Korreferentzia-ebazpena euskaraz idatzitako testuetan

Ander Soraluze eta **Olatz Arregi** eta **Xabier Arregi** eta **Arantza Díaz de Ilarraza**

IXA taldea. Euskal Herriko Unibertsitatea.

Laburpena

Artikulu honetan euskarazko korreferentziak ebazteko sistema baten garapena azaltzen da. Lehenik eta behin, azterketa linguistiko batean oinarritutako aipamen-detektatzaile automatikoa aurkezten dugu. Sistema hori erregelatan oinarritutakoa da, eta egoera finituko teknologia erabiliz inplementatu da. Behin testuko aipamenak detektatuta, beraien artean gertatzen diren korreferentzia-erlazioak ebazten dituen sistema ere garatu dugu. Horretarako, ingeleserako diseinatu den sistema eraginkor bat oinarritzat hartu, eta euskararen ezaugarrietara egokitu da. Egokitze-lan hori deskribatzen da artikuluaren bigarren partean.

Hitz gakoak: aipamen-detekzioa, korreferentzia-ebazpena, hizkuntzaren prozesamendua

Abstract

This paper presents the first steps in the development of a Basque coreference resolution system. Firstly, we describe a mention detector system based on a linguistic study of the nature of mentions. The mention detector is rule-based and has been implemented using finite state technology. The system identifies mentions that are potential candidates to be part of coreference chains in Basque written texts. Finally, we describe the process of adapting a state-of-the-art English coreference resolution system to Basque.

Keywords: mention detection, coreference resolution, natural language processing

1 Sarrera eta motibazioa

Testu bateko bi espresio testualek objektu berbera adierazi edo erreferentziatzen dutenean, bi espresio horien artean korreferentzia-erlazio bat dagoela esan ohi da. Testu batean ager daitezkeen espresio testual horien arteko korreferentzia-erlazioak ebaztea helburu duen atazari korreferentzia-ebazpena deritzo.

Ataza honetan sarritan erabiltzen diren bi termino *entitatea* eta *aipamena* dira. Entitate bat mundu errealeko objektua edo objektu multzoa dela esaten da; aipamena, aldiz, entitate bati erreferentzia egiten dion espresio testuala da (Doddington *et al.*, 2004).

Ohikoa da korreferentzia-ebazpena bi azpi-ataza nagusitan banatzea: aipamenen detekzioa, batetik, eta erreferentzien ebazpena, bestetik (Pradhan *et al.*, 2011). Lehenbizi testuko aipamenak detektatzen dira, eta, ondoren erabakitzen da zein aipamenek egiten dioten erreferentzia entitate berari.

Azaldutako terminoak modu argiagoan ulertzeko ikus dezagun adibide bat.

(1) [Miguel Indurain] erretiratu zenean, [[hura] ordezkatuko zuen pertsona bat] bilatu nahian zebiltzan.

Goiko adibidean, kortxete artean hiru aipamen ikus ditzakegu, [Miguel Indurain], [hura] eta [hura ordezkatuko duen pertsona bat]. Garbi ikusten da [Miguel Indurain] eta [hura] aipamenek mundu errealeko objektu berbera erreferentziatzen dutela, beraz, korreferenteak direla esan dezakegu. Aipamen-detekzioan egin beharrekoa, hiru aipamenak zuzen identifikatzea da. Korreferenzia-ebazpenean berriz, [Miguel Indurain] eta [hura] lotu egin behar dira, mundu errealeko pertsona bera adierazten dute eta.

Gaur egun, korreferentzia-ebazpen automatikoa gakotzat har dezakegu testuak ulertu ahal izateko (Recasens, 2010); ondorioz, behar-beharrezkoa da diskurtsoaren ulerkuntza sakona eskatzen duten Lengoaia Naturalaren Prozesamenduko (NLP) hainbat atazatan; adibidez, informazioaren erauzketan, testuen laburpenean, galderak erantzuteko sistemetan, itzulpen automatikoan, sentimenduen analisian edota irakurketa automatikoan.

2 Arloko egoera eta ikerketaren helburuak

Azken bi hamarkadetan garrantzi handia eman zaio korreferentzia-ebazpenari eta gai honen inguruan zentratutako kongresu ugari antolatu dira. *Message Understanding Conference* (MUC-6, 1995; MUC-7, 1998) kongresuetan korreferentziaren inguruko ataza espezifikoak antolatu ziren. *The Automatic Content Extraction (ACE)* kongresuan aurredefinitutako entitate multzo baten arteko erlazioak identifikatzen saiatu ziren (Doddington *et al.*, 2004).

Azken urteetako kongresuei dagokienez, *SemEval-2010 Task 1* atazan korreferentzia-ebazpena gauzatu behar zen hizkuntza desberdinetan (Recasens *et al.*, 2010). Hurrengo urtean, *CoNLL-2011 Shared Task* atazan (Pradhan *et al.*, 2011), parte-hartzaileek Ontonotes corpusean (Pradhan *et al.*, 2007) ebatzi behar izan zuten korreferentzia ingeleserako, eta *CoNLL 2012 Shared Task 2* atazak (Pradhan *et al.*, 2012) ingelesaren, txineraren eta arabieraren gaineko korreferentzia-ebazpena egitea eskatzen zuen.

Aipamen-detekzioak berebiziko garrantzia du korreferentzia-ebazpenerako sistemetan. Hala adierazi da hainbat lanetan: Uryupina autoreak (2008) dio bere korreferentzia-ebazpenerako sistemaren estaldura-erroreen % 35 aipamen-detekzioan identifikatu gabeko aipamenen ondorioz dela, eta, horretaz gain, Uryupina berak (2010) gehitzen du doitasun erroreen % 20 gaizki identifikatutako aipamenengatik dela. Stoyanov *et al.* autoreek (2009), berriz, korreferentzia-ebazpenerako sistemetan erabiltzen diren aipamen-detektatzaileen hobekuntzak artearen egoera nabarmen hobetuko lukeela ondorioztatzen dute.

Aipamen-detektatzaileak garatzeko orduan erabilitako teknologiari dagokionez, bi joera nagusi nabarmendu daitezke: batetik, erregelatan oinarritutako sistemak ditugu, bestetik, ikasketa automatikoan oinarritutakoak. Ikasketa automatikoan oinarritutako sistemek doitasun eta estaldura balio orekatuak lortu ohi dituzte; erregelatan oinarritutakoek, aldiz, estaldura balio hobeak lortzen dituzte. Aipamenen detekzioan oso garrantzitsua da estaldura balio onak lortzea, bestela, detektatu gabe utzitako aipamenak ezin dira-eta aurreragoko pausoetan berreskuratu.

Ikasketa automatikoko sistemak errazago moldatzen dira beste hizkuntzetarako, baita garatuak izan ez direnetarako ere. Hala nabarmendu zen *CoNLL 2011 Shared Task* eta *CoNLL 2012 Shared task* atazetan, non, Uryupina eta Moschitti autoreek (2013) dioten moduan, erregelatan oinarritutako sistemek portaera okerragoa izan zuten ingelesetik arabierara edo txinerara egokitzean.

Ezaguna da euskara bezalako hizkuntza batek urri dituela baliabide linguistikoak, hizkuntza handi eta ahaltsuen aldean. Hori dela eta, korreferentzia-ebazpena bezalako atazetarako tresna eraginkorrak garatzea erronka handia da. Korreferentzia-ebazpenerako lehen sistemak ingeleserako sortuak izan ziren. Hala ere, azken urteetan ingelesaz gain beste hizkuntza batzuetarako sortuak izan diren sistemak ere baditugu. Adibidez, *SemEval-2010 Task 1* atazan katalana, nederlandera, alemana eta italiera bezalako hizkuntzetarako sistemak aurkeztu ziren (Recasens *et al.*, 2010). Lehen esan bezala, *CoNLL 2012 Shared Task* atazan txinera eta arabiera gehitu ziren. Hizkuntza gutxituei dagokienez, hasi dira korreferentzia-ebazpenerako sistemak agertzen. Azken urteotako lanetan ikus dezakegu, hungariera (Miháltz, 2008), poloniera (Ogrodniczuk eta Kopeć, 2011), txekiera (Nguy *et al.*, 2009) edo hindi hizkuntzetarako (Sobha *et al.*, 2011) sistemak garatu direla.

Lan honen helburua euskaraz idatzita dauden testuetan korreferentzia ebatziko duen sistema bat garatzea da. Hau lortzeko, lehenik aipamen-detektatzaile eraginkor bat garatuko da, eta ondoren detektatzaile hori korreferentzia-ebazpenerako sisteman integratuko da.

3 Ikerketaren muina

Euskarazko korreferentziak ebazteko tresna automatikoa sortzeko bidean, izenen (edo, zabalago adierazita, izenen funtzioa betetzen duten egituren) korreferentziak landuko ditugu lehenbizi. Gerorako utziko dugu gertaeren edota diskurtso mailako korreferentzien ebazpena.

Atal honetan, lehenik euskarazko testuetan ager daitezkeen aipamen motak azalduko ditugu. Ondoren, mota horietan oinarrituta, aipamenak automatikoki identifikatzeko garatu dugun tresnaren oinarriak eta ebaluazioa azalduko dira. Azkenik, aipamen-detektatzailearen irteera erabiltzen duen korreferentzia-ebazpenerako sistemaren nondik norakoak aurkeztuko ditugu.

3.1 Aipamenen azterketa

Aipamen-detektatzailea garatzen hasi aurretik, ezinbestekoa da jakitea zein espresio testual kontsideratu behar diren aipamentzat eta zein ez. Horiek osatzen dute, azken batean, korreferentetzat har daitezkeen elementuen

unibertsoa. Esana dugu izenen funtzioa betetzen duten egiturei erreparatuko diegula, baina horien guztien artean izaera erreferentziala duten espresioak hartuko ditugu aipamentzat. Ondorengo zerrendan adierazten dugu zein egitura linguistikok betetzen duten izen-aipamena izateko erreferentzialtasunaren baldintza.

- **Izen-kate arruntak:** Burutzat izen arrunta duten egiturak .

  (1) [Langileak] haserre daude hartutako erabakiarekin.

- **Izen bereziak:** Burutzat izen berezia duten egiturak.

  (2) [Clinton] itxaropentsu agertu zen kazetarien aurrean.

- **Izenordainak:** Izenordain pertsonal guztiak aipamentzat hartzen ditugu. Hala ere, euskaraz, determinatza-ile erakusleek hirugarren pertsonako izenordain gisa jokatzen dute (Laka, 1996). Anbiguotasun honi aurre egiteko, izenordain moduan erabiltzen diren erakusleak aipamen moduan markatzen ditugu (3. adibidea).

  (3) LDPko buruek Mori hautatu zuten apirilean Keizo Obuchi orduko lehen ministroa ordezkatzeko, [hark] tronbosia izan ostean.

- **Posesiboak:** Izenordain posesiboak aipamentzat hartzen dira, (4. adibidea).

  (4) Ekisoainek Granollersera joatea nahiago du, [bere] emazteak lagunak baititu bertan.

- **Aditz-izenak:** Aditzetatik eratorritako zenbait formak izenaren funtzioa betetzen dute, eta horrelakoak aditz-izenak direla esan ohi da. Aditz-izen horiek osatzen dituzten izen-kateak aipamen gisa etiketatuko ditugu.

  (5) [Europar Batasunaren zabaltze honek] arazo asko konpontzera behartuko ditu.

- **Postposizio lokuzioetako aipamenak:** Postposizio-lokuzioen kasuan, postposizio beregainaren aurreko izena hartuko dugu kontuan aipamentzat, izenaren lemari erantsirik ageri zaizkion atzizkiak barne. 6. adibidean *Athleticen* izango litzateke aipamena, *-ren aurka* postposizioari lotuta doana.

  (6) Moreno eta Vlatko Djolonga [Athleticen] aurka jokatzeko moduan daude.

- **Mendeko perpausa duten izen-kateak:** Izen-kate hauetan burua izen arrunta izango da eta izen horrek modifikatzaile gisa jokatuko duen mendeko perpaus bat izango du. 7. adibidean ikusten dugun bezala, mendeko perpausak (*Oslon hasitako*) izena (*prozesua*) osatu edo zehaztu egiten du, eta horregatik, mendeko perpausa ere aipamenaren barnean sartu dugu. Mendeko perpausen artean erlatibozkoak oso ohikoak dira, 8. adibidekoa bezala.

  (7) [Oslon hasitako prozesua] gaur bukatuko da.

  (8) [Antimisilen inguruan Pentagonoa atontzen ari den sistema] aurkeztu dute.

- **Elipsia:** Euskaraz, bada maila morfosintaktikoan ematen den elipsia. Elipsi mota hauek ere aipamentzat hartuko ditugu.

  (9) [Bigarren sailkatu zenak], segundo bakarra kendu zion.

- **Koordinazioa:** Izen-kate koordinatuak juntagailu (*eta, edo, edota, nahiz...*) baten bitartez lotzen diren izen-kateak dira. Koordinazio kasu hauetan, juntagailuaren ezker (*Xabier Mikel Errekondo*) eta eskuin (*Alvaro Jauregi*) aldeko osagaiak aipamentzat hartuko ditugu, bai eta egitura osoa (*Xabier Mikel Errekondo eta Alvaro Jauregi*) ere.

  (10) [[Xabier Mikel Errekondo] eta [Alvaro Jauregi]] kontratuak berritzear daude.

- **Lekuzko adberbioak:** Orokorrean adberbioek izaera erreferentzialik ez duten arren, lekuzko adberbioek badute izaera hori, aurretik aipatutako lekuren bat erreferentziatzen baitute. Ondorioz, lekuzko adberbioak aipamentzat hartu ditugu. 11. adibidean ikus dezakegu *han* lekuzko adberbioak, esaldian lehenago agertzen den *Biarritzera* aipamenari egiten diola erreferentzia.

  (11) Futbol jokalariak Biarritzera joatekoak ziren, [han] festa antolatuta baitzuten.

3.2 Aipamen-detektatzailea

Aipamenen detekzioa egiteko, komenigarria da tratatu nahi diren testuen aurreprozesaketa bat egitea. Aurreprozesaketa egiteko, IXA taldean sortutako analisi-katea erabiltzen dugu. Kate honetan, hainbat tresna erabiltzen dira, bakoitzak bere zeregin zehatza duelarik. Tresna horiek, testuen analisi sakona egiten dute urrats desberdinetan banatuta (tokenizazioa, analisi morfologikoa...). Kate horretako tresnetatik honako hauek erabiltzen dira aipamen-detekzioan: analizatzaile morfologikoa, lematizatzailea, hitz-anitzeko unitateen detektatzailea, entitate izendunen ezagutzaile eta sailkatzailea, izen- eta aditz-kateen detektatzailea, eta perpaus-detektatzailea. Aurreprozesaketan erabilitako tresna hauek guztiak Lengoaia Naturalaren Prozesamendurako tresna orokorrak dira eta beraz zehazki ez dira aipamenen detekzioa egiteko sortuak izan. Ondorioz, tresna hauen emaitzetatik lortzen diren aipamenen mugak ez dira erabat zuzenak, eta doitze-lana eskatzen dute. Horretarako 34 erregela definitu dira eskuz eta erregela horiek konpilatuz 12 Egoera Finituko Transduktoreak (*Finite State Transducers, FST*) lortu dira. Egoera Finituko Teknologia erabiliz datu multzo handiak azkar eta memoria gutxi kontsumituz prozesa daitezke. *Foma* (Hulden, 2009), automata eta transduktoreekin lan egiteko aukera eskaintzen duen kode irekiko tresna, erabili dugu erregelak definitu eta transduktoreak lortzeko.

Gure FSTek analisi-katetik jasotako informazio linguistikoa erabiltzen dute aipamenak eta beraien mugak identifikatzeko. Garatutako FSTak 3.1. azpiatalean deskribatutako aipamen motak identifikatzeaz arduratzen dira. FST hauen inguruko informazio gehiago (Soraluze *et al.*, 2012) artikuluan aurki daiteke.

Sistema garatu eta ebaluatzeko erabili den corpusa, testu-bilduma, EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) izan da (Aduriz *et al.*, 2006). Corpus honen helburua, euskararako Lengoaia Naturalaren Prozesamendurako tresnak garatzeko orduan erreferentzia-corpusa izatea da. *Euskaldunon Egunkarian* argitaratutako berrien bilduma bat da, 300.000 hitzez osatua dagoena. Corpusa eskuz etiketatua izan da analisi maila ezberdinetan (morfologia, sintaxia, izen- eta aditz-kateak...). Berriki, aipamenak eta korreferentzia-kateak ere etiketatuak izan dira. Aipamen-detektatzailea garatzeko EPEC corpusaren zati bat erabili da, 46.383 hitzez osatua eta 12.792 aipamen dituena.

Sistema automatiko bat garatzen denean, komenigarria izaten da ebaluazio bat egitea, sistemaren portaera aztertu eta dituen gabeziez jabetzeko. Aipamenen detekzioa ebaluatzeko, nazioartean ezagunak diren bi parekatze-metodo erabili ditugu: *Lenient Matching* eta *Strict Matching*. *Lenient Matching* parekatzean aipamen bat zuzena dela kontsideratzen da, automatikoki detektatu den aipamenaren mugak urre-patroiaren (eskuz etiketatu den aipamenaren) mugen barnean badaude eta burua (*head word*) ere aipamenaren barnean kokatzen bada (Kummerfeld *et al.*, 2011). Hala ere, parekatze-metodo zorrotzagoak aplikatu izan dira. Adibidez, *CoNLL-2011 Shared Task*-en (Pradhan *et al.*, 2011), *Strict Matching* metodoa erabili zen. Metodo honen arabera, aipamen bat zuzena dela kontsideratzen da baldin eta soilik baldin urrezko aipamenaren berdina bada. Parekatze-metodo bakoitza hobeto ulertzeko hona hemen adibide bat.

- Esaldia: Argentinan DINAK egindako krimenak ikertuko dituzte.
- Eskuz etiketatutako aipamena: [Argentinan DINAK egindako krimenak]
- Aipamen-detektatzaileak proposatutako aipamena: [DINAK egindako krimenak]

Sistemaren erantzuna eskuz etiketatutako zatiarekin konparatzen da bi parekatze-metodoak erabiliz, eta *Lenient Matching* protokoloak dio aipamena zuzena dela, eskuz etiketatutako aipamenaren mugak ez dituelako gainditzen eta gainera aipamenaren burua (*krimenak*) mugen barruan dagoelako. *Strict Matching* protokoloa erabiliz, aldiz, proposatutako aipamena ez litzateke zuzentzat hartuko, eskuz etiketatutakoaren berdin-berdina izateko *Argentinan* hitza falta baitzaio.

1 Table: Aipamen-detektatzaileak lortutako emaitzak.

	P	R	F_1
SM	76.85	78.59	**77.58**
LM	81.96	83.97	**82.81**

1. taulan ikus daitekeenez, *Strict Matching (SM)* neurriaren arabera, aipamen-detektzaileak 78.59eko estaldura (R) lortzen du, hau da identifikatu beharreko aipamen guztien artean % 78.59 identifikatu ditu ondo, eta 76.85eko doitasuna (P), identifikatu dituen aipamen horien artean % 76.85 zuzenak dira. F_1 neurriak doitasuna eta estalduraren arteko batazbeztekoa adierazten du. 77.58koa da *Strict Matching (SM)* neurria erabiltzean. *Lenient Matching (LM)* parekatze-metodoarekin, emaitzak hobexeagoak dira, 83.97ko estaldura, 81.96ko doitasuna eta 82.51eko F_1 balioa.

3.3 Korreferentzia-ebazpenerako sistema

Euskararako korreferentzia-sistema garatzeko, ingeleserako garatu zen eta *CoNLL-2011 shared task* atazan (Pradhan *et al.*, 2011) emaitzarik onenak lortu zituen Stanfordeko unibertsitatean garatutako korreferentzia-ebazpenerako sistema (Lee *et al.*, 2013) egokitu dugu. Sistema egokitzen hasi aurretik, gure corpusean ebaluatu da eta lortu den CoNLL F_1 balioa (Pradhan *et al.*, 2014) 48.67koa izan da.

Sistema hau erregelatan oinarritutakoa da, eta oinarrian 10 *bahe* (korreferentzia-ebazpenerako modulu espezifiko) erabiltzen ditu. 10 bahe horiek banan-banan aplikatzen dira, doitasun handiena lortzen dutenak aurrenik eta doitasun baxuagokoak ondoren. Planteamendu honi esker, lehenengo baheetan erabaki ziurrak hartzen dira (doitasun handia) eta ondorengoetan ez hain ziurrak, estaldura hobetuz baina batzuetan doitasuna kaltetuz. Sistema Honakoak dira ingeleserako sistemak dituen baheak:

1. ***Speaker identification*** bahea: Estilo zuzenean agertzen diren esaldietan hizlariak identifikatu eta dagozkien izenordainekin lotzen ditu.

   (12) "[Niretzat] [oso urte txarra] izan zen [jaitsi ginenekoa], [[nire] karrerako tristeena]".
   Korreferentzia-katea: [Niretzat] -> [nire].

2. ***Exact String Match*** bahea: Bi aipamen guztiz berdinak badira, korreferentzia-kate berean elkartzen ditu.

   (13) [Milosevic presidentea] [Argentinara] joan da. [Milosevic presidentea] atzerrian dabil.
   Korreferentzia-katea: [Milosevic presidentea] -> [Milosevic presidentea]

3. ***Relaxed String Match*** bahea: Bi aipamen korreferentzia-kate berean biltzen ditu, aipamenetatik mendeko perpausak ezabatu ostean gelditzen diren aipamenen buruak berdinak badira. 14. adibidean, [Estatu Batuetako presidentea den Bill Clinton] aipamenari mendeko perpausa kenduko bagenio [Bill Clinton] aipamena lortuko genuke, eta aipamen hori ondoren agertzen den [Bill Clinton] aipamenaren berdina da, beraz, baheak korreferentzia-kate berean biltzen ditu.

   (14) [Estatu Batuetako presidentea den Bill Clinton] itxaropentsu agertu zen [kazetarien] aurrean. [Bill Clinton] Irakeko gerraz mintzatu zen.
   Korreferentzia-katea: [Estatu Batuetako presidentea den Bill Clinton] -> [Bill Clinton]

4. ***Precise Constructs*** bahea: Bi aipamen korreferenteak direla ebazten du baldin eta aposizio-egituran daude (15. adibidea), predikazio-egituran daude (16) edo akronimoak dira (17).

   (15) [Ibarretxe], [EAEko lehendakaria], etorri da.
   Korreferentzia-katea: [Ibarretxe] -> [EAEko lehendakaria]

   (16) [Ibarretxe] [lehendakaria] da.
   Korreferentzia-katea: [Ibarretxe] -> [lehendakaria]

   (17) [Euskal Autonomi Erkidegoa] [hiru probintziaz] osatua dago. [Aurten] [bisitari ugari] izan dira [[EAE] ikustera etorri direnak].
   Korreferentzia-katea: [Euskal Autonomi Erkidegoa] -> [EAE]

5. ***Strict Head Match A*** bahea: Ondorengo hiru murriztapen hauek aldi berean betetzen badira, bi aipamenak korreferentetzat hartzen ditu:
   - *Entity Head Match*: Aipamenaren buruak, aurrekari hautagaiaren (testuan lehenago agertu den aipamena) entitateko aipamenen bururen batekin parekatu behar du.
   - *Word Inclusion*: Entitate bateko *stopword*[1] ez diren hitz guztiak, aurrekariaren entitatean ere egon behar dute.
   - *Compatible Modifiers Only*: Aipamen baten modifikatzaile guztiak aurrekari hautagaiak ere izan behar ditu. Modifikatzaile gisa, izen arruntak eta adjektiboak kontsideratzen dira.

   18. adibideko aipamenei begiratzen badiegu, [Madrileko Auzitegi Gorenera] eta [Auzitegi Gorenera] aipamenek korreferenteak izateko hiru baldintzak betetzen dituzte. Bi aipamenen buruak *Auzitegi* dira, beraz lehen baldintza betetzen da. Stopwordak ez diren hitz guztiak, *Auzitegi* eta *Gorenera* badaude aurrekariaren entitatean, hortaz, bigarren baldintza ere betetzen da. Azkenik, bi aipamenek ez dute elkarrekin bateraezinak diren modifikatzailerik, *Madrileko* izen berezia baita eta ez izen arrunta. Ondorioz, bi aipamenak korrefentzia-kate berean biltzen dira. Aldiz, [kotxe gorrian] eta [kotxe berdean] aipamenak

[1] Esanahirik gabeko hitzak, hala nola, puntuazio-ikurrak, artikuluak...

ez dira korreferentetzat kontsideratuko. Ez baitira bigarren baldintza (berdean hitza ez dago aurrekariaren entitatean) eta hirugarrena (berdean eta gorrian hitzak modifikatzaile bateraezinak dira) betetzen.

(18) [Garzon epailea] [kotxe gorrian] iritsi zen [Madrileko Auzitegi Gorenera]. [Auziperatua] berriz [kotxe berdean] heldu zen [Auzitegi Gorenera].
Korreferentzia-katea: [Madrileko Auzitegi Gorenera] -> [Auzitegi Gorenera].
Ez dira korreferenteak: [kotxe berdean] eta [kotxe gorrian].

6. ***Strict Head Match B*** bahea: *Strict Head Match A* bahearen aldaera bat da. Bi aipamen korreferentetzat kontsideratuko dira, *Entity Head Match* eta *Word Inclusion* murriztapenak betetzen badira. Bahe honetan *Compatible Modifiers Only* murriztapena ez da aplikatzen.
7. ***Strict Head Match C*** bahea: Bahe hau ere *Strict Head Match A* bahearen aldaera bat da. Kasu honetan, bi aipamen korreferentzia-kate berean bilduko dira baldin eta *Entity Head Match* eta *Compatible Modifiers Only* murriztapenak betetzen badira. *Word Inclusion* murriztapena ez da aplikatzen.
8. ***Proper Head Word Match*** bahea: Burutzat izen berezi berdina duten bi aipamen korreferentetzat hartzen ditu, baldin eta ez badute kokapen-bateraezintasunik, ezta bateraezintasun numerikorik ere. Bahe honek 19. adibideko [Taman Salam] eta [Salam] aipamenak korreferentzia-kate berean bilduko lituzke, bi aipamenen burua izen berezi berdina (*Salam*) delako eta bien artean ez dagoelako ez kokapen-bateraezintasunik ez eta numerikorik ere. Aldiz, ez lituzke bilduko [Libano] eta [Libano hegoaldean], nahiz eta bien buruak izen berezi berdinak izan (*Libano*), kokapen bateraezintasun bat dute, bigarren aipamenak *hegoaldea* adierazten du eta lehenengoak ez.

(19) [Tamam Salam] arduratuta dago [[Libano] [azken egunetan] jasaten ari den atentatuak] dela eta. [Salam] [[[Libano hegoaldean] izandako atentatuak] sortutako kalteak] ikusten izan da.
Korreferentzia-katea: [Taman Salam] -> [Salam].
Ez dira korreferenteak: [Libano] eta [Libano hegoaldean].

9. ***Relaxed Head Match*** bahea: Lehen aipatutako *Entity Head Match* murriztapena erlaxatzen du, eta aipamen baten burua aurrekari hautagaiaren entitateko hitz baten berdina izatea eskatzen du soilik. Erlaxazio horri esker, adibidez, [Baltasar] aipamena hiru aipamen hauez [Garzon], [epailea], [Audientzia Nazionaleko epaile Baltasar Garzon] osatutako entitatearekin biltzen du. Bahearen doitasuna mantentzeko, *Entity Head Match* murriztapen erlaxatuaz gain *Word Inclusion* murriztapena ere bete behar da.
Korreferentzia-katea: [Garzon] -> [epailea] -> [Audientzia Nazionaleko epaile Baltasar Garzon] -> [Baltasar]
10. ***Pronoun Resolution*** bahea: Izenordainei dagokien korreferentzia-katea identifikatzeaz arduratzen da. Izenordainak eta aurrekariak honako murriztapenak bete behar dituzte korreferentetzat kontsideratzeko: numero-, genero- eta pertsona-ezaugarri berdinak izan behar dituzte, biek bizidun edo biek bizigabe izan behar dute, aurrekaria entitate izenduna bada, entitatearen motak bat etorri behar du izenordainaren motarekin, eta izenordainaren eta aurrekarien arteko distantziak 3 esaldikoa baino txikiagoa izan behar du.

(20) [Palaciosentzat], [ordura arteko [[bere] idoloak]] [malkoei] ezin eutsiz ikustea izan zen [gogorrena].
Korreferentzia-katea: [Palaciosentzat] -> [bere].

Sistemaren arkitektura guztiz modularra izanik, erraz integra daitezke korreferentzia-ebazpenerako bahe berriak; hori dela eta, ingelesa ez beste hizkuntzetarako ere nahiko erraz egokitzen ahal da. Dakigunez, ingelesa eta euskara ezaugarri desberdineko hizkuntzak dira. Euskara hizkuntza eranskaria eta ordena librekoa da, ingelesa aldiz ez. Ondorioz, sistemaren egokitzapenean euskararen ezaugarriak kontuan hartu behar izan dira, eta bahe batzuk egokitu eta bahe berriak sortu.

Sistemak IXA taldeko analisi-katearekin prozesatutako testuak eta garatutako aipamen-detektatzailearen irteera jasotzen ditu sarrera moduan. Horiek erabiliz, gai da euskarazko testuetako korreferentzia-erlazioak identifikatzeko.

Egokitzapen-prozesuan morfologiaren erabilera sakonagoa egiten dugu, hizkuntza eranskariaren ezaugarriak hobeto lantzeko. Adibidez,*Exact Match* baheak [Milosevicek], [Milosevic], [Milosevici]... bezalako aipamenak ez lituzke korreferentzia-kate berean lotuko. Nahiz eta beraien lemak (*Milosevic*) berdinak izan, *Exact Match* baheak guztiz berdinak diren aipamenak lotzen ditu, hau da, forma berdina duten aipamenak. Ingelesez kasu horiek aipamenaren mugetatik kanpo dauden preposizio bidez adieraziko lirateke, adibidez, “to [Milosevic]”, “with [Milosevic]”... beraz, formak begiratzea nahikoa da, baina euskara bezalako hizkuntza eranskari batean ez.

Arazo hau dela eta, *Exact Match* egokitu dugu eta *Exact Morphology Match* izena eman diogu.

11. ***Exact Morphology Match*** bahea: Bi aipamen korreferentzia-katea berean elkartzen ditu beraien lemak, numeroa eta mugatasuna berdinak badira.

    (21) [Txori polita] ikusi du [gizonak]. [Txori politak] txio egin dio.
    Korreferentzia-katea: [Txori polita] -> [Txori politak]

    (22) [Txori politak] ikusi ditu [gizonak]. [Txori politak] txio egin dio.
    Ez dira korreferenteak: [Txori politak] -> [Txori politak].

Inplementatu dugun bahe honek, 21. adibideko [Txori polita] eta [Txori politak] aipamenak korreferentzia-kate berean lotuko lituzke. Deklinabidea kenduta *txori polit* berdina da bi aipamenetan, eta biak singular mugatuak dira. Aldiz, ez lituzke 22. adibideko [Txori politak] eta [Txori politak] lotuko, nahiz eta bi aipamenak berdinak izan, bai formaz bai lemaz, lehenengoa plural mugatua da eta bigarrena berriz singular mugatua, alegia, beraien numeroak ez datoz bat.

Ondoren, bahe bakoitza euskarara egokitzeko egin ditugun moldaketak azalduko dira.

Relaxed String Match baheari dagokionez, bahea egokitu dugu euskarazko mendeko perpausak identifikatu eta bi aipamenen konparaketa egiteko, bati mendeko perpausa kendu ostean. Kasu honetan ere, bi aipamenak konparatzerako orduan, lemaren erabilera egiten da formaren ordez, eta numeroa eta mugatasunaren murriztapenak betetzea eskatzen da bi aipamenak korreferentetzat hartzeko.

Precise Constructs bahean ere egin ditugu aldaketak. Euskaraz aposizio eta predikazio egiturak detektatzeko tresnak integratu ditugu bahe honetan.

Strict Head Match A, B eta C baheak egokitzeko *Entity Head Match* murriztapena aldatu dugu, aipamenen buruen formak begiratu ordez, aipamenen buruen lemak, numeroa eta mugatasuna konparatzen dira.

Proper Head Word Match bahean aipamen-buruen lemak konparatzen dira egokitu dugun bahe berrian, ingelesekoan aldiz formak.

Relaxed Head Match bahean ere, aipamen-buruen formak begiratu ordez, buruen lemak eta aipamenen numeroa eta mugatasuna konparatzen dira, bi aipamen korreferenteak diren edo ez erabakitzeko.

Pronoun Resolution bahean ere egin ditugu aldaketak. Bahe honetan, zehazki, izenordainak identifikatzeko sistemak erabiltzen duen zerrenda euskaratu dugu. Horretaz gain, izen bizidunen eta bizigabeen zerrenda egokitu da euskararako.

Egokitzapenarekin amaitzeko, sisteman integratu dugun *Ellipsis* bahe berria azalduko dugu. 3.1 azpiatalean aipatu dugun moduan, euskaraz bada maila morfosintaktikoan ematen den elipsia, eta horrelako egiturak aipamentzat hartzen ditugu. Ingeleserako sistemak ez dauka aipamen mota horiek tratatzeko inolako baherik, ingelesez ez baita fenomeno hori gertatzen. Euskaraz, ordea, aipamen eliptiko horiek bere aurrekariarekin lotzeko beharra ikusi dugu, eta hori dela-eta bahe berria sortu da. Bahe horri esker, 23. adibidean agertzen den [kalitate handikoak] bezalako aipamen eliptikoak bere aurrekariarekin, kasu honetan [Marcelo Nicola eta Walter Guiñazu], lotzeko gai gara.

(23) [[Marcelo Nicola] eta [Walter Guiñazu]] [oso gazteak] ziren arren, [kalitate handikoak] ziren, eta [etorkizun oparoa] zuten.
Korreferentzia-katea: [Marcelo Nicola eta Walter Guiñazu] -> [kalitate handikoak]

Behin korreferentzia-sisteman proposatutako aldaketak eginda, honen ebaluazio bat egin dugu. Lortutako emaitzak 2. taulan ikus daitezke. Erabilitako metrikak korreferentzia-ebazpena ebaluatzeko erabiltzen diren ohiko metrikak dira (Pradhan *et al.*, 2014).

Metrika	R	P	F_1
MUC	36.63	44.34	40.11
B^3	58.34	64.08	61.08
$CEAF_m$	58.52	60.00	59.25
$CEAF_e$	60.99	58.71	59.83
BLANC	39.13	47.64	42.44
CONLL			53.67

2 Table: Korreferentzia-sistemaren emaitzak.

Lortutako emaitzak ezin dira beste hizkuntzetan egin diren lanekin konparatu, baina sistemaren egokitzapena egin aurretik lortutako emaitzarekin (48.67) alderatuta, 5 puntuko hobekuntza lortu da.

4 Ondorioak

Euskarazko testuetan korreferentzia-ebazpena egiteko gai den sistema bat garatu dugu. Horretarako, lehenik aipamen-detektatzaile bat sortu dugu, aipamenen azterketa linguistiko batean oinarrituta eta FST teknologia erabiliz. Lortutako emaitzak onak izan dira.

Ondoren, korreferentzia-ebazpenerako sistema eraginkor bat euskararako egokitu dugu, hizkuntzaren ezaugarriak kontuan hartuta. Korreferentzia-sistema honek automatikoki detektatutako aipamenak erabiltzen ditu korreferentzia-erlazioak sortzeko. Horrela, testu soiletik hasita analisi-urrats guztiak automatikoki egiten dira korreferentzia-kateak lortu arte.

Sistemaren garapenean, EPEC corpusa ere aberastu dugu, aipamenak eta korreferentzia-erlazioak etiketatuz.

5 Etorkizunerako planteatzen den norabidea

Etorkizunean, aipamen-detektatzailearen ebaluazio kualitatibo bat egin nahi dugu, egiten dituen erroreak aztertu eta horiek konpontzeko irtenbideak topatzeko.

Korreferentzia-ebazpenerako sistemari dagokionez, egokitzapen prozesuarekin jarraitu nahi dugu, baheak hobetuz eta beharra balego berriak sortuz. Hobetze-bide horretan, beharrezkoa ikusten dugu korreferentzia-ebazpenerako sistemaren ebaluazio kuantitatibo nahiz kualitatibo bat egitea, erroreak aztertu eta beraien kausa jakiteak hobekuntzarako bide berriak eskainiko baitizkigu. Korreferentzia-erlazio zailenak ebazteko ikasketa automatikoko teknikak erabiltzeko aukera ere aztertzen ari gara.

Hizkuntza arteko korreferentzia-ebazpenaren bidea ere urratu nahi dugu, hau da, corpus paraleloak (testu berdinak hizkuntza desberdinetan eta parekatuta dituzten corpusak) erabiliz, hizkuntza batean ebazten diren korreferentzia-erlazioek beste hizkuntza batekoak ebazteko lagundu dezaketen aztertu nahi dugu.

Erreferentziak

Aduriz, Itziar, Maxux Aranzabe, Jose Mari Arriola, Maite Atutxa, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa, eta Ruben Urizar. 2006. Methodology and Steps towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic Levels for the Automatic Processing. 1–15. Rodopi. Book series: Language and Computers.

Doddington, George, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, eta Ralph Weischedel. 2004. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. In *Proceedings of Language Resources and Evaluation Conference*, (LREC 2004), 837–840.

Hulden, Mans. 2009. Foma: a Finite-state Compiler and Library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, (EACL 2009), 29–32, Stroudsburg, PA, USA.

Kummerfeld, Jonathan K., Mohit Bansal, David Burkett, eta Dan Klein. 2011. Mention Detection: Heuristics for the OntoNotes Annotations. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), 102–106, Stroudsburg, PA, USA.

Laka, Itziar, 1996. A Brief Grammar of Euskara, the Basque Language. http://www.ehu.es/grammar. University of the Basque Country.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, eta Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Computational Linguistics* 39.

Mihálz, Márton. 2008. Knowledge-based Coreference Resolution for Hungarian. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

MUC-6. 1995. Coreference Task Definition (v2.3, 8 Sep 95). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, 335–344, Columbia, Maryland, USA.

MUC-7. 1998. Coreference Task Definition (v3.0, 13 Jul 97). In *Proceedings of the 7th Message Understanding Conference (MUC-7)*, Fairfax, Virginia, USA.

Ngụy, Giang Linh, Václav Novák, eta Zdeněk Żabokrtský. 2009. Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. In *Proceedings of the SIGDIAL 2009 Conference*, p. 276–285, London, UK. Association for Computational Linguistics, Association for Computational Linguistics.

Ogrodniczuk, Maciej, eta Mateusz Kopeć. 2011. End-to-end Coreference Resolution Baseline System for Polish. In *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, ed. by Zygmunt Vetulani, 167–171, Oznań, Poland.

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, eta Michael Strube. 2014. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 30–35. Association for Computational Linguistics.

——, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, eta Yuchen Zhang. 2012. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

——, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, eta Nianwen Xue. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), 1–27, Stroudsburg, PA, USA.

Pradhan, Sameer S., Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, eta Ralph Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. In *Proceedings of the International Conference on Semantic Computing*, (ICSC 2007), 517–526, Washington, DC, USA. IEEE Computer Society.

Recasens, Marta, 2010. *Coreference: Theory, Annotation, Resolution and Evaluation*. University of Barcelona tesia.

——, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, eta Yannick Versley. 2010. SemEval-2010 task 1: Coreference Resolution in Multiple Languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), 1–8, Stroudsburg, PA, USA.

Sobha, Lalitha Devi, Rk Rao Pattabhi, R. Vijay Sundar Ram, Cs. Malarkodi, eta A. Akilandeswari. 2011. Hybrid Approach for Coreference Resolution. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, (CoNLL 2011), 93–96, Stroudsburg, PA, USA.

Soraluze, Ander, Olatz Arregi, Xabier Arregi, Klara Ceberio, eta Arantza Díaz de Ilarraza. 2012. Mention Detection: First Steps in the Development of a Basque Correference Resolution System. In *KONVENS 2012, The 11th Conference on Natural Language Processing*, Vienna, Austria.

Stoyanov, Veselin, Nathan Gilbert, Claire Cardie, eta Ellen Riloff. 2009. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 656–664, Suntec, Singapore.

Uryupina, Olga. 2008. Error Analysis for Learning-based Coreference Resolution. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

——. 2010. Corry: A System for Coreference Resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 100–103, Uppsala, Sweden. Association for Computational Linguistics.

——, eta Alessandro Moschitti. 2013. Multilingual Mention Detection for Coreference Resolution. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 100–108, Nagoya, Japan.

6 Eskerrak eta oharrak

Lan hau Euskara Errektoreordetzaren (UPV/EHU) doktore-aurreko beka batek, Ber2Tek proiektuak (IE12-33) eta IXA taldea, A motako ikertalde finkatuari (IT344-10) EJk emandako diru laguntzak finantziatua izan da.