edo aipamen guztiak itzultzen ditu, eta aipamen hauei ezagutza baseko erlazio etiketa jarri. Pauso hau entrenamenduko datu multzoa sortzeko egiten da.

Testeko datu multzoa sortzeko helburu entitatea bakarrik pasatzen diogu motoreari, honek entitate hau agertzen den aipamen guztiak itzultzeko. Aipamen berdinean agertzen diren beste entitateak eta hitzak dira erlazio desberdinetarako betegarri potentzialak. Eskuratzen ditugun aipamen hauetan, bi entitateak esaldi berean agertzen dira.

Aipatu beharra dago oraingoan ez daukagula entrenamenduko datu multzoan aipamen negatiborik, hau da, ikasketa gainbegiratuan erabilitako *NIL* etiketa duten aipamenik. Markatutako entitate eta betegarri guztien artean dago erlazio bat gutxienez.

Aipatu berri ditugun aipamenak *UBC2* eta *UBC3* exekuzioetan erabiltzen dira. *UBC1* exekuzioarentzat lortutako aipamenetan ez da esaldi bereizketarik egiten, bi entitateen artean puntu bat (.) egon daiteke, honek bi entitateak testuinguru desberdinetan egotearen arriskua ekartzen du, entrenamenduan sailkatzailea nahasteko arriskuarekin.

[3]Stanford NLP Group, http://nlp.stanford.edu/
[4]http://lucene.apache.org/

Aipamen guztiak lortu ondoren, entrenamenduko datu multzoa prest dago. Aipamenen aurreprozesaketa sistema gainbegiratuaren berdina da, hitz bakoitzaren lema, kategoria gramatikala eta entitate izen mota eskuratuz.

3.3.2 Ezaugarrien erauzketa

Aipamenak aurreprozesatu ondoren, entrenamenduko datu multzoari dagokionez, ikasketarako sortutako ezaugarriak 3.2.2 azpiataleko berdinak dira. Ezaugarrien adibide bat dugu, 3.6 taulan.

Testeko datu multzoko helburu entitateen testuinguruak sortzeko, hauen aurretik eta atzetik dauden 30 hitzak hartzen ditugu. Jarraian inguruko entitateak antzeman, eta helburu entitatearekin batera testuinguruak sortzen dira. Amaitzeko, ezaugarriak erauzten ditugu.

Testuinguruan aurkitutako entitate mota bakoitzeko betegarri potentzialak sortu ditugu, eta ezaugarriak normal prozesatu, ezaugarriak entrenamenduko datu multzoko berdinak dira.

3.3.3 Sailkatzailearen doiketa

Ataza honetako garapen faseak eta ikasketa gainbegiratuarenak helburu desberdinetan oinarritzen dira. Ikasketa gainbegiratuan, egitura berdina daukate garapeneko datu multzoak eta testeko datu multzoak: dokumentu bakoitzean bi entitate eta hauen arteko erlazioak markatuta daude, sistemak garapenekoak ikasteko erabiltzen ditu eta testekoak asmatu behar ditu. Bai garapen fasean, bai testean, 100 helburu entitateri buruzko informazioa erauzi behar dugu.

Garapen faseko helburu entitateak 2010eko atazako testeko entitateak ziren. Testeko ebaluaziorako berriz, 2011koak, parte hartu genuen urtekoak. D eranskinean daude garapen faserako (D.1 taula) eta testerako (D.2) erabilitako helburu entitateen zerrendak.

Azkenean, garapen faseko sistemarako sortutako 2010eko garapeneko multzoa sortzeko, testeko pauso berdinak jarraitu behar dira, aldatzen den gauza bakarra helburu entitateen zerrenda da. Garapen fasean ez da egiaztapen gurutzaturik erabili.

Optimizazio prozesua ondorengo urratsetan definituta dago:

ENTITATE NAGUSIA - *erlazioa* - BETEGARRIA : Dominican University - `org:city_of_headquarters` - River Forest
AIPAMENA: *...courses at* **Dominican University** *in the Chicago suburb of* **River Forest** *shortly before...*

EZKERRA	NE1	ERDIALDEA	NE2	ESKUINA
[]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[]
[at/IN]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[]
[courses/NN at/IN]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[]
[courses/NN at/IN]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[shortly/RB]
[courses/NN at/IN]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[shortly/RB before/IN]
[at/IN]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[shortly/RB before/IN]
[at/IN]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[shortly/RB]
[]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[shortly/RB]
[]	ORG/ENTITY	in/IN the/DT Chicago/NN suburb/NN of/IN	LOC/FILLER	[shortly/RB before/IN]

EZAUGARRI MOTA	BALIOA
WBF	in
WBL	of
WBO	the Chicago suburb
BM1F	courses
BM1L	at
AM2F	shortly
AM2L	before
DIR	ENTFILL
DIST	5
WM1	Dominican_University
HM1	Dominican
HM1	University
PM1	NN
EM1	ORG
WM2	River_Forest
HM2	River
HM2	Forest
PM2	NN
EM2	LOC

3.6 taula – Urruneko gainbegiraketan erabilitako ezaugarriak. Lehenengo 9 ilarak Mintz *et al.*, 2009 lanekoak dira, ondorengo 7ak Zhou *et al.*, 2005 lanekoak, eta beste guztiak Surdeanuren lanekoak.

1. Aukeratu kostu parametro egokiena.

2. Aukeratu pixu atalase egokiena, exekuzioaren arabera dagokion inferentzia metodoarekin.

Sailkatzailearen optimizazioa

Ikasketa prozesurako, oraingoan ere SMV^{perf} sailkatzailea erabiltzen dugu. Berriz ere kostu parametroa (C) optimizatu dugu eta sistemak itzultzen duen emaitza onenarekin gelditu. Probatutako kostu balioak beste behin 0.01 eta 20 artekoak dira.

Sailkatzailea entrenatzean, pertsonei buruzko aipamenak eta erakundeei buruzkoak ez dira nahastu. Horrela, sailkatzaileak testeko aipamenak jasotzean, helburu entitate mota badakienez, demagun pertsona dela, aipamenak aztertzean pertsonei buruz ikasitakoaren arabera egiten ditu iragarpenak, erakundeei buruzkoak alde batera utzita.

Inferentzia

Atal honen hasieran aipatu dugun bezala, ataza honetan 2 inferentzia metodo desberdinekin esperimentatu dugu. Inferentzia metodo hauek aurreko ataleko ikasketa gainbegiratuko sistemaren inferentzia metodoan oinarritzen dira. Oraingoan ere, pixu atalase bat daukagu zehaztuta, -1 eta 2 arteko atalase bat, non atalase hau baino pixu gutxiago duten betegarri potentzialak alde batera uzten diren. Aukeratutako atalasea ebaluatzean, emaitza onena ematen duena da.

Inferentziarekin hasi aurretik, kontuan hartu behar dugu helburu entitate bakoitzeko hainbat aipamen ditugula eskuratuta, eta aipamen hauen barnean hainbat betegarri potentzial daudela. Betegarri potentzial bakoitza hainbat alditan agertu daiteke aipamen desberdinetan, eta aipamen bakoitzean pixu desberdina edukiko du erlazio bakoitzeko.

Lehenbiziko inferentzian, *UBC1* eta *UBC2* exekuzioetan erabilia, erlazio bakoitzarentzat pixu altuena daukan betegarria itzuli aurretik, betegarri potentzial hau erlazio motarekin bateragarria ote zen ziurtatzen dugu; hala bada, orduan betegarri hau itzultzen dugu; ez bada, orduan betegarri hau baztertu eta bigarren pixu altuena daukan betegarriaren bateragarritasuna aztertzen dugu, eta horrela jarraitzen dugu bateragarria zen betegarri batekin ematen dugun arte. Betegarri hori aurkitzean, bere pixua atalase batena

baino altuagoa bada, orduan betegarri hau itzultzen du sistemak ebaluatzeko. Betegarri posibleak pixuaren arabera ordenatuta daude.

Bigarren inferentzia teknika aurrekoaren hedapen bat da, hau *UBC3* exekuzioari aplikatu diogu. Inferentzia honetan, helburu entitate bakoitzaren betegarri potentzialak biltzen ditugu, berdin izanda pixu hau positibo edo negatiboa den, erlazio bakoitzarentzat daukaten pixuen batera. Jarrian betegarri bakoitzaren bateragarritasuna aztertzen dugu erlazio bakoitzeko, eta bateragarriak direnekin gelditu. Ondoren betegarri bakoitzaren pixuaren batuketa egiten dugu, eta lehenengo 3 batura altuena daukaten betegarriekin gelditu. Azkenik, batura horien balioak atalase baten baliotik gora baldin badaude, betegarri hauek aukeratzen ditugu erlazio bakoitzaren balio posible bezala eta hauek itzuli ebaluatzeko.

Bi inferentzietan, betegarri potentzial bakar batek ere ez baldin baditu baldintzak betetzen, orduan ez dugu erantzunik itzultzen.

3.3.4 Emaitzak

Ebaluaziorako *"Slot Filling"* atazaren ebaluatzaile ofiziala erabiltzen dugu. Ebaluatzaile honek hain ohikoak diren doitasuna, estaldura eta F1 neurria kalkulatzen ditu. Oraingoan ez dago beste neurri osagarririk, hori dela eta optimizazioa F1 neurriaren arabera egin dugu. Kontuan izan betegarriek ez dutela zertan erlazio guztiekiko informazioa eduki behar.

Emaitzak ebaluatzean, betegarri bakoitza corpuseko zein dokumentutatik lortu dugun espezifikatu behar zaio ebaluatzaileari. Urre patroian dokumentu batzuk zerrendatuta daude erlazio bakoitzeko, aurretik eskuz konprobatuta dago dokumentu horietan betegarriak eta erlazioak bat egiten dutela testuinguruarekin. Guk pasatako dokumentuak zerrenda horretako batekin bat egiten badu, erantzuna ontzat hartzen da. Behin baino gehiagotan gertatu ohi da betegarri bat zuzena izatea, baina hau lortu dugun dokumentua ontzat ez ematea, aipamenaren dokumentua zuzena ez delako.

Ebaluatzaileak aukerako parametro berezi bat dauka: *anydoc* parametroa. Parametro hau erabilita, ebaluatzaileak ez du kontuan hartzen zein dokumentutatik eskuratu dugun betegarria. Gure optimizazioak parametro hau erabilita egin ditugu. Hala ere, antolatzaileek emaitza ofizialak argitaratzean aukera hau ez dute erabiltzen.

Orora, 3 exekuzio bidali genituen:

	Exekuzioak		
Optimizatutako parametroak	**UBC1**	**UBC2**	**UBC3**
Kostu parametroa	5	1	1
Pixu atalasea	-0.5	-0.5	-0.25

3.7 taula – Garapen fasean exekuzio bakoitzarentzat F1 neurri optimoena lortzeko optimizatu diren parametroen balioak.

- *UBC1* exekuzioan, datu multzoko aipamenetako entitateak bi esaldi desberdinetan egon daitezke. Erabilitako inferentzia teknika sinpleena da, pixu altuenak dituzten betegarri potentzialak eta erlazio mota bateragarriak diren baino ez da konprobatzen.

- *UBC2* exekuzioan, datu multzoko aipamenetako entitateak esaldi berdinean daude. Erabilitako inferentzia metodoa *UBC1*eko berdina da.

- *UBC3* exekuzioan, entitateak esaldi berdinean daude. Inferentzia mota desberdin bat erabili da oraingoan, non erlazioaren eta betegarriaren arteko bateragarritasuna konprobatzeaz gain, hauen pixuak batu dira eta pixu altuena zuten lehenengo 3 betegarriak itzuli.

D eranskineko D.1 taulan, garapen faseko helburu entitateen zerrenda dago (2010eko edizioan erabili ziren testeko entitateak). Eranskin bereko D.2 taulan berriz, testeko faseko helburu entitateak daude.

3.7 taulak garapen fasean F1 neurriarentzat lortutako parametro optimoen balioak laburtzen ditu, exekuzio bakoitzarentzat, *anydoc* aukera erabiliz. *UBC2* eta *UBC3*ren kostu parametroa berdina da testuinguruaren balio berdina erabili dugulako, bi exekuzio hauen artean inferentzia da aldatzen den gauza bakarra.

3.8 taulan ditugu exekuzio bakoitzaren garapen faseko emaitzak, *anydoc* erabiliz eta hau gabe. Azken lerroan. 2010eko atazan lortutako batezbestekoa F1 neurria adierazten da; emaitza ofizialak *anydoc* erabili gabe kalkulatu zirenez, parametro hau erabiliz lortutako batezbestekoaren daturik ez dago.

Taularen lehenengo zatian *anydoc* parametroa erabili dugu ebaluaziorako, ikus dezakegunez, *UBC1* exekuzioan lortu ditugu emaitza baxuenak, baina ez dira hain baxuak beste biekin konparatuta, altuena berriz *UBC2* exekuzioan lortu da, %1 baino gehiagoko aldearekin. Doitasun aldetik, altuena

Exekuzioa	Doitasuna	Estaldura	F1 neurria
UBC1	4.46	7.87	5.69
UBC2	5.92	**8.04**	**6.82**
UBC3	**6.30**	5.65	5.96
Emaitzak *anydoc* gabe			
UBC1	1.71	**3.04**	2.19
UBC2	2.14	2.87	2.45
UBC3	**2.66**	2.54	**2.60**
2010eko batezbesteko F1 neurria:			**10.54**

3.8 taula – Urruneko gainbegiraketa sistemaren garapen faseko emaitzak exekuzio bakoitzarentzat, *anydoc* parametroa erabilita eta erabili gabe.

UBC3 exekuzioan lortu dugu eta baxuena *UBC1* exekuzioan. Estaldura altuena *UBC2* exekuzioan eta baxuena *UBC3* exekuzioan. Gogoan izanda exekuzio bakoitzaren ezaugarriak zeintzuk diren, argi ikusten da entitateen arteko erlazioak bilatu nahi baditugu, biak esaldi berean egon behar direla. Inferentziaren aldetik berriz, bigarren inferentzia teknikak, *UBC3*n erabilitakoak, doitasunari mesede handia egin dio baina estaldura aldetik jaitsiera nabarmena da.

anydoc parametroa ez badugu erabiltzen, emaitzek beherakada handia dute, espero bezala. Emaitza altuena *UBC3* exekuzioak lortu du eta baxuena *UBC1*k, hiru exekuzioen artean desberdintasun gutxi dago F1 neurriaren aldetik. Doitasun altuena *UBC3*k dauka orain ere eta baxuena *UBC1*k. Estalduran egon dira aldaketak, *UBC1*k dauka altuena eta ez *UBC2*k. Emaitzak oso baxuak dira, guk lortutako F1 neurriak oso urruti daude 2010eko atazan lortutako F1 neurriaren batezbestekotik.

Gure sistemak emandako erantzunen artean, baliteke horietako asko zuzenak izatea, baina aurretik ezagutza basean zegoenez informazio hori, hauek ez dira ez zuzentzat ez okertzat hartzen. Kontuan hartu ataza honen helburua ezagutza baseak aberastea dela, ez aurretik genuen edozein informazio eskuratzea. Gauza berdina gerta liteke testeko ebaluazioan.

3.9 taulan testeko ebaluazioan bilatu beharreko erlazio kopuruak daude zerrendatuta. Erlazio askorentzat informazio asko erauzi behar dugu, *per:title*, *per:employee_of* eta *org:alternate_names* erlazioentzat adibidez. Beste erlazio batzuentzat, hala nola *per:country_of_birth* eta *org:founded*, infor-

mazio gutxi eskuratu behar dugu, 50 pertsonetatik batentzat bakarrik bilatu jaioterria, beste 49entzat berriz ez, informazio hori aurretik bazegoelako ezagutza basean.

3.10 taulan testeko emaitza ofizialak ditugu. Garapen faseko emaitzak ikusita, espero genuen emaitza oso baxuak lortzea eta hala izan da. Emaitza onena *UBC3* exekuzioan lortu dugu, eta txarrena *UBC1* exekuzioan, baina ez dago alde handirik bien artean. Doitasun altuena *UBC2*k lortu du. Estaldura altuena berriz, *UBC3*k lortu du, garapen fasean baxuena ematen zuen eta oraingoan altuena izan da. Azken lerroan 2011ko atazako ebaluazio neurri bakoitzaren batezbesteko emaitzak daude, guk lortutakoak oso urruti daude edizio horretan lortu ziren batezbesteko emaitzekiko.

Gure sistemak akats asko ditu, hori argi eta garbi ikusten da aurreko taulak ikusita. Akats horiek zeintzuk izan ziren jakiteko errore analisi bat egin dugu. Analisi honen helburua ez da bakarrik gure sistemaren gabeziak detektatzea, urruneko gainbegiraketak dituen beste hainbat ahulgune detektatu eta zuzentzea ere.

3.3.5 Errore analisia

Nahiz eta hasieratik emaitza txarrak espero izan urruneko gainbegiraketan, lortutako emaitza hauek oso urruti daude egokitzat hartzeko. Egoera honek gure sistema nola eginda dagoen aztertzera eramaten gaitu. Azpiatal honetan, aurkitutako errore guztiak zerrendatuko ditugu eta hauek konpontzeko proposamenak egin.

Aipamen positibo falta

Erlazio konkretu batzuentzat aipamen gutxi eskuratu dira corpusetik, horrek ikasketa asko zailtzen du erlazio hauentzat, baita testeko aipamenak aztertzean erantzunak iragartzea ere. *per:date_of_death*, *per:siblings* eta *org:website* erlazioek 10 aipamen edo gutxiago dituzte. Beste erlazio batzuentzat berriz, ez dago inongo aipamenik datu multzoan. Beraz, ezinezkoa da erlazio hauei buruzko iragarpenak lortzea, sailkatzailearentzat existituko ez balira bezala da. Erlazio hauek *per:cause_of_death*, *per:charges*, *per:other_family* eta *org:shareholders* dira.

Pertsona erlazioak	**Guztira**	**Erakunde erlazioak**	**Guztira**
per:age	13	org:alternate_names	81
per:alternate_names	39	org:city_of_headquarters	15
per:cause_of_death	2	org:country_of_headquarters	14
per:charges	14	org:founded	3
per:children	17	org:founded_by	4
per:cities_of_residence	10	org:member_of	7
per:city_of_birth	7	org:members	9
per:city_of_death	1	org:number_of_employees/members	4
per:countries_of_residence	14	org:parents	20
per:country_of_birth	1	org:political/religious_affiliation	1
per:country_of_death	1	org:shareholders	20
per:date_of_birth	3	org:stateorprovince_of_headquarters	13
per:date_of_death	3	org:subsidiaries	37
per:employee_of	65	org:top_members/employees	77
per:member_of	50	org:website	11
per:origin	14		
per:other_family	6		
per:parents	1		
per:religion	5		
per:schools_attended	15		
per:siblings	5		
per:spouse	8		
per:stateorprovince_of_birth	2		
per:stateorprovinces_of_residence	1		
per:title	173		
Guztira	475	**Guztira**	316

3.9 taula – Erlazio kopurua TAC-KBP 2011ko urre patroian.

Exekuzioa	Doitasuna	Estaldura	F1 neurria
UBC1	4.45	2.96	3.55
UBC2	**5.36**	2.85	3.72
UBC3	4.74	**3.28**	**3.87**
2011ko batezbestekoa	**16.50**	**10.31**	**12.69**

3.10 taula – Urruneko gainbegiraketa sistemaren test emaitzak exekuzio bakoitzarentzat, *anydoc* parametroa erabili gabe eta 2011ko atazako batezbesteko emaitzekin batera.

Aipamen negatiborik ez

Ikasketa gainbegiratuan aipamen batzuk izan ditugu, non bi entitateren artean ez dagoen erlaziorik. Aipamen hauek *NIL* etiketa daukate eta ikasketa prozesuan oso baliagarriak direla erakutsi dugu, aipamen berriak ebaluatzean bereiztu ahal izateko noiz zeuden aipamenen bi entitateak erlazionatuta eta noiz ez.

Ataza honetan, urruneko gainbegiraketa sistemarako, ez daukagu aipamen negatiborik. Hori dela eta, sistemak betegarri potentzial desberdinak detektatu eta iragarpenak egiterakoan, beti ezarri beharko die erlazio bat, ikasi duenaren arabera bi entitate beti daudelako erlazionatuta, inoiz ezin izango du esan bi entitateren artean erlaziorik ez dagoenik.

Aipamen positibo zaratatsuak

Bilaketa motoreak corpusetik erauzitako hainbat aipamen ez dira egokiak ikasketa automatikorako. Urruneko gainbegiraketaren arabera, ezagutza base batean bi entitate erlazionatuta badaude, orduan bi entitateek parte hartzen duten edozein aipamenak bien arteko erlazioa adieraziko du. Baina hau ez da beti horrela gertatzen, kasu asko topatu ditugu non bi entitateak aipamen berean dauden, baina testuinguruaren arabera erlazio hori ematen ez den, beraz aipamen horretan bi entitateak ez daude erlazionatuta. Honek azkenean ikasketa prozesuan sailkatzaileak gaizki ikastera eramaten du.

Har dezagun adibide bezala <Chrissie Hynde - *per:city_of_birth* - Akron> tripleta. Urruneko gainbegiraketaren arabera Chrissie Hynde eta Akron entitateak agertzen diren esaldi orok Chrissie Akronen jaio zela esaten du, baina ondorengo aipamenean argi ikusten da hau ez dela beti horrela gertatzen:

- (...) Singer **Chrissie Hynde** took a ride on an **Akron** city bus to show her (...)

Adibide honetan bi entitateek parte hartzen dute, baina inon ez du ezer esaten jaioterriari buruz, bere herrian egun konkretu batean egin zuen ekintza bati buruz baizik. Adibide honek sailkatzailea nahastuko du, jaioterriari buruzkoak ez diren testuinguruak jaioterriei buruzkoak direla pentsatuz, eta ondorioz, erlazio berriak detektatzean betegarri okerrak itzuliz.

Sistemen eraginkortasuna eta kalitatea hobetzeko, teknika eta heuristiko berriak ikertu behar dira, aipamen zaratatsuak detektatu eta hauek filtratzeko.

3.4 Ondorioak

Kapitulu honetan, erlazioen erauzketarako urruneko gainbegiraketa sistema baten lehen urratsak eman ditugu. Aldez aurretik, ikasketa gainbegiratuan oinarritutako sistema bat garatu dugu, eskuz etiketatutako ACE 2005 corpusa erabiliz.

Ikasketa gainbegiratuko sistemak eta urruneko gainbegiraketa sistemak oso antzekoak dira elkarrekiko. Desberdintasun nabariena corpusa etiketatzeko modua da, lehenengoan etiketatze prozesua eskuzkoa da, bigarrenean berriz, automatikoa. Aipamenen aurreprozesaketa, ezaugarrien erauzketa, sailkatzailearen entrenamendua eta inferentzia prozesuak bateragarriak dira bi sistemetan.

Urruneko gainbegiraketa sistemari dagokionez, ezagutza baseen aberasketan oinarritzen den TAC 2011ko *Slot Filling* atazan parte hartu dugu. Sistema honetan, aipamenak erauzten ditugu corpus batetik. Aipamen hauek elkarrekiko erlazionatuta dauden entitate bikoteak dituzte, bien arteko erlazioa ezagutza base batean zehaztuta baitago. Aipamen guztiak eskuratu ondoren, ikasketa gainbegiratuko sisteman erabilitako ezaugarri, optimizazio teknika eta inferentzia metodo berdinak aplikatzen dizkiogu urruneko gainbegiraketa sistemari.

Lortutako emaitzak egokitzat hartzeko urruti daude. Gure errore analisiak erakusten du gure sistemak emaitzak txarrak ematearen arrazoietako bat aipamen negatiboen falta dela. Beste arrazoietako bat erlazio jakin batzuei buruzko aipamenen urritasuna da, erlazio hauen ikasketa zailduz.

Baina akatsen artean garrantzitsuena, gure datu multzoan topatutako aipamen zaratatsu kantitate handia da. Gure datu multzoko aipamen askotan

parte hartzen duten entitateen arteko erlazioak eta testuinguruak ez dute bat egiten. Adibide zaratatsuek ikasketa prozesua asko okertzen dute, sailkatzailea nahastuz, iragarpen okerrak eginez, eta ondorioz sistemaren eraginkortasuna kolokan jarriz. Eraginkortasuna hobetzeko, aipamen zaratatsuak kendu behar ditugu.

Ondorengo kapituluan aipamen zaratatsuak detektatu eta hauek ezabatzeko teknika desberdinak aztertzen ditugu. Artearen egoeran dagoen sistema batekin egiten ditugu esperimentuak, konparatu ahal izateko gure heuristikoak benetan eraginkorrak diren ala ez. Aukeratutako sistema Stanford Unibertsitateak garatutako urruneko gainbegiraketa sistema da, sistema honek Mintz *et al.*-en algoritmoa hobetzen du.

4

Aipamen zaratatsuen garbiketa

Kapitulu honetan, datu multzo bateko aipamen desberdinak aztertuko ditugu, aipamen zaratatsuak detektatuz eta hauen sailkapen bat eginez. Beste aipamen sorta bat hartuta, datu multzo berdinean dauden zarata mota desberdinen estimazio estatistikoak aterako ditugu. Ondoren zarata filtratzen duten heuristiko desberdinak azalduko ditugu, eta jarraian metodo hauek aplikatuz lortzen ditugun emaitzak eztabaidatu.

Esperimentuak artearen egoeran dagoen sistema batekin egingo ditugu, guk aurreko atalean garatutako sistema erabili ordez, heuristikoen eraginkortasuna hobeto neurtzeko. Aplikatutako heuristikoak tuplen maiztasunean, elkarrekiko informazio puntualean eta erlazioen zentroideetan oinarritzen dira. Heuristiko hauek informazio erauzketa ereduren emaitzak esanguratsuki hobetuko dituzte, eta erakutsiko dute aipamen zaratatsuak topatzea eta filtratzea beharrezkoa dela urruneko gainbegiraketa sistemen eraginkortasuna hobetzeko.

4.1 Sarrera

Komunitatea urruneko gainbegiraketan sortzen diren aipamen zaratatsuen arazoari aurre egiten saiatzen ari da era desberdinetan. Gure proposamena aipamen zaratatsuak ikasketa prozesuaren aurretik detektatzea da, eta ondoren datu multzotik ezabatzea. Horretarako heuristiko sinple batzuk garatu ditugu.

Urruneko gainbegiraketan, aipamen batean topatzen ditugun bi entitate ezagutza base batean erlazionatuta badaude, aipamen horri ezagutza basean adierazten den erlazioa jartzen diogu. Atal honetako esperimentuetan erabilitako ezagutza basea Freebase[1] da. Hona hemen tripleta batzuk, bi entitate eta hauen arteko erlazioarekin[2]:

- <Albert Einstein - *education* - University of Zurich>
- <Austria - *capital* - Vienna>
- <Steven Spielberg - *director_of* - Saving Private Ryan>

Urruneko gainbegiraketaren arabera, goiko edozein tripleta hartuta, esaldi batean bi entitateak agertzen badira, esaldi horren testuinguruak esplizituki bi entitateen arteko erlazioa adieraziko du. Hona hemen goiko tripleten adibide batzuk, non bi entitateen (letra lodiz) arteko erlazioak argiak diren:

- **Albert Einstein** was awarded a PhD by the **University of Zurich**.
- **Albert Einstein** earned a doctorate from the **University of Zurich**.
- **Vienna**, the capital of **Austria**.
- Dodging **Vienna's** tourist traps and making the most of the **Austrian** capital on a tight (...)
- Allison co-produced the Academy Award-winning **Saving Private Ryan**, directed by **Steven Spielberg** (...)
- When **Steven Spielberg** made **Saving Private Ryan** he aimed to portray (...)

Baina suposizio hau ez da beti zuzena, ondorengo adibideetan ez dago inongo erlaziorik bi entitateen artean, testuinguruaren arabera:

- **Albert Einstein** became an assistant professor at the **University of Zurich**.

[1] www.freebase.com

[2] Irakurketa errazteko, etiketa intuitiboak erabiliko ditugu Freebaseeko erlazio izenentzat, hala nola */education/education/student* ordez *education* etiketa, */location/country/capital* ordez *capital*, eta */film/director/film* ordez *film* etiketa.

- The city of **Vienna** is situated on the Danube in the eastern part of **Austria**.
- Mark Greczmiel, interviews **Steven Spielberg** shortly before the release of **Saving Private Ryan** in 1998.

Lehenengo esaldian, testuinguruak ez du ezer esaten Einsteinen hezkuntzari buruz, lan bizitzari buruz baizik. Bigarren esaldian, Viena hiria Austrian dagoela argi uzten da, baina hiri austriar bat dela dio, ez du hiriburua denik inon esaten. Hirugarren esaldian, ulertzen da Steven Spielberg jaunak *Saving Private Ryan* filmean parte hartu zuela, baina ez da bere eginkizuna zehazten, zuzendaria izan daiteke, aktore nagusia, edo filmaren ekoizlea.

Urruneko gainbegiraketak erlazio etiketa okerra duten 3 aipamen hauek erabiltzen ditu ikasketa prozesurako. Horrek sailkatzailea nahastea besterik ez du lortzen, inferentzia prozesuan erlazio eta betegarri okerrak iragarriz eta emaitza txarrak lortuz.

Aipamen zaratatsuak oso ugariak dira urruneko gainbegiraketa aplikatzen den sistemen datu multzoetan, eta garrantzitsua da aipamen zaratatsu hauek detektatzea eta ezabatzea. Ondorengo atalean datu multzo bateko aipamen asko eskuz aztertu eta hemen aurki genitzakeen zarata mota desberdinen sailkapen bat dugu.

4.2 Zarata azterketa

Riedel *et al.*en laneko datu multzoa (2010) da aipamen zaratatsuak aztertzeko aukeratu duguna. Horretarako auskazko hainbat aipamen jaso eta hauek banaka aztertzen ditugu. Ondorengoak dira aurkitu ditugun zarata mota desberdinak:

1. **Testuinguru okerra:** Zarata mota hau ohikoena da. Aurreko kapituluko errore analisian (3.3.5 atala) eta kapitulu honetako sarreran jarri ditugun adibideak mota honetakoak dira. Ezagutza basean bi entitate erlazionatuta badaude, hauek agertzen diren aipamen guztiak erlazio horrekin markatuta daude, baina badira aipamen asko non testuinguruaren arabera erlazio hori ez den gauzatzen.

   Zarata mota honen egoera bat datu multzoko ondorengo tripletan aurki dezakegu, <Jerrold Nadler - *born_in*[3] - Brooklyn>. Corpusetik ondo-

[3] /people/person/place_of_birth

rengo aipamenak eskuratu dira eta biei ezarri zaie *born_in* erlazioa, baina bi aipamenetatik batek bakarrik dauka zerikusia Nadlerren jaiotzaz:

- **Nadler** was born in **Brooklyn**, New York City.
- (...) Representative **Jerrold Nadler**, a Democrat who represents parts of Manhattan and **Brooklyn**, (...)

Lehenengo esaldiak argi adierazten du Nadler Brooklynen jaio zela; bigarrenak berriz, Manhattan eta Brooklyn ordezkatzen dituen politikaria dela. Sistemak bi aipamenak erlazionatutzat bezala hartzen ditu, eta ondorioz, bigarren esaldiarekiko antzekoak diren aipamenak ebaluatzean, hauei *born_in* etiketa jarri bi entitateen artean erlazio hau ez denean ematen.

2. **Ezagutza base osatu gabea:** Bi entitateren artean ez dagoenean ezagutza basean inongo erlaziorik zehaztuta, baina aipameneko testuinguruak erlazio jakin bat esplizituki aipatzen duenean gertatzen da fenomeno hau.

   Aztertutako adibideen artean, <Brazil - Celso Amorim> tupla topatu dugu datu multzoan. Bi entitateen artean inongo erlaziorik ez dagoela jartzen du, baina aipamena irakurrita, testuinguruak argi uzten du erlazio bat badagoela bi entitateen artean:

   - **Celso Amorim**, the **Brazilian** foreign minister, said the (...)

   Celso Amorim eta Brasil elkarrekiko erlazionatuta daude, bi entitate hauek dituen aipamen batek *country_minister* bezalako erlazio bat jaso beharko luke, baina datu multzoan kontrakoa adierazten da. Fenomeno hau zergatik gertatu den ulertzeko Freebase ezagutza baseko Celso Amorimen orrialdea[4] aztertu dugu, bertan politikaria dela espezifikatzen da, baina gobernuan eduki dituen karguei buruz ez du ezer esaten. Beraz, ezagutza base honen arabera, bi entitateen artean ez dago inongo erlaziorik. 4.1 irudiak Freebaseen agertzen den informazioa erakusten du.

[4] http://www.freebase.com/m/03xxdx

4.1 irudia – Celso Amorimi buruzko agerpena Freebaseen.

Zarata mota hau ere kaltegarria da sistemarentzat, erlazio esplizitu eta argia duten aipamenak jasotzean, hauek erlazionatu gabe daudela iragartzen duelako. Horrez gain, sailkatzailea nahastu besterik ez da egiten, antzeko bi aipamen jasotzen baditu non batean erlazio bat dagoen eta bestean ez.

3. **Erlazio anitza:** Kasu batzuetan, bi entitatek erlazio bat baino gehiago dute. Erlazio hauek "erlazio etiketa anitzak" (*multi-label*, Hoffmann *et al.* 2011) dira, datu multzoetan erlazio bat baino gehiago daukatelako markatuta. Kasu hauetan, urruneko gainbegiraketaren arabera, aipamen bakoitzeko testuinguruak bi erlazioak adierazten ditu aldi berean. Zoritxarrez, aipamenaren testuinguruari erlazio bakar bat dagokio, eta beste erlazioaren marka soberan dago.

   Jarri dezagun adibide bat egoera hobeto ulertzeko, adibide bezala, <Rupert Murdock - News Corporation> tupla. Bi entitate hauek erlazio bat baino gehiago dute elkarrekiko: *founder*[5] eta *top_member*[6]. Urruneko gainbegiraketaren oinarriak direla eta, ondorengo bi aipamenek bi erlazioak dituzte:

   - **News Corporation** was founded by **Rupert Murdock**.
   - **Rupert Murdock** is the CEO of **News Corporation**.

[5] /business/company/founders
[6] /business/asset_ownership/owner

Erlazio mota	Testuinguru okerra	Ezagutza base osatu gabea	Erlazio anitza
Erlazio bakarra	%28	-	-
Erlaziorik ez	-	%11	-
Erlazio anitza	%60	-	%15

4.1 taula – (Riedel *et al.* 2010) laneko datu multzoko aipamen zaratatsu mota desberdinen estimazioen taula.

Lehenengo esaldian ematen den erlazioa *founder* da, ez du inon esaten Murdock une honetan enpresaburua denik. Kontrako kasua gertatzen da bigarren esaldian, honen erlazioa *top_member* da, ez du inon esaten Murdock enpresaren sortzailea denik.

Horrelako kasuek ere ondorio larriak ekartzen dituzte sistemarentzat. Antzeko aipamenak jasotzean ebaluatzeko, erlazio bat baino gehiago daudela interpretatzen da, errealitatean bakarra dagoenean.

Zarata mota hauek urruneko gainbegiraketa sistemen eraginkortasuna jaisten dute. Sortzen diren datu multzoak hain handiak izanik, ezin ditugu aipamen zaratatsu guztiak eskuz, banaka ezabatu.

Jakiteko aipamen zaratatsuen hedapena noraino iristen zen, Riedelen datu multzoko hainbat aipamen berriro aztertu ditugu eskuz. Analisirako, erlazio bakarra duten auskazko 100 aipamen eskuratu ditugu, erlaziorik gabeko beste 100 aipamen, eta erlazio etiketa anitzak diren beste 100 aipamen. Aipamen bakoitza banaka analizatzen dugu, eta zaratatsuak zarata motaren arabera sailkatu.

4.1 taulak analisiaren emaitza laburtzen du. Erlazio bakarra duten aipamenetatik, %28ak testuinguru okerra dauka, hau da, aipamenaren testuinguruak ez dauka markatutako erlazioarekin inongo zerikusirik. Erlaziorik ez duten aipamenentzat, %11 dira testuinguruaren arabera erlazio bat adierazten dutenak, baina ezagutza basean erlazio hori ez dagoenez erlaziorik markatuta ez dutenak. %60 dira erlazio anitzetan testuinguru okerra dutenak, tupla bakoitzaren erlazio bakar batek ere ez duelako testuinguruarekin bat egiten; %15ean bakarrik topatzen dugu testuinguruarekin bat egiten duen erlazio bat tupla bakoitzeko, baina beste erlazioek ez.

Datu multzoak 91373 aipamen negatibo ditu, etiketa bakarreko 2330 aipamen, eta etiketa anitzeko 26587 aipamen. Zarata analisiaren ondoren, erlazio

bat edo gehiago adierazten duten aipamenen %29a bakarrik garbiak direla estimatzen dugu. Aipamen negatiboak ere kontuan hartuta, %74 aipamen garbi daudela estimatzen da.

Erlazio bat adierazten duten adibide garbien kantitatea oso txikia da, laurdena baino pixkat gehiago, ez dira kasu bereziak. Hainbeste adibide zaratatsurekin, normala da sistemen eraginkortasuna kolokan jartzea eta emaitza txarrak lortzea, horregatik lortu behar dugu ahalik eta aipamen gehien filtratzea, sailkatzailea ez dadin nahastu eta ahalik eta iragarpen gehien zuzenak izateko.

4.3 Esperimentazio ingurunea

Zarata garbiketarako proposatuko ditugun metodoak sistemekiko independenteak dira, hau da, urruneko gainbegiraketa bidez sortutako edozein erlazio erauzketa sisteman aplikatu daiteke, behar ditugun gauza bakarrak entrenamenduko datu multzo bat eta bertan erabilitako tuplak dira.

Esperimentuak egiteko erabilitako corpusa Riedel *et al.*ek sortu zutena da, corpus hau beste hainbat ikertzailek erabili dute: Hoffmann *et al.*-ek 2011n, eta Surdeanu *et al.*-ek 2012an adibidez. Corpus honek Freebase ezagutza basea erabiltzen du urruneko gainbegiraketa iturri bezala, eta Sandhausen lanerako (2008) sortutako corpusa, New York Times-en (NYT) oinarritua (ikus 2.1.2 atala).

Datu multzoa bi zatitan banatuta badago: entrenamendu zatia eta testeko zatia. 4.2 taulan zati bakoitzak dituen tupla kantitatea eta aipamen kantitatea zehazten dugu, tupla eta aipamen positibo eta negatiboak bereiztuz. Azken bi lerroek partizio bakoitzean dauden tupla eta aipamen kopuru totala erakusten dute.

Hasierako negatiboen %5 bakarrik argitaratu zuten corpusaren egileek. Honek aipamen positiboen hirukoitza esan nahi du. Proportzio desorekatu hau bat dator aurreko ataleko ondorioekin: adibide negatiboen erabilera beharrezkoa da sistemaren emaitzak hobetzeko.

E eranskinean datu multzo hau osatzen duten erlazioen zerrenda aurki dezake irakurleak. Datu multzo honi buruz gehiago jakin nahi izanez gero, ikus artearen egoerako 2.1.2 atala.

Datu multzoa		Entrenamendua	Testa
Positiboak	**Tuplak**	4700	1950
	Aipamenak	34811	6444
Negatiboak	**Tuplak**	63596	94917
	Aipamenak	91373	166004
Guztira	**Tuplak**	67946	96678
	Aipamenak	120290	171360

4.2 taula – Datu multzoaren estatistikak.

4.3.1 Datu multzoko ezaugarriak

Ikasketa prozesurako entrenatu genituen ezaugarriak ez genituen guk prozesatu, aurretik Riedel eta bere taldeak erauzitakoak erabili ditugu. 4.3 taulan datu multzoko tripleta bat, honen aipamen bat eta aipamenaren ezaugarriak daude. *sourceGuid* eta *destGuid* erlazioan parte hartzen duten entitateak dira, hauek Freebaseeko identifikatzailearekin agertzen dira[7]; *sourceGuid* entitate nagusia da, eta *destGuid* betegarriaren papera egiten duena. *relType* bi entitateen arteko erlazioa da.

Datu multzoan tripleta honen 5 aipamen daude, taula horretan horietako aipamen bat jarri dugu eta honen ezaugarriak erakutsi. Aipamena *sentence* lerroan dago, eta NYT corpuseko *filename* fitxategitik eskuratu da. *sourceId* eta *destId* atributuek bi entitateekin zerikusia dute, baina ez dakigu zein den beraien egitekoa ezaugarri hauetan, metadatuetako informazioa delakoan gaude, Riedel *et al.* 2010 artikuluan ez omen dituzte ezaugarriak ondo deskribatzen.

Ikasketa prozesuan erabiltzen diren ezaugarri desberdinak *feature* lerroetan daude. Ezaugarri garrantzitsuenak deskribatuko ditugu orain:

- Bi entitateen izenak adierazten dira.
- *inverse_true* ezaugarriak betegarria aipamenean entitate nagusiaren aurretik datorrela adierazten du, kontrako kasuan *inverse_false* adierazten da.
- Bi entitateen artean dauden hitzen ordez, *LONG* agertzen da, hitz gehiegi daudelako bien artean; gutxi daudenean hitz formak jartzen

[7]Identifikatzaile hauek zaharkituta daude, gaur egun beste batzuk erabiltzen dira.

sourceGuid	/guid/9202a8c04000641f8000000006e70346
destGuid	/guid/9202a8c04000641f800000000036f9b
relType	/broadcast/content/location
mention	
filename	/m/vinci8/data1/riedel/projects/relation/kb/nyt1/docstore/nyt-2005-2006-joint/1692706.xml.pb
sourceId	-2147483591
destId	-2147483588
feature	ORGANIZATION->LOCATION
feature	inverse_true|LOCATION|*LONG*|ORGANIZATION
feature	inverse_true|of|LOCATION|*LONG*|ORGANIZATION|meant
feature	inverse_true|Spirits of|LOCATION|*LONG*|ORGANIZATION|meant to
feature	inverse_true|LOCATION|IN DT NNP NNP NNP "CC PRP VBP TO VB VBN IN NNP NNP DT JJ NNS TO VB WP|ORGANIZATION
feature	inverse_true|of|LOCATION|IN DT NNP NNP NNP "CC PRP VBP TO VB VBN IN NNP NNP DT JJ NNS TO VB WP|ORGANIZATION|meant
feature	inverse_true|Spirits of|LOCATION|IN DT NNP NNP NNP "CC PRP VBP TO VB VBN IN NNP NNP DT JJ NNS TO VB WP|ORGANIZATION|meant to
feature	str:ORGANIZATION|[SBJ]->meant[OBJ]->know[ADV]->been[VC]->have[OBJ]->have[NMOD] ->announcer[NULL]<-for[NMOD]<-Spirits[PMOD]<--of[NMOD]<-|LOCATION|[NMOD]->St.
feature	dep:ORGANIZATION|[SBJ]->[OBJ]->[ADV]->[VC]->[OBJ]->[NMOD]->[NULL]<-[NMOD]<-[PMOD]<-[NMOD]<-|LOCATION|[NMOD]-> dir:ORGANIZATION|->->->->->-<-<-<-<-|LOCATION|->
feature	str:ORGANIZATION|[SBJ]->meant[OBJ]->know[ADV]->been[VC]->have[OBJ] ->have[NMOD]->announcer[NULL]<-for[NMOD]<-Spirits[PMOD]<--of[NMOD]<-|LOCATION|[NMOD]->of
feature	dep:ORGANIZATION|[SBJ]->[OBJ]->[ADV]->[VC]->[OBJ]->[NMOD]->[NULL]<-[NMOD]<-[PMOD]<-[NMOD]<-|LOCATION|[NMOD]-> dir:ORGANIZATION|->->->->->-<-<-<-<-|LOCATION|->
feature	str:ORGANIZATION|[SBJ]->meant[OBJ]->know[ADV]->been[VC]->have[OBJ]- >have[NMOD]->announcer[NULL]<-for[NMOD]<-Spirits[PMOD]<--of[NMOD]<-|LOCATION
sentence	From a business standpoint , the Cardinals did the right thing , "said Bob Costas , who worked for KMOX from 1974 to 1992 , first as the announcer for the Spirits of St. Louis of the American Basketball Association "But you have to have been around St. Louis these many years to know what KMOX meant to the city , not just the Cardinals .

4.3 taula – Riedel datu multzoko aipamen bat eta honen ezaugarriak.

dituzte. Ezaugarri hauetaz gain, beste ezaugarri batzuetan tarteko hitz hauen kategoria gramatikalak agertzen dira.

- Ezaugarriek entitateen aurretik eta atzetik datozen hitzekin konbinaketa desberdinak egiten dituzte, lehenengo inguruko hitzik gabe, ondoren inguruko hitz bat (*of, meant*), eta azkenik inguruko bi hitzak (*Spirits of, meant to*).

- Gainontzekok ezaugarriek hitzen arteko dependentzia bidezko erlazio sintaktikoak adierazten dute. Bi entitateen dependentzia bidea erakusten dute.

4.4 Zarata garbiketarako metodo desberdinak

Datu multzotik aipamen zaratatsuak ezabatzeko hiru heuristiko desberdin garatu ditugu. Tuplak kentzen ditugu hauen aipamen kantitatearen arabera, hauen elkarrekiko informazio puntualaren arabera, eta erlazio aipamen guztien zentroideen eta aipamen bakoitzaren antzekotasunaren arabera. Heuristiko hauek testuinguru okerra eta erlazio anitzeko zarata motak antzemateko aplikatzen dira.

Heuristiko bakoitzaren parametro optimoak lortu ondoren, garbiketa hobetzeko, hiru metodoen arteko konbinaketa desberdinak egiten ditugu, zarata garbiketa are gehiago hobetzeko.

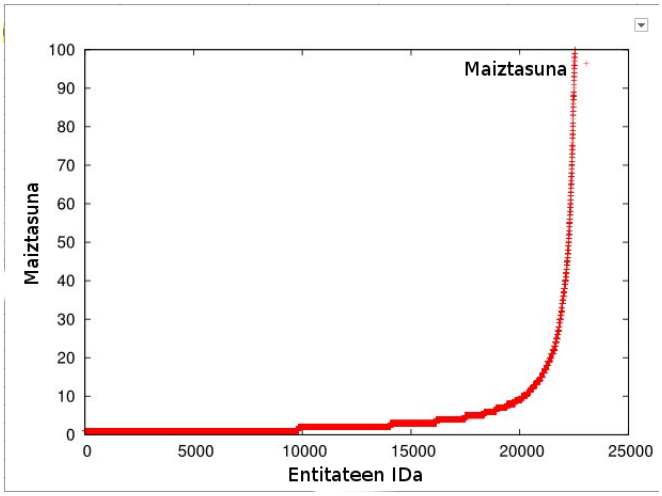
Metodo hauek datu multzoarekiko eta sistemarekiko independenteak dira, eta ez da inongo eskuzko etiketaziorik erabiltzen.

4.4.1 Tuplen maiztasuna

Lehenengo heuristikoan, aipamen gehiegi dituzten tuplek aipamen zaratatsu asko edukitzeko aukera gehien dituztela uste dugu, beraz aipamen kantitate finko bat baino gehiago dituzten tuplak ezabatzen ditugu. Heuristiko honetan, tupla positiboak eta negatiboak ezabatzen dira, guk ezarritako aipamen kopuru maximo bat baino gehiago baldin badituzte. Heuristiko honek **testuinguru okerra** duen aipamen zaratatsuetan jartzen du arreta.

4.2 irudian garapeneko datu multzoko tupla eta aipamen kopuruen grafikoa daukagu[8]. Grafiko honetan ikusten da orokorrean tuplek bat eta hogei

[8]Grafiko hau ez dago guztiz osatuta, 100 aipamen baino gutxiago dituzten tuplak baka-



4.2 irudia – Tupla eta aipamen kopuruen grafikoa, 100 aipamen arte. Garapeneko datu multzoa.

aipamen inguru dituztela, hemendik aurrera aipamen gehiago edukitzea ez dirudi oso ohikoa. Hala ere, badira hainbat tupla aipamen kopuru oso altuekin, 200 aipamen, 500, 1000 eta 2000 ingurukoak. Kasu hauek oso bereziak dira, eta 4.2 irudiko grafikoan jartzen baditugu, hau zeharo distortsionatzen da, errealitatearen irudi okerra erakutsiz.

Tupla hauetako batzuk hartu eta hauen aipamen batzuk aztertu ditugu. Aipamen asko garbiak dira, hau da, testuinguruak markatutako erlazioa irudikatzen dute, baina badira beste hainbat aipamen non kontrakoa gertatzen den, markatutako erlazioa ez da testuinguruan ematen. Orokorrean aipamen zaratatsu gehiago aurkitzen ditugu garbiak baino.

Aipamen gutxiago dituzten tupla batzuen aipamenak aztertuta, egoera desberdina da: baziren tuplak aipamen garbiekin bakarrik, aipamen gehienak garbiak, aipamen gehienak zaratatsuak, edo aipamen guztiak zaratatsuak.

Analisi guzti horietatik, ondoko hipotesia egiten dugu: zenbat eta aipamen gehiago eduki tupla batek, orduan eta aipamen zaratatsu gehiago sortzeko aukerak ditu.

rrik hartu ditugu kontuan. Gutxi batzuk izan dira kanpoan gelditu direnak, errealitatearen irudikapena errazteko helburuz.

Garapen fasean[9] aipamen kopuru maximo desberdinekin egiten ditugu esperimentuak. Aukeratutako aipamen kopuru maximoen balioa 90 da, F1 neurriaren balio altuena eskuratzen dugulako, beraz 90 aipamen baino gehiago dituzten tupla guztiak ezabatzen dira.

Ezabatutako tupla kantitatea positiboen %40 inguru da, eta negatiboen %15 inguru datu multzo osotik. Honek esan nahi du tupla positibo askok aipamen kantitate handia dutela negatiboekin konparatuta, negatiboen ia gehienek aipamen gutxi dituztenean. Guztira, datu multzo osoko tuplen %17 inguru ezabatzen dira.

Hona hemen ezabatutako tupla baten adibidea: <European Union - *has_location*[10] - Brussels>. Tupla honek 95 aipamen ditu, haietako batzuk aipamen garbiak dira, adibidez:

- The **European Union** is headquartered in **Brussels**.

Baina badira aipamen zaratatsuak ere:

- The **European Union** foreign policy chief, Javier Solana, said Monday in **Brussels** that (...)
- At an emergency **European Union** meeting of interior and justice ministers in **Brussels** on Wednesday, (...)

Bi adibide hauetan inon ez du esplizituki aipatzen Brusela Europako Batasunean dagoenik, eta ondorioz erlazio erauzketa sistema gainbegiratua okertu daiteke. Lehenengo adibidean Europar Batasuneko buruetako batek, Solanak hitzaldi bat Bruselan eman zuela aipatzen du; bigarrenak EBko ministro desberdinek Bruselan bilera bat egin zutela. Heuristiko honek tupla honen aipamen guztiak ezabatzen ditu.

Hemendik aurrera, laburtzeko, kapituluan *MF* deituko diogu heuristiko honi.

4.4.2 Elkarrekiko informazio puntuala

Bigarren heuristikoak entitate tuplen eta erlazio etiketaren arteko elkarrekiko informazio puntualaren (*Pointwise Mutual Information, PMI*) balioa

[9] Atal honetan, garapeneko esperimentuak entrenamendu partizioari 3 zatitako egiaztapen gurutzatua aplikatzen da.

[10] /location/location/contains

kalkulatzen du. Elkarrekiko informazio puntualak (EIP) bi entitate independenteren artean, hauen agerpena datu multzoan informatiboa den edo ez jakin dezakegu. Heuristiko honek ere **testuinguru okerra** duen aipamen zaratatsuetan jartzen du arreta.

Gure hipotesiaren arabera, entitate bikote bat askotan azaltzen bada aipamen desberdinetan, eta gutxiagotan independenteki, hauek korrelatuta daude ziurrenik. Bestalde, entitate tupla baten agerpen maiztasuna, bi entitateak indibidualki agertzeko duten maiztasunarekin konparatzen badugu, eta hau oso baxua bada, orduan aipamen horiek zaratatsuak dira.

Tupla bakoitzaren EIP balioa lortu ondoren, hau atalase baten azpitik baldin badago, tupla honek aipamen zaratatsuak dituela kontsideratzen dugu, eta tupla hauek kendu.

Enpirikoki, gure sistema askoz eraginkorragoa da EIP baxua duten tupla positiboak bakarrik kentzen baditugu, eta negatibo guztiak mantentzen baditugu, kontuan hartu gabe hauen EIP balioa.

Gure sistemak eraginkortasun onena EIPrako 2.3ko atalasea erabilita lortzen du, hau baino baxuagoa duten tuplak eta hauen aipamen guztiak kenduta. Guztira tupla positiboen %8 inguru ezabatzen dira, datu multzo osoaren %1 inguru.

Heuristiko hau Min *et al.*-en lanekoan (2012) inspiratuta dago eta erabilitako formula ondorengoa da, non $entitatea_1$ entitate nagusia den eta $entitatea_2$ entitate nagusiaren erlazioaren betegarria:

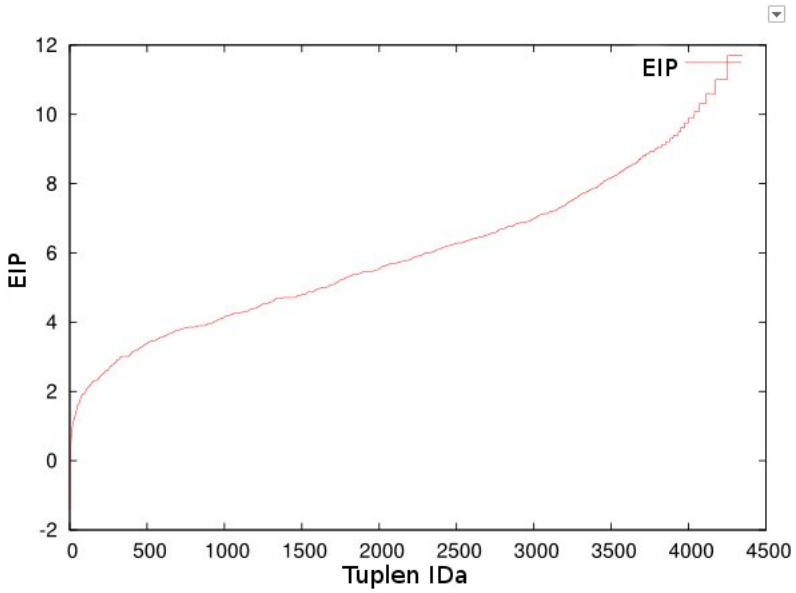
$$eip(entitatea_1, entitatea_2) = log \frac{p(entitatea_1, entitatea_2)}{p(entitatea_1)p(entitatea_2)} \tag{4.1}$$

$p(entitatea_X)$ entitate batek datu multzoan agertzeko duen probabilitatea da, eta $p(entitatea_X, entitatea_Y)$ bi entitate aipamen berean agertzeko duten probabilitatea.

4.3 irudiko grafikoa garapeneko datu multzoko tupla positibo guztien elkarrekiko informazio puntuala lortuz sortu dugu. X ardatzeko hasierako tuplek dute elkarrekiko informazio puntual txikiena eta ondorioz datu multzotik ezabatzen dira.

Hurbilketa honek ondorengo tupla ezabatzen du: <Natasha Richardson - *place_of_death*[11] - Manhattan>. Tupla honek aipamen bakarra dauka:

[11] /people/deceased_person/place_of_death



4.3 irudia – Tupla positiboen elkarrekiko informazio puntualaren grafikoa. Garapeneko datu multzoa.

- (...) **Natasha Richardson** will read from 'Anna Christie,' (...) at a dinner at the Yale Club in **Manhattan** on Monday night.

Aipamen honek ez du inola ere *place_of_death* erlazioa erakusten, hau da, nahiz eta Natasha Richardson Manhattanen hil, aipamen hau ez dago gertaera horrekin erlazionatuta, aktoresak Manhattanen izandako afari batetaz ari delako.

Hemendik aurrera, laburtzeko berriz, kapituluan *PMI* deituko diogu heuristiko honi.

4.4.3 Aipamenen zentroideak

Heuristiko honetan, etiketa berdina duten aipamen guztien zentroideak kalkulatzen ditugu, eta zentroidearekiko antzekoenak diren aipamenekin gelditu. Gure hipotesiaren arabera, zentroidetik urrutien dauden aipamenak dira zaratatsuenak.

Esperimentu honetarako, aipamen bakoitza bektoretzat hartzen dugu, eta bektore honen ezaugarriak espazioaren dimentsioak osatzen dituzte. Urruneko gainbegiraketa sistemak erabiltzen dituen ezaugarri berdinak erabiltzen

ditugu bektoreak sortzeko, maiztasuna erabiliz ezaugarriaren balio gisa. Ondorengo ekuazioan ikus dezakezue nola sortzen dugun zentroidea:

$$\vec{Z_i} = \left(\frac{ezaug_1}{aipamenak_i}, \frac{ezaug_2}{aipamenak_i}, \ldots, \frac{ezaug_N}{aipamenak_i} \right) \tag{4.2}$$

i erlazioaren identifikatzailea da, $aipamenak_i$ erlazio hori duten aipamen kopurua, $ezaug_j$ zenbat aldiz j ezaugarria agertu den adierazten du, eta azkenik Z_i erlazioaren zentroidea da.

Zentroide bat eta beste edozein aipamenen arteko antzekotasuna kosinua erabiliz kalkulatzen dugu (Z zentroidea, A aipamena):

$$kosinua(Z, A) = \frac{\vec{Z} \cdot \vec{A}}{\sqrt{\vec{Z} \cdot \vec{Z}} \cdot \sqrt{\vec{A} \cdot \vec{A}}} \tag{4.3}$$

4.4 irudian dugu heuristiko honen errepresentazio bat. Etiketa berdineko aipamen guztietatik, zentroidetik gertuen daudenen ehuneko kantitate bat hartzen dugu, beste guztiak kenduz.

Garapen fasean, gure sistemak erlazio guztietarako antzekoenak ziren aipamenen %90a mantenduz lortzen ditu emaitza onenak. Heuristiko hau ez dugu erlazionatu gabeko aipamenetan aplikatzen, hauek mantenduta emaitzak askoz hobeak direlako kenduta baino.

Guztira, tupla positiboen %10a ezabatzen dira, datu multzo osotik %1 inguru.

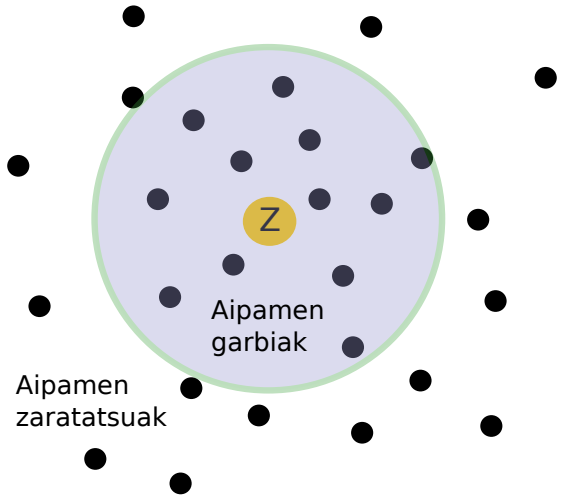
Adibide bezala, har dezagun *company_founders* erlazioa[12]. Zentroidea lortu ondoren, ondorengoa da antzekotasun handiena duen aipamena, <Steve Case - AOL> tuplarena:

- (...) its majority shareholder is **Steve Case**, the founder of **AOL**.

Kanpoan gelditu diren aipamenak aztertuta, <Thom Mayne - Morphosis> tuplaren ondorengo aipamena da horietako bat:

- Ms. Tsien and Mr. Williams were chosen after a competition that began with 24 teams of architects and was narrowed to two finalists, **Thom Mayne's Morphosis** being the other.

[12] /business/company/founders



4.4 irudia – Erlazio berdineko aipamenak eta hauen zentroidea grafikoki errepresentatuta. Puntu beltzak aipamenak dira, eta erdiko puntua (Z) aipamenen zentroidea.

Aipamen honek ez du inon esplizituki esaten Thom Mayne Morphosisen sortzailea denik, jabea da bai, baina jabea izateak ez du zertan esan nahi berak sortu zuenik.

Aurretik aipatu ditugun bi heuristikoek, tupla baten aipamen guztiak ezabatzen dituzte datu multzotik, aipamen guztiak zaratatsuak izango balira bezala. Heuristiko honek tupla baten aipamenak bitan banatzen ditu, alde batean zaratatsuak bezala detektatu direnak eta beste aldean aipamen garbiak, eta zaratatsuak bakarrik ezabatu, tupla horren aipamen guztiak ezabatu ordez.

Heuristiko honek abantaila bat dauka beste bi metodoekiko: erlazio bat baino gehiago duten tuplekin lan egin dezakegu, erlazio bakoitza indibidualki tratatuz, ondorioz, jatorriz erlazio anitza den aipamen bat erlazio bati atxikituta geratuko da kasu onenean. Kasu okerrenean, bi aipamenak bi erlazioetan, atalasearen mugaren barnean gelditzen badira, erlazio anitz bezala mantenduko dira.

Hemendik aurrera atal honetan MC deituko diogu heuristiko honi, laburtzeko baita ere.

4.4.4 Heuristikoen arteko konbinaketa ereduak

Aurreko hiru heuristikoak konbinatzen dituen eredu desberdinekin hainbat froga egiten ditugu, konbinaketa hauek edozein ordenetan eginez. Garapen fasean, 4.5.1 atalean erakutsiko dugun bezala, desberdinak dira *Mintz** eta *Mintz++* algoritmoetarako (ikus 2.3.1 atala).

Lehenengorako (*Mintz**), ondorengo ordena da egokiena: PMI heuristikoarekin hasten gara garbiketa egiten, ondoren MF heuristikoa aplikatuz, eta amaitzeko gelditzen diren aipamenekin zentroideetan oinarritutako heuristikoa. Bigarrenerako (*Mintz++*), konbinaketa onena PMI aplikatuz eta ondoren MF aplikatuz lortzen dugu. Azken honetan, zentroideen metodoa ondoren aplikatzeak ez du inongo hobekuntzarik ekarri.

4.5 Emaitzak

Esan bezala, Riedel eta bere kideek sortutako datu multzoa erabiltzen dugu esperimentuak egiteko. Guk sortutako ezaugarriak erabili ordez, beraiek sortutako berdinak erabiltzen ditugu aurreko atalean erabilitako heuristikoentzako eta sailkatzaileak entrenatzeko.

Garapenerako datu multzoa 3 zatitako egiaztapen gurutzatua erabiliz sortu da, Surdeanu *et al.*-en lanean bezala, erabili zuten sistema berdinarekin (ikus 2.3.1 atala). Testeko esperimentuak egiteko, konbinaketa eredu onenak erabiltzen dira, heuristiko bakoitza bakarka aplikatu ordez. Heuristikoak aplikatutako sistema eta jatorrizkoaren arteko konparaketa doitasun/estaldura kurben bidez egiten da.

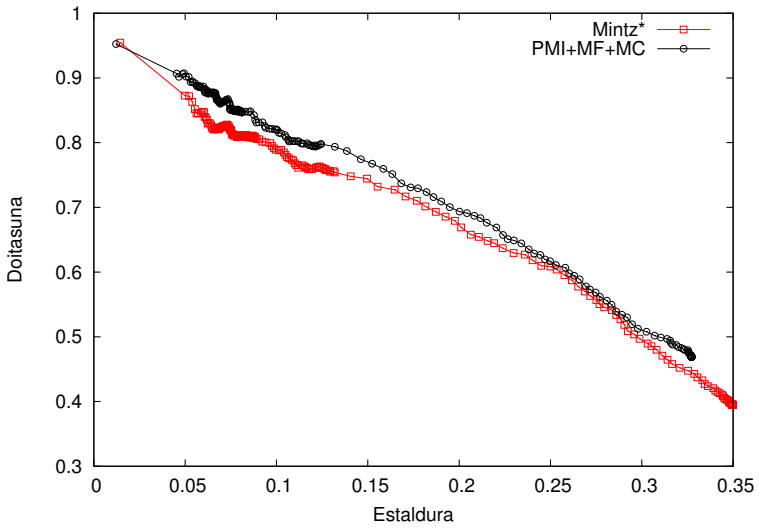
Esangura estatistikoak neurtzeko, Berg-Kirkpatrick *et al.*-ek proposatutako metodoak (2012) erabiltzen ditugu, egiaztatuz aipamenen filtraziorako heuristikoek emandako hobekuntzak esanguratsuak direla. Metodo hauek ondorioztatzen dute esperimentu guztien emaitzak estatistikoki esanguratsuak direla.

4.5.1 Garapen fasearen emaitzak

Hasierako esperimentu guztiak Surdeanu eta bere kideen laneko *Mintz** sistema erabiliz egiten dira, sailkatzailea eredu bakar batekin entrenatuta.

Heuristikoa	Doitasuna	Estaldura	F1
Mintz*	39.44	34.98	37.07
MF 90	44.49	33.19	**38.01** †
PMI 2.3	40.64	34.49	37.31 †
MC 90%	40.31	34.81	37.33 †
PMI+MF+MC	46.36	32.72	**38.53** †

4.4 taula – Garapen faseko esperimentuak *Mintz** erabiliz, heuristiko bakoitzaren emaitzekin eta konbinaketa onenarekin. † agertzen diren emaitzek *Mintz** sistemarekiko hobekuntzak estatistikoki esanguratsuak direla adierazten dute ($p < 0.05$).



4.5 irudia – *Mintz** sistemaren doitasun/estaldura kurbak. Garapen fasea. Lerro beltza gure filtrazio eredua da.

Mintz* entrenamendua

4.4 taulan heuristiko bakoitzarekin lortutako emaitzak ditugu. Metodo bakoitza banaka exekutatuz gero, emaitza onenak *MF* heuristikoarekin lortzen dira (4.4.1 atala), non gure sistemaren F1 neurriak %1eko igoera duen. Beste bi heuristikoek berriz, *PMI*k (4.4.2 atala) eta *MC*k (4.4.3 atala), oinarri lerroarekiko hobekuntza txikiak dituzte.

Konbinaketa ereduak erabilita, emaitza onenak *PMI* heuristikoa *MF*rekin batera konbinatzean lortzen dugu, eta ondoren *MC* aplikatuz, ia %1.5 puntuko hobekuntza hobetuz, eredu honi *PMI+MF+MC* deitu diogu. Gure sistemak doitasuna asko hobetzen du esperimentu bakoitzean, ia %7ko igoerarekin; estaldura berriz, kasu guztietan jaisten da, jaitsiera gehiena konbinaketa ereduan izan du, hau %2a baino gehiagokoa izan da. Hau esperogarria da, aipamen zaratatsuak filtratu ondoren, entrenamenduko datu multzoak adibide positibo gutxiago dituelako, ondorioz sailkapen eredua askoz kontserbadoreagoa bihurtzen da erlazio etiketak iragartzean. Taulan ikus daitekeen bezala, emaitzak guztiak oinarri lerroarekiko esanguratsuak izan dira.

4.5 irudian garapen faseko oinarri lerroaren eta garbiketa sistema onenaren arteko doitasun/estaldura kurbak dauzkagu. Irudiak argi uzten du gure hurbilketa jatorrizko sistemarena baino eraginkorragoa dela, maila guztietan oinarri lerrotik gorago dago.

Mintz++ entrenamendua

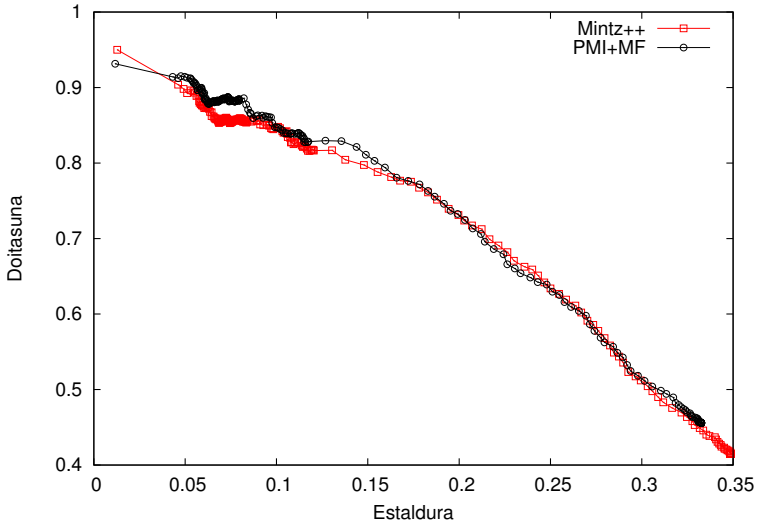
Heuristiko berdinak aplikatzen ditugu *Mintz++* sisteman eta hauek optimizatu. Parametro onenak ondorengoak dira: *MF*rentzat 180 aipamen gehienez eta *PMI*rentzat gutxienezko atalasea 2.4. Zoritxarrez, *MC* heuristikoak ez du inongo hobekuntzarik ekarri oraingoan. Amaitzeko, *PMI* heuristikoa *MF* heuristikoarekin batera konbinatzen dugu emaitzak hobetzeko.

4.5 taulan ikus ditzakegu emaitza guztiak. *MF* heuristikoak eta *PMI+MF* konbinaketak itzulitako F1 neurriaren balioa berdina da, hala ere *PMI+MF* sistemak besteak baino doitasun hobea dauka, %4ko igoerarekin oinarri lerroarekiko (*Mintz++*) eta *MF* heuristikoarekiko %1eko aldearekin. Bi sistemek %1 baino gehiagoko estaldura jaitsiera izan dute, baina bi sistemen arteko estaldura desberdintasuna, %0.4, ez da doitasunarena bezain handia. Doitasunaren aldetik dagoen hobekuntza handia dela eta, konbinaketa eredua hobea dela erabaki dugu eta ondorioz, testeko ebaluaketa eredu hau erabilita egin. Taulan ikus daitekeen bezala, emaitza guztiak oinarri lerroarekiko esanguratsuak izan dira oraingoan ere.

4.6 irudiak *Mintz++* sisteman oinarritutako doitasun/estaldura kurbak ditu, oinarri lerroarenak eta filtratze sistema onenarena. Filtratutako datuak entrenatu dituen ereduak orokorrean sistema originalak baino eraginkortasun hobea dauka, baina desberdintasunak ez dira aurreko ereduarenak bezain adierazgarriak. Honek adierazten du banakako bost ereduen konbinaketa ereduek, hala nola *Mintz++* ereduan inplementatutakoa, urruneko gainbe-

Heuristikoa	Doitasuna	Estaldura	F1
Mintz++	41.45	34.85	37.86
MF 180	44.48	33.65	**38.45** †
PMI 2.4	42.97	34.00	37.95 †
PMI+MF	**45.57**	33.25	**38.45** †

4.5 taula – Garapen faseko esperimentuak *Mintz++* erabiliz, heuristiko bakoitzaren emaitzekin eta konbinaketa onenarekin. † agertzen diren emaitzek *Mintz++* sistemarekiko hobekuntzak estatistikoki esanguratsuak direla adierazten dute ($p < 0.05$).

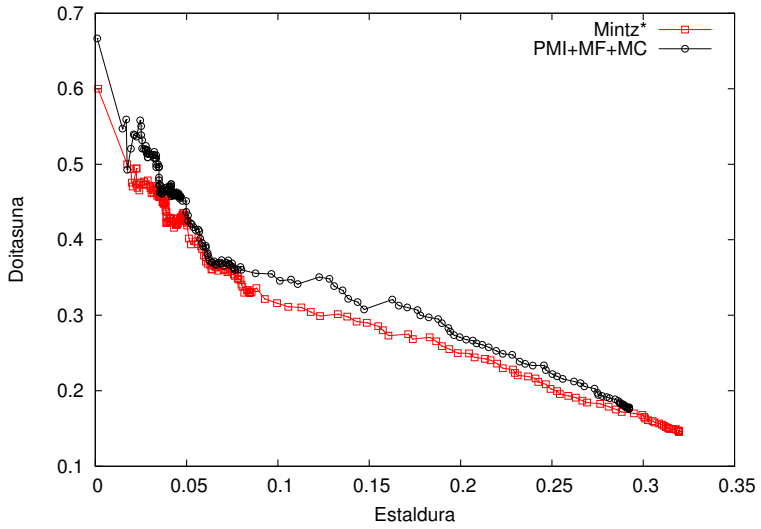


4.6 irudia – *Mintz++* sistemaren doitasun/estaldura kurbak. Garapen fasea. Lerro beltza gure filtrazio eredua da.

giraketak sortutako zarataz errekuperatzeko gai direla. Hala ere, estrategia hauek garatzeko oso garestiak dira, gure heuristiko sinpleekin konparatuta, non datuak txanda bakar batean filtratzen ditugun.

Heuristikoa	Doitasuna	Estaldura	F1
Mintz*	14.57	31.95	20.02
PMI+MF+MC	17.64	29.23	**22.00** †

4.6 taula – Test faseko esperimentuak *Mintz** erabiliz, garapen fasean lortutako konbinaketa onenarekin. † ikurrak hobekuntza *Mintz** sistemarekiko estatistiko esanguratsua dela adierazten du ($p < 0.05$).



4.7 irudia – *Mintz** sistemaren doitasun/estaldura kurbak. Test fasea. Lerro beltza gure filtrazio eredua da.

4.5.2 Testeko emaitzak

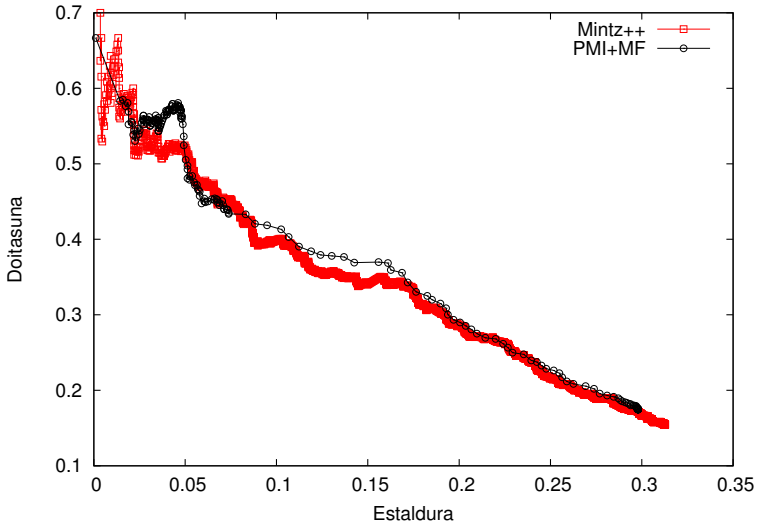
Testeko datu multzoa ebaluatzeko, garapen faseko konbinaketa ereduak erabiltzen ditugu bakarrik *Mintz** eta *Mintz++* ereduetarako, garapen fasean eredu bakoitzerako lortutako parametro optimo berdinekin.

Mintz* ebaluazioa

*Mintz** ereduko testeko emaitzak 4.6 taulan ditugu ikusgai. Konbinaketa ereduaren F1 neurriak 2 puntu inguruko igoera bat dauka oinarri lerroarekiko, eta %3ko igoera doitasunean. Garapen fasean bezala, estaldura jaitsi

Heuristikoa	Doitasuna	Estaldura	F1
Mintz++	15.43	31.28	20.67
PMI+MF	17.48	29.79	**22.03** †

4.7 taula – Test faseko esperimentuak *Mintz++* erabiliz, garapen fasean lortutako konbinaketa onenarekin. † ikurrak hobekuntza *Mintz++* sistemarekiko estatistiko esanguratsua dela adierazten du ($p < 0.05$).



4.8 irudia – *Mintz++* sistemaren doitasun/estaldura kurbak. Test fasea. Lerro beltza gure filtrazio eredua da.

da, %2.5 puntu baino gehiagokoa izan da jaitsiera. Taulan ikusten den bezala, testeko fasean heuristikoen konbinaketaren emaitzak oinarri lerroarekiko esanguratsuak izaten jarraitzen dute.

4.7 irudian bi sistemen arteko doitasun/estaldura kurbak dauzkagu. Garapen fasean bezala, oraingoan ere gure filtrazio sistemaren kurba maila guztietan oinarri lerrotik gorago dago.

Mintz++ ebaluazioa

4.7 taulan, garapen fasean *Mintz++* sistemarentzat lortutako filtrazio eredu onenaren emaitzak dauzkagu test partiziorako. Kasu honetan, garapen

fasean gertatu den bezala, igoera ez da horren nabaria, %1.3koa bakarrik. Doitasunaren igoera %2koa da, eta estaldura orain ere jaitsi da, baina jaitsiera txikiagoa izan da, %1.5 bakarrik.

Beste behin ere, taulak argi uzten du testeko fasean heuristikoen konbinaketaren emaitzak oinarri lerroarekiko esanguratsuak izan direla.

4.8 irudiak test faseko *Mintz++* sisteman oinarritutako doitasun/estaldura kurbak ditu. Garapen fasean bezala, desberdintasunak ez dira hain adierazgarriak.

4.6 Ondorioak

Kapitulu honetan, Riedel *et al.*-en datu multzoko ausazko hainbat aipamen analizatu ditugu, eta bertatik hiru zarata mota kategorizatu.

Datu multzoek barnean duten zarata kantitate altuak motibatuta, atal honetan hiru metodo sinple eta trinko aurkeztu ditugu zarata garbitzeko. Heuristiko hauek ez dute datu multzoetan inongo eskuzko etiketatzerik erabiltzen, eta domeinu eta sistemarekiko independenteak dira. Hiru heuristiko hauen artean konbinaketa desberdinak egin ditugu, eta emaitza onenak ematen dituen konbinaketa erabili bukaerako emaitzak eskuratzeko.

Esperimentuak Stanford Unibertsitateak garatutako urruneko gainbegiraketa sistema erabiliz egin ditugu. Sistema hau artearen egoeran dago eta Mintz *et al.*-en algoritmo originala hobetzen dituen bi aldaera ditu. Sistema hau erabiliz, gure heuristikoen eraginkortasuna ebaluatzeko gai izan gara.

Heuristiko hauek, batez ere beraien konbinazioek, Stanford Unibertsitateak garatutako bi oinarri lerro trinkoren emaitzak hobetzen ditu, urruneko gainbegiraketaren eraginkortasuna hobetzeko aipamen zaratatsuak detektatzea eta ezabatzea beharrezkoa dela frogatuz.

Hurrengo ataletan, gertaeren erauzketan zentratuko gara, urruneko gainbegiraketa ataza honetara moldatuz. Horretarako lurrikarei buruzko ezagutza base bat eta datu multzo bat sortuko ditugu.

5

Gertaeren erauzketarako datu multzo baten sorkuntza

Kapitulu honetan eta ondorengoan gertaeren erauzketarekin egingo dugu lan. Urruneko gainbegiraketarekin jarraituko dugu, erlazioak erauzi ordez gertaerak erauziz, horretarako lurrikarei buruzko ezagutza base bat eta txioetan oinarritutako datu multzo bat sortuko ditugu. Atal honetan, esperimentuetarako behar dugun materialaren sorkuntzarako jarraitu diren pausoak zehaztuko dira, esperimentuak eta emaitzen analisiak hurrengo atalerako utziz.

5.1 Sarrera

Urruneko gainbegiraketak arrakasta handia eduki du erlazio erauzketan. Bada garaia bere potentziala beste hainbat arlotan frogatzeko, gertaeren erauzketan adibidez.

Horrez gain, normalean urruneko gainbegiraketarako erabiltzen diren corpusak berri agentzien dokumentuek osatzen dituzte. Beste aukera bezala sare sozialak erabiltzea proposatzen dugu, paradigma beste eremu batera eramateko. Guk aukeratu dugun sare soziala Twitter da.

Esperimentuak egiteko aukeratu dugun domeinua lurrikarena da. Horretarako Wikipediako infotauletan oinarritutako ezagutza base bat sortu dugu. Ezagutza base honetan dauden lurrikarei buruzko ahalik eta txio gehien berreskuratzen ditugu Twitterretik esperimentuetako datu multzoa sortzeko.

Sortutako datu multzoa eta ezagutza basea txikiak dira, aurreko kapituluetako esperimentuetan eduki ditugun baliabideekin alderatuta. Honek urruneko gainbegiraketari buruzko analisi sakonago bat egiten laguntzen digu. Analisi hauen bidez, paradigman sortzen diren hainbat arazo berri detektatu eta hauek konpontzeko proposamenak egiteko aukera dago.

Gertaeren erauzketari buruzko lana bi kapitulutan banatu dugu. Kapitulu honetan erabiliko ditugun baliabideei buruzko sorkuntzan eta detaileetan zentratzen gara. Esperimentuak, analisiak eta hobekuntzak hurrengo kapituluan lantzen ditugu.

Hasteko, kapitulu honetan ezagutza basea nola sortu dugun azalduko dugu. Jarraian, Twitterretik aipamenak nola eskuratu ditugun azalduko dugu, hauen garbiketa eta normalizazio prozesuak ere aipatuz. Ondoren, ezagutza basean dauden argumentu guztiei buruzko informazioa adierazten duten aipamen guztien eskuzko etiketatze prozesua azalduko dugu. Amaitzeko, atal honetan egindako ekarpenak laburtuko dira.

5.1.1 Erronkak

Denbora errealean lurreko leku desberdinetan gertatzen diren albisteei buruzko informazio iturri ona bilakatu da Twitter, gertaeren erauzketarako alternatiba bat izateko motibatuz, ohiko albiste artikuluen ordez.

Egunkari, aldizkari eta beste informazio iturriekin lan egiten dugunean, hizkera formala eta garbia dugu (ez beti). Txioekin lan egiteak berriz erronka gogorra dakar: hizkera kolokiala, sintaxi eta diskurtso zaratatsua, emandako informazioa okerra izatea, eta abar. Gertaera baten propietate guztiak ez ditugu txio bakar batean aurkituko, hauen karaktere muga dela eta[1]. Nahiz eta txio batek informazio gutxi eduki, ikasketa posible da, hainbat eta hainbat txio aurkitu ditzakegulako gertaera berdinari erreferentzia egiten.

Twitterren erabilerak potentzial handia dauka: posible da baita ezagutza base batean ez dagoen informazioa txioetan topatzea, tesi honetako ondorengo kapituluan ikusiko dugun bezala.

Lan honetarako aukeratu dugun domeinua **lurrikarena** da, magnitudea, lurrikararen kokapena, hildakoen kopurua, ordua eta beste hainbat informaziorekin.

Lan hau ez da Twitterren erabilera proposatzen duen lehenengoa gertaerei buruzko informazioa erauzteko urruneko gainbegiraketaren bidez. Ben-

[1]Txio bakoitzean gehienez 140 karaktere onartzen dira.

Date	Time (UTC)	Place	Lat.	Long.	Fatalities	Magnitude	Comments
September 2, 2009	07:55	Java, Indonesia *see* 2009 West Java earthquake	−7.809	107.259	79	7.0	M_w (USGS) Centred 100km SSW of Bandung, Java, Indonesia, at a depth of 46.2km.
April 4, 2010	22:40	Baja California, Mexico *see* 2010 Baja California earthquake	32.259	−115.287	4	7.2	M_w (USGS) Centred 50km SSE of Mexicali, Baja California, Mexico, at a depth of 10km.

5.1 irudia – Wikipediako lurrikaren taulako bi lurrikaren adibidea.

son *et al.*-en lanean (2011), artista famatu batek egingo dituen emanaldiei buruzko informazioa erauzten du, baina lan hori argumentu bati buruzko informazioa lortzeaz bakarrik mugatzen da. Atal honetan zehar ikusiko dugun bezala, gure lanak 20 argumentu desberdinei buruzko informazioa eskuratzen saiatzen da.

5.2 Lurrikarei buruzko ezagutza basearen sorkuntza Wikipedian oinarrituta

Ezagutza basea sortzeko, ingeleseko Wikipediako XXI. mendeko lurrikara garrantzitsuenen webgunera[2] jo dugu. 2009ko hasiera eta 2013ko uztailaren 7aren arteko lurrikarak aukeratu ditugu. Webgune honetara jo dugu egunean 2013ko uztailaren 7ko lurrikara zen erregistratutako azkena. 5.1 irudiak taula hauek zernolako itxura duten erakusten digu, bi lurrikara desberdinekin.

Guztira 108 lurrikara ditu ezagutza baseak. 5.1 taulan zerrendatuta daude argumentu guztiak, argumentu mota, hauen aipamen kopurua ezagutza basean, eta euskaraz zer adierazten duten azalduz.

Ezagutza basea webgune horretako tauletako informazioa bilduz sortu da. Lurrikara batzuek berezko orrialdea dute Wikipedian, eta ezagutza basea are aberatsagoa egiteko, orrialde hauetara jo genuen infotauletako informazioa biltzeko. 5.2 irudian ikus ditzakegu 5.1 irudiko bi lurrikaren artikuluen infotaulak. 5.2 taulan irudiko lurrikaren informazioa dugu, ezagutza basean agertzen den moduan.

Kasu batzuetan, lurrikara guztien orrialdean agertzen diren tauletako informazioa eta lurrikararen orrialdeko infotaulako informazioa ez datoz bat. Horrelakoetan, orrialde orokorreko taulako balioak hartu dira kontutan, in-

[2] https://en.wikipedia.org/wiki/List_of_21st-century_earthquakes

Arg. izena	**Arg. mota**	**Aipamen kopurua**	**Esanahia euskaraz eta azalpen labur bat**
date	E	108	Eguna
time	D	108	Ordua
country	L	108	Estatua
region	L	77	Herrialdea
city	L	77	Hiria
latitude	Z	108	Latitudea
longitude	Z	108	Longitudea
dead	Z	71	Hildako kopurua
injured	Z	39	Zauritu kopurua
missing	Z	8	Desagertuen kopurua
magnitude	Z	108	Magnitudea
depth	Z	99	Sakonera, kilometrotan
affected-country(*)	L	37	Kaltetutako estatuak (lurrikara nagusiarekiko desberdinak)
affected-region(*)	L	4	Kaltetutako herrialdeak (lurrikara nagusiarekiko desberdinak)
landslides	B	8	Lubiziak, bai ala ez?
tsunami	B	10	Tsunamiak, bai ala ez?
aftershocks	Z	20	Erreplikak, zenbat?
foreshocks	Z	3	Aurrelurrikarak, zenbat?
duration	D	7	Lurrikararen iraupena
peak-acceleration	Z	8	Azelerazio sismikoa, galetan
GUZTIRA		1116	

5.1 taula – Lurrikarei buruzko ezagutza basearen informazioa, argumentuen izena, mota, bakoitzak duen aipamen kopurua ezagutza basean, eta argumentuaren esanahia euskaraz. (*) marka duten argumentuak balio anitzak (*multi-valued*) dira. *E* motakoek egunak adierazten dituzte, *D* motakoek denbora adierazpenak, *L* dutenak lekuentzako dira, *Z* zenbakizko aipamenak dira, eta *B* dutenak balio boolearrak dituztenak (bai ala ez).

fotauletakoaren ordez. Adibidez, 5.2 irudian bi lurrikaren infotaulak ditugu, eta 5.2 taulak ezagutza basean dagoen bi lurrikarei buruzko informazioa dakar; 5.1 irudiko taulen informazioa aztertzen badugu, konturatzen gara koordenatu geografikoen balioak ez datozela bat: ezagutza basean Javako lurrikaran latitudea eta longitudea -7.809 eta 107.259 dira, infotaulan berriz, -7.778 da latitudea eta 107.328 longitudea. Nahiz eta balioak inoiz ez izan berdinak, beti dira oso antzekoak, zenbaki hamartar batzuk bakarrik dira desberdinak.

5.1 taula berriro aztertzen badugu, ikusten da ezagutza baseko argumentuen banaketa oso irregularra dela. Ordua, denbora, lekuak, magnitudea eta koordenatu geografikoak adierazten dituzten argumentuek lurrikara guztietan parte hartzen dute. Badira argumentu marjinal asko, aipamen gutxi dituztenak eta ondorioz lurrikara gutxitan parte hartzen dutenak: desagertutako pertsonak, kaltetutako herrialdeak, lubiziak, tsunamiak, erreplika kopurua, aurrelurrikara kopurua, iraupena eta azelerazio sismikoa.

Bost argumentu mota desberdin daude: eguna adierazten dutenak, denbora adierazpenak, zenbakizko balioak dituztenak, lekuak adierazten dutenak, edo balio boolearrak (bai edo ez) dituztenak.

5.3 Lurrikarei buruzko aipamenak Twitterretik eskuratzen

Lurrikarei buruzko txioak eskuratzeko, Topsy Labs[3] enpresaren baliabideak[4] erabili ditugu. Enpresa hau sare sozialen edukien bilaketan eta analisian jarduten da. Twitterrekin elkarlanean aritzen da, 2006tik aurrera izan diren mila milioi txioak indexatuz, eta tresna komertzial desberdinak egiten ditu txioen bilaketak eta analisiak egin ahal izateko. Baliabide batzuk dituzte publikoki eskuragarri Twitterren bilaketak egiteko, guk behar izan dugunerako nahikoa.

Lurrikara bakoitzeko bilaketak egitean, *earthquake* hitz gakoa erabili eta ezagutza basean agertzen zen kokapena (hiriak, herrialdeak eta estatuak) zehazten ditugu, lurrikara gertatu baino egun bat lehenago eta hortik 7 egun geroago idatzitako txioak eskuratuz.

[3] http://topsy.com/
[4] http://api.topsy.com/doc/

5.2 irudia – 2 lurrikarei buruzko infotaulak, 1) Java, Indonesiako lurrikara batena. 2) Baja California, Mexikoko lurrikara batena.

Lurrikara gertatu baino egun bat lehenagoko txioak eskuratzeak arrazoi bat dauka: ordu eremuak[5]. Txioak ez daude geolokalizatuta, eta honek denboraren normalizatzea ezinezkoa egiten du. Gai honekin kontu handiz ibili behar gara lan honetarako denborari buruzko informazioa lortzeko, kontuan hartu behar da txiolariak ez direla profesionalak, eta txiokatzean lurrikara gertatutako unea aipatzeko beren bizilekuko ordua erabiliko dutela denbora estandarraren ordez. Adibidez, lurrikara bat ordu unibertsalaren (*GMT*, Greenwichekoa) arabera goizeko 6etan gertatu bada Japonian, japoniar ba-

[5] http://eu.wikipedia.org/wiki/Ordu-eremu

Arg. izena	Java, Indonesia	Baja California, Mexiko
date	September 02, 2009	April 04, 2010
time	07:55:00	22:40:00
country	Indonesia	Mexico
region	Java	Baja California
latitude	-7.809	32.128
longitude	107.259	-115.303
dead	79	4
injured	-	100
magnitude	7	7.2
depth	49	10
affected-country	-	United States
landslides	-	yes
tsunami	yes	-
duration	-	00:01:48

5.2 taula – Ezagutza baseko 2 adibide eta haiei buruzko informazioa. 1) Java, Indonesiako lurrikara batena. 2) Baja California, Mexikoko lurrikara batena.

tentzat eguerdiko 2tan gertatu da, eta Kalifornian bizi den batentzat berriz, aurreko eguneko gaueko 10etan; 2 pertsona hauek lurrikara horri buruz emango diguten informazioak kontraesanak izango ditu denborarekiko.

Kokapenaren arabera egitean bilaketa, hiriari ematen diogu lehentasuna (informazio hau baldin badago), eta gutxienez 30 txio eskuratu behar ditugu. Horrela izan ezean, orduan herrialdea zehaztuta egiten dugu bilaketa. Oraindik 30 txio baino gutxiago baldin baditugu, estatua zehazten dugu. Hasieratik 30 txio erauzten baditugu, orduan bilaketa amaitutzat ematen dugu, azken batean, hiria zehaztean aukera handiak daude txioan bertan herrialdea eta estatuari buruzko aipamenak edukitzeko.

Txio guztiak ingelesez eskuratzea da gure asmoa, baina badira txio gutxi batzuk beste hizkuntzatan, hala nola alemaneraz, frantsesez, gaztelaniaz, japonieraz, eta abar. Txio hauek *earthquake* hitz gakoa daukate idatzita, nahiz eta edukia ingelesa ez den beste hizkuntza batean izan. Txio hauek datu multzoan mantendu dira.

Zoritxarrez, lortutako txio asko lurrikaren erreplikei buruzkoak dira. Erreplikak lurrikara nagusiaren ondoren gertatutako beste lurrikara batzuk dira, hauek epizentrotik gertu daude eta normalean nagusiak baino magnitude

txikiagoa dute. Erreplikak direla eta, txioetako informazioak eta ezagutza baseko informazioak kontraesanak izango dituzte, sistemaren entrenamendu prozesua nahastuz eta ebaluatzean emaitza okerrak itzuliz, eta hortaz, tratamendu berria egin dugu.

5.3.1 Erreplikei buruzko aipamenak antzematen eta ezabatzen

Erreplikak gertaera desberdin bezala tratatu ditugu, eta ezagutza basean zeuden lurrikarei buruzko txioak bakarrik edukitzearren, metodo zorrotz bat aplikatu genuen erreplikei buruzko txioak baztertzeko. Heuristiko hau txio desberdinetan aipatzen diren denbora adierazpenetan oinarritzen da:

1. Txioetan lurrikara bakoitzeko lortutako lehen denbora adierazpena gordetzen dugu. *ordua:minutua* patroia erabili da denbora adierazpen hauek antzemateko. Segundoak ez ditugu kontuan hartzen.

2. Geroagoko txio batean agertzen den denbora adierazpena lehenengoarekiko desberdina bada, bai ordua bai minutua, txio hau erreplika bati buruz ari dela ulertzen dugu. Txio hau eta ondoren datozen beste guztiak kentzen ditugu, hemendik aurrera jasoko ditugun txioak erreplika horri edo beste batzuei buruzkoak izango direlakoan. Ordua lurrikara nagusiko orduarekiko desberdina bada baina minutua berdina, orduan lurrikara nagusitzat hartzen dugu txio hau, ordu eremu desberdin batean dagoen pertsona batek txiokatu duelakoan.

Adibidez, 2010eko otsailaren 26an, 20:31etan Japoniako Ryukyu Irletan lurrikara bat gertatu zen. Lurrikara honi erreplika detektatzeko heuristikoa aplikatzean, ondorengo txioa aurkitu zuen sistemak:

- A 4.9 earthquake occurred in Ryukyu Islands, Japan on February 27, 2010 10:33:21 AM at epicenter.

Ezagutza basean lurrikara honen denbora desberdina da (*20:31 vs 10:33*), minutuek ez dute bat egiten, hori dela eta txio hau erreplikatzat hartu da, hau kendu, eta ondoren zetozen txio guztiak ere ezabatu ditugu automatikoki.

Zoritxarrez ez ditugu beti txio egokiak ezabatzen. Batzuetan kendutako txioaren ondoren lurrikara nagusiari buruzko txio gehiago daude, eta hauek ere kentzen dira. Beste kasu batzuetan berriz, erreplika bati buruzko txioak

Lurrikara mota	Kendutakoak	Mantenduak	Guztira
Nagusia	105	773	878
Erreplika	193	83	276
Guztira	298	856	**1152**

5.3 taula – Lagin batean heuristikoak detektatutako eta ezabatutako erreplikei buruzko estatistikak.

antzeman aurretik, baziren erreplikei buruzko beste txio batzuk, baina ez zutenez denbora adierazpenik ezin izan genituen detektatu.

Ondorengo txioa 2011ko urriaren 23an Van hiri kurduan gertatu zen erreplika bati buruzkoa da, honen magnitudea 5.5ekoa da, gure ezagutza basearen arabera lurrikara nagusiarena 7.1 izan zen, baina ez dagoenez erreplikaren ordua zehaztuta, gure heuristikoak ezin izan du erreplika hau antzeman:

- 5.5 quake hits #Van again reports Turkish #Earthquake Center after approving the scale of the previous tremor at 7.3.

Ausaz aukeratutako 10 lurrikaren lagin bat hartu dugu, ezabatutako eta mantendutako txioei buruzko balorazio bat egiteko. Lagin hau 1152 txioz osatuta dago. 5.3 taulan ikus dezakegu lagin honi heuristikoa aplikatu eta gero mantendu eta kendu diren lurrikara motak. Taularen estatistikak datu multzo osoan aplikatuta, gure estimazioen arabera, hasierako datu multzo honetan %76a dira lurrikara nagusiei buruzko txioak, gelditzen den %24a erreplikei buruzko txioak izanik. Hasierako datu multzoko txioen %26a ezabatu dira, erreplikei buruzko txioak direlakoan, eta ezabatutako txio horietatik %35a lurrikara nagusiari buruzko txioak dira. Mantendu ziren txioetatik, %10a erreplikei buruzkoak dira.

Bukaerako datu multzoak 108 lurrikara desberdin ditu eta guztira 7841 txio desberdin, batezbesteko 72 txio lurrikara bakoitzeko, gehienez 654 txio eta gutxienez 2 edukiz. 19 lurrikarek 10 txio baino gutxiago dituzte. Txioetan batezbesteko bi argumentu ditugu.

5.3.2 Aipamenen aurreprozesaketa

Erreplikak kendu ondoren, sistemaren eraginkortasuna hobetzeko eta urruneko gainbegiraketaren bidez etiketatzea ahalik eta gehien errazteko, txio guztietako testuak normalizatzen ditugu.

Aurreko kapituluetan aipatu dugun bezala, urruneko gainbegiraketak ezagutza baseetan agertzen den informazio zehatza bilatzen du corpusetan, ondoren informazio hori entitate nagusi edo betegarri bezala markatzeko, esaldi bakoitzaren tokenak aztertu ondoren.

Gogoratu behar dugu Twitterreko hizkera eta berri agentziek albisteak ematean duten hizkerak eta sintaxiak ez duela inongo zerikusirik, lehenengoan sintaxia oso zaratatsua izan daiteke (txiolariaren arabera) eta bigarrenean berriz oso garbia da. Sintaxi zaratatsua dela eta, askotan interesgarria eta komenigarria den informazio asko galtzen dugu, hori dela eta zerbait egin behar dugu ahalik eta aipamen gehien etiketatzeko. Beste kasu batzuetan, aipamen batzuk era desberdin batean adierazita daude, baina ezagutza basean dugun informazioarekiko baliokideak dira.

Aurreprozesaketa baldintza hauek ez dira ezagutza basearekiko informazio berdina duten aipamenetan bakarrik aplikatu, aurkitu ditugun kasu guztietan baizik.

Denbora adierazpenen normalizazioa

Alde batetik denbora adierazpenak ditugu, hauek modu askotara adierazi daitezke: *"January 1, 2010"* (hau ezagutza baseko formatua da), *"1st of January, 2010"*, *"2010/1/11"*,... Nahiz eta hauek baliokideak izan, urruneko gainbegiraketak ez ditu aipamen hauek etiketatuko. Hauek normalizatzea komeni da, bai ezagutza basean, bai txioetan, bietan aurki ditzakegun adierazpenek bat egin eta etiketatu ahal izateko. Egunak normalizatzeaz gain, orduak ere normalizatzen ditugu.

Denbora adierazpenak normalizatzeko, Stanford CoreNLP[6] sistemaren SUTime tresna erabili dugu, sistema honek ingeleseko denbora adierazpenak antzematen ditu eta hauek normalizatu, *"January 1, 2010"* adierazpena *"2010-1-1"* bihurtuz. Ezagutza basea tresna honen bidez ere normalizatu da.

Askotan txioetako denbora adierazpenek eguna eta hilabetea bakarrik aipatzen dituzte (*"January 1*), normalizazioa honela utziz: *"XXXX-1-1"*, nahiz eta urtea ez markatu, hauek normalizatzea ere ondo etorriko zaizkigu etorkizuneko esperimentuetarako. Beste kasu askotan, asteko eguna idatzita badator, adibidez *"Friday, January 1, 2010*, normalizatu eta gero adierazpena honela gelditzen da: *"2010-1-1-WXX-5"*, 5 zenbakiak ostirala adierazten du. Hala ere badira kasu asko normalizatzaileak detektatu ez dituenak, adi-

[6] http://nlp.stanford.edu/software/corenlp.shtml

bidez *"aug-27"* ez du denbora adierazpen bezala detektatu, antzeman izan balu, *"2012-8-27"* bihurtuko zen.

Orduak normalizatzean, *20:43:15* bezalako adierazpenak *T20:43:15* bezala gelditzen dira. Normalizatutako denbora adierazpen guztiek "T" bat daramate hasieran. Batzuetan Segundoak ez daude markatuta txioetan (*20:43*), kasu hauetan normalizatzaileak bi zero eransten ditu: *T20:43:00*.

Zenbakizko argumentuen normalizazioa

5.3.1 azpiatalean, estatistikak ateratzeko lagin bat aztertzean, aipamen zaratatsu asko aurkitu ditugu txio desberdinetan, esperimentuetako ikasketa prozesurako nabarmena den informazioa, urruneko gainbegiraketa algoritmoak detektatuko ez duena. Kasu desberdinak bildu ditugu eta garbiketa algoritmo bat martxan jarri.

Garbiketarako baldintza desberdinak erabili dira, argumentu motak eta honen edukiak garrantzi handia dute garbiketa prozesuan. Baldintza batzuk argumentu bakar batentzat pentsatuta daude, beste batzuk berriz, hainbat argumentutarako balio dute.

Hasteko, magnitudeei buruzko ondorengo aipamenak ez ditu urruneko gainbegiraketak antzemango 7.3ko magnitudea dagoen lurrikaretan: *7.3-magnitude* eta *M7.3*. Hauen detekzioa posible egiteko, aipamen hauek formatu hauetara pasa ditugu: *7.3 - magnitude* eta *M 7.3*; honi esker, urruneko gainbegiraketan posible da aipamen hauek markatzea.

Magnitudeak alde batera utzita, sakonerari buruzko argumentuak kilometrotan ditu irudikatuta balio guztiak ezagutza basean. Orokorrean txioetako sakontasuna ere kilometrotan dago; salbuespen batzuk badira, miliak bezala, baina ez dira oso ohikoak. Nahiz eta ia txio denak ingelesez izan, orokorrean erabilitako luzera sistema metrikoa da, hasieran pentsatu genuen ezagutza baseko informazioa miliatara pasatzea, baina hau ikusita erabaki genuen ez zela beharrezkoa; ondorioz ondorengo baldintzak kilometroetarako bakarrik aplikatu dira. Sakontasunari buruzko aipamen asko *35km* edo *35kms* bezala daude idatzita; urruneko gainbegiraketak hauek antzemateko, karaktere kateak zenbakikoetatik banatu ditugu, horrela utziz: *35 km* eta *35 kms*. Kontuan izan kilometroak aipatzen diren testuinguruak ez duela zertan sakonerari buruz izan behar, hiri garrantzitsu batetik epizentrora dagoen distantziarena izan liteke baita.

Ezagutza basean zenbakizko balio guztiak zenbakiz adierazita daude, eta txioetan hala aurkitzea espero dugu, baina zoritxarrez ez da beti horrela ger-

tatzen, zenbaki asko letraz adierazita daudelako: *two, five* eta *seven* adibidez. Aipamen hauek detektatzeko heuristiko bat garatu dugu, hauek zenbakitara pasatzeko, *2, 5* eta *7* bezala agertzeko txioetan. Egoera hauek hildako, zauritu eta desagertutako pertsonei buruzko aipamenetan askotan gertatzen dira, oso ohikoa da *"five people injured"* bezalako aipamenak aurkitzea, eta gure urruneko gainbegiraketa sistemarentzat egokiagoa da *"5 people injured"* bezala agertzea.

Amaitzeko, koordenatu geografikoetan jartzen dugu arreta. Hauei buruzko adierazpenak oso anbiguoak dira, eta gure sistemak hauek ahalik eta gehien detektatzea interesatzen zaigunez, zerbait egin behar dugu hauen antzematea posible izateko. Ezagutza basean, latituderako, balio positiboa erabiltzen da iparraldea markatzeko, eta negatiboa hegoalderako; longituderako, positiboa ekialdearentzat eta negatiboa mendebaldearentzat. Koordenatu geografikoak markatzeko anotazio desberdin bat erabili daiteke, norabidea zehaztuz; adibidez *-73.45* longitudeak hegoaldea adierazten du, baina hau *73.45 S* bezala adieraz daiteke, negatiboak hegoaldea dela adierazten duen moduan, negatiboa *S* batekin ordezkatu dezakegu, eta iparraldea denean *N* bat erantsiz edo zenbakia balio positiboarekin mantenduz, baina guri zenbakia bakarrik agertzea interesatzen zaigu. Txioetan ondorengo aipamenak aurkitu ditugu: *73.45 S*, *73.45 °S* edo *73.45 °S*, kasu guzti hauek *-73.45* baliora bihurtu dira.

F eranskinean aipamenen garbiketarako jarraitutako irizpideei buruzko laburpen eskematizatu bat aurki dezake irakurleak.

5.4 Aipamenen eskuzko etiketatzea

Gure gertaera erauzketa sistemak sortzen dizkigun erronkak sakonki aztertzeko, txio guztiak eskuz etiketatu ditugu eta hauekin datu multzo berri bat sortu. Etiketatutako informazioa, ezagutza basekoa izateaz gain, bat egiten ez duen informazioa ere etiketatu dugu, hala nola data eta denboren bariazioak, antzeko koordenatu geografikoen balioak, hildako eta zaurituen kantitate kopuru desberdinak, eta beste hainbat. Guztira 17672 aipamen etiketatu ditugu.

Eskuzko etiketatzearen helburuak ondorengoak dira:

- **Urruneko gainbegiraketaren errendimendu sabaia zehaztu:** honela gure sistemak lortu lezakeen emaitza maximoa zein izango litzatekeen jakin dezakegu.

- **Atazaren arazoak hobeto ezagutu:** honela gure sistemaren, eta ataza osoaren ahulguneei konponbideren bat bilatu ahalko diegu.

Ondoren etiketatu gabeko txio bat ikus dezakegu, eskuz etiketatutako bertsioarekin batera:

- Earthquake in Honduras. So strong it was felt in Guatemala as well. 7.1 offshore atlantic.
- Earthquake in *<country>Honduras</country>*. So strong it was felt in *<affected-country>Guatemala</affected-country>* as well. *<magnitude>7.1</magnitude>* offshore atlantic.

5.4 taulan, argumentu bakoitzeko eskuz etiketatu diren aipamen kopuruak agertzen dira. Taula hau 5.1 taularekin batera aztertzen badugu, lekuei erreferentzia egiten dieten aipamen asko etiketatu dira, baita magnitudeei buruzko aipamenak ere. Hauek lurrikara bakoitzerako informazioa dute ezagutza basean, koordenatu geografikoak ezik, nahiz eta informazio hau beti eduki, oso gutxitan aipatzen dira txioetan; orduak eta egunak ere ez dira lekuak eta magnitudeak bezain beste aipatzen, baina hauei buruzko aipamen asko etiketatu dira.

Badira argumentu marjinal asko: desagertutako pertsonak, kaltetutako herrialdeak, lubiziak, erreplikak, iraupena, aurrelurrikarak eta azelerazio sismikoari buruzkoak; hauei buruko oso aipamen gutxi etiketatu dira edo ez da ezer etiketatu. Nahiz eta ezagutza basean tsunamiei buruzko informazio gutxi egon, hauei buruzko aipamen asko etiketatu dira.

Irakurleak eskuzko etiketatzerako jarraitutako irizpideei buruz gehiago jakin nahi badu, G eranskinean aurki dezake informazio gehiago.

Eskuzko etiketatzearen kalitatea neurtzeko, beste aditu batek etiketatu du datu multzoaren %5a. Adostasun maila oso handia izan da, ondorengo balioekin: %90 ITA eta %85 Fleiss Kappa (Artstein eta Poesio 2008). Adituetako batek aipamen gutxi batzuk despistez etiketatu ez izanak eragin ditu desadostasun gehienak.

Argumentuaren izena	**Eskuz etiketatutako aipamenak**
date	706
time	589
country	6327
region	2663
city	1723
latitude	28
longitude	28
dead	984
injured	192
missing	18
magnitude	3403
depth (km)	313
affected-country	357
affected-region	36
landslides	9
tsunami	273
aftershocks	22
foreshocks	-
duration	1
peak-acceleration	-
GUZTIRA	17672

5.4 taula – Argumentu bakoitzeko eskuz etiketatutako aipamenak.

5.5 Ondorioak

Atal honetan txioetan oinarritutako gertaera erauzketa sistema bat sortzeko hastapenak egin ditugu, aukeratutako domeinua lurrikarak izanik. Lehenengo urratsa Wikipedian oinarritutako lurrikarei buruzko ezagutza base bat sortzea izan da. Jarraian ezagutza baseko lurrikara bakoitzeko jendeak egindako txioak eskuratu ditugu Twitterretik. Eskuratu ondoren, erreplikak detektatu ditugu, hauek kendu, eta gainontzeko txioen edukia normalizatu.

Atal honetan egindako ekarpenak ondorengoak dira:

1. Azken urteetako lurrikarei buruzko ezagutza base bat sortu dugu, Wikipediako informazioan oinarrituta. Ezagutza base hau publikoa da eta edonork erabili dezake norberaren ikerketetarako. Ezagutza base hau urruneko gainbegiraketan urre patroi bezala erabiltzeko da.

2. Ezagutza basean zehaztutako lurrikarei buruzko hainbat txio eskuratu ditugu. Txio hauek ere publiko egin ditugu.

3. Txio hauen denbora adierazpenak normalizatu ditugu eta ondoren eskuz etiketatu argumentu guztiei buruzko aipamenak. Hauek ere publikoak dira[7], ikerlariek nahi izanez gero urre patroi bezala erabiltzeko.

4. Guk dakigula, ez dago beste datu multzorik non ezagutza basea, corpus gordina eta eskuz anotatutakoa biltzen dituenik.

Ondorengo atalean gure gertaera erauzketa sistemarekin egindako etiketatze prozesuaz, esperimentu eta analisiei buruz hitz egingo dugu, horretarako eskuzko etiketatzea oso ondo etorriko zaigu. Analisi horietatik urruneko gainbegiraketarako hobekuntza batzuk proposatuko ditugu.

[7] https://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1427215138/publikoak/earthquake-kb-dataset.zip

6

Gertaeren erauzketa urruneko gainbegiraketaren bidez

Atal honetan, aurreko atalean sortutako datu multzoa automatikoki etiketatuko dugu. Jarraian, baliabide hauekin egindako esperimentuak deskribatuko ditugu, eta errore analisi batzuk egin. Analisiaren ondoren, emaitzak hobetzeko teknika batzuk proposatuko ditugu, bakoitzari atal bat eskainiz: etiketatze automatikoaren aldaera bat, eta ezaugarri globalen erabilera. Teknika hauetaz gain, emaitza errealistagoak ematen dituen oinarrizko ebaluazio metriken aldaera bat proposatuko dugu. Amaitzeko, ondorioak eta ekarpenak eztabaidatuko ditugu.

6.1 Sarrera

Aurreko atalean, lurrikarei buruzko ezagutza base bat eta datu multzo bat sortu ditugu. 108 lurrikara ditu ezagutza baseak, eta 8000 txio inguru esperimentuak egiteko. Ezagutza basean 20 argumentu desberdinei buruzko informazioa dago eskuragarri.

Badugu datu multzoaren beste bertsio bat, non txio bakoitzean argumentu desberdinei buruzko informazioa eskuz etiketatu den. Etiketatze honen helburua analisiak egiteko besterik ez da, urruneko gainbegiraketan sortzen diren beste arazo batzuk detektatzeko, eta hauek konpontzeko proposamenak egiteko. Proposatuko ditugun aldaketa hauek urruneko gainbegiraketaren arlo guztietan aplikatzeko balioko dute, ez bakarrik gertaeren erauzketarako.

Azterketaren ondoren proposatzen dugun aldaketa bat ebaluazio sistema aldatzean datza, sistemak induzitutako balioak eta ezagutza basean daudenak oso antzekoak direnean hauek ontzat hartuz. Beste proposamen bat etiketatze prozesuan dago, lurrikara konkretu bateko txio bateko informazioa ezagutza basean dagoen balio batekiko antzekoa bada, orduan hau ere etiketatuz, ez ezagutza baseko balio zehatzak bakarrik, orain arte bezala. Amaitzeko, lurrikara bakoitzaren txioak bildu eta elkarren arteko ezaugarri globalak erabiliz egiten ditugu esperimentuak, azaldu berri ditugun hobekuntzekin konbinatuz. Emaitzek erakutsiko dute proposamen hauek urruneko gainbegiraketan oinarritutako sistemak hobetzen dituztela.

6.2 Aipamenen etiketatze automatikoa urruneko gainberaketa erabiliz

Gogora ditzagun urruneko gainbegiraketaren pauso nagusiak erlazio erauzketa sistema batean (ikus 3. kapituluko esperimentuak):

- Ezagutza basean elkarren artean agertzen diren entitateak hartzen ditugu, eta hauek aipatzen dituzten esaldiak erauzi corpus batetik.
- Esaldi (edo aipamen) horretan bi entitateen arteko erlazioa markatzen dira, ezagutza basean agertzen den erlazio berdina zehaztuz.
- Esaldiak aurreprozesatzen ditugu, hitz bakoitzaren lema, kategoria gramatikala eta entitate izen mota lortuz.
- Aipamen bakoitzarentzako ezaugarriak erauzten ditugu.
- Sailkatzailea entrenatzen dugu aipamen bakoitzaren ezaugarriekin.
- Testerako, entrenamenduan parte hartu ez duten entitateei buruzko aipamen desberdinak eskuratzen ditugu corpusetik. Aipamen hauetan entitate horrekin erlazionatuta egon daitezkeen beste entitate batzuk daude. Sailkatzaileak entitate hauek aztertu eta elkarren artean zein erlazio dagoen iragartzen saiatzen da.
- Sailkatzaileak iragarritako erlazioak ebaluatzen ditugu.

Oraingoan, gertaera erauzketarekin jarraitutako pausoak oso antzekoak dira. Orain arte, bi entitateren artean erlazio bat zehaztuta baldin bazegoen ezagutza base batean, bi entitate hauek zituen edozein esaldik erlazio hori adierazten zutela esaten genuen. Oraingoan, gertaeraren izena ez denez esplizitua (batzuetan baliteke izatea baina ez da ohikoa), gertaera konkretu bati buruzko aipamenetan topatzen dugun hitz batek ezagutza baseko argumentu baten balioarekin bat egiten badu, hitz horrek argumentu hori adieraziko du nola edo hala. Etiketatze prozesuan, lurrikara bakoitzeko txioak hartu, lurrikara horri buruzko informazioa ezagutza basetik hartu, eta etiketatze zehatza egiten dugu.

Bi argumentu mota identifikatu ditugu: balio binarioa (*yes/no*) dutenak (*landslides* eta *tsunami*), eta beste motatako balioak dituztenak. Automatikoki etiketatzeko jarraitutako heuristikoak ondorengoak dira:

- **Balio boolearren etiketatzea:** ezagutza basean argumentu bitarrak *yes* balioarekin markatuta baditu, argumentuaren izena duten hitzak bilatzen ditugu (adibidez, *tsunami* edo *tsunamis*), eta hauek topatzean etiketatu (<*tsunami*>*tsunami*</*tsunami*>).

- **Gainontzeko balioen etiketatzea:** lurrikara baten argumentua ezagutza basean agertzen den balio zehatza bilatzen dugu txioetan, eta aurkitu ondoren etiketatu.

  - **Balioen desanbiguazioa:** argumentu bat baino gehiagok balio berdina baldin badute lurrikara konkretu baterako, adibidez 7 bai *magnitude* bai *dead* argumentuentzat, orduan ezagutza basean bietatik ohikoena den argumentua aukeratzen dugu.
  - **Traolak (#):** Gure urruneko gainbegiraketa sistemak traolak dituzten kokapenak ere etiketatzen ditu, <*country*>#*Honduras* </*country*> adibidez.

Adibide bezala, har dezagun 2009ko maiatzaren 28an Hondurasen gertatutako lurrikara. Hona hemen etiketatu nahi dugun txioa[1], eta txioaren edukia automatikoki etiketatua:

- Earthquake in Honduras. So strong it was felt in Guatemala as well. 7.1 offshore atlantic.

[1] Adibide hau 5.4 atalean eskuzko etiketatzean erabili dugun berdina da.

- Earthquake in <*country*>*Honduras*</*country*>. So strong it was felt in <*affected-country*>*Guatemala*</*affected-country*> as well. 7.1 offshore atlantic.

Txioa aztertzen badugu, konturatzen gara lurrikararen magnitudea, 7.1, etiketatu gabe dagoela. Ezagutza basean magnitudea 7.3 da, horregatik ez dugu magnitudea etiketatu adibide honetan.

Guztira 13562 aipamen etiketatu ditu sistemak. 6.1 taulan ditugu urruneko gainbegiraketa bidezko etiketatzearen estatistikak, eskuzko etiketatzearekin batera.

Etiketatzearen analisia

Ikusgarria da denbora adierazpena adierazten duten argumentuen artean dagoen aldea bi etiketatzeak konparatuta, ikusi dugu nola normalizatzailea ez den perfektua, eta adierazpen asko ezin izan dituenez normalizatu hauek urruneko gainbegiraketan kanpoan gelditu dira; ezin ditugu ahaztu erreplikei buruzko txioak, detektatu ez diren horiek eta datu multzoan mantendu diren horiek.

Koordenatu geografikoak oso gutxitan azaltzen dira txioetan, nahiz eta ezagutza basean lurrikara bakoitzak berea eduki, eta gainera, ia agerpen guztietan ez dute ezagutza baseko informazioarekin bat egiten.

Hildako eta zaurituen kopurua baita ere oso desberdina da, kontuan hartu balio hauek oso dinamikoak direla eta denbora aurrera doala eguneratzen direla.

Magnitudeak aztertuta, hemen desadostasun handia aurki dezakegu, txiolariek informazio okerra dutelako edo denbora aurrera doala informazio zehatzagoa jaso dutelako. Eskuzko etiketatzeari esker, jakin dugu badirela lurrikara asko non sarrien agertzen zen magnitudeen balioak ez zuen ezagutza basekoarekin bat egiten, nahiz eta desberdintasuna hamartar batzuena bakarrik izan.

affected-country argumentuan deigarria da urruneko gainbegiraketan aipamen gehiago etiketatuta egotea eskuzkoan baino. Honek azalpen erraza dauka: kasu askotan herrialdea bera aipatzen da, baina testuinguruak ez du inon aipatzen lurrikarak eraginik izan duenik bertan. Badira ere kasu asko non beste herrialde batean lurrikarak eragina izan duen, baina ezagutza basean herrialde hori ez den inon zehaztu.

Tsunamiekin ere gauza berdina gertatzen da, aipamen gehiago automatikoki etiketatuta; txio askok tsunami alerta baten berri ematen dute, baina

Argumentuaren izena	**Eskuzko etiketatzea**	**Urruneko gainbegiraketa**
date	706	291
time	589	378
country	6327	6294
region	2663	2598
city	1723	1426
latitude	28	2
longitude	28	4
dead	984	143
injured	192	22
missing	18	-
magnitude	3403	933
depth	313	27
affected-country	357	436
affected-region	36	-
landslides	9	7
tsunami	273	408
aftershocks	22	5
foreshocks	-	6
duration	1	-
peak-acceleration	-	-
GUZTIRA	17672	13562

6.1 taula – Argumentu bakoitzeko etiketatutako aipamenak. Erdiko zutabean eskuzko etiketatzean zenbat aipamen etiketatu diren azaltzen da (5.4 atala). Eskuinekoan berriz, urruneko gainbegiraketa sistemak etiketatutako aipamenak (6.2 atala).

horrek ez du esan nahi derrigorrez tsunamiak egon direnik, geroago aurkituko dugu txioren bat hau konfirmatzen duena, orduan bai etiketatu dezakegula, bestela ez.

6.3 Esperimentazio ingurunea

Atal honetan, gure gertaera erauzketa sistemarekin egindako oinarrizko esperimentua azaltzen dugu, aipamenen aurreprozesaketa, ezaugarrien erauzketa, sistemaren entrenamendua eta inferentzia prozesua deskribatuz. Amaitzeko, lortutako emaitzak eztabaidatzen ditugu.

6.3.1 Aurreprozesaketa eta entrenamendua

Txio bakoitzaren ezaugarriak sortzeko Stanfordeko CoreNLP tresna erabili dugu beste behin ere. Horretarako txio bakoitza tokenizatu, eta hitz bakoitzeko bere lema, kategoria gramatikala eta entitate izen mota lortzen ditugu.

Adibide bezala, har dezagun 6.2 taulan ikus dezakegun txioa etiketatua, 2010eko apirilaren 4an Mexikoko Baja California herrialdean gertatutakoa. Taula horretan ikus dezakegu txioa eta honen ezaugarriak, sailkatzailearen entrenamendurako prest.

Lurrikaren datu multzoa kronologikoki hartu eta lehenengo %75a entrenamenduko partiziorako aukeratu ditugu, beste %25a testeatzeko utziz. Entrenamendurako 81 lurrikara desberdin daude, 6078 txiokin; testerako berriz, 27 lurrikara 1763 txiokin. Garapeneko esperimentu guztiak bost partiziotako egiaztapen gurutzatuaren bidez egin dira, partizio bakoitzean ausaz aukeratutako 15 lurrikara ingururekin. Partizio bakoitzeko txio kopurua oso desberdina da, batean 585 txio aurki ditzakegu, eta beste batean 2229 txio.

6.3.2 Sailkapena eta ezaugarriak

Sistemaren entrenamendurako erabilitako sailkatzailea CoreNLP tresnako BAE sailkatzailea (ikus 2.5.2 atala) da. Entrenatzeko erabili den teknika hau CoreNLP tresnak entitate izenen ezagutzarako (*Named Entity Recognition, NER*) erabiltzen duen teknikaren oso antzekoa da (Finkel *et al.* 2005). Kategoriatzat entitate izena hartu ordez, argumentua jasotzeko moldatu dugu sailkatzailearen kodea, entitate izena beste ezaugarri bat gehiago bezala onartu dezan.

Update : Earthquake in <region>Baja California</region> , <country>Mexico</country> upgraded to <magnitude>7.2</magnitude> magnitude , from 6.9 - USGS

Hitza	**Lema**	**POS**	**NER**	**Etiketa**
Update	Update	NNP	O	O
:	:	:	O	O
Earthquake	earthquake	NN	O	O
in	in	IN	O	O
Baja	Baja	NNP	LOCATION	region
California	California	NNP	LOCATION	region
,	,	,	O	O
Mexico	Mexico	NNP	LOCATION	country
upgraded	upgrade	VBN	O	O
to	to	TO	O	O
7.2	7.2	CD	NUMBER	magnitude
magnitude	magnitude	NN	O	O
,	,	,	O	O
from	from	IN	O	O
6.9	6.9	CD	NUMBER	O
-	-	:	O	O
USGS	usg	NN	ORGANIZATION	O

6.2 taula – Txio baten aurreprozesaketa. Erabilitako txioa taularen gainean dago. Txioan hitzak banandu dira eta bakoitzarentzat bere lema, kategoria gramatikala (*POS*) eta entitate izen motaren (*NER*) balioak lortu. Azken zutabean hitzari dagokion etiketa dago, urruneko gainbegiraketa bidez sistemak etiketatutakoa, edo guk eskuz etiketatutakoa.

6.3 taulan, 6.2 taulan erabilitako adibideko hitz batzuen ezaugarrien erauzketa dago. Erauzketa hau CoreNLP tresnak egiten du sailkatzailea martxan jartzean. Ezaugarriak antolatzeko orduan, aipamenen hitzak jasotzean, azken hitzetik hasten da eta lehenengo hitzarekin amaitu, entrenatzean hitzen alderantzizko ordena erabiliz.

Berez, sailkatzaile honek ezaugarrien arteko 3 *clique* desberdin sortzen ditu, *clique* bakoitzerako norbere ezaugarriak sortzen ditu, guk bidalitakoetan oinarrituta. Lehenengo *clique*ak une horretan dagoen elementuarekin (gure kasuan hitza) sortzen da, bigarrena une horretako elementua eta aurretik zetorren elementuarekin, eta hirugarrena aurretik zetozen bi elementuak eta une horretako elementuaren artean. Hona hemen clique bakoitzaren deskribapena:

- *0garren cliqueak* uneko hitzaren ezaugarriekin egiten du jolas, bertan daude kodetuta hitza bera, honen lema, kategoria gramatikala eta ikasketa prozesurako beste hainbat elementurekin batera. *Clique* honetako ezaugarri guztiek *"—C"* marka daukate, *"current word"* edo "uneko hitza" bezala ulertu genezake hau. Aurreko eta ondorengo hitzen informazioa ere barne dator.

- *1go cliqueak* uneko hitza eta aurretik daukan hitzaren artean informazioa tratatzeko kodeketa berezi bat egiten du ikasketarako. *Clique* honetako ezaugarri guztiek *"—CpC"* marka daukate, *"previous and current word"* edo "aurreko eta uneko hitza" bezala ulertu genezake hau.

- *2garren cliqueak* uneko hitza, aurretik dagoen hitza eta bi posizio atzerago dagoen hitza erabiltzen ditu kodeketa berezi honetan. *"—CpCp2C"* marka daukate *clique* honetako ezaugarriek.

Azpiatal honen hasieran entitate izen motak onartzeko sailkatzailea moldatu dugula komentatu dugu. Ondorengo lerroetan dugu moldaketa azalduta, tartean taulako adibideak barne jarri ditugu, azalpena errazteko:

- Entitate izen mota hitz bakoitzaren *0garren cliquean* dago zehaztuta.

- "HITZA-ENTITATE_IZEN_MOTA-NER|C" formatua erabiltzen da uneko hitzerako:
  - *USGS* hitzaren *0. cliquean* → *USGS-ORGANIZATION-NER|C*
  - *6.9* hitzaren *0. cliquean* → *6.9-NUMBER-NER|C*

Hitza	*#clique*	**Ezaugarriak**
USGS	0	[.....-PCNTYPE—C, #¡U#|C, ...-PW_CTYPE|C, ...-PCTYPE|C, #GS>#|C, USGS–W-PW|C, -NTYPE|C, ...-CNTYPE|C, :-NTAG|C, #¡US#|C, –NW|C, USGS-ORGANIZATION-NER|C, NO-OCCURRENCE-PATTERN|C, -PTAG|C, -PW|C, usg-LEM|C, NNS-TAG|C, ...–NW_CTYPE|C, –O-NNER|C, 6.9...-NNW_CTYPE|C, USGS—W-NW|C, #SGS¿#|C, #¡USGS¿#|C, #USGS¿#|C, USGS-WORD|C, #¡USG#|C, ...-PPW_CTYPE|C, #S¿#|C, #¡USGS#|C, -TYPE|C, –PNER|C, -PTYPE|C]
	1	[USGS-PSEQW|CpC, PSEQ|CpC, –TYPES|CpC, -NNS-TS|CpC, -PSEQcS|CpC, -PSEQpW|CpC, -PSEQpDS|CpC, -PSEQpS|CpC, -PSEQcDS|CpC, null-TNS1|CpC, -USGS-PSEQW2|CpC, –PSEQpcDS|CpC, -TPS2|CpC, –PSEQpcS|CpC]
	2	[null-null-null-TYPETYPES|CpCp2C, PPSEQ|CpCp2C, –NNS-TTS|CpCp2C]
-	0	[–O-NER|C, ...-PCNTYPE|C, USGS-ORGANIZATION-PNER|C, USGS-PW|C, ...-PCTYPE|C, #->#|C, -NTYPE|C, ...-CNTYPE|C, NNS-PTAG|C, NO-OCCURRENCE-PATTERN|C, –USGS-W-PW|C, #<-#|C, #<->#|C, :-TAG|C, 6.9-NUMBER-NNER|C, USGS...-PW_CTYPE|C, from...-NNW_CTYPE|C, –WORD|C, –6.9-W-NW|C, –LEM|C, 6.9-NW|C, CD-NTAG|C, ...6.9-NW_CTYPE|C, -TYPE|C, -PTYPE|C]
	1	[–PSEQW|CpC, PSEQ|CpC, –TYPES|CpC, -PSEQcS|CpC, USGS—PSEQW2|CpC, -PSEQpDS|CpC, -PSEQpS|CpC, -PSEQcDS|CpC, null-TNS1|CpC, USGS-PSEQpW|CpC, –PSEQpcDS|CpC, NNS-:-TS|CpC, -TPS2|CpC, –PSEQpcS|CpC]
	2	[-NNS-:-TTS|CpCp2C, null-null-null-TYPETYPES|CpCp2C, PPSEQ|CpCp2C]
6.9	0	[.....-PCNTYPE|C, –PW|C, #<6.#|C, 6.9-WORD|C, :-PTAG|C, ...-PCTYPE|C, 6.9-LEM|C, -NTYPE|C, ...-CNTYPE|C, #9>#|C, #<6.9#|C, 6.9-from-W-NW|C, NO-OCCURRENCE-PATTERN|C, USGS...-PPW_CTYPE|C, #<6.9>#|C, #.9>#|C, 6.9—W-PW|C, CD-TAG|C, IN-NTAG|C, from-NW|C, 6.9-NUMBER-NER|C, -...-PW_CTYPE|C, –O-PNER|C, #<6#|C, #6.9>#|C, ...from-NW_CTYPE|C, from-O-NNER|C, -TYPE|C, -PTYPE|C]
	1	[–6.9-PSEQW2|CpC, PSEQ|CpC, :-CD-TS|CpC, –TYPES|CpC, -PSEQcS|CpC, -PSEQpDS|CpC, -QpS|CpC, 6.9-PSEQW|CpC, -PSEQcDS|CpC, null-TNS1|CpC, –PSEQpW|CpC, –PSEQpcDS|CpC, -TPS2|CpC, –PSEQpcS|CpC]
	2	[null-null-null-TYPETYPES|CpCp2C, NNS-:-CD-TTS|CpCp2C, PPSEQ|CpCp2C]
from	0	[.....-PCNTYPE|C, from-LEM|C, from-6.9-W-PW|C, magnitude...-NNW_CTYPE|C, 6.9-NUMBER-PNER|C, 6.9...-PW_CTYPE|C, ...-PCTYPE|C, #rom>#|C, ,-NTAG|C, IN-TAG|C, #<fro#|C, -NTYPE|C, ...-CNTYPE|C, ,-O-NNER|C, CD-PTAG|C, NO-OCCURRENCE-PATTERN|C, 6.9-PW|C, from-,-W-NW|C, -...-PPW_CTYPE|C, #m>#|C, #<f#|C, #om>#|C, #<from>#|C, #<fr#|C, ...,-NW_CTYPE|C, from-O-NER|C, from-WORD|C, #<from#|C, ,-NW|C, #from>#|C, -TYPE|C, -PTYPE|C]
	1	[from-PSEQW|CpC, PSEQ|CpC, –TYPES|CpC, -PSEQcS|CpC, 6.9-PSEQpW|CpC, -PSEQpDS|CpC, -PSEQpS|CpC, -PSEQcDS|CpC, null-TNS1|CpC, CD-IN-TS|CpC, 6.9-from-PSEQW2|CpC, –PSEQpcDS|CpC, -TPS2|CpC, –PSEQpcS|CpC]
	2	[:-CD-IN-TTS|CpCp2C, null-null-null-TYPETYPES|CpCp2C, PPSEQ|CpCp2C]
		...

6.3 taula – Erauzitako ezaugarriak BAE sinple bat erabiliz.

- Aurretik datorren hitzaren entitate izen motarako, "HITZA-ENTITATE_IZEN_MOTA-PNER|C" formatua:
  - *"-"* hitzaren *0. cliquean* → *USGS-ORGANIZATION-PNER|C*
  - *from* hitzaren *0. cliquean* → *6.9-NUMBER-PNER|C*
- Aurretik datorren hitzaren entitate izen motarako, "HITZA-ENTITATE_IZEN_MOTA-NNER|C" formatua:
  - *"-"* hitzaren *0. cliquean* → *6.9-NUMBER-NNER|C*
  - *6.9* hitzaren *0. cliquean* → *from-O-NNER|C*

6.3.3 Inferentzia

BAE etiketatzaile sekuentzialak emandako irteeraren arabera, ezagutza baseko adierak induzitu behar dira. Horretarako hurbilketa sinple bat erabiltzen dugu, non lurrikara bakoitza eta argumentu bakoitza independenteki tratatzen ditugun: lurrikara baten txio guztiak hartuta, BAEk emandako iragarpenen arabera, argumentu bakoitzerako *NoisyOr* balio altuena duen aipamen potentziala aukeratzen dugu.

NoisyOr metodoa egokia da kategorizaziorako, ereduen Konfiantza (zenbat eta probabilitate altuagoa, orduan eta puntuazio altuagoa) eta jarioa (zenbat eta aipamen gehiago etiketa baterako iragarri, orduan eta altuagoa izango da etiketaren puntuazioa) ondo orekatzen dituelako:

$$NoisyOr(a, aipamen) = 1 - \prod_{w \in W} (1 - w) \tag{6.1}$$

a argumentuaren izena da eta *aipamen* argumentuaren aipamen potentziala. BAE sailkatzaileak aipamen bakoitzari iragarpen probabilitate (edo pixu) bat ematen dio (w) argumentu bakoitzeko. Wk aipamenaren iragarpen probabilitate guztiak multzokatzen ditu, a argumenturako. Formula hau Surdeanu *et al.*-en lanetik (2012) hartu dugu. Nahiz eta oso ohikoa ez izan, kasu batzuetan *NoisyOr* balio berdina lortzen dugu bi aipamen potentzialentzat, biek balio altuena badute, orduan ausaz aukeratzen dugu emaitza.

Balio anitzeko argumentuentzat (*affected-country, affected-region*), etiketatzaile sekuentzialak iragarritako balio guztiak itzultzen ditugu. Hitz bat baino gehiago daukaten aipamenentzat, adibidez *New Zealand*, hitz guztien pixuen batezbestekoa kalkulatzen dugu.

Beste aukerak

Esperimentuetarako erabili dugun inferentzia *NoisyOr* da. Formula honetaz gain, honen aldaera desberdinak ere probatu ditugu, emaitzak hobetzeko asmoz, baina zoritxarrez hauek ez dituzte emaitzak hobetu.

Hasteko, hipotesi bat dugu, non **beranduen argitaratutako txioek informazio eguneratua eta zehatzagoa** duten. Argumentu batzuk denboran zehar aldatzen dira, informazio zehatzagoa emanez: hildako kopuru zehatzagoa, zauritu kopurua, sakontasuna eta beste hainbatetan ezagutza basearekiko oso antzekoa izanik. Inferentzia teknika hau erabiltzeko, *NoisyOr* formulan ondorengo moldaketa egin dugu, t aldagaiak aipamen potentziala zenbatgarren txioan azaldu den adierazten du, eta T aldagaiak aipamen potentziala agertzen den txio kopurua:

$$NoisyOr(a, aipamen) = 1 - \prod_{w \in W} (1 - w * \frac{t}{T}) \tag{6.2}$$

Emaitzek kontrakoa erakutsi dute: egun eta ordu okerra, magnitude eta sakonera gaizki, eta abar. Aurreko kapituluan erreplikei buruz eztabaidatu dugu, hauek nola detektatu eta kendu ditugun azalduz, baina gogoan eduki behar dugu ez ditugula denak detektatzen eta asko ezabatu gabe gelditzen direla. Nahiz eta kasu batzuetan eraginkorra izan, orokorrean sistemak induzitzen dituen erantzun gehienak errepliketatik datoz eta ez dute inola ere ez bat egiten ezagutza baseko informazioarekin.

Proposatutako beste alternatiba bat, lortutako aipamen potentzialak **argumentu horrentzako tipikoak ote diren edo ez** neurtzea da. Adibidez, magnitudea lortzeko induzitutako aipamen baten balioa 0.1 bada, hau argi eta garbi ezin da izan, orduan baztertu aipamen hau.

Inferentzia honetan, zenbakizko balioak bakarrik onartzen dituzten argumentuekin egiten dugu lan, horretarako aipamen bakoitzaren *Pvalue* balioa eskuratzen dugu. Izan bedi $\bar{x}$ argumentuaren aipamen guztien batezbestekoa, σ desbiazio tipikoa, n induzitutako aipamen kopurua, eta *aipamen* aipamena bera; aipamen bakoitzaren *Pvalue* lortu aurretik, hauen *Ztest* lortu behar dugu, hona hemen bi balioak lortzeko formulak[2]:

[2]Formula hauek R programazio lengoaiaren funtzioak erabiliz aplikatu dira. *pnorm* funtzioa banaketa normala lortzeko erabiltzen da.

$$Ztest = \frac{\bar{x} - aipamen}{\sigma + \sqrt{n}} \tag{6.3}$$

$$Pvalue = 2 * pnorm(-|Ztest|) \tag{6.4}$$

Bi modu desberdin daude *Ztest* eta *Pvalue* lortzeko: (1) induzitutako aipamen guztien balioak erabiliz, eta (2) argumentu horrek ezagutza basean dituen balio guztiak erabiliz. Hauen balioak eskuratuta, indukziorako hiru formula desberdin erabili ditugu, bi *NoisyOr* formularen bariazioak dira, eta hirugarrenak *Pvalue* altuena duen aipamena itzultzen du:

$$NoisyOr(a, aipamen) = 1 - \prod_{w \in W} (1 - w * Pvalue) \tag{6.5}$$

$$NoisyOr(a, aipamen) = 1 - \prod_{w \in W} (1 - Pvalue) \tag{6.6}$$

$$max(Pvalue) \tag{6.7}$$

Zoritxarrez, emaitzak ez doaz hobera. Kasu askotan urruneko gainbegiraketa sistemak hobekuntzak ditu, baina aldi berean eskuzko etiketatzearena okerrera doa, edo alderantziz. Ez dago kasu bakar ere ez non biak hobera doazenik. Saiatzen gara muturreko aipamenak (*outliers*) kentzen, baina alferrik.

Pvalue aplikatzearen porrotaren ondoren, **aipamenen arteko antzekotasunaren konparaketan** oinarritzen diren heuristiko batzuk probatu ditugu, bai aipamen potentzialen arteko antzekotasunean, baita urre patroiko (ezagutza baseko) balioen arteko antzekotasunean. Esperimentu hauek ere zenbakizko argumentutan aplikatu dira. Ondorengo formulan, a uneko argumentua da, *aipamen* da aipamen potentziala, F aipamen potentzialen zerrenda eta n aipamen potentzial kopurua, edo beste era batera esanda, Fren tamaina.

$$SimSum(a, aipamen) = \frac{\sum_{f \in F}(1 - \frac{|aipamen - f|}{f}) * freq(aipamen)}{n} \tag{6.8}$$

*Pvalue*rekin bezala, *SimSum*-ekin ere *NoisyOr*en hiru bariazioekin egin ditugu probak, aurretik aipatutako bi formula berdinekin, *Pvalue* ordez *SimSum* erabiliz. Esperimentu hauek ere porrot egin dute, nahiz eta muturreko datuak kenduta ere probak egin.

Honen alternatiba bezala, beste formula sinpleago batekin saiatu gara, non aipamen potentzial bakoitza aipamen guztien (bai potentzialak bai urre aipamenak) bataz bestekoarekin konparatzen dugun. Ondorengo formulan, $\bar{x}$ aipamen guztien batezbestekoa da, *aipamen* aztertzen ari garen aipamena, eta n aipamen kopurua:

$$Sim(aipamen) = 1 - \frac{aipamen - \bar{x}}{n * \bar{x}} \tag{6.9}$$

Honekin ere *NoisyOr* formularen aldaera desberdinak aplikatu ditugu, baina emaitzak ez dira oraingoan ere hobetu. Hainbat porroten ostean, *NoisyOr* formula tradizionala erabiltzea erabaki dugu aipamenak induzitzeko.

6.3.4 Oinarrizko emaitzak

Gure sistemaren emaitzak ebaluatzeko, *Slot Filling* (ikus 2.1.2 atala) atazaren ebaluatzaile ofiziala erabili dugu doitasuna, estaldura eta F1 neurria kalkulatzeko. Ebaluatzaile honek sistemak induzitutako aipamenak ezagutza basekoekin konparatzen ditu.

Ebaluatzaile honek baliokidetasun klaseak onartzen ditu, honi esker ezagutza basean *region* argumentuarentzat *Guerrero* badugu eta sistemak *Guerrero State* induzitu badu, emaitza hau ere ontzat hartzen dugu baliokidetasuna dela eta. Twitterren hainbestetan erabiltzen diren traolak (#) ere kontuan hartzeko erabili ditugu klase hauek, traolak dituzten estatu (*#Japan*), herrialde eta hiriak ere ontzat hartzeko.

Ezagutza basea potentzialki osatu gabe dago, hau da, lurrikara konkretu batzuentzat ez dago argumentu batzuei buruzko inongo informaziorik; induzitutako aipamenen artean, batzuk argumentu hauei badagokie, hauek zigortuko ditu ebaluatzaileak, ezin dugunean jakin hauek zuzenak diren edo ez. Arazo hau leuntzeko, sistemak anotatutako aipamenak eskuzko anotazioekin konparatzen ditugu. Ataza honetarako, CoNLL 2003ko *Named Entity Recognition*[3] atazaren ebaluatzaile ofiziala erabili dugu. Ebaluatzaile honek ez ditu induzitutako aipamenen kalitatea neurtzen, etiketatze sekuentzialaren kalitatea txioak anotatzean baizik. Nahiz eta induzitutako argumentu baten aipamena ezagutza basean ez egon, eskuz argumentu horrekin etiketatu bada, ziur egon gaitezke sistemaren indukzio metodoa ona dela eta txioaren testuinguruarekin bat egiten duela.

[3] Atazaren webgune ofiziala: http://www.cnts.ua.ac.be/conll2003/ner/

		Zehaztasuna	Doitasuna	Estaldura	F1
Aipamen etiketatzea	**EE-BAE**	**93.52**	**92.88**	**58.53**	**71.81**
	UG-BAE	89.83	90.39	33.74	49.13
Argumentuen betetzea	**EE-BAE**	-	44.13	**26.14**	**32.83**
	UG-BAE	-	**53.05**	21.97	31.07

6.4 taula – Garapena: Aipamen etiketatzerako emaitzak CoNLL ebaluatzailea erabilita (lehenengo bi errenkadak), eta argumentuen betetzearen emaitzak KBP ebaluatzailea erabilita (azken bi errenkadak), EE-BAE eskuzko etiketatzearen entrenamenduarentzat eta UG-BAE urruneko gainbegiraketaren entrenamenduarentzat.

6.4 taulan ditugu garapenerako erabilitako datu multzoen emaitzak, urruneko gainbegiraketan oinarritutako gertaera erauzketa sistemarentzat (UG-BAE), eta eskuzko etiketatze sistemarentzat (EE-BAE). Aipamenen etiketatzean, urruneko gainbegiraketa sistemak estalduran errendimendua galtzen du. Argumentuen betetzerako berriz, nahiz eta eskuzko etiketatzeak F1 altuagoa izan, urruneko gainbegiraketak doitasun gehiago dauka (baina estaldura txikiagoa), urruneko gainbegiraketa sistemak eskuzko etiketatzeak baino aipamen gutxiago itzultzen dituelako (328 *vs* 469). Eskuzko etiketatzearen doitasun faltak bere arrazoia dauka: datu multzoan ezagutza basean ez dauden hainbat aipamen daude etiketatuta, hau dela eta iragarritako aipamen asko faltsuak dira, eta ondorioz doitasuna jaisten da. Azpimarratzekoa da zein F1 antzekoa hartzen duten biek. Eszenatoki honetan eskuzko anotazioak ez du inbertitutako esfortzua justifikatzen. Ebaluazio erlaxatuarekin ikusiko dugu hau ez dela guztiz horrela (ikus 6.5).

Aipamenen etiketatzea aztertuz, urruneko gainbegiraketa bidez induzitutako datu multzoak orokorrean aipamenak zuzen etiketatzen dituela ikusten dugu, baina nahiz eta eskuzkoak bezain beste doitasun eduki, estaldura handia galtzen du. Urruneko gainbegiraketa bidezko etiketatze prozesua oso zorrotza da, ezagutza baseko informazio zehatza bakarrik etiketatzen du, eskuzkoak berriz lurrikarari buruzko informazio guztia dauka etiketatuta.

Argumentuen betetzeak argumentuz argumentu lortutako emaitzak jakin nahi izatekotan, H eranskineko H.1 atalean aurki ditzake irakurleak ebaluaketa honen emaitzak, H.1 taulan. Orokorrean, eskuzko etiketatzeak urruneko gainbegiraketak baino emaitza hobeak lortzen ditu argumentuz argumentu, salbuespen bakarrak magnitudeak eta tsunamiak dira. Eskuzkoan koordena-

tu geografiko, zauritu eta sakontasunari buruz aipamen gehiago etiketatuta daudenez, urruneko gainbegiraketaren bidez lortu ez dugun informazioa eskuratu ahal izan dugu. Desberdintasun gutxi dago eskuzko etiketatzearen eta urruneko gainbegiraketaren artean, kontuan hartu behar dugu ezagutza basearen aurka egiten dugula ebaluaketa, hori dela eta, nahiz eta eskuzkoak induzitutako balioak txioaren testuinguruarekiko zuzenak izan, ezagutza basearekin bat egiten ez badute ez dira zuzentzat hartuko, hori gertatu da magnitudeekin adibidez, ezagutza basearekiko antzeko balioak induzitu dira, hamartar batzuen desberdintasunarekin (adibidez 7.2 *vs* 7.3), horregatik lortu dugu emaitza txarragoa magnitudeentzat eskuzkoan.

6.4 Errore analisia

UG-BAE sistemaren mugak ulertzeko eta hauei aurre egiteko, ondorengo analisiak egin ditugu eskuz:

- Urruneko gainbegiraketarako etiketatze automatikoaren kalitatea
- EE-BAE eta UG-BAE sistemek induzitutako aipamenen kalitatea

6.4.1 Urruneko gainbegiraketako etiketatzearen analisia

Urruneko gainbegiraketaren etiketatze prozesuak 13562 aipamen ekoiztu ditu. Etiketatze prozesuak hitz batzuk gaizki etiketatu ditzake (positibo faltsuak) eta aipamen egokiak alde batera utzi (negatibo faltsuak).

Analisi honetan, CoNLL ebaluatzailearen bidez, urruneko gainbegiraketaren etiketatzearen kalitatea eskuzko etiketatzearen aurka ebaluatzen dugu. Ebaluazio honetan ez dugu BAE sailkatzailerik entrenatu. 6.5 taulan ikus dezakegu ebaluazio honen emaitza, gure sistemaren ebaluaketa prozesuak lan ona egiten duela argi ikusten da, eta etiketatutako informazio gehiena zuzena da, baina garrantzitsua den informazioa etiketatzean huts egiten du.

Positibo faltsuak

Positibo faltsuak doitasunaren akatsetan oinarritzen dira, 6.5 taulan ikusten da oso doitasun altua dugula, hala eta guztiz ere komeni da jakitea zeintzuk izan diren hanka sartzeak.

Urruneko gainbegiraketa *vs* **eskuzko etiketatzea**			
Zehaztasuna	**Doitasuna**	**Estaldura**	**F1**
95.79	97.36	72.55	83.14

6.5 taula – Urruneko gainbegiraketaren etiketatzearen ebaluaketa, eskuzko etiketatzearen aurka, BAErik erabili gabe. Garapen fasea.

Positibo faltsuen artean, 3 bereizketa egiten ditugu, sakonago aztertzeko:

1. **Eskuzko etiketatzean etiketarik ez duten hitzak, baina urruneko gainbegiraketan bai:** Hauek positibo faltsuen %2a dira, tsunamiak etiketatzean gertatu da gehienbat, ezagutza basean tsunami bat egon dela zehazten delako etiketatu da, baina txioko testuinguruak ez du esaten egon denik. Normalean tsunami alertak dira, hauetako batzuk azkenean gertatzen dira eta beste batzuk ez. Beste kasuak hildako, zauritu, sakonera edo aurrelurrikarenak izan dira.

2. **Bi etiketatze prozesuetan etiketatutako hitzak, baina etiketek ez dutenean bat egiten:** Hauek positibo faltsuen beste %2a izan dira, eta kasu guztietan zenbakizko balioekin gertatu da, adibidez eskuzkoan hildako kopurua zehazten duelako dagokion etiketa jarri zaiolako, baina ezagutza basean balio horrek sakonera adierazten du eta urruneko gainbegiraketan sakonera bezala espezifikatu delako.

3. **Eskuzko etiketatzean etiketa bat duten hitzak, baina urruneko gainbegiraketak ez duenean etiketarik jarri:** Azken kasu hau positibo faltsuen %96ari dagokio, testuinguruaren arabera etiketa hori jarri behar dugulako baina ezagutza-baseak eta hitzak ez dutelako bat egiten. Kasu ohikoenak hildako kopuruak eta magnitudeekin izan dira. Zaurituak, egunak, orduak eta koordenatu geografikoak dira beste hainbat kasu.

Emaitza hau garrantzitsua da: erlazio erauzketan ez bezala, non arazo nagusia etiketatze automatikoetako (Riedel *et al.* 2010) positibo faltsu kopurua den (%30 inguru), frogatu dugu oraingoan huts egindako etiketatzeak direla oztopo nagusia. Desberdintasuna garbi uzteko, gertaera erauzketarako

komeni da domeinuarekin lotuta dauden txioekin lan egitea. Hau beste gertaera erauzketa domeinutara orokortzen da, hegazkin istripuetara adibidez (Reschke *et al.* 2014), non gertaeraren testuingurua zehaztasunez laburtu daitekeen hitz gako gutxi batzuekin, adibidez, hegaldi zenbakia eta eguna hegazkin istripuetarako.

6.1 taulako argumentuen arteko desberdintasunak emaitza hauek konfirmatzen dizkigu. Orokorrean, urruneko gainbegiraketarekin lortutako aipamen kopurua eskuzkoan baino askoz txikiagoa izanik, argumentu "hedatuenetan" kontzentratzen da, aipamen gutxi dituzten argumentuetan baino, hildakoentzat adibidez, horregatik lortzen da estaldura txikia UG-BAErentzat. Orain arteko erlazio erauzketa eredu gehienak, hala nola Hoffmann *et al.*-ena (2011), Surdeanu *et al.*-rena (2012), edo guk tesi honetan orain arte egindakoak bezala, positibo faltsuek sartutako zarata helbideratu dugu hobekuntzak lortzeko asmokin; gure oraingo lana berriz, negatibo faltsuen kantitatea murriztean datza hobekuntzak lortzeko.

Negatibo faltsuak

Negatibo faltsuak estalduraren akatsetan oinarritzen dira, 6.5 taulan ikusten da estaldura altua dugula, baina nahiko baxua doitasunarekin konparatuta.

Ulertzeko zergatik dauden hainbeste negatibo faltsu, ausaz aukeratutako 100 txio analizatzen ditugu eskuz. Bertatik, urruneko gainbegiraketak etiketatu ez dituen 46 aipamen hartu.

24 kasutan, aipamenaren zenbakizko balioa oso antzekoa da ezagutza basean dagoenarekiko, adibidez 7.1 ordez 7.3 magnituderako.

20 kasutan, txioetako informazioa okerra zen, baliteke informazioa zaharkituta egotea (60 hildako txioan, 4000 inguru ezagutza basean), edo txiolaria gaizki informatuta egon izana, adibidez, txio batek 2010eko otsailean lurrikara bat Txilen izan zela dio, baina errealitatean Argentinan gertatu zen[4], Txileko mugatik oso gertu.

Gelditzen diren bi kasuetan berriz, ezagutza basean ez dago inongo informaziorik aipamen horri buruz, adibidez lurrikara baten sakonera ezagutza-basean ez dagoenean zehaztuta baina bai txio batean[5], ezin dugu jakin jarritako informazioa zuzena edo okerra den.

[4]Txioa: "The 8.8 - magnitude quake was centered in Chile , but was felt in Argentina http://bit.ly/c9AlKR"

[5]Txioa: "#Earthquake M 7.0 - Ryukyu Islands, Japan T20:31:27 UTC , 25.95 128.40 depth: **22 km** http://bit.ly/aq9Vxa Local tsunami alert issued"

Eskuzko etiketatzea *vs* **ezagutza basea**		
Doitasuna	**Estaldura**	**F1**
57.74	40.02	47.28

6.6 taula – Eskuzko etiketatzearen ebaluaketa, ezagutza basearen aurka, BAErik erabili gabe. Garapen fasea.

Analisi honek frogatzen du, nahiz eta guzti hau domeinu zail bateko ataza desafiatzaile bat izan, posible dugula adibide negatibo faltsu gehienak urruneko gainbegiraketa sistema batekin helbideratu, ezagutza baseko aipamenen antzeko balioak etiketatzen dituen algoritmo bat sortzen badugu.

6.4.2 Aipamenen indukzioen analisia

Analisi honek BAEek induzitutako aipamenen emaitzak aztertzen ditu. Kasu askotan, txio askok argumentu berdinerako bat ez datozen aipamenak dakartzate. Adibidez, 2010eko otsaileko Ryukyuko (Japonia) lurrikaran, 6.9 eta 7.3 arteko balioen berri ematen dute txioek[6]. Zorionez lurrikara honetarako, *NoisyOr* indukzio algoritmoak magnitude zuzena itzultzen du (7.0).

Txioetako informazioa eta ezagutza basekoaren arteko erlazioak aztertzeko, indukzio algoritmo berdina aplikatzen dugu eskuz etiketatutako entrenamenduko txioei, sailkatzailea entrenatu gabe, etiketatutako aipamen guztiak bilduz eta hauei *NoisyOr* aplikatuz[7]. Hemen gure gain hartzen dugu aipamen guztiak zuzenak direla. 6.6 taulan ikus dezakegu eszenatoki honen emaitza, non 549 aipamen induzitu diren, UG-BAE (333) eta EE-BAEk (470) baino gehiago. Doitasunari esker konfirmatzen da txioetako informazio kantitate bat okerra zela, eta zehaztasunak argi uzten du ezagutza baseko informazio guztia ez dagoela txioetan.

Aurretik espero genuen estaldura txikia lortzea, txioek orokorrean ez dituztelako aipatzen aipamen guztiak (*latitude* eta *longitude* adibidez), baina

[6]Emandako magnitudeak eta hauen maiztasunak (parentesi artean) hauek dira: 6.9 (3), 7.0 (7), 7.1 (1) eta 7.3 (1).

[7]Aipamen guztiei pixu (w) berdina ematen zaie, 0.9

	Eskuzkoa	EE-BAE	UG-BAE
Gaizki	47.8% (32)	56.7% (38)	55.2% (37)
EBrekiko antzekoak	31.3% (21)	32.8% (22)	25.4% (17)
EBn ez	19.4% (13)	10.4% (7)	19.4% (13)

6.7 taula – 67 aipamen okerren analisia txioen eskuzko etiketatzea induzituta (Eskuzkoa), eskuzko etiketatzearen BAE entrenatutik induzituta (EE-BAE), eta urruneko gainbegiraketa bidez etiketatutako BAE entrenatuaren indukzioak (UG-BAE). Parentesi artean, 67 horietatik zenbat izan diren okerrak. EB laburdurak "ezagutza basea"adierazten du.

doitasun txiki hori, UG-BAErena (ikus 6.4 taula) baino pixkat altuagoa bakarrik, ez genuen inola ere ez espero.

Analisiarekin amaitzeko, *Slot Filling* atazako ebaluatzailearekin lortutako emaitzak aztertzen ditugu, gaizki zeuden erantzunetatik ausaz 67 hartuz. Analisia eskuzko etiketatzean, UG-BAEn eta EE-BAEn egin ditugu. Erantzun oker horiek 3 kategoriatan sailkatu ditugu:

1. Guztiz gaizki daudenak.
2. Ezagutza baseko balioarekiko desberdinak direnak baina antzekoak.
3. Lurrikara horrekiko ezagutza basean agertzen ez diren argumentuak.

6.7 taulak laburbiltzen ditu hiru sistemen emaitzak. Gure estimazioen arabera, 3 sistemetarako lortu diren emaitza okerretatik, erdia guztiz gaizki daude. Gelditzen den %50etik erdia baino gehiago ezagutza basean aurki ditzakegun balioekiko oso antzekoak dira. Beste guztietarako, lurrikara konkretu batzuen argumentuen balioak falta dira ezagutza basean, ez daukagu informazio hori, ondorioz ezin dugu zehaztu ea sistemak induzitutako balioa zuzena edo okerra den.

6.5 Ebaluazio erlaxatua

Aurreko ataleko analisian ikusi dugu induzitutako aipamen asko ezagutza basearekiko oso antzekoak direla eta ebaluatzaileak txartzat hartzen dituela. Iragarritako balioei inongo punturik eman ordez, ebaluatzaileari aldaketa batzuk egiten dizkiogu, kontuan hartu ditzan sistemak iragarritako eta urrezko aipamenen arteko balioen antzekotasuna.

6.5.1 Ebaluazio erlaxatua zenbakizko balioentzat

Demagun lurrikara konkretu batentzat sistemak 7.2ko magnitudea induzitu duela, urre patroiaren arabera, hau da, ezagutza basearen arabera, 7.3 denean. Kontuan izanda balitekeela ezagutza basea gaizki egotea, edo txiolariak balioa jartzeko unean 7.2 izatea informazio ofiziala, bi balioak oso antzekoak izanda, partzialki ontzak hartu beharko genuke. Horretarako, sistemak emaitza hauei puntu bat eman ordez (orain arte, sistemak puntu bat ematen zion emaitza zuzen bakoitzari, eta zero puntu okerrei), 0 eta 1 artean egongo den bien arteko antzekotasun balioa ematen diogu.

Bi balio hauen arteko antzekotasuna neurtzeko, ondorengo formula erabiltzen dugu, non a iragarritako aipamena den eta g ezagutza baseko balioa; ekuazio honek 0 itzultzen du kenketaren emaitza 0 baino txikiagoa bada:

$$sim(a, g) = max\left(1 - \frac{|a - g|}{g}, 0\right) \tag{6.10}$$

7.3 eta 7.2 magnitudeen arteko antzekotasuna 0.98koa da; iragarritako magnitudea 14.6 balitz, orduan bien arteko antzekotasun balioa 0 izango zen.

Ebaluazio erlaxatu hau *longitude, latitude, dead, injured, missing, magnitude, depth, aftershocks, foreshocks* eta *peak-acceleration* argumentuei aplikatzen zaie.

6.5.2 Ebaluazio erlaxatua lekuentzat

Gainontzeko argumentuentzat, antzekotasun funtzioak itzulitako balioak diskretuak dira, 1 (proposatutako balioa zuzena da) edo 0 (okerra) balioak emanez.

Lekuei buruzko argumentuentzat, GeoNames[8] erabiltzen dugu tokien koordenatu geografikoak eskuratzeko, proposatutako balioak eta ezagutza basekoak bat egiten ez dutenean.

Koordenatu geografikoen bidez, bi lekuen arteko distantzia kalkulatzen dugu, *Haversine* formula[9] erabiliz. Formula hau nabigazioan erabiltzen da, esfera batean bi punturen arteko distantzia kalkulatzeko, longitudeak eta latitudeak erabiliz. Formula honetan erabiltzen diren ekuazioak ondorengoak dira:

[8] http://www.geonames.org/
[9] http://en.wikipedia.org/wiki/Haversine_formula

$$\Delta\varphi = \varphi_2 - \varphi_1 \tag{6.11}$$

$$\Delta\lambda = \lambda_2 - \lambda_1 \tag{6.12}$$

$$a = \sin^2\left(\frac{\Delta\varphi}{2}\right) + \cos\varphi_1 * \cos\varphi_2 * \sin^2\left(\frac{\Delta\lambda}{2}\right) \tag{6.13}$$

$$c = 2 * \arctan\left(\sqrt{a}, \sqrt{1-a}\right) \tag{6.14}$$

$$d = R * c \tag{6.15}$$

R lurraren erradioa da (6371 km), φ latitudea radianetan eta λ longitudea radianetan. Koordenatuak radianetara pasa behar dira ekuazioak ondo funtzionatzeko.

Estatu, herrialde eta hirien batezbestekoa tamainan oinarrituta, induzitutako tokiak ontzat hartzen ditugu urre patroikoarekiko duten distantzia ondorengoetara edo gertuago badaude: 500 km estatuentzat, 50 km estatuentzat, eta 10 km hirientzat.

Ebaluazio erlaxatu hau *country, region, city, affected-country* eta *affected-region* argumentuei aplikatzen zaie. Sistemak iragarritako *affected-country*, ezagutza basean *country*ren berdina bada, orduan okertzat hartzen da.

6.5.3 Ebaluazio erlaxatua gainontzeko argumentuentzat

Lurrikararen unearen argumentuarentzat (*time*), zuzentzat hartzen ditugu baldin eta lortutako balioa eta urre patroieko balioaren artean, gehienez 5 minutuko desberdintasuna baldin badago. Sistemak lurrikara gertatu den ordua *07:52:00* dela induzitu badu, eta ezagutza baseko balioa *07:55:00* bada, orduan emandako erantzuna ontzat hartzen da.

Lurrikaren iraupenerako berriz (*duration*), gehienez 10 segundoko desberdintasuna baldin badago hartzen dira ontzat.

Lurrikara gertatu den egunerako (*date*), proposatutako balioaren eta urre patroiko balioaren artean gehienez egun bateko desberdintasuna onartzen dugu[10].

[10]Proposatutako muga hauek beste domeinuetan aldatuko lirateke, baina hauek doitzea garrantzirik gabekoa da.

		Doitasuna	Estaldura	F1
Argumentuen betetzea	**EE-BAE**	64.18	**38.01**	**47.74**
	UG-BAE	**67.41**	27.92	39.48
	Eskuzkoa	73.42	50.89	60.12

6.8 taula – Ebaluazio erlaxatuaren emaitzak, garapeneko fasean.

Orain arte, SF atazaren ebaluatzaileak benetako positiboen (BP) balioa puntu batekin gehitu du iragarpen bat ezagutza basearekiko parekoa bazen. Guk proposatutako ebaluatzaile erlaxatuan, BPen balioa antzekotasun balioarekin gehitzen da, ondorioz, doitasuna eta estalduraren formulak ondorengoak dira (SYS iragarritako aipamenentzat, $GOLD$ urre patroiko aipamenentzat):

$$doitasuna = \frac{\sum sim(x, g)}{SYS} \tag{6.16}$$

$$estaldura = \frac{\sum sim(x, g)}{GOLD} \tag{6.17}$$

6.5.4 Emaitzak

6.8 taulan ikus ditzakegu argumentuen betetzearen emaitzak ebaluazio erlaxatua erabilita. Espero bezala, emaitzak asko igo dira lehenengo esperimentuak lortutakoekin konparatuz gero (ikus 6.4 taula). Doitasunaren desberdintasuna EE-BAE eta UG-BAEren artean jaitsi da, honek erakusten du EE-BAEk iragarritako aipamen asko ezagutza basekoaren oso antzekoak direla, UG-BAErenak ez bezala. Hala ere, estalduraren desberdintasuna oraindik oso handia da, ondorengo atalean landuko dugu hau. Oraingoan F1 neurrien desberdintasunak eskuzko anotazioaren eta automatikoaren artean nabariak dira, eskuzko etiketatzeak suposatu duen esfortzua justifikatuz. Eskuzko etiketatzeko aipamenak induzitzean, emaitzak ere asko igo dira. Emaitza hauek guk espero ditugun errendimendu sabaitik (60.12) gertu daude, argi utziz proposatutako ebaluazio erlaxatua egokiagoa dela ataza honetarako.

Eranskinetako H.2 taulan, ebaluazio erlaxatua erabili duten argumentuen emaitzak ditugu, hauek banaka ebaluatuta. Argumentu guztientzako igo dira emaitzak. Bi sistemetan asko igo da magnitudeen emaitza, igoera altu honek argi uzten du induzitutako balioak ezagutza basearekiko oso antzekoak direla,

eta desberdintasun txiki batengatik ez direla aurretik ontzat eman. Beste argumentuetan igoera txikia dago, horrek esan nahi du induzitutako balioak ez direla ezagutza baseko balioen hain antzekoak, batzuk asko hurbiltzen dira (sakontasunak ia 20 puntuko igoera eduki du), baina emaitzen bi herenak gutxienez urruti daude.

Ebaluazio hau oso bidezkoa dela frogatzen dugu hemen. Magnitudeen kasuan induzitutako balio gehienak antzekoak direnez, asko hurbiltzen dira 100eko F1 neurrira, antzekotasunik ez duten balioek eragotzi dute 100era iristea. Gainontzeko argumentuetan, nahiz eta balio asko oso antzekoak izan, gehienak asko urruntzen dira ezagutza basearen balioarekiko eta horregatik jarraitu dute F1 txikiarekin. Ondorioz, ebaluazio erlaxatuarekin antzekoak diren balioak saritzen ditugu, eta antzekoak ez direnak, orain arte bezala, ez dute inongo saririk jasotzen.

Ondorengo hobekuntza eta esperimentuetan, garrantzi handiena ebaluazio erlaxatuaren emaitzei emango diegu.

6.6 Urruneko gainbegiraketa hurbila

Aurreko atalean erakutsi dugu txio askok aipamen desberdinei buruzko datuak dituztela baina desberdinak ezagutza basearekiko. Beraz, aipamen hauek ez ditu urruneko gainbegiraketa tradizionalak etiketatzen, honek txioetako eta ezagutza baseko testuak %100 berdinak izatea eskatzen duelako. Honek esan nahi du UG-BAE entrenatu beharko dugula datu kantitate txikiekin, benetan eskuragarri ditugunak baino (6.2 ataleko adibidean, 7.1 magnitudea ez dugu eskuragarri). Atal honetan erakutsiko dugu nola hobetu ditzakegun urruneko gainbegiraketako emaitzak, ezagutza basearekiko antzekoak diren balioak etiketatuz.

Txioak etiketatzeko proposatzen dugun sistemak, lurrikara konkretu bakoitzaren txioetan, ezagutza basean duen argumentu baten balioaren antzeko hitz bat aurkitzen duen bakoitzean, aipamen hau eta ezagutza basean agertzen diren aipamenen arteko antzekotasuna neurtzen du, 6.5 atalean azaldu ditugun formula desberdinen bidez, helburu argumentuaren arabera. Antzekotasuna lortu ondoren, honek guk zehaztutako baldintzak betetzen baditu, sistemak argumentu horren etiketa jarriko dio aipamenari.

Zenbakizko aipamenen artean gertatzen dira bat etortze gehienak. Kasu batean hitz batek etiketa bat baino gehiago baldin baditu, orduan antzekotasun handiena duen etiketa aukeratzen du algoritmoak; eta etiketa guztiek

antzekotasun berdina baldin badute, orduan garapenerako datu multzoan maiztasun gehien duen etiketarekin gelditzen da.

Urruneko gainbegiraketa hurbila ebaluazio erlaxatuan parte hartzen duten argumentu berdinei aplikatzen zaie.

6.6.1 Emaitzak

Hasteko, **garapenean** lortutako emaitzak emango ditugu. 6.9 taulak ebaluazio zorrotz eta erlaxaturako emaitzak laburtzen ditu. Ebaluazio erlaxatuan urruneko gainbegiraketa hurbila erabiltzen duten sistemek hobetzera jotzen dute nabarmen, batez ere estaldura aldetik (27.92 *vs* 37.60). Sistema berriak entrenamendurako aipamen gehiago ditu etiketa bakoitzeko, hori dela eta, zenbakizko argumentu bakoitzarentzat aipamen potentzial gehiago induzitzen dira, hauen artean lehia handiagoa da eta hobeto asmatzen du.

6.1 irudiak garapeneko datu multzoaren doitasun/estaldura kurbak (ikus 2.6.2 atala). erakusten ditu, beste hiru atalase desberdinen emaitzekin batera (0.97, 0.95 eta 0.5) ebaluazio erlaxaturako, argi eta garbi ikus daiteke hiru atalaseak UG-BAE oinarrizko sistemak baino askoz hobeto dabiltzala.

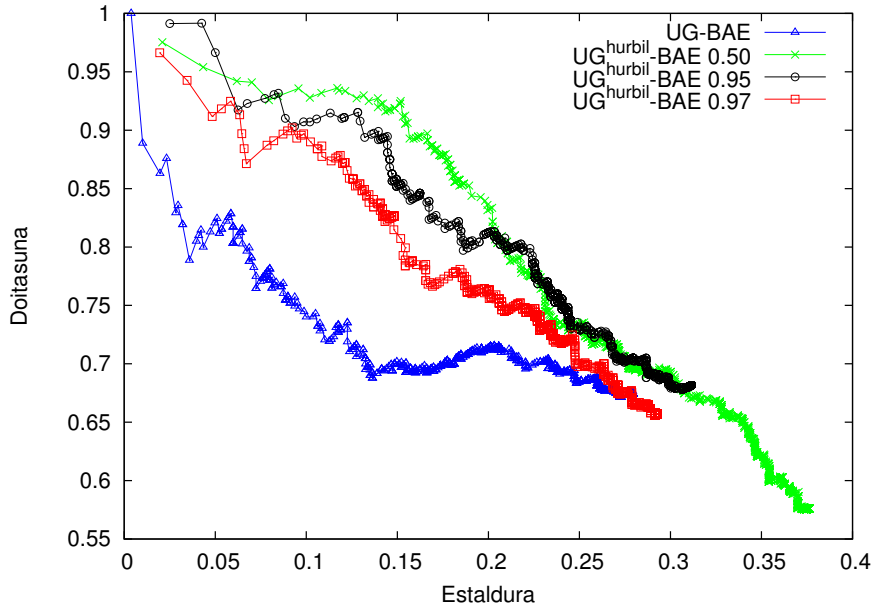
Garapen faseko UG^{hurbil}-BAE 0.95 eta UG-BAEren emaitzak ikusteko argumentuz argumentu, bi ebaluazioekin, ikus H.2 eranskineko taulak. Ebaluazio zorrotzean, H.3 taula, ikus dezakegu orokorrean argumentu gehienek igoera txikiak eduki dituztela; baina beste batzuetan, hirietan, tsunamietan, eta denboran, jaitsiera nabarmenak egon dira. Denborarentzako erorketa handia egon da, 16 puntu ingurukoa.

Ebaluazio erlaxatua aplikatuz gero, H.4 taulan ikus dezakegu nola zenbakizko argumentu gehienen emaitzak gora egin duten, zaurituenak ezik. Oraingoan induzitutako balioak aurretik emandakoak baino askoz antzekoagoak izanik ezagutza basearekiko, magnitudea 100% puntuetatik oso gertu dago, eta lehenengo aldiz, urruneko gainbegiraketan, sakontasunari buruzko emaitza batzuk lortzen ditugu, hauek zero absolutua izan gabe.

Joera positibo hau **testeko** datu multzoak baieztatzen du. 6.10 taulak testeko emaitzak erakusten dizkigu, hauek erakusten dute oinarrizko sistemarekiko bi ebaluazio teknikek erakutsitako hobekuntzak estatistikoki esanguratsuak direla, erlaxatuan igoera handia edukiz estalduraren aldetik (21.35 *vs* 32.73) eta ondorioz F1 neurrian (32.54 *vs* 42.15). 6.2 irudian aurki ditzakegu dagokien doitasun/estaldura kurbak, hauek ere erakusten dute hobekuntza hauek estaldura puntu guztiekin lortu direla, erlaxatuan eta zorrotzean.

GARAPENA	Zorrotza			Erlaxatua		
Sistema	Doit.	Est.	F1	Doit.	Est.	F1
UG-BAE	**53.05**	21.97	**31.07**	67.41	27.92	39.48
UG^{hurbil}-BAE 0.97	45.89	20.45	28.29	65.71	29.29	40.52 †
UG^{hurbil}-BAE 0.95	48.07	21.67	30.15	**68.15**	31.15	42.75 †
UG^{hurbil}-BAE 0.5	34.23	**22.35**	27.04	57.60	**37.60**	**45.50** †

6.9 taula – Garapen faseko oinarrizkoa (UG-BAE) eta UG hurbila (UG^{hurbil}-BAE) emaitzak, ebaluazio zorrotz eta erlaxatuarekin. † agertzen diren emaitzek oinarrizko sistemarekiko hobekuntzak estatistikoki esanguratsuak direla adierazten dute ($p < 0.05$).



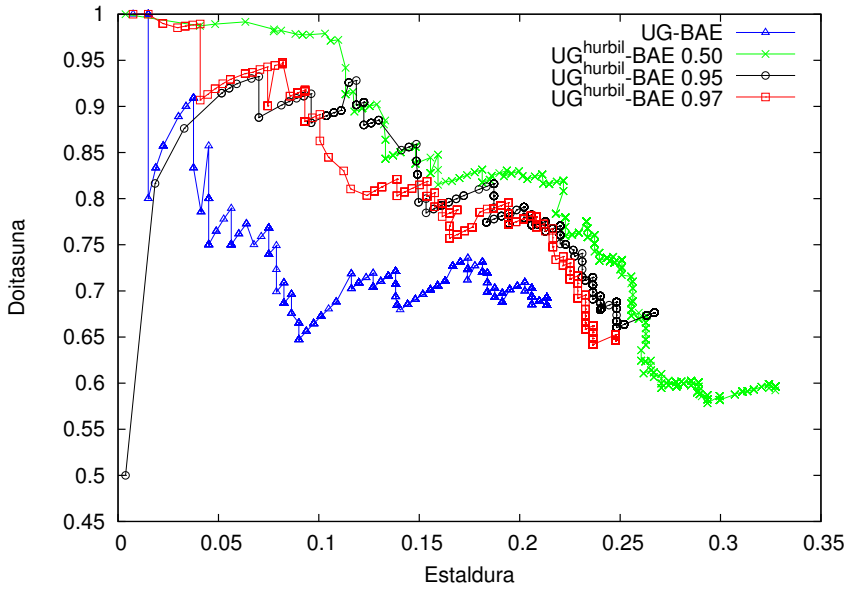
6.1 irudia – Garapeneko doitasun/estaldura kurbak ohiko urruneko gainbegiraketarako eta etiketatze hurbilerako.

6.7 Ezaugarrien agregazioa

UG-BAEk induzitutako emaitzak aztertzean (6.4.2 atala), gertaera erauzketa kolokan jartzen duen txio anbiguo bat aurkitu dugu. Txio hori Peruko lurrikara bati buruzkoa da, baina BAEk "Indonesia" induzitu du (letra lodiz induzitutako estatua, azpimarratuta estatu zuzena):

TESTA	Zorrotza			Erlaxatua		
Sistema	**Doit.**	**Est.**	**F1**	**Doit.**	**Est.**	**F1**
UG-BAE	**50.60**	17.79	24.07	68.42	21.35	32.54
UG^{hurbil}-BAE 0.97	47.06	18.04	26.08	64.60	24.77	35.81 †
UG^{hurbil}-BAE 0.95	46.67	**18.42**	**26.41**	**70.75**	27.85	39.95 †
UG^{hurbil}-BAE 0.50	33.33	**18.42**	23.73	59.22	**32.73**	**42.15** †

6.10 taula – Test faseko oinarrizkoa (UG-BAE) eta UG hurbila (UG^{hurbil}-BAE) emaitzak, ebaluazio zorrotz eta erlaxatuarekin. † agertzen diren emaitzek oinarrizko sistemarekiko hobekuntzak estatistikoki esanguratsuak direla adierazten dute ($p < 0.05$).



6.2 irudia – Testeko doitasun/estaldura kurbak ohiko urruneko gainbegiraketarako eta etiketatze hurbilerako.

- DTN **Indonesia**: Peru Earthquake Destroys Homes, Injures 100 (...)

Txioetan testuinguru falta handia dagoela eta gertatu anbiguotasun kasu hau, eta ziurrenik beste hainbat ere. Hau konpontzeko, erlazio erauzketako lan batzuetan inspiratutako metodo batzuekin egin dugu lan, non entitate nagusiaren aipamenak partekatuta sailkatzen diren, aipamen horien ezaugarriak partekatuz. Hau Mintz *et al.*-en lanean ere egiten dute.

Ideia hau honela egokitzen diogu gure etiketatze sekuentzialezko gertaera erauzketa sistemari:

1. Ezaugarrien espazioan gehiegizko egokitzea saihesteko, aipamen potentzial izateko aukera gehien eta anbiguotasuna eduki dezaketen hitzak identifikatzen ditugu, hala nola leku, egun eta denbora entitateak (lurrikara gertatutako unea eta iraupena). Hauek argumentu mota bat baino gehiagorekin sailkatzeko arriskua dago, lekuzko entitate batek adibidez "hiri", "estatu" eta antzeko argumentu baten marka jasotzeko arriskua dagoelako.

   Zenbakizko entitateak alde batera utzi ditugu, argumentu desberdinen arteko ezaugarri potentzialen arteko talka saihesteko. Adibide bezala, 2012an Zohanen gertatutako lurrikaran, magnitudea eta sakontasuna 5.6koak izan ziren. Hasierako esperimentuek hipotesi hau baieztatu dute: garapen fasean ezaugarrien agregazioak ez ditu emaitzak hobetu zenbakizko argumentuekin.

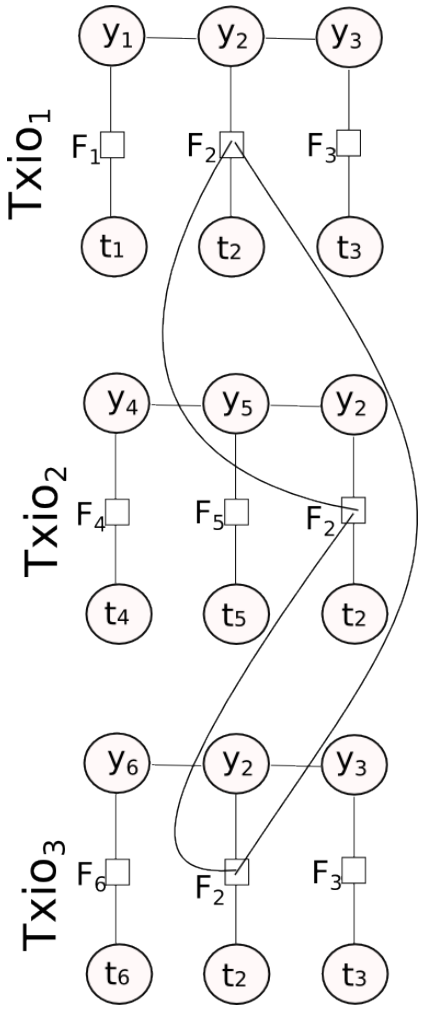
2. Lurrikara bakoitzeko eta hitz bakoitzeko, goiko puntuko baldintzak betetzen dituzten hitzak hartu, eta hauen ezaugarriak partekatzen ditugu lurrikara berdinari dagokien txio guztien artean. Adibide bezala, har ditzagun ondorengo txioak, atal honen hasieran jarri dugun txioarekin batera:

   - 6.9 magnitude earthquake rocks Peru.
   - U.S.G.S. reports 6.9 Earthquake in Peru. NO TSUNAMI threat to Hawaii.

   Baldintza hauekin, ziurtasun handia dugu lekuak adierazten dituzten hitzetan aplikatzen dela ezaugarrien agregazioa. Efektu berdina lortzen dugu epe eta denbora adierazpen bezala markatu diren entitateekin.

   *Peru*ri buruzko 3 aipamenak partekatutako ezaugarri berdinekin sailkatuko dira, adibidez *previous-word* ezaugarriarentzat ondorengo 3 balioak edukiz: *:, rocks* eta *in*. Paradigma honek %60 ezaugarri gehiago sortu ditu UG-BAE oinarrizko sistemarekin alderatuta.

6.3 irudian ikus dezakegu nola antolatzen den BAEa ezaugarrien agregazioa erabiliz. Adibide honetan, t_2 hitza txio guztietan agertzen da, eta honek entitate izen mota bat daukala zehaztu da aurreprozesaketan; hitz



6.3 irudia – BAEaren egitura ezaugarrien agregazioa erabiliz.

honi ezaugarrien agregazioa aplikatu zaionez, txio bakoitzean dituen ezaugarriak elkarri lotuta daude. Ikusten denez, ezaugarrien agregazioa aplikatzeko ez da beharrezkoa txio bat baino gehiagotan egotea, t_3 hitza adibidez bi txiotan agertzen da, eta ez zaio aplikatu, honen entitate izen mota *"O"* delako, ondorioz ez ditu ezaugarrien agregazioaren baldintzak betetzen.

6.7.1 Emaitzak

6.11 taulako lehenengo errenkadek **garapeneko** ezaugarrien agregazioen eredua laburtzen dute. Nabarmendu beharra dago ezaugarrien agregazioak (UG^{agreg}-BAE) doitasunean lortzen duen hobekuntza, bai zorrotzean (53.05 *vs* 60.48), bai erlaxatuan (67.41 *vs* 74.06). Hobekuntza hauek ere sendoak dira estalduran bi ebaluazio metrikentzat (21.97 *vs* 25.12 zorrotza eta 27.92 *vs* 30.76 erlaxatua) eta F1 neurriarentzat (31.07 *vs* 35.50 zorrotza eta 39.48 *vs* 43.47 erlaxatua).

Ebaluazio erlaxatuaren emaitzek argi adierazten dute ezaugarrien agregazioa osagarria dela etiketatze hurbilarekiko, biak erabiltzen dituen urruneko gainbegiraketa ereduak askoz hobeto jokatzen du bakoitza bere kabuz erabiltzen dituen ereduak baino, hobekuntza estalduran oso nabaria da (27.92 *vs* 40.02), ondorioz F1 neurrian ere (39.48 *vs* 48.84). Zorrotzean berriz, konbinaketak huts egin du, baina erakutsi dugun bezala, ebaluazio erlaxatua egokiagoa da. 6.4 irudiak doitasun/estaldura kurbak ebaluazio erlaxaturako, sistema hauetan hobekuntza estaldura maila guztietan dagoela argi utziz.

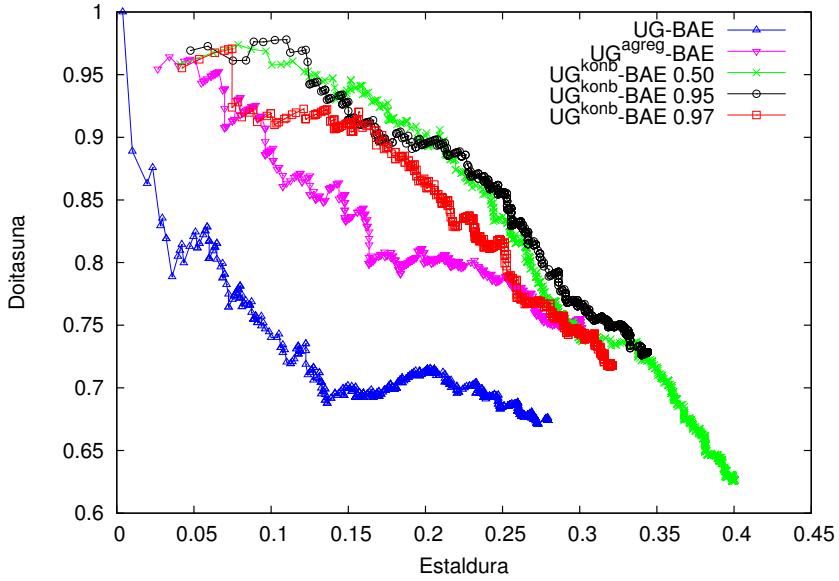
Garapen faseko UG^{agreg}-BAE eta UG-BAEren emaitzak ikusteko argumentuz argumentu, bi ebaluazioekin, ikus H.3 eranskina. H.5 taulan ebaluazio zorrotzaren emaitzak ditugu, ezaugarrien agregazioa erabiltzen duen sistema eta oinarrizko sistemaren emaitzak konparatuz. Argumentu gehienen emaitzak igo dira, hirietarako igoera aparta eduki dugu adibidez, eta herrialdeak %100era asko hurbiltzen dira. Lurrikararen orduarentzat igoeratxo bat egon da, egunarentzat berriz jaitsiera txiki bat. Hemen ikusten da ezaugarriak agregatuz BAEk hobeto tratatzen dituela leku entitateak.

H.6 taulan berriz, argumentuen emaitzak ditugu ebaluazio erlaxatua erabiliz. Ebaluazio zorrotzean, egunarentzat emaitza txikiagoa lortu dugu agregaziorako, baina oraingoan emaitza hobea lortu dugu, nahiz eta oso nabaria ez izan. Denbora berriz, mantendu da, baina urruneko gainbegiraketa tradizionalak aurre hartu dio gutxigatik ezaugarrien agregazioa erabiltzen duen sistemari. Desberdintasun gutxi dago zenbakizko argumentuen artean, non zaurituen emaitza berdina izan den, eta magnitudeak ia parekoak dira. Honen bidez ondorioztatu dezakegu ezaugarrien agregazioa ondo datorkigula batez ere lekuei buruzko entitateak desanbiguatzeko, eta onuragarria izan da denbora adierazpenentzat.

UG^{konb}-BAE 0.95enak jakiteko, ikus H.4 eranskina. H.7 taula aztertuz gero, oraingoan herrialdeei buruzko argumentua inoiz baino gertuago dago %100etik. Egunentzat eta eskualdeentzat igoera izan dugu ere. Hala ere,

GARAPENA	Zorrotza			Erlaxatua		
Sistema	**Doit.**	**Est.**	**F1**	**Doit.**	**Est.**	**F1**
UG-BAE	53.05	21.97	31.07	67.41	27.92	39.48
UG^{agreg}-BAE	**60.48**	25.12	**35.50** †	**74.06**	30.76	43.47 †
UG^{konb}-BAE 0.97	52.54	23.48	32.46	71.78	32.08	44.34 †
UG^{konb}-BAE 0.95	53.74	**25.38**	34.47 †	72.84	34.40	46.73 †
UG^{konb}-BAE 0.50	39.32	25.13	30.66	62.64	**40.02**	**48.84** †
EE-BAE	44.13	26.14	32.83	64.18	38.01	47.74

6.11 taula – Garapen faseko oinarrizko urruneko gainbegiraketa (UG-BAE), urruneko gainbegiraketa ezaugarrien agregazioarekin (UG^{agreg}-BAE), eta ezaugarrien agregazioa etiketatze hurbilarekin konbinatzen duen eredua (UG^{konb}-BAE 0.95), ebaluazio zorrotz eta erlaxatuarekin. † ikurrak hobekuntza UG-BAErekiko estatistiko esanguratsua dela adierazten du ($p < 0.05$). BAEren eskuzko etiketatzearen entrenamenduaren emaitzak (EE-BAE) jarri ditugu errendimendu sabaia adierazteko.



6.4 irudia – Garapeneko doitasun/estaldura kurbak ohiko urruneko gainbegiraketarako, ezaugarrien agregaziorako, eta ezaugarrien agregazioa etiketatze hurbilarekin konbinatzen dituen eredurako.

denbora eta hirientzat jaitsierak egon dira. Zenbakien kasuan, hildakoentzat igoera handiagoa egon da, magnitudeak igoera gutxi izan du berriz. Emaitza hauek ikusita, esan genezake ezaugarrien agregazioak urruneko gainbegiraketa hurbilarekin konbinatzeak ez duela onurarik ekartzen, baina ebaluazio erlaxatuaren bidez frogatzen dugu benetan onuragarria dela. Dagoeneko gehienezko errendimendu sabaiaren %88ra iritsi gara ataza honetan (ikus 6.8 taula, *"eskuzkoa"*), errendimendu sabaia BAEk entrenatutako eskuzko etiketatzearekin lortzen dugu (EE-BAE).

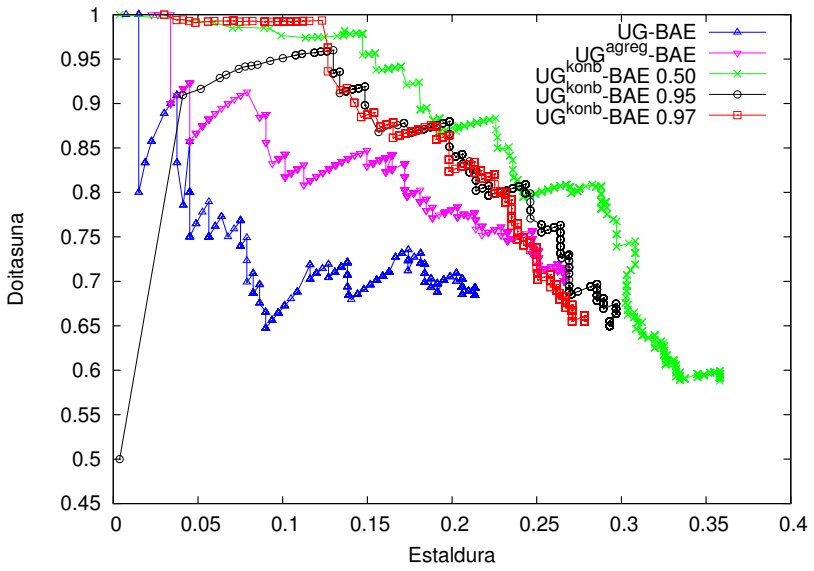
Ebaluazio erlaxatuko emaitzak zenbakizko argumentuentzat H.8 taulan daude, hemen ere emaitzek gora egin dute orokorrean, egunentzat 18 puntutako igoera edukiz adibidez, eta eskualdeentzat 20 puntutakoa. Bukaerako F1 neurritik desberdintasun handia dago oraingoan.

Ezaugarrien agregazioaren eragin positiboa **testeko** datu multzoan ikus dezakegu. 6.12 taulak estatistikoki esanguratsuak diren hobekuntzak erakusten dizkigu ebaluazio zorrotzerako eta erlaxaturako, bai estalduran (21.97 *vs* 25.38 zorrotzerako, eta 27.92 *vs* 40.02 erlaxaturako) bai F1 neurrian (31.07 *vs* 35.50 zorrotzerako, eta 39.48 *vs* 48.84 erlaxaturako). Ezaugarrien agregazioari esker, doitasuna asko igo da zorrotzean (50.60 *vs* 55.44), erlaxatuan berriz ez da oso nabarmena izan (68.42 *vs* 70.10). 6.5 irudian ikus ditzakegu testeko doitasun/estaldura kurbak ebaluazio erlaxaturako, bertan ikus daiteke UG^{agreg}-BAE sistemaren hobekuntza estaldura hobeari esker dela, eta estaldura puntu gehienentzat, doitasuna UG-BAE sistemarena bezain altua dela.

Ebaluazio erlaxatuarekin bakarrik, %19ko hobekuntza lortzen dugu. Berriz ere, bi hobekuntzak bateratuz banakako ereduak baino askoz hobeto dabil. Guztia kontuan hartuta, eredu konbinatuak F1 erlaxatua %33ko hobetzea dakar UG-BAE oinarrizko sistemarekiko.

TESTA	Zorrotza			Erlaxatua		
Sistema	**Doit.**	**Est.**	**F1**	**Doit.**	**Est.**	**F1**
UG-BAE	50.60	15.79	24.07	68.42	21.35	32.54
UG^{agreg}-BAE	**55.44**	**21.95**	**30.52** †	**70.10**	26.62	38.58 †
UG^{konb}-BAE 0.97	50.44	21.43	30.08 †	65.53	27.84	39.07 †
UG^{konb}-BAE 0.95	47.50	21.23	29.53 †	69.20	31.22	43.03 †
UG^{konb}-BAE 0.50	35.80	21.80	27.10 †	58.80	**35.81**	**44.51** †
EE-BAE	47.65	26.69	34.21	66.86	37.45	48.01

6.12 taula – Testeko oinarrizko urruneko gainbegiraketa (UG-BAE), urruneko gainbegiraketa ezaugarrien agregazioarekin (UG^{agreg}-BAE), eta ezaugarrien agregazioa etiketatze hurbilarekin konbinatzen duen eredua (UG^{konb}-BAE 0.95), ebaluazio zorrotz eta erlaxatuarekin. † ikurrak hobekuntza UG-BAErekiko estatistiko esanguratsua dela adierazten du ($p < 0.05$). BAEren eskuzko etiketatzearen testeko emaitzak (EE-BAE) jarri ditugu errendimendu sabaia adierazteko.



6.5 irudia – Testeko doitasun/estaldura kurbak ohiko urruneko gainbegiraketarako, ezaugarrien agregaziorako, eta ezaugarrien agregazioa etiketatze hurbilarekin konbinatzen dituen eredurako.

6.8 Ondorioak

Kapitulu honetan, urruneko gainbegiraketan oinarritutako gertaera erauzketa sistema bat garatu dugu. Gure sistemak gertaera konplexuak erauzten ditu, 20 argumentu desberdinei buruzko informazioa emanez.

Hasteko, urruneko gainbegiraketa bidez automatikoki etiketatutako datu multzoa analizatu dugu, eskuzko etiketatzearekin konparatuz. Estimatzen dugu urruneko gainbegiraketak etiketatutako %83a testuinguruarekiko zuzena dela. Eskuzko etiketatzea ezagutza basearen aurka ebaluatzean, %47ak bakarrik egiten du bat urre patroiarekin.

Ikusi dugu ere iragarritako informazioaren %55ak, ez duenean ezagutza basearekin bat egiten, honekiko oso antzekoa dela. Baita iragarritako erantzunen %20ak ez duela ezagutza basean inongo informaziorik argumentu horri buruz, txioetan ezagutza basean agertzen ez den informazioa topatu dezakegula frogatuz. Horretarako, ebaluazio erlaxatu bat proposatzen dugu, non antzekoak diren iragarpenak partzialki zuzentzat hartzen diren. Ebaluazio erlaxatuak emaitzak hobetzen ditu, espero bezala, eta sistemaren ebaluazioa errealistagoa bihurtzen du, batez ere kontuan hartuta ezagutza baseko informazioa batzuetan ez dela zehatza.

Urruneko gainbegiraketarako aldaera berri bat proposatu dugu ere, non ezagutza basearekiko antzekoak diren txioetako aipamenak ere etiketatzen ditugun; honi "etiketatzen hurbila" deitu diogu. Gainera, ezaugarrien agregazio global bat aplikatu dugu, hauek aberasgarriagoak izateko, lurrikara berdinekoak diren txioen artean informazioa elkarbanatuz. Gure emaitzek %19ko hobekuntza esanguratsua dute. Horrez gain, ekarpen hauek osagarriak dira: etiketatze hurbila eta ezaugarrien agregazioa konbinatzen dituen eredu bat erabiltzean, hauek indibidualki aplikatuta baino askoz eraginkortasun hobea lortzen da, %33ko hobekuntza lortuz oinarri lerroarekiko. Gure emaitzak eraginkortasunaren sabaiarekiko (gure sistemak eskuzko etiketatzea entrenatuz lortzen duen emaitza) %88ra hurbiltzen dira, urruneko gainbegiraketa gertaera konplexuen erauzketarako egokia dela frogatuz.

7

Ondorioak eta etorkizuneko lerroak

Lehenengo kapituluan informazio erauzketaren sarrera bat egin dugu, eta honen bi ataza nagusi definitu: erlazio erauzketa eta gertaera erauzketa; tesi hau bi ataza hauetan zentratu da ikerketak egiteko. Erlazio erauzketaren bidez, testuetatik entitateen arteko erlazioei buruzko informazioa eskuratu dezakegula komentatu dugu, hauekin ezagutza baseen edukia aberastu ahal izateko.

Erlazio erauzketarako erabiltzen den hurbilketa bat urruneko gainbegiraketa da. Hurbilketa honek erlazio erauzketan erabiltzen diren ikasketa gainbegiratu eta ez-gainbegiratuaren abantailak aprobetxatzen ditu, eta erlazio erauzketarako geroz eta gehiago erabiltzen da ezagutza baseak aberasteko. Algoritmo honen helburu nagusia corpusen eskuzko etiketatze automatikoa saihestea da, hau automatiko bihurtuz, eskuzko etiketatzeak suposatzen duen kostu eta denbora ekiditeko asmoz. Sarreran aipatu bezala, urruneko gainbegiraketaren arabera, esaldi berean agertzen diren bi entitateren artean erlazio bat zehaztuta badago ezagutza base batean, esaldi horrek erlazio hori adierazten du nola edo hala.

Zoritxarrez urruneko gainbegiraketak arazo asko ditu, eta hauek konpontzea komenigarria da, sistemen eraginkortasuna hobetzeko. Tesi honen helburua arazo gehienei konponbideak bilatzea izan da, hurbilketaren kalitatea eta sistema hauen errendimendua hobetzeko. Horrez gain, hurbilketa ia erabili ez den ataza batean aplikatu dugu: gertaeren erauzketan.

7.1 Ekarpenak

Atal honetan lan honen ekarpen nagusiak zerrendatzen ditugu, aldi berean tesian eskaini diegun kapitulua adieraziz.

Urruneko gainbegiraketaren zailtasun nagusiak aztertu ditugu (*Kapitulu guztiak*)

Ondorengo zailtasun nagusiak esperimentu desberdinen bitartez aztertu ditugu:

1. Arazoak ikasketa prozesuan erlazio batzuentzat aipamen positibo gutxi ditugulako. (*3. kapitulua*)

2. Aipamen negatiboen beharra ikasketa prozesurako. (*3. kapitulua*)

3. Sistemaren nahasmena aipamen zaratatsuak direla eta. (*4. kapitulua*)

4. Desadostasun txikiak corpuseko eta ezagutza baseko informazioaren artean, hauek elkarren artean oso antzekoak izanda. (*6. kapitulua*)

Urruneko gainbegiraketan topatu ditzakegun zarata mota desberdinak analizatu ditugu (*4. kapitulua*)

Riedel *et al.*ek garatutako datu multzoko ausazko aipamenak aztertu ondoren, hiru zarata mota desberdin kategorizatu ditugu:

1. **Testuinguru okerra:** aipamen positiboetan, elkarri erlazionatuta dauden bi entitate aipamen berean agertzen badira, baina testuinguruak ez duenean erlazio hori adierazten. Zarata mota ohikoena da.

2. **Ezagutza base osatu gabea:** testuinguruak bi entitateen artean erlazio bat adierazten duenean, baina ezagutza basean ez denez erlazio hori ageri, aipamena negatibotzat hartuz.

3. **Erlazio anitza:** ezagutza baseak bi entitateren artean erlazio bat baino gehiago markatuta dituenean, baina aipameneko testuinguruaren arabera erlazio bakarra (edo agian bat ere ez) onartzen denean.

Aipamen positiboetako testuinguru okerrak ezabatzeko heuristikoak proposatu ditugu (*4. kapitulua*)

Aipamen zaratatsuak antzeman eta ezabatzen dituzten hiru heuristiko aurkeztu ditugu, urruneko gainbegiraketa sistemen eraginkortasuna hobetzeko:

1. **Aipamen gehiegi dituzten tripletak:** aipamen gehiegi dituzten tripletek aipamen zaratatsu gehiegi izan ohi dituzte, guk atalase jakin batetik gorako tripletak ezabatu ditugu.

2. **Tuplen eta erlazioen arteko elkarrekintza informazio puntuala (EIP):** EIP baxua duten tuplek aipamen zaratatsuak edukitzeko joera dute, guk atalase jakin batetik behera dauden tripletak ezabatu ditugu.

3. **Aipamenak eta hauen erlazioen zentroideekiko antzekotasuna:** dagokien erlazioaren zentroidetik urruti dauden aipamenak zaratatsuak dira, atalase jakin bat baino urrutiago dauden aipamenak ezabatu ditugu.

Hiru heuristiko hauen konbinaketa batek Stanford Unibertsitatean garatutako bi oinarri lerro trinko garaitu dituzte.

Urruneko gainbegiraketarako ebaluazio erlaxatu bat proposatu dugu (*6. kapitulua*)

Ebaluazio erlaxatu bat proposatu dugu, non iragarritako balioak urre patroiarekiko oso antzekoak badira, hauek partzialki zuzentzat hartzen ditugun. Ebaluazio honek erlazio erauzketa sistemen errendimendua hobeto islatzen du.

Urruneko gainbegiraketarako hedapen bat proposatu dugu antzeko balioak etiketatzeko (*6. kapitulua*)

Urruneko gainbegiraketak aipamenak erlazionatutzat etiketatzen ditu, baldin eta soilik baldin esaldian topatutako informazioa eta ezagutza basekoa zehazki berdinak badira. Antzekoak diren aipamenak ez dira erlazionatutzat hartzen, baina aipamen horiek ikasketa prozesurako patroi onak izan ditzakete.

Gure proposamena ezagutza basearekiko antzekoak diren aipamenak aipamen positibotzat hartzea izan da, entrenamenduko datu multzoaren kalitatea hobetzeko. Honek sistemaren eraginkortasuna modu esanguratsuan hobetzen du.

Gertaeren erauzketarako ezagutza base bat eta datu multzo bat sortu ditugu (*5. kapitulua*)

Urruneko gainbegiraketa gertaeren erauzketan frogatzeko, lurrikarei buruzko ezagutza base bat sortu dugu Wikipediatik erauzitako informazioarekin. Ezagutza base honek 20 argumentu desberdinei buruzko informazioa dauka. Gure ezagutza baseko lurrikarei buruzko txioak erauzi ditugu Twitterretik, txio hauekin datu multzo bat sortuz. Ezagutza base hau eta datu multzo hau 6. kapituluko esperimentuetan erabili ditugu.

Datu multzoko informazio garrantzitsua eskuz etiketatu dugu, non testuinguruak ezagutza baseko argumentuarekin bat egiten duen. Eskuzko etiketatze hau oso baliagarria izan da geroagoko analisietarako.

Guk dakigula, hau da urruneko gainbegiraketako datu multzo bakarra, non aipamenen eskuzko etiketatzea eskuragarri dagoen aldi berean. Ezagutza basea eta datu multzoa, eskuzko anotazioekin batera, publikoki daude eskuragarri.

Urruneko gainbegiraketa gertaera konplexutan aplikatu dugu (*5. eta 6. kapituluak*)

Esperimentuak aipatu berri ditugun ezagutza base eta datu multzoetan aplikatu ditugu, urruneko gainbegiraketa gertaeren erauzketarako baliagarria dela erakutsiz. Orain arte, hurbilketa gertaeren erauzketan aplikatzen saiatu diren oro 1 edo 2 argumenturekin bakarrik egin dute lan, gure sistemak berriz, 20 argumenturekin lan egiten du.

Urruneko gainbegiraketa sare sozialetan aplikatu dugu (*5. eta 6. kapituluak*)

Gertaeren erauzketan egin ditugun esperimentuetan erabilitako datu multzoa Twitterreko txioak erabiliz sortuta dago, tradizionalki erabiltzen diren berri agentzien dokumentuen testu zatien ordez. Berri agentzien dokumentuak jasoak, konplexuak eta anbiguotasunik gabekoak dira. Bien bitartean, sare sozialetako testuak informalak, anbiguoak eta zaratatsuak dira. Gure

esperimentuek erakutsi dute urruneko gainbegiraketak mikroblogetatik ondo egiten duela lan ezagutza baseetako informazioa aberasteko.

7.2 Ondorioak

Orain aurkezten ditugun ondorioak aurreko atalean aurkeztutako ekarpenetatik eratorriak dira.

Urruneko gainbegiraketa sistema baten garapena (*3. kapitulua*)

Kapitulu honetan, erlazioen erauzketarako urruneko gainbegiraketa sistema baten lehen urratsak eman ditugu. Aldez aurretik, ikasketa gainbegiratuan oinarritutako sistema bat garatu dugu, eskuz etiketatutako ACE 2005 corpusa erabiliz.

Ikasketa gainbegiratuko sistemak eta urruneko gainbegiraketa sistemak oso antzekoak dira elkarrekiko. Desberdintasun nabariena corpusa etiketatzeko modua da, lehenengoan etiketatze prozesua eskuzkoa da, bigarrenean berriz, automatikoa. Aipamenen aurreprozesaketa, ezaugarrien erauzketa, sailkatzailearen entrenamendua eta inferentzia prozesuak bateragarriak dira bi sistemetan.

Urruneko gainbegiraketa sistemari dagokionez, ezagutza baseen aberasketan oinarritzen den TAC 2011ko *Slot Filling* atazan parte hartu dugu. Sistema honetan, aipamenak erauzten ditugu corpus batetik. Aipamen hauek elkarrekiko erlazionatuta dauden entitate bikoteak dituzte, bien arteko erlazioa ezagutza base batean zehaztuta baitago. Aipamen guztiak eskuratu ondoren, ikasketa gainbegiratuko sisteman erabilitako ezaugarri, optimizazio teknika eta inferentzia metodo berdinak aplikatzen dizkiogu urruneko gainbegiraketa sistemari.

Lortutako emaitzak egokitzat hartzeko urruti daude. Gure errore analisiak erakusten du gure sistemak emaitzak txarrak ematearen arrazoietako bat aipamen negatiboen falta dela. Beste arrazoietako bat erlazio jakin batzuei buruzko aipamenen urritasuna da, erlazio hauen ikasketa zailduz.

Baina akatsen artean garrantzitsuena, gure datu multzoan topatutako aipamen zaratatsu kantitate handia da. Gure datu multzoko aipamen askotan parte hartzen duten entitateen arteko erlazioak eta testuinguruak ez dute bat egiten. Adibide zaratatsuek ikasketa prozesua asko okertzen dute, sailkatzailea nahastuz, iragarpen okerrak eginez, eta ondorioz sistemaren eragin-

kortasuna kolokan jarriz. Eraginkortasuna hobetzeko, aipamen zaratatsuak kendu behar ditugu.

Aipamen zaratatsuen garbiketa (*4. kapitulua*)

Kapitulu honetan, Riedel *et al.*-en datu multzoko ausazko hainbat aipamen analizatu ditugu, eta bertatik hiru zarata mota kategorizatu.

Datu multzoek barnean duten zarata kantitate altuak motibatuta, atal honetan hiru metodo sinple eta trinko aurkeztu ditugu zarata garbitzeko. Heuristiko hauek ez dute datu multzoetan inongo eskuzko etiketatzerik erabiltzen, eta domeinu eta sistemarekiko independenteak dira. Hiru heuristiko hauen artean konbinaketa desberdinak egin ditugu, eta emaitza onenak ematen dituen konbinaketa erabili bukaerako emaitzak eskuratzeko.

Esperimentuak Stanford Unibertsitateak garatutako urruneko gainbegiraketa sistema erabiliz egin ditugu. Sistema hau artearen egoeran dago eta Mintz *et al.*-en algoritmo originala hobetzen dituen bi aldaera ditu. Sistema hau erabiliz, gure heuristikoen eraginkortasuna ebaluatzeko gai izan gara.

Heuristiko hauek, batez ere beraien konbinazioek, Stanford Unibertsitateak garatutako bi oinarri lerro trinkoren emaitzak hobetzen ditu, urruneko gainbegiraketaren eraginkortasuna hobetzeko aipamen zaratatsuak detektatzea eta ezabatzea beharrezkoa dela frogatuz.

Gertaera erauzketarako datu multzo baten sorkuntza (*5. kapitulua*)

Kapitulu honetan, txioetan oinarritutako gertaera erauzketa sistema bat sortzeko lehenengo hastapenak egin ditugu. Aukeratutako domeinua lurrikarena da. Lehenengo urratsa Wikipedian aurki ditzakegun lurrikaren artikuluetatik ezagutza base bat sortzea izan da. Geroago, ezagutza basean ditugun lurrikarei buruzko txioak eskuratu ditugu Twitterretik, eta erreplikei buruzko txioak kendu.

Ezagutza baseak 108 lurrikara desberdinei buruzko informazioa dauka, 20 argumentu desberdinekin. Datu multzoak lurrikara hauei buruzko txioz osatuta dago, guztira 7841 txio ditugu. Txioetako aipamenak eskuz etiketatu ditugu, geroago egin ditugun analisietarako lagungarria izanik.

Ezagutza basea eta txioen datu multzoa, eskuzko etiketatzearekin batera, publikoki eskuragarri daude[1].

[1] https://ixa.si.ehu.es/Ixa/Argitalpenak/Tesiak/1427215138/publikoak/earthquake-kb-dataset.zip

Gertaeren erauzketa urruneko gainbegiraketaren bidez (*6. kapitulua*)

Kapitulu honetan, urruneko gainbegiraketan oinarritutako gertaera erauzketa sistema bat garatu dugu. Gure sistemak gertaera konplexuak erauzten ditu, 20 argumentu desberdinei buruzko informazioa emanez.

Hasteko, urruneko gainbegiraketa bidez automatikoki etiketatutako datu multzoa analizatu dugu, eskuzko etiketatzearekin konparatuz. Estimatzen dugu urruneko gainbegiraketak etiketatutako %83a testuinguruarekiko zuzena dela. Eskuzko etiketatzea ezagutza basearen aurka ebaluatzean, %47ak bakarrik egiten du bat urre patroiarekin.

Ikusi dugu ere iragarritako informazioaren %55ak, ez duenean ezagutza basearekin bat egiten, honekiko oso antzekoa dela. Baita iragarritako erantzunen %20ak ez duela ezagutza basean inongo informaziorik argumentu horri buruz, txioetan ezagutza basean agertzen ez den informazioa topatu dezakegula frogatuz. Horretarako, ebaluazio erlaxatu bat proposatzen dugu, non antzekoak diren iragarpenak partzialki zuzentzat hartzen diren. Ebaluazio erlaxatuak emaitzak hobetzen ditu, espero bezala, eta sistemaren ebaluazioa errealistagoa bihurtzen du, batez ere kontuan hartuta ezagutza baseko informazioa batzuetan ez dela zehatza.

Urruneko gainbegiraketarako aldaera berri bat proposatu dugu ere, non ezagutza basearekiko antzekoak diren txioetako aipamenak ere etiketatzen ditugun; honi "etiketatze hurbila" deitu diogu. Gainera, ezaugarrien agregazio global bat aplikatu dugu, hauek aberasgarriagoak izateko, lurrikara berdinekoak diren txioen artean informazioa elkarbanatuz. Gure emaitzek %19ko hobekuntza esanguratsua dute. Horrez gain, ekarpen hauek osagarriak dira: etiketatze hurbila eta ezaugarrien agregazioa konbinatzen dituen eredu bat erabiltzean, hauek indibidualki aplikatuta baino askoz eraginkortasun hobea lortzen da, %33ko hobekuntza lortuz oinarri lerroarekiko. Gure emaitzak eraginkortasunaren sabaiarekiko (gure sistemak eskuzko etiketatzea entrenatuz lortzen duen emaitza) %88ra hurbiltzen dira, urruneko gainbegiraketa gertaera konplexuen erauzketarako egokia dela frogatuz.

7.3 Etorkizuneko lerroak

Lan honetatik ikerketa lerro batzuk atera dira, geroago aztertzeko. Atal honetan etorkizunean landu nahiko genituzkeen esperimentu eta ikerkuntza bideak jorratzen ditugu.

Aipamen zaratatsuak beste datu multzo batzuetatik garbitu

Gure heuristikoak datu multzo bakarrean frogatu ditugu. Interesgarria litzateke hauek beste datu multzo batean frogatzea. Stanford Unibertsitateko lengoaia naturalaren prozesamenduko taldeak[2] sortutako KBP datu multzoa erabili genezake.

Gertaeren erauzketa beste domeinu batean

Gure gertaera erauzketa sistema beste eremu batzuetan frogatzea gustatuko litzaiguke, berri agentzien dokumentuak erabiliz baita ere. Adibide bezala, Reschke *et al.*-ek (2014) sortutako datu multzoa eta ezagutza basea erabili genezake.

Lurrikarei buruzko informazioa denbora errealean erauzten duen sistema bat garatu

Saiatu ahal gintezke lurrikarei buruzko txioak erauzten denbora errealean. Era honetan, lurrikarei buruzko informazioa eman genezake gertatu diren unean.

Urruneko gainbegiraketa eleanitza

Behin gure urruneko gainbegiraketa sistemak emaitza onak lortzen dituela, pauso bat aurrera eman genezake eta euskararako lehenengo urruneko gainbegiraketa sistema sortu. Hizkuntza honen baliabideak txikiak dira, ingelesa bezalako beste hizkuntzekin konparatuta. Nahiz eta ezagutza basea eta corpusa mugatuak izan, frogatu genezake zernolako errendimendua duen hurbilketak baliabide urriko hizkuntzetan. Hau aukera aparta litzateke urruneko gainbegiraketa sistema eleanitzak ere ikertzeko.

[2] http://nlp.stanford.edu/

Bibliografia

Agichtein E. eta Gravano L. Snowball: Extracting relations from large plain-text collections. *Proceedings of the Fifth ACM Conference on Digital Libraries*, 85–94, 2000. ISBN 1-58113-231-X.

Angeli G., Chaganty A., Chang A., Reschke K., Tibshirani J., Wu J.Y., Bastani O., Siilats K., eta Manning C.D. Stanford's 2013 kbp system. *TAC-KBP*, 2013.

Angeli G., Tibshirani J., Wu J.Y., eta Manning C.D. Combining distant and partial supervision for relation extraction. *EMNLP*, October 2014.

Artstein R. eta Poesio M. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008. ISSN 0891-2017.

Babych B. *Information Extraction Technology in Machine Translation: IE methods for improving and evaluating MT quality*. Doktoretza-tesia, University of Leeds, 2005.

Bach N. eta Badaskar S. A Review of Relation Extraction. URL http://www.cs.cmu.edu/~{}nbach/papers/A-survey-on-Relation-Extraction.pdf. unpublished, 2007.

Banko M., Cafarella M.J., Soderland S., Broadhead M., eta Etzioni O. Open information extraction from the web. *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, 2670–2676, 2007.

Banko M. eta Etzioni O. The tradeoffs between open and traditional relation extraction. *Proceedings of ACL-08: HLT*, 28–36, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Becker H., Naaman M., eta Gravano L. Beyond trending topics: Real-world event identification on twitter. *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*, 2011.

Benson E., Haghighi A., eta Barzilay R. Event discovery in social media feeds. *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*, 2011.

Berg-Kirkpatrick T., Burkett D., eta Klein D. An empirical investigation of statistical significance in nlp. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, 995–1005, 2012.

Brin S. Extracting patterns and relations from the world wide web. *Selected Papers from the International Workshop on The World Wide Web and Databases*, WebDB '98, 172–183, London, UK, 1999. ISBN 3-540-65890-4.

Casillas A., de Ilarraza A.D., Gojenola K., Oronoz M., eta Rigau G. Using kybots for extracting events in biomedical texts. *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, 138–142, 2011. ISBN 9781937284091.

Chun H.w., Hwang Y.s., eta Rim H.C. Unsupervised event extraction from biomedical literature using co-occurrence information and basic patterns. *Proceedings of the First International Joint Conference on Natural Language Processing*, IJCNLP'04, 777–786, 2005. ISBN 3-540-24475-1, 978-3-540-24475-2.

Craven M. eta Kumlien J. Constructing biological knowledge bases by extracting information from text sources. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, 77–86, 1999. ISBN 1-57735-083-9.

Dong X., Gabrilovich E., Heitz G., Horn W., Lao N., Murphy K., Strohmann T., Sun S., eta Zhang W. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. *Proceedings of the 20th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, KDD '14, 601–610, 2014. ISBN 978-1-4503-2956-9.

Etzioni O., Cafarella M., Downey D., Popescu A.M., Shaked T., Soderland S., Weld D.S., eta Yates A. Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intell.*, 165(1):91–134, June 2005. ISSN 0004-3702.

Fader A., Soderland S., eta Etzioni O. Identifying relations for open information extraction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 1535–1545, 2011. ISBN 978-1-937284-11-4.

Fan M., Zhao D., Zhou Q., Liu Z., Zheng T.F., eta Chang E.Y. Distant supervision for relation extraction with matrix completion. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 839–849, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Fernandez I. *Euskarazko Entitate-Izenak: identifikazioa, sailkapena, itzulpena eta desanbiguazioa*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2012.

Finkel J.R., Grenager T., eta Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 363–370, 2005.

Giuliano C., Lavelli A., eta Romano L. Exploiting shallow linguistic information for relation extraction from biomedical literature. *In Proc. EACL 2006*, 2006.

Grishman R., Westbrook D., eta Meyers A. NYU's English ACE 2005 system description. *ACE 2005*, 2005.

Gurrutxaga A. *Idiomatikotasunaren karakterizazio automatikoa: izena+aditza konbinazioak*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2014.

Hoffmann R., Zhang C., Ling X., Zettlemoyer L., eta Weld D.S. Knowledge-based weak supervision for information extraction of overlapping relations.

Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2011.

Hong Y., Zhang J., Ma B., Yao J., Zhou G., eta Zhu Q. Using cross-entity inference to improve event extraction. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 1127–1136, 2011. ISBN 978-1-932432-87-9.

Huang R. eta Riloff E. Bootstrapped training of event extraction classifiers. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, 286–295, 2012. ISBN 978-1-937284-19-0.

Ji H. eta Grishman R. Refining Event Extraction through Unsupervised Cross-document Inference. *Proceedings of ACL-08*, 2008.

Ji H. eta Grishman R. Knowledge base population: Successful approaches and challenges. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 1148–1158, 2011. ISBN 978-1-932432-87-9.

Klebanov B.B., Knight K., eta Marcu D. Text simplification for information-seeking applications. *On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science*, 735–747. Springer Verlag, 2004.

Koch M., Gilmer J., Soderland S., eta Weld S.D. Type-aware distantly supervised relation extraction with linked arguments. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1891–1901. Association for Computational Linguistics, 2014.

Lafferty J.D., McCallum A., eta Pereira F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289, 2001. ISBN 1-55860-778-1.

Li J., Ritter A., eta Hovy E. Weakly supervised user profile extraction from twitter. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 165–174. Association for Computational Linguistics, 2014.

Li Q., Ji H., eta Huang L. Joint event extraction via structured prediction with global features. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 73–82, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Li Y., Zaragoza H., Herbrich R., Shawe-Taylor J., eta Kandola J.S. The perceptron algorithm with uneven margins. *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, 379–386, 2002. ISBN 1-55860-873-7.

Liao S. eta Grishman R. Filtered ranking for bootstrapping in event extraction. *COLING*, 680–688. Tsinghua University Press, 2010a.

Liao S. eta Grishman R. Using document level cross-event inference to improve event extraction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 789–797, Uppsala, Sweden, July 2010b. Association for Computational Linguistics.

Lin D. eta Pantel P. Dirt - discovery of inference rules from text. *In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 323–328, 2001.

Lopez de Lacalle O. eta Lapata M. Unsupervised relation extraction with general domain knowledge. *EMNLP*, 415–425. ACL, 2013. ISBN 978-1-937284-97-8.

Lu W. eta Roth D. Automatic event extraction with structured preference modeling. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, 835–844, 2012.

Manning C.D., Raghavan P., eta Schütze H. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

Mihalcea R. eta Tarau P. TextRank: Bringing order into texts. *Proceedings of EMNLP-04and the 2004 Conference on Empirical Methods in Natural Language Processing*, July 2004.

Min B., Grishman R., Wan L., Wang C., eta Gondek D. Distant supervision for relation extraction with an incomplete knowledge base. *In Proceedings*

of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL, 2013.

Min B., Li X., Grishman R., eta Ang S. New york university 2012 system for kbp slot filling. *Proceedings of the Fifth Text Analysis Conference (TAC 2012)*. National Institute of Standards and Technology (NIST), November 2012a.

Min B., Li X., Grishman R., eta Ang S. New york university 2012 system for kbp slot filling. *Proceedings of the Fifth Text Analysis Conference (TAC 2012)*. National Institute of Standards and Technology (NIST), 2012b.

Mintz M., Bills S., Snow R., eta Jurafsky D. Distant supervision for relation extraction without labeled data. *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*, 2009.

Miyanishi K. eta Ohkawa T. A method of extracting sentences containing protein function information from articles by iterative learning with feature update. *Computational Intelligence Methods for Bioinformatics and Biostatistics*, 81–94. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-38341-0.

Nivre J., Hall J., eta Nilsson J. Memory-based dependency parsing. *Proceedings of CoNLL-2004*, 49–56. Boston, MA, USA, 2004.

Petrović S., Osborne M., eta Lavrenko V. Streaming first story detection with application to twitter. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 181–189, Los Angeles, California, June 2010. Association for Computational Linguistics.

Popescu A.M., Pennacchiotti M., eta Paranjpe D. Extracting events and event descriptions from twitter. *Proceedings of the 20th International Conference on World Wide Web*, 2011.

Pronoza E., Yagunova E., Volskaya S., eta Lyashin A. Restaurant information extraction (including opinion mining elements) for the recommendation system. *Human-Inspired Computing and Its Applications - 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I*, 201–220, 2014.

Ravichandran D. eta Hovy E. Learning surface text patterns for a question answering system. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 41–47, 2002.

Reschke K., Jankowiak M., Surdeanu M., Manning C.D., eta Jurafsky D. Event extraction using distant supervision. *Proceedings of LREC*, 2014.

Riedel S., Yao L., eta McCallum A. Modeling relations and their mentions without labeled text. *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*, 2010.

Riedel S., Yao L., eta McCallum A. Latent relation representations for universal schemas. *CoRR*, abs/1301.4293, 2013.

Rifkin R. eta Klautau A. In defense of one-vs-all classification. *J. Mach. Learn. Res.*, 5:101–141, December 2004. ISSN 1532-4435.

Riloff E. eta Jones R. Learning dictionaries for information extraction by multi-level bootstrapping. *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, 474–479, 1999. ISBN 0-262-51106-1.

Ritter A., Mausam, Etzioni O., eta Clark S. Open domain event extraction from twitter. *Proceedings of KDD*, 2012.

Roth B., Barth T., Wiegand M., Singh M., eta Klakow D. Effective slot filling based on shallow distant supervision methods. *CoRR*, 2013.

Rozenfeld B. eta Feldman R. Self-supervised relation extraction from the web. *Knowl. Inf. Syst.*, 17(1):17–33, October 2008. ISSN 0219-1377.

Rusu D., Hodson J., eta Kimball A. Unsupervised techniques for extracting and clustering complex events in news. *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, 26–34, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.

Sakaki T., Okazaki M., eta Matsuo Y. Earthquake shakes twitter users: Real-time event detection by social sensors. *Proceedings of the 19th International*

Conference on World Wide Web, WWW '10, 851–860, 2010. ISBN 978-1-60558-799-8.

Sandhaus E. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 2008.

Soraluze A., Arregi O., Arregi X., Ceberio K., eta de Ilarraza A.D. Mention detection: First steps in the development of a basque coreference resolution system. *Proceedings of KONVENS*, 128–163, 2012.

Sun A., Grishman R., Xu W., eta Min B. New york university 2011 system for kbp slot filling. *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*. National Institute of Standards and Technology (NIST), November 2011.

Surdeanu M. eta Ciaramita M. Robust information extraction with perceptrons. *Proceedings of the NIST 2007 Automatic Content Extraction Workshop (ACE07)*, March 2007.

Surdeanu M., Gupta S., Bauer J., McClosky D., Chang A.X., Spitkovsky V.I., eta Manning C.D. Stanford's distantly-supervised slot-filling system. *Proceedings of the TAC-KBP 2011 Workshop*, 2011.

Surdeanu M., McClosky D., Tibshirani J., Bauer J., Chang A.X., Spitkovsky V.I., eta Manning C.D. A simple distant supervision approach for the tac-kbp slot filling task. *Proceedings of the TAC-KBP 2010 Workshop*, 2010.

Surdeanu M., Tibshirani J., Nallapati R., eta Manning C.D. Multi-instance multi-label learning for relation extraction. *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*, 2012.

Takamatsu S., Sato I., eta Nakagawa H. Reducing wrong labels in distant supervision for relation extraction. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, 721–729, 2012.

Urizar R. *Euskal lokuzioen tratamendu konputazionala*. Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2012.

Wang J., Xu Q., Lin H., Yang Z., eta Li Y. Semi-supervised method for biomedical event extraction. *Proteome Science*, 11(1), 2013.

Weston J., Bordes A., Yakhnenko O., eta Usunier N. Connecting language and knowledge bases with embedding models for relation extraction. *CoRR*, 2013.

Wick M., Rohanimanesh K., Culotta A., eta Mccallum A. Samplerank: Learning preference from atomic gradients. *In NIPS WS on Advances in Ranking*, 2009.

Witten I.H. eta Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000. ISBN 1-55860-552-5.

Wu F. eta Weld D.S. Open information extraction using wikipedia. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, 118–127, 2010.

Xu Y., Kim M.Y., Quinn K., Goebel R., eta Barbosa D. Open information extraction with tree kernels. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 868–877, Atlanta, Georgia, June 2013.

Yao L., Riedel S., eta McCallum A. Collective cross-document relation extraction without labelled data. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, 1013–1023, 2010.

Yao L., Riedel S., eta McCallum A. Unsupervised relation discovery with sense disambiguation. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, 712–720, 2012.

Zhou G., Su J., Zhang J., eta Zhang M. Exploring various knowledge in relation extraction. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 427–434, 2005.

Zhou G., Zhang M., Ji D., eta Zhu Q. Tree Kernel-Based Relation Extraction with Context-Sensitive Structured Parse Tree Information. *Proceedings of*

the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 728–736, 2007.

Zhu J., Nie Z., Liu X., Zhang B., eta Wen J.R. Statsnowball: A statistical approach to extracting entity relationships. *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 101–110, 2009. ISBN 978-1-60558-487-4.

Glosategia

Aipamen (*Mention*)

Entitate edo gertaera baten agerpena dokumentuan.

Atalase (*Threshold*)

Gutxieneko maila adierazten duen parametroa. Tesi honetan orokorrean aldagai baten balioa atalase batena baino handiagoa bada, orduan ontzat hartzen dugu.

Baldintzazko Ausazko Eremu (*Conditional random field*)

Etiketatze sekuentzialean oinarritzen den eredu probabilistikoa.

Baliokidetasun klase (*Equivalence class*)

Betegarri baliokideen multzoa, gauza bera modu desberdinean adierazten duten betegarri bi edo gehiagok osatzen dute.

Betegarri (*Filler*)

Erlazio bateko objektua, atributua edo bigarren entitatea.

Datu multzo (*Dataset*)

Datu bilduma, erregistro moduan edukitzeko edo datuen beraien azterketa sakonagoa egiteko. Datu multzoak erabilgarriak izan daitezen, kalkuluak eta azterketak behar bezala egiteko, modu ordenatu batean azaldu behar dira.

Egiaztapen gurutzatua (*Cross validation*)

Analisi estatistiko baten emaitzak ebaluatzeko teknika, partizio desberdinen ebaluazio neurrien batezbestekoa aritmetikoa kalkulatuz.

Elkarrekiko informazio puntual (*Pointwise mutual information*)

Aldagai batek beste bati buruz ematen duen informazioa neurtzeko teknika.

Entitate (*Entity*)

Existitzen den edozer gauza, eta izen bat daukana, hala nola pertsonak, erakundeak, tokiak,...

Entitate izenen desanbiguazioa (*Entity linking*)

Entitate izenak dokumentu batetik erauzi eta entitate hauek ezagutza base batekin lotzen dituen ataza.

Erlazio erauzketa (*Relation extraction*)

Dokumentu batean agertzen diren entitateen artean erlaziorik dagoen, eta egotekotan, zein motatako erlazioa den zehazten duen ataza.

Erlazio etiketa anitz (*Multi-label relations*)

Bi entitateren artean erlazio bat baino gehiago dagoenean.

Erregresio logistiko (*Logistic regression*)

Gertakizun baten probabilitatea aurresateko erabiltzen den erregresio teknika, aldagai independente zenbaitetan oinarrituta kurba logistiko bat egokituz.

Errendimendu sabai (*Performance ceiling*)

Esperimentu batean gehienez lortu dezakegun aurreikusitako emaitza.

Ezagutza base (*Knowledge base*)

Informazio konplexu egituratu eta egituratu gabea ordenagailuetan gordetzeko teknologia.

Gertaera erauzketa (*Event extraction*)

Dokumentu batean gertaerak aurkitu, eta hauei buruz ahalik eta informazio gehien lortzeko prozesua.

Hitz huts (*Stop words*)

Esanahirik ez duten hitzak, hala nola artikuluak, izenordainak, juntagailuak,...

Hitzen adiera desanbiguazioa (*Word sense disambiguation*)

Perpaus jakin batean hitz polisemiko batek duen adiera antzematean datzan prozesua.

Informazio erauzketa (*Information extraction*)

Ordenagailu batek dokumentuetatik informazio zehatz batzuk lortzeko helburu duen ataza.

Infotaula (*Infobox*)

Wikipediako entitate baten orrialde batean, eskuinean entitate horri buruzko informazio laburtua jasotzen duen taula.

Izendegi (*Gazetteer*)

Arlo jakin batean erabiltzen diren hitzen zerrenda.

Klase anitzeko erregresio logistiko (*Multiclass logistic regression*)

Klase bat baino gehiago, edo bi emaitza diskretu baino gehiago dagoenean erabiltzen den erregresio logistikoko metodoa.

Markoven eredu ezkutua (*Hidden markov model*)

Ikusgai diren parametroen bidez, ezkutuan dauden parametroak zeintzuk diren asmatzeko eredua.

Muturreko datu (*Outlier*)

Estatistikan, gainerako datuen balioetatik urrun kokatzen diren datuak.

Perzeptroi (*Perceptron*)

Sare neuronaletan eta ikasketa automatikoan erabiltzen den algoritmo gainbegiratua.

Rol semantikoen etiketatzea (*Semantic role labeling*)

Testuetan argumentu semantikoak bilatu eta hauen rolak zehazteko ataza. Argumentuak esaldiaren predikatua edo aditzarekin lotuta daude.

Sostengu bektoreen makina (*Support vector machine*)

Datuak eta ereduak aztertzeko erabiltzen den ikasketa algoritmo gainbegiratua, sailkapen eta erregresio analisirako erabiltzen da.

Tupla (*Tuple*)

Entitate bikotea.

Tripleta (*Triplet*)

Entitate bikotea eta hauen arteko erlazioa.

Txio (*Tweet*)

Twitter sare sozialean, pertsona (edo txiolari) batek idazten duena gero komunitate osoarekin edo sarean dituen jarraitzaileekin partekatzeko.

Urre etiketatze (*Gold annotation*)

Zuzentzat hartzen den etiketatutako datu multzoa.

Urre patroi (*Gold Standard*)

Ebaluatzeko garaian, ustez erantzun zuzenak direnak, adituek eskuz sortuak.

Urruneko gainbegiraketa (*Distant supervision*)

Informazio erauzketako paradigma bat. Aipamen berean ezagutza base batean batera parte hartzen duten entitateak agertzen direnean, biak erlazionatuta daudela aintzat hartzen da.

WordNet

Ingeleseko hitz eta adierei buruzko informazioa duen ezagutza base lexikal bat. Izen, aditz, adjektibo eta adberbioak aurkitzen dira bertan *synset* delakoen arabera antolatuta eta hainbat erlazio semantikorekin lotuta.

A

ACE corpuseko entitate eta erlazioen XML kodea etiketatuta

Har ditzagun bi entitate: *Khoser River* eta *Iraq*. Bien arteko erlazioa *Part-Whole* da eta azpimota *Geographical*. A.1 kode zatian aurki dezakegu bi entitateen arteko erlazioa XML bidez etiketatuta. Entitate bakoitzak bere identifikatzailea dauka, identifikatzaile hau dokumentu osoan erabiliko da, beste erlazioetan zehaztean eta entitate honi buruzko informazioa ematean.

A.2 kodeak adibidean jarritako lehenengo entitatearen (*Khoser River*) informazioa dakar, entitate izen mota *Location* da eta azpimota *Water-Body*. Entitate hau dokumentuan hainbat aldiz aipatzen da, era desberdinetan; kode honetan entitate honi erreferentzia egiten dioten lau adibide desberdin ditugu, dokumentuko posizio desberdinekin batera, eta bakoitza bere identifikatzailearekin, hirugarren entitate aipamena da aurreko adibideko erlazioan erabili dena.

A.3 kodeak adibidean jarritako bigarren entitatearen (*Iraq*) informazioa dakar, entitate izen mota *Location* da eta azpimota *Region-General*. Lehenengo aipamena da goiko adibideko erlazioan erabili dena.

```
<relation ID="CNN_IP_20030404.1600.00-2-R40" TYPE="PART-WHOLE"
        SUBTYPE="Geographical" TENSE="Unspecified" MODALITY="Asserted">
  <relation_argument REFID="CNN_IP_20030404.1600.00-2-E16" ROLE="Arg-1"/>
  <relation_argument REFID="CNN_IP_20030404.1600.00-2-E115" ROLE="Arg-2"/>
  <relation_mention ID="CNN_IP_20030404.1600.00-2-R40-1"
        LEXICALCONDITION="Preposition">
    <extent>
      <charseq START="2728" END="2760">the Khoser River in northern
        Iraq</charseq>
    </extent>
    <relation_mention_argument REFID="CNN_IP_20030404.1600.00-2-E16-77"
        ROLE="Arg-1">
      <extent>
        <charseq START="2728" END="2743">the Khoser River</charseq>
      </extent>
    </relation_mention_argument>
    <relation_mention_argument REFID="CNN_IP_20030404.1600.00-2-E115-78"
        ROLE="Arg-2">
      <extent>
        <charseq START="2748" END="2760">northern Iraq</charseq>
      </extent>
    </relation_mention_argument>
  </relation_mention>
</relation>
```

Listing A.1 – Erlazio baten adibidea etiketatuta.

```
<entity ID="CNN_IP_20030404.1600.00-2-E16" TYPE="LOC" SUBTYPE="Water-Body"
    CLASS="SPC">
  <entity_mention ID="CNN_IP_20030404.1600.00-2-E16-20" TYPE="NOM"
        LDCTYPE="NOM">
    <extent>
      <charseq START="1065" END="1099">the river on the main road to
        Mosul</charseq>
    </extent>
    <head>
      <charseq START="1069" END="1073">river</charseq>
    </head>
  </entity_mention>
  <entity_mention ID="CNN_IP_20030404.1600.00-2-E16-35" TYPE="NAM"
        LDCTYPE="NAM">
    <extent>
      <charseq START="1602" END="1617">the Khoser River</charseq>
    </extent>
    <head>
      <charseq START="1606" END="1617">Khoser River</charseq>
    </head>
  </entity_mention>
  <entity_mention ID="CNN_IP_20030404.1600.00-2-E16-77" TYPE="NAM"
        LDCTYPE="NAM">
    <extent>
      <charseq START="2728" END="2743">the Khoser River</charseq>
    </extent>
    <head>
      <charseq START="2732" END="2743">Khoser River</charseq>
    </head>
  </entity_mention>
  <entity_attributes>
    <name NAME="Khoser River">
      <charseq START="1606" END="1617">Khoser River</charseq>
    </name>
    <name NAME="Khoser River">
      <charseq START="2732" END="2743">Khoser River</charseq>
    </name>
  </entity_attributes>
</entity>
```

Listing A.2 – Erlazioko lehenengo entitatea etiketatuta, dokumentuan dituen aipamen guztiekin.

```
<entity ID="CNN_IP_20030404.1600.00-2-E115" TYPE="LOC"
        SUBTYPE="Region-General" CLASS="SPC">
  <entity_mention ID="CNN_IP_20030404.1600.00-2-E115-78" TYPE="NOM"
        LDCTYPE="BAR">
    <extent>
      <charseq START="2748" END="2760">northern Iraq</charseq>
    </extent>
    <head>
      <charseq START="2757" END="2760">Iraq</charseq>
    </head>
  </entity_mention>
  <entity_mention ID="CNN_IP_20030404.1600.00-2-E115-194" TYPE="NOM"
        LDCTYPE="BAR">
    <extent>
      <charseq START="337" END="349">northern Iraq</charseq>
    </extent>
    <head>
      <charseq START="346" END="349">Iraq</charseq>
    </head>
  </entity_mention>
</entity>
```

Listing A.3 – Erlazioko bigarren entitatea etiketatuta, dokumentuan dituen aipamen guztiekin.

B

ACEeko ebaluatzaile ofiziala

Hona hemen ACEeko ebaluatzailea azaltzen duen dokumentuaren irudia.

The primary output from ace-eval is a conditional error analysis. The conditions analyzed are shown in the header. Each output record in this analysis shows the error statistics for the selected condition, which is identified in the first field of the record. The record contains 4 sets of statistics. These are namely 1) the raw count statistics, 2) the normalized count statistics, 3) the cost/value statistics based on the model of system output value, and 4) the unconditioned cost/value statistics. An example is shown in the figure below.

```
======== entity scoring ========

Entity Detection and Recognition statistics:

 ref     _________Count___________     ___________Count_(%)__________      _______________Cost_(%)_______________      __Unconditioned_Cost_(%)_
 entity  Ent   Detection     Rec       Detection    Rec    Unweighted      Detection    Rec   Value     Value-based       Max      Detection     Rec
 type    Tot     FA  Miss    Err        FA  Miss    Err   Pre--Rec--F       FA  Miss    Err    (%)     Pre--Rec--F      Value     FA   Miss     Err
    FAC   285    20    40     44       7.0  14.0   15.4   75.8 70.5 73.1    1.3   8.9   10.1   79.7    87.7 81.0 84.2     5.61   0.07   0.50   0.57
    GPE   822    26    17     66       3.2   2.1    8.0   88.9 89.9 89.4    0.1   1.4    6.1   92.4    93.7 92.5 93.1    30.97   0.02   0.42   1.90
    LOC   220    25    49     61      11.4  22.3   27.7   56.1 50.0 52.9    3.0  15.5   15.9   65.6    78.4 68.6 73.2     4.83   0.15   0.75   0.77
    ORG   598    39    61    117       6.5  10.2   19.6   72.9 70.2 71.6    1.5   7.6   12.0   78.8    85.5 80.3 82.8    16.75   0.26   1.28   2.02
    PER  2380   215   132    514       9.0   5.5   21.6   70.4 72.9 71.6    0.6   2.0    9.3   88.0    89.9 88.6 89.3    38.67   0.22   0.79   3.61
    VEH   124     4     7     19       3.2   5.6   15.3   81.0 79.0 80.0    0.0   2.0   11.4   86.5    88.3 86.5 87.4     2.11   0.00   0.04   0.24
    WEA   142    15    16     34      10.6  11.3   23.9   65.2 64.8 65.0    2.3   6.1   12.3   79.3    84.8 81.6 83.2     1.06   0.02   0.06   0.13
  total  4571   344   322    855       7.5   7.0   18.7   73.9 74.3 74.1    0.7   3.9    9.2   86.2    89.7 86.9 88.3   100.00   0.75   3.85   9.24
```

Here is a brief description of the fields in each of these records:

1) The count statistics:
   - `Tot` - the number of (reference) elements for the selected condition
   - `FA` - the number of spurious (false alarm) objects output by the system
   - `Miss` - the number of (reference) objects that the system missed
   - `Err` - the number of objects that the system detected but misrecognized

2) The count percentage statistics. These are simply the raw counts divided by the number of reference elements for the selected condition. Included are the precision, recall, and F-measure based on these percentage counts.

3) The cost/value statistics. These costs and values are normalized by the total value of the reference elements for the selected condition.
   - `FA` - the cost of spurious (false alarm) objects output by the system
   - `Miss` - the cost of objects that the system missed
   - `Err` - the cost of objects that the system detected, due to errors in misrecognizing object attributes[1]
   - `Value` - the value of system output relative to maximum achievable value (= 100% - FA% - Miss% - Err%)

   Included are the precision, recall and F-measure based on these percentage costs.

4) The unconditioned cost statistics. These statistics are the same as the cost/value statistics in 3), but here they are normalized by the total value for all reference objects rather than by the value of the reference objects for the selected condition.

[1] Note that some errors which are counted in the count statistics incur no cost in the cost statistics. For example, errors in recognizing the tense attribute of relations and events incur no cost penalty under the current default value model, but these errors are counted nonetheless in the error count statistics.

C

TAC-KBP ezagutza baseko entitate baten adibidea

C.1 kode zatian, Sarah Michelle Gellar aktoresari buruzko informazioa aurki dezakegu. Hasierako lerroek entitatea, entitate izen mota (kasu honetan pertsona) eta entitatearen identifikatzailea ezagutza basean zeintzuk diren zehazten dute.

Ondorengo lerroek entitateari buruzko informazio desberdina ematen dute, informazio mota eta honen balioarekin. Balio batzuek identifikatzaileak dituzte, *New York City*k adibidez, horrek esan nahi du hiri horri buruzko informazioa aurki dezakegula ezagutza basean, identifikatzaile berdinarekin. Hauen bidez sortzen ditugu tripletak, adibidez:

- Sarah Michelle Gellar - *birthplace* - New York City

Amaitzeko, aktoreari buruzko Wikipediako artikulua dakar ezagutza baseak, *CDATA* formatuan.

Ezagutza base hau ere Wikipedian oinarrituta dago. Gellar aktoresaren infotaula C.1 irudian dago, ezagutza basearekin konparatu nahi izanez gero.

```
<entity wiki_title="Sarah_Michelle_Gellar" type="PER" id="E0339589"
      name="Sarah Michelle Gellar">
<facts class="Infobox Actor">
<fact name="othername">Sarah Michelle Prinze</fact>
<fact name="birthdate">April 14, 1977 (1977-04-14) (age_32)</fact>
<fact name="birthplace"><link entity_id="E0787907">New York City</link>,
      <link>New York</link>, <link entity_id="E0679687">USA</link></fact>
<fact name="birthname">Sarah Michelle Gellar</fact>
<fact name="spouse"><link entity_id="E0727066">Freddie Prinze, Jr.</link>
      (2002 - present)</fact>
<fact name="occupation">actress</fact>
<fact name="yearsactive">1981 - present</fact>
<fact name="emmyawards"><link>Outstanding Younger Actress - Drama
      Series</link>
1995 <link entity_id="E0609464">All My Children</link></fact>
<fact name="awards"><link>Saturn Award for Best Actress on Television</link>
      1999 <link entity_id="E0141768">Buffy the Vampire Slayer</link></fact>
</facts>
<wiki_text><![CDATA[Sarah Michelle Gellar
Sarah Michelle Prinze, (born April 14, 1977) better known by her birth name
of Sarah Michelle Gellar, is an American actress. She is best known for her
role as the character Buffy Summers in the acclaimed television series Buffy
the Vampire Slayer, for which she won in total six Teen Choice Awards, and
the Saturn Award for Best Genre TV Actress and received a Golden Globe Award
nomination. The show set up Gellar as an icon and star in the industry from
the first season of the show. She won a Daytime Emmy Award for her role in
All My Children as character Kendall Hart.
(...)
]]></wiki_text>
</entity>
```

Listing C.1 – TAC-KBP ezagutza baseko entitate baten adibidea.

Born	April 14, 1977 (age 37)[1] New York City, New York, U.S.
Residence	Los Angeles, California, U.S.
Occupation	Actress, producer
Years active	1981–present
Spouse(s)	Freddie Prinze, Jr. (m. 2002)
Children	2

C.1 irudia – Sarah Michelle Gellarri buruzko infotaula Wikipedian.

D

Slot Filling atazako helburu entitateen zerrendak

Eranskin honetan bi taula ditugu, garapen faseko eta testeko faseko helburu entitateekin.

D.1 taulak 2011ko garapen faseko helburu entitateak ditu, 2010ean testerako erabili ziren entitateak.

D.2 taulak berriz, testeko helburu entitateak.

Pertsonak	**Erakundeak**
George Boyd	Chelsea Library
Erika Rose	Crown Prosecution Service
Donald Wildmon	Ontario Human Rights Commission
Frankie Delgado	Haifa University
Chris Ivery	Babyshambles
Molly Malaney	Samsung
Mark Buse	Manila Economic and Cultural Office
Charles Wuorinen	Madoff Securities
Kendra Wilkinson	Opera National de Paris
Paul Kim	Nuclear Supplier Group
Chad White	Paralyzed Veterans of America
Lewis Hamilton	Ownit Mortgage Solutions
Alice Dellal	National Union for the Total Independence of Angola
Chante Moore	Pennsylvania Supreme Court
Susan Boyle	Old Lane Partners
Alexandra Burke	Option One Mortgage
Ellen Degeneres	National Republican Campaign Committee
Julian Bond	Professional Rodeo Cowboys Association
Raul Castro	New Hampshire Institute of Politics
Ali Fedotowsky	Massachusetts House of Representatives
Bryan Fuller	Institute for Diversity and Ethics in Sport
Chris Dodd	Inter-American Press Association
Adam Senn	Jackson Hewitt
Hector Elizondo	National Beef Packing Co.
Simon Cowell	National Military Family Association
Olivia Palermo	Multidisciplinary Association for Psychedelic Studies
Richard Lindzen	Metro Atlanta Chamber of Commerce
Kelly Cutrone	Myanmar Timber Enterprise
Brian McFadden	National Museum of Women in the Arts
Trista Sutter	Pentax Corp.
Paul Sculfor	New South Wales Rugby Union
Brandon McInerney	Bernama
Steve McPherson	Northwood University
Michael Sandy	National Center for Disaster Preparedness
Sean Preston	Northland Church
Spencer Pratt	Project Islamic Hope
Ezra Levant	Pamela Martin & Associates
Beyonce Knowles	Nottingham-Spirk Design Associates
Jamie Spears	National Wenchuan Earthquake Expert Committee
Jake Pavelka	Illinois Tool Works
Hugo Chavez	Independent Steelworkers Union
Melanie Fiona	Jakarta Globe
Holly Montag	Jewish National Fund
David Banda	North Phoenix Baptist Church
Kate Gosselin	New Jerusalem Foundation
Ali Larijani	Noordhoff Craniofacial Foundation
Robert Morgenthau	National Christmas Tree Association
Michael Johns	Norris Comprehensive Cancer Center
Carolyn Maloney	Nitschmann Middle School
Jeremy Hooper	National Red Cross

D.1 taula – Garapen faseko helburu entitateak.

Pertsonak	Erakundeak
Abdul Jalil Jan	AQR Capital Management
Abdul Karim al-Khawinay	Army Medical Command
Abdul Rahim Noor	BNC Mortgage
Abdullahi Hassan Garweyne	Badr Organization
Abu Zubaydah	Defensor Sporting
Ahmad Qattan	Denso
Al Hubbard	DirecTV
Alberto Gonzales	EnergySolutions
Barbara Boxer	FMCSA
Barry Goldwater	FMR Corp.
Berthold Huber	Federal Aviation Administration
Bill McAllister	Federal Election Commission
Bob Dillinger	FirstGroup
Bryan Baldwin	Fyffes PLC
Chen Zhu	GMAC's Residential Capital LLC
Christopher Bentley	German ARD
Dan Abrams	Ghent University
Danny Glover	Golden Baseball League
David Gregory	Highland Capital Partners
Fraser Robinson	Hindalco Industries
George Roy Hill	Hong Kong Disneyland
George Sheldon	IPSCO
Herb Gibson	IRGC-QF
Hu Sheng	Itochu
Hussein Kamel	Jacksonville Jaguars
James B. Stewart	KENNEDY SPACE CENTER
Jerome Robbins	Konica Minolta
John Kerry	Le Journal du Dimanche
John Negroponte	Le Point
Johnny Knoxville	MBIA
Juliette Binoche	MF Global
Justin Theroux	Markit
Lindsay M. Hayes	Marywood University
Lindsey Hord	Medina hospital
Lou Ferrara	Millonarios
M. Enkhbold	NEC Corp.
Mamoor Khan	NFC South
Manuel Barcena	National Action Network
Marco Contiero	New South Wales Waratahs
Mia Farrow	Philadelphia Inquirer
Mohamed ElBaradei	Red Sox
Murat Kurnaz	Royal Caribbean
Ospel	SLDN
Paavo Nurmi	Sadia
Paul Watson	Spelman College
Remy Ma	TACC
Richard Perle	Taiwan Research Institute
Roy Scheider	The Tax Policy Center
Sean Parker	Tyco Healthcare
Sean Ross	Zagat

D.2 taula – Testeko helburu entitateak.

E

Riedel datu multzoko erlazioen zerrenda

Freebase ezagutza basean oinarritutako Riedel *et al.*, 2010 datu multzoan 7 entitate izen mota desberdinek hartzen dute parte:

- Pertsonak (*people*)
- Erakundeak (*business*)
- Lekuak (*location*)
- Denbora (*time*)
- Filmeak (*film*)
- Kirolak (*sports*)
- Telebista saioak (*broadcast*)

Guztira 52 erlazio desberdin aurki ditzakegu, E.1 taulan zerrendatuta.

Entitate motak	**Erlazioak**
Telebista saioak (*broadcast*)	/broadcast/content/location /broadcast/producer/location
Erakundeak (*business*)	/business/business_location/parent_company /business/company/advisors /business/company/founders /business/company/locations /business/company/major_shareholders /business/company/place_founded /business/company_advisor/companies_advised /business/person/company /business/shopping_center/owner /business/shopping_center_owner/shopping_centers_owned
Filmeak (*film*)	/film/film/featured_film_locations /film/film_festival/location /film/film_location/featured_in_films
Lekuak (*location*)	/location/administrative_division/country /location/br_state/capital /location/cn_province/capital /location/country/administrative_divisions /location/country/capital /location/de_state/capital /location/fr_region/capital /location/in_state/administrative_capital /location/in_state/judicial_capital /location/in_state/legislative_capital /location/it_region/capital /location/jp_prefecture/capital /location/location/contains /location/mx_state/capital /location/neighborhood/neighborhood_of /location/province/capital /location/us_county/county_seat /location/us_state/capital
Pertsonak (*people*)	/people/deceased_person/place_of_burial /people/deceased_person/place_of_death /people/ethnicity/geographic_distribution /people/ethnicity/included_in_group /people/ethnicity/includes_groups /people/family/country /people/family/members /people/person/children /people/person/ethnicity /people/person/nationality /people/person/place_lived /people/person/place_of_birth /people/person/profession /people/person/religion /people/place_of_interment/interred_here /people/profession/people_with_this_profession
Kirolak (*sports*)	/sports/sports_team/location
Denbora (*time*)	/time/event/locations
NA (aipamen negatiboentzat)	

E.1 taula – Freebase ezagutza basean oinarrituta, datu multzoan dauden erlazio ezberdinen zerrenda.

F

Txioetako aipamenen normalizazioa, jarraitutako irizpideak

Ondorengoak dira garbiketa prozesurako erabilitako baldintzak:

- Magnitudeari buruzko adierazpenen normalizazioa:
  - *7.3-magnitude* → 7.3 - magnitude
  - *M7.3* → M 7.3
- Sakontasunari buruzko adierazpenen normalizazioa:
  - *35km* → 35 km
- Zenbakiak hitzez adieraztean, dagokien bihurketa zenbakizko balioetara:
  - *seven* → 7
  - *dozens* → 100
  - *hundreds* → 1000
  - *thousands* → 10000

- Koordenatu geografikoen adierazpenen normalizazioa:
  - *73.45 S* → -73.45
  - *73.45 ° S* → -73.45
  - *73.45 ° S* → -73.45
  - *73.45 N* → 73.45
  - *73.45 ° N* → 73.45
  - *73.45 ° N* → 73.45
  - *73.45 E* → 73.45
  - *73.45 ° E* → 73.45
  - *73.45 ° E* → 73.45
  - *73.45 W* → -73.45
  - *73.45 ° W* → -73.45
  - *73.45 ° W* → -73.45

G

Txioen eskuzko etiketatzea, jarraitutako irizpideak

Eskuzko etiketatzerako jarraitutako irizpideak ondorengoak dira:

- Gaztelaniaz eta alemanez dauden txioak ere etiketatu dira, gainontzeko hizkuntzetakoak ez.
- Denbora adierazpen guztiak ez dira normalizazio prozesuan normalizatu, txiolariak oso modu arraroan idatzi eta normalizatzailea ez delako gai izan hauek detektatu eta normalizatzeko. Hala ere hauek etiketatu ditugu.
- Txio batzuetan, duela urte asko gertatutako lurrikara bati buruzko informazioa ematen dute, lurrikara hau kasualitatez ezagutza baseko lurrikara baten leku berean gertatu zela eta, kasu horietan ez dugu ezer etiketatu.
- Bi lurrikara gertatu badira aldi berean bi leku desberdinetan, helburuko lurrikarari buruzko informazioa bakarrik etiketatu dugu, bestearen informazioa alde batera utzita. Ondorengo adibidea Indian gertatutako lurrikara bati dagokio, eta aldi berean Japonian, baina txio hau Indiako lurrikarari dagokionez, Japoniako informazioa ez dugu etiketatu:
  - An earthquake in Japan injures 43 people , while a separate quake in <country>India</country> 's <region>Andaman Islands</region> ...

- Sakontasunari buruzko aipamenetan, *km* adierazpena ere sartu dugu etiketaren barruan (<depth>20 km</depth>). Aurrerago konturatu ginen hau egitea ez zela erabaki ona izan.

- Hildako, zauritu eta desagertuei buruzko aipamenetan, informazio orokorra bakarrik etiketatu da, osagarria ez. Adibidez:

  – ... <dead>589</dead> people died , including 56 students ...

- Ordu eremuekin kontu handia izan dugu, Afghanistan bezalako estatuetan ordu eremuak estandarrarekiko 4 ordu eta erdiko (+4:30) desberdintasuna duelako, 4 edo 5 orduren ordez. Txiolari batzuek lurrikara gertatu den unea adierazteko, epizentroko momentu zehatza aipatzen dute. Kasu hauek oso urriak dira, eta gure urruneko gainbegiraketa sistemak ez ditu egoera hauek kontuan hartzen etiketatzerakoan.

Irizpide horiez gain, txioetako aipamen asko ere eskuz konpondu ditugu, etiketatu eta gero ikasketa prozesua garbia izateko. Ondorengo zerrendak zein argumentutan gertatu den adierazten du, konponketa horien adibide batzuk emanez:

- Magnitudeak:

  – *7 . 1* → <magnitude>7.1</magnitude>

  – *magnitude-7 .5* → magnitude - <magnitude>7.5</magnitude>

- Denbora adierazpenak:

  – *:2009 -11 -08 T19 :41:44* → <date>2009-11-08</date> <time>T19:41:44</time>

- Beste batzuk:

  – *more than 300ppls are lost* → more than <missing>300</missing> ppls are lost

H

Gertaera erauzketaren emaitzak, argumentuz argumentu

Eranskin honetan, 6. kapituluko garapen faseko esperimentu bakoitzean argumentuak banaka ebaluatuta lortutako emaitzak agertzen dira.

Erabilitako laburdurak:

- *Doit.* → Doitasuna.
- *Est.* → Estaldura.
- *F1* → F1 neurria.
- *#trip.* → Tripleta kopurua.

H.1 Eskuzko etiketatzea *vs* urruneko gainbegiraketa, oinarrizko sistema

Atal honetan eskuzko etiketatzea eta etiketatze automatikoaren emaitzak konparatzen dituen bi taula ditugu.

1. Lehenengoak argumentu bakoitzean banaka ebaluatuta lortutako emaitzak erakusten dizkigu, ebaluazio zorrotza erabiliz.
2. Bigarrenak berriz, ebaluazio erlaxatuaren emaitzak erakusten dizkigu.

	NoisyOR ebaluazio zorrotza					
	Eskuzko etiketatzea			**Urruneko gainbegiraketa**		
	Doit.	**Est.**	**F1**	**Doit.**	**Est.**	**F1**
Dena batera	44.13	26.14	32.83	53.05	21.97	31.07
Argumentua						
date	55.55	43.21	48.61	80.00	34.57	48.27
time	68.85	51.58	59.15	72.34	41.97	53.12
country	80.82	72.84	76.62	78.08	70.37	74.02
region	63.15	40.68	49.48	59.37	32.20	41.76
city	50	33.33	40	37.50	25.00	30.00
latitude	100.00	3.70	7.14	0/0	0/81	NaN
longitude	100.00	3.70	7.14	0/0	0/81	NaN
dead	15	11.11	12.76	21.43	5.55	8.82
injured	13.33	6.89	9.09	0/5	0/29	NaN
missing	0/2	0/4	NaN	0/0	0/4	NaN
magnitude	25	24.69	24.84	32.25	24.69	27.97
depth	5	2.77	3.57	0/0	0/72	NaN
affected-country	17.65	9.09	12.00	13.04	9.09	10.71
affected-region	0/3	0/2	NaN	0/0	0/2	NaN
landslides	100.00	14.28	25.00	100.00	14.28	25.00
tsunami	8.69	28.57	13.33	23.07	85.71	36.36
aftershocks	0/0	0/12	NaN	0/0	0/12	NaN
peak-acceleration	0/0	0/7	NaN	0/0	0/7	NaN
duration	0/0	0/5	NaN	0/0	0/5	NaN
foreshocks	0/0	0/3	NaN	0/2	0/3	NaN

H.1 taula – Eskuzko etiketatzeko eta urruneko gainbegiraketako argumentuen emaitzak banaka ebaluatuta, ebaluazio zorrotza erabiliz.

	NoisyOR ebaluazio erlaxatua					
	Eskuzko etiketatzea			**Urruneko gainbegiraketa**		
	Doit.	**Est.**	**F1**	**Doit.**	**Est.**	**F1**
Dena batera	64.18	38.01	47.74	67.41	27.92	39.48
Argumentua						
date	73.01	56.79	63.89	85.71	37.04	51.72
time	66.56	50.12	57.18	75.32	43.70	55.31
country	82.19	74.07	77.92	79.45	71.60	75.32
region	64.86	40.68	50.00	59.37	32.20	41.76
city	57.15	33.33	42.10	37.50	25.00	30.00
longitude	100.00	3.70	7.14	0/0	0/81	NaN
latitude	100.00	3.70	7.14	0/0	0/81	NaN
dead	38.13	28.24	32.45	30.95	8.02	12.74
injured	42.22	21.84	28.79	35.32	6.09	10.39
missing	23.39	11.69	15.59	0/0	0/4	NaN
magnitude	95.56	94.38	94.97	96.16	73.61	83.39
depth	40.74	29.10	22.63	0/0	0/72	NaN
affected-country	17.65	9.09	12.00	13.04	9.09	10.71
affected-region	0/3	0/2	NaN	0/0	0/2	NaN
aftershocks	0/0	0/12	NaN	0/0	0/12	NaN
peak-acceleration	0/0	0/7	NaN	0/0	0/7	NaN
duration	0/0	0/5	NaN	0/0	0/5	NaN
foreshocks	0/0	0/3	NaN	0/2	0/3	NaN

H.2 taula – Eskuzko etiketatzeko eta urruneko gainbegiraketako argumentuen emaitzak banaka ebaluatuta, ebaluazio erlaxatua erabiliz. *Landslides* eta *tsunami* argumentuenak ebaluazio zorrotzaren berdinak dira, hauek jakin nahi izanez gero ikus H.1 taula.

H.2 Urruneko gainbegiraketa tradizionala *vs* urruneko gainbegiraketa hurbila

Atal honetan beste bi taula ditugu, bata ebaluazio zorrotzerako (H.3 taula), bestea ebaluazio erlaxaturako (H.4 taula), urruneko gainbegiraketa tradizionala eta urruneko gainbegiraketa hurbila sistemen artean konparatzeko.

Ebaluazio zorrotza								
	UG$^{\text{hurbil}}$-BAE 0.95				**UG-BAE**			
Argumentua	**#trip.**	**Doit.**	**Est.**	**F1**	**#trip.**	**Doit.**	**Est.**	**F1**
affected-country	21	14.28	9.09	11.11	23	13.04	9.09	10.71
affected-region	0	0	0	0	0	0	0	0
aftershocks	0	0	0	0	0	0	0	0
city	7	28.57	16.66	21.05	8	37.50	25.00	30.00
country	74	81.08	74.07	77.42	73	78.08	70.37	74.02
date	52	65.38	41.97	51.13	35	80.00	34.57	48.27
dead	23	17.39	7.41	10.39	14	21.43	5.55	8.82
depth	3	0	0	0	0	0	0	0
duration	5	0	0	0	0	0	0	0
foreshocks	3	0	0	0	2	0	0	0
injured	4	0	0	0	5	0	0	0
landslides	1	100.00	14.28	25.00	1	100.00	14.29	25.00
latitude	8	0	0	0	0	0	0	0
longitude	5	0	0	0	0	0	0	0
magnitude	76	31.58	29.63	30.57	62	32.26	24.69	27.97
missing	0	0	0	0	0	0	0	0
peak-acceleration	0	0	0	0	0	0	0	0
region	33	63.63	35.59	45.65	32	59.37	32.20	41.76
time	22	86.36	23.46	36.89	47	72.34	41.97	53.12
tsunami	30	20.00	85.71	32.43	26	23.07	85.71	36.36
Dena batera	362	48.06	21.97	30.15	328	53.05	21.97	31.07

H.3 taula – Urruneko gainbegiraketa tradizionaleko eta urruneko gainbegiraketa hurbileko argumentuak banaka ebaluatuta, ebaluazio zorrotza erabiliz.

Ebaluazio erlaxatua								
	UG$^{\text{hurbil}}$-BAE 0.95				**UG-BAE**			
Argumentua	**#trip**	**Doit.**	**Est.**	**F1**	**#trip**	**Doit.**	**Est.**	**F1**
affected-country	21	14.28	9.09	11.11	23	13.04	9.09	10.71
affected-region	0	0	0	0	0	0	0	0
aftershocks	0	0	0	0	0	0	0	0
city	7	28.57	16.66	21.05	8	37.50	25.00	30.00
country	74	82.43	75.31	78.71	73	79.45	71.60	75.32
date	52	88.46	56.79	69.17	35	85.71	37.04	51.72
dead	23	23.17	9.87	13.84	14	30.95	8.02	12.74
depth	3	66.36	2.76	5.31	0	0	0	0
duration	5	0	0	0	0	0	0	0
foreshocks	3	0	0	0	2	0	0	0
injured	4	32.28	4.45	7.82	5	35.32	6.09	10.38
latitude	8	27.46	2.71	4.93	0	0	0	0
longitude	5	23.10	1.42	2.68	0	0	0	0
magnitude	76	97.03	91.04	93.94	62	96.16	73.61	83.39
missing	0	0	0	0	0	0	0	0
peak-acceleration	0	0	0	0	0	0	0	0
region	33	69.70	38.98	50.00	32	59.37	32.20	41.76
time	22	86.36	23.45	36.89	47	75.32	43.70	55.31
Dena batera	362	68.15	31.15	42.75	328	67.41	27.92	39.48

H.4 taula – Urruneko gainbegiraketa tradizionaleko eta urruneko gainbegiraketa hurbileko argumentuak banaka ebaluatuta, ebaluazio erlaxatua erabiliz. *Landslides* eta *tsunami* argumentuen emaitzak ebaluazio zorrotzaren berdinak dira, hauek H.3 taulan daude ikusgai.

H.3 Urruneko gainbegiraketa tradizionala *vs* ezaugarrien agregazioa

Atal honetan bi taula ditugu ere, bata ebaluazio zorrotzerako (H.5 taula), bestea ebaluazio erlaxaturako (H.6 taula), urruneko gainbegiraketa tradizionala eta ezaugarrien agregazioa erabiltzen duen sistemaren artean konparatzeko.

Ebaluazio zorrotza								
	UG$^{\text{agreg}}$-BAE				**UG-BAE**			
Argumentua	**#trip**	**Doit.**	**Est.**	**F1**	**#trip**	**Doit.**	**Est.**	**F1**
affected-country	21	28.57	18.18	22.22	23	13.04	9.09	10.71
affected-region	0	0	0	0	0	0	0	0
aftershocks	0	0	0	0	0	0	0	0
city	6	66.66	33.33	44.44	8	37.5	25.00	30.00
country	74	94.59	86.42	90.32	73	78.08	70.37	74.02
date	33	81.81	33.33	47.37	35	80.00	34.57	48.27
dead	15	26.67	7.41	11.59	14	21.43	5.55	8.82
depth	0	0	0	0	0	0	0	0
duration	0	0	0	0	0	0	0	0
foreshocks	2	0	0	0	2	0	0	0
injured	5	0	0	0	5	0	0	0
landslides	1	100.00	14.29	25.00	1	100.00	14.29	25.00
latitude	0	0	0	0	0	0	0	0
longitude	0	0	0	0	0	0	0	0
magnitude	62	33.87	25.92	29.37	62	32.26	24.69	27.97
missing	0	0	0	0	0	0	0	0
peak-acceleration	0	0	0	0	0	0	0	0
region	42	64.28	45.76	53.46	32	59.37	32.20	41.76
time	41	80.48	40.74	54.10	47	72.34	41.97	53.12
tsunami	27	22.22	85.71	35.29	26	23.07	85.71	36.36
Dena batera	329	60.48	25.12	35.50	328	53.05	21.97	31.07

H.5 taula – Urruneko gainbegiraketa tradizionalaren eta ezaugarrien agregazioaren argumentuak banaka ebaluatuta, ebaluazio zorrotza erabiliz.

Ebaluazio erlaxatua								
	UG$^{\text{agreg}}$-BAE				**UG-BAE**			
Argumentua	**#trip**	**Doit.**	**Est.**	**F1**	**#trip**	**Doit.**	**Est.**	**F1**
affected-country	21	28.57	18.18	22.22	23	13.04	9.09	10.71
affected-region	0	0	0	0	0	0	0	0
aftershocks	0	0	0	0	0	0	0	0
city	6	66.66	33.33	44.44	8	37.50	25.00	30.00
country	74	94.59	86.42	90.32	73	79.45	71.60	75.32
date	33	90.91	37.04	52.63	35	85.71	37.04	51.72
dead	15	34.89	9.69	15.17	14	30.95	8.02	12.74
depth	0	0	0	0	0	0	0	0
duration	0	0	0	0	0	0	0	0
foreshocks	2	0	0	0	2	0	0	0
injured	5	35.32	6.09	10.38	5	35.32	6.09	10.38
latitude	0	0	0	0	0	0	0	0
longitude	0	0	0	0	0	0	0	0
magnitude	62	96.22	73.65	83.44	62	96.16	73.61	83.39
missing	0	0	0	0	0	0	0	0
peak-acceleration	0	0	0	0	0	0	0	0
region	42	64.28	45.76	53.46	32	59.37	32.20	41.76
time	41	80.49	40.74	54.10	47	75.32	43.70	55.31
Dena batera	329	74.06	30.76	43.47	328	67.41	27.92	39.48

H.6 taula – Urruneko gainbegiraketa tradizionalaren eta ezaugarrien agregazioaren argumentuak banaka ebaluatuta, ebaluazio erlaxatua erabiliz. *Landslides* eta *tsunami* argumentuen emaitzak ebaluazio zorrotzaren berdinak dira, hauek H.5 taulan daude ikusgai.

H.4 Urruneko gainbegiraketa tradizionala *vs* urruneko gainbegiraketa hurbila eta ezaugarrien agregazioa

Amaitzeko, ebaluazio zorrotzerako (H.7 taula) eta ebaluazio erlaxaturako (H.8 taula), urruneko gainbegiraketa tradizionala eta ezaugarrien agregazioa hurbilarekin batera konbinatzen duen sistemaren emaitzen taulak dauzkagu.

Ebaluazio zorrotza								
	UG^{konb}-BAE 0.95				**UG-BAE**			
Argumentua	**#trip**	**Doit.**	**Est.**	**F1**	**#trip**	**Doit.**	**Est.**	**F1**
affected-country	18	22.22	12.12	15.68	23	13.04	9.09	10.71
affected-region	0	0	0	0	0	0	0	0
aftershocks	0	0	0	0	0	0	0	0
city	5	40.00	16.67	23.53	8	37.5	25.00	30.00
country	76	97.37	91.36	94.26	73	78.08	70.37	74.02
date	51	66.66	41.97	51.51	35	80.00	34.57	48.27
dead	22	22.73	9.26	13.16	14	21.43	5.55	8.82
depth	3	0	0	0	0	0	0	0
duration	0	0	0	0	0	0	0	0
foreshocks	4	0	0	0	2	0	0	0
injured	4	0	0	0	5	0	0	0
landslides	1	100.00	14.29	25.00	1	100.00	14.29	25.00
latitude	8	0	0	0	0	0	0	0
longitude	6	0	0	0	0	0	0	0
magnitude	76	30.26	28.39	29.30	62	32.26	24.69	27.97
missing	0	0	0	0	0	0	0	0
peak-acceleration	0	0	0	0	0	0	0	0
region	45	66.67	50.85	57.69	32	59.37	32.20	41.76
time	24	91.66	27.16	41.90	47	72.34	41.97	53.12
tsunami	31	19.35	85.71	31.58	26	23.07	85.71	36.36
Dena batera	374	53.74	25.38	34.47	328	53.05	21.97	31.07

H.7 taula – Urruneko gainbegiraketa tradizionalaren eta ezaugarrien agregazioa hurbilarekin batera konbinatzen duen sistemaren argumentuak banaka ebaluatuta, ebaluazio zorrotza erabiliz.

Ebaluazio erlaxatua								
	UG^{konb}-BAE 0.95				**UG-BAE**			
Argumentua	**#trip**	**Doit.**	**Est.**	**F1**	**#trip**	**Doit.**	**Est.**	**F1**
affected-country	18	16.66	9.09	11.76	23	13.04	9.09	10.71
affected-region	0	0	0	0	0	0	0	0
aftershocks	0	0	0	0	0	0	0	0
city	5	40.00	16.67	23.53	8	37.50	25.00	30.00
country	76	97.37	91.36	94.26	73	79.45	71.60	75.32
date	51	90.19	56.79	69.70	35	85.71	37.04	51.72
dead	22	27.71	11.29	16.04	14	30.95	8.02	12.74
depth	3	66.36	2.76	5.31	0	0	0	0
duration	0	0	0	0	0	0	0	0
foreshocks	4	0	0	0	2	0	0	0
injured	4	32.28	4.45	7.82	5	35.32	6.09	10.38
latitude	8	27.46	2.71	4.93	0	0	0	0
longitude	6	19.25	1.42	2.65	0	0	0	0
magnitude	76	96.99	91.00	93.90	62	96.16	73.61	83.39
missing	0	0	0	0	0	0	0	0
peak-acceleration	0	0	0	0	0	0	0	0
region	45	71.11	54.24	61.54	32	59.37	32.20	41.76
time	24	91.67	27.16	41.90	47	75.32	43.70	55.31
Dena batera	374	72.84	34.40	46.73	328	67.41	27.92	39.48

H.8 taula – Urruneko gainbegiraketa tradizionalaren eta ezaugarrien agregazioa hurbilarekin batera konbinatzen duen sistemaren argumentuak banaka ebaluatuta, ebaluazio erlaxatua erabiliz. *Landslides* eta *tsunami* argumentuen emaitzak ebaluazio zorrotzaren berdinak dira, hauek H.7 taulan daude ikusgai.