EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Doktorego-tesia

Korreferentzia-ebazpena euskarazko testuetan

Ander Soraluze Irureta

2017

EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Korreferentzia-ebazpena euskarazko testuetan

Ander Soraluze Iruretak Olatz Arregi Uriarte eta Xabier Arregi Iparragirreren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2017ko maiatza.

A todos nos gusta que lo bueno dure siempre, pero luego dura lo que tiene que durar, normalmente, más de lo debido.

Francis Diéz

Eskerrak

Lehenik eta behin eskerrak eman nahi nizkieke tesi-lan honen zuzendariak izan diren Olatz eta Xabierri, bidea irekitzeagatik, bidean gidatzeagatik, konfiantzagatik eta pazientziagatik. Umorea eta lana uztartzen bikainak izateagatik eta tunel luze eta ilunen amaieran beti argia dagoela erakusteagatik. Eskerrik asko Donostiako *aitatxo* eta *amatxo*. Arantza zuri ere mila esker, Xabierren hutsunea hain ongi betetzeagatik.

Ixakide guztiei, tesi hau gauzatzeko eman dizkidazuen gomendio eta laguntza guztiagatik eta hasiera hasieratik, bederatzi urte jada IXA taldean hasi nintzela, taldearen parte sentiarazteagatik.

318 bulegoko lankideei, berogailuaren tenperatura aukeratzeko sortutako eztabaidatxoez aparte oso giro ona dagoelako bulegoan. Eskerrik asko pasillora bidaltzearen mehatxua ez betetzeagatik. *Gustura nago zuekin, jada ez dut bulego txikira joan nahi!*

I am also very grateful to Massimo Poesio and the whole Language and Computation Group at School for Computer Science and Electronic Engineering (CSEE) of the Essex University. You were exceptional hosts during my stay as a visitor researcher in the United Kingdom. It was an unforgettable experience.

Donostiako kuadrillatxoari lanetik kanpo pasatu ditugun eta oraindik pasatzeko dauzkagun momentu zoragarri guztiengatik. Noizko hurrengo kazuela? Hor doakizue proposamen bat: *1, 2, Aspes, 5, 6, Fregona, 8, 9, Aldatu izarak!, 11, Hija dejale dejale, 13, Elvira. Ueeeeeeeee. Txin txin!*

Azpeitiko lagunak ere ezin ahaztu. Donostiara bizitzera etorri nintzenetik

nahi baino gutxiago egon gara elkarrekin, ea hemendik aurrera gehiago egoteko aukera daukagun. Eskerrik asko Azpeitira joandako bakoitzean eskaini dizkidazuen tartetxoengatik eta tesia bukatzeko sartu didazuen presioagatik. *Ez dut tesiko kurtsorik errepikatu!*

Donostian egindako lagun berri guztiei, urte hauetan guztietan zehar etxean bezala sentitu naiz zuen alboan.

Ostegun gauetako pintxopotelariei ere mila esker, laneko buruhausteak ahaztu eta astea ondo amaitzeko erremediorik onena izan zaretelako. *Gora Matia kaleko pintxopoteak!*

Gracias a las camareras del Pako's por haberme cuidado tan bien.

Bukatzeko, nola ez familia eskertu nahiko nuke. Aita, Ama, Jon eta Leire. Eskerrik asko bizitzan hartu ditudan erabaki guztietan zuen babesa eskaintzeagatik eta une zailetan alboan egoteagatik. Hurrengo bazkaria nire kontu! Ama, hau bukatzean ez naiz *rector* izango *doctor* baizik.

Aita nahiz eta orain dela gutxi utzi gintuzun, beti egongo zara gure artean. Mila mila esker gugatik egin duzun guztiagatik. Tesi hau zuri eskaini nahi dizut. Ah eta bihar pasako naiz *sin falta* zu bisitatzera!

Esker instituzionalak

Euskal Herriko Unibertsitateko Euskara Errektoreordetzari ikerketa-lan hau egiteko emandako ikertzaileak prestatzeko bekarengatik.

Laburpena

Gaur egun, korreferentzia-ebazpen automatikoa gakotzat har dezakegu testuak ulertu ahal izateko; ondorioz, behar-beharrezkoa da diskurtsoaren ulerkuntza sakona eskatzen duten Lengoaia Naturalaren Prozesamenduko (NLP) hainbat atazatan.

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.

Tesi-lan hau, hizkuntzalaritza konputazionalaren arloan kokatzen da eta euskaraz idatzitako testuen korreferentzia-ebazpen automatikoa du helburu, zehazkiago esanda, euskarazko korreferentzia-ebazpen automatikoa gauzatzeko dagoen baliabide eta tresnen hutsunea betetzea du helburu.

Tesi-lan honetan, lehenik euskarazko testuetan ager daitezkeen espresio testualak automatikoki identifikatzeko garatu dugun erregelatan oinarritutako tresna azaltzen da.

Ondoren, Stanfordeko unibertsitatean ingeleserako diseinatu den erregelatan oinarritutako korreferentzia-ebazpenerako sistema euskararen ezaugarrietara nola egokitu den eta ezagutza-base semantikoak erabiliz nola hobetu dugun aurkezten da.

Bukatzeko, ikasketa automatikoan oinarritzen den BART korreferentzia-ebazpenerako sistema euskarara egokitzeko eta hobetzeko egindako lana azaltzen da.

Gaien aurkibidea

SARRERA

1 Tesi-lanaren aurkezpen orokorra

1.1 Sarrera eta motibazioa

Tesi lan hau, hizkuntzalaritza konputazionalaren arloan kokatzen da eta euskaraz idatzitako testuen korreferentzia-ebazpen automatikoa du helburu.

Hizkuntzalaritza konputazionalean honela defini genezake korreferentzia-ebazpena:

> "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."

Honako esaldian, adibidez:

(1) Nazio Batuen Erakundea izan zen bitartekari eta hark hartu zuen prozesuaren ardura.

Nazio Batuen Erakundea, *bitartekari* eta *hark* espresio testualen artean korreferentzia-erlazioa dagoela ikus daiteke, hiruek Nazio Batuen Erakundeari egiten baitiote erreferentzia.

Azken urteetan, hizkuntzalaritza konputazionalean lan ugarik izan dute korreferentzia-ebazpena aztergai, aipatuenetakoak Mitkov (2002), Recasens and Martí (2010) eta Poesio *et al.* (2016) dira.

Euskararen kasua hizpide dutenen artean Ceberio *et al.* (2008), Soraluze *et al.* (2015b) edota Garcia Azkoaga (2016) lanak aipa ditzakegu.

Ikuspuntu konputazionaletik korreferentzia-ebazpena terminoa *Message Understanding Conference* (MUC-6, 1995) konferentzian zehaztu zen lehen aldiz. Hizkuntzalaritza konputazionalean korreferentzia terminoa anaforaren sinonimotzat hartzen da. Hala ere, erabaki horrek eztabaida teoriko ugari (van Deemter and Kibble, 1995) sortu ditu.

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 (McCarthy and Lehnert, 1995), testuen laburpenean (Steinberger *et al.*, 2007), galdera-erantzuteko sistemetan (Vicedo and Ferrández, 2006), itzulpen automatikoan (Peral *et al.*, 1999), sentimenduen analisian (Nicolov *et al.*, 2008) edota irakurketa automatikoan (Poon *et al.*, 2010).

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

Azaldutako terminoak modu argiagoan ulertzeko, azter dezagun sakonago lehengo adibidea.

(2) [Nazio Batuen Erakundea] izan zen [bitartekari] eta [hark] hartu zuen [prozesuaren ardura].

Adibide horretan, kortxete artean lau aipamen ikus ditzakegu, [Nazio Batuen Erakundea], [bitartekari], [hark] eta [prozesuaren ardura]. Garbi ikusten da [Nazio Batuen Erakundea], [bitartekari] eta [hark] aipamenek mundu errealeko objektu berbera erreferentziatzen dutela, hau da, entitate bera erreferentziatzen dute, nahiz eta horretarako espresio testual desberdinak erabili, beraz, korreferenteak direla esan dezakegu.

Ohikoa da korreferentzia-ebazpena bi azpi-ataza nagusitan banatzea: aipamenen detekzioa, batetik, eta erreferentzien ebazpena, bestetik (Pradhan *et al.*, 2011). Hau da, lehenbizi testuko aipamenak detektatzen dira, eta, ondoren erabakitzen da zein aipamenek egiten dioten erreferentzia entitate berari. 2. adibidearekin jarraituz, aipamen-detekzioan egin beharrekoa lau aipamenak zuzen identifikatzea da. Korreferenzia-ebazpenean berriz, [Nazio Batuen Erakundea], [bitartekari] eta [hark] aipamenak lotu egin behar dira, mundu errealeko entitate bera adierazten dute eta.

Korreferentzia-ebazpena gauzatzeak, adibidez, itzulpen automatikoan nola lagun dezakeen jabetzeko, pentsa dezagun 2. adibidea ingelesera itzuli nahi

dugula. Hona hemen itzulpen zuzen posible bat:

(3) [The United Nations Organisation] was [a mediator] and [it] took charge of [the process].

Itxura batean itzulpenak ez du konplexutasun handirik erakusten, hala ere, *hark* izenordainaren itzulpenak badu bere zailtasuna. Gizakiok, adibidea irakurtzean, erraz ulertzeko gai gara *hark* izenordaina *Nazio Batuen Erakundea* izen berezi bizigabea ez errepikatzeagatik erabili dela, eta ondorioz, ingeleseko *it* hirugarren pertsona singular bizigabeetarako erabiltzen den izenordainaz itzuli behar dugula. Itzulpen automatikoa gauzatzean ordea, itzultzaile automatikoak zalantza izango luke *hark* izenordaina *she*, *he* edo *it* izenordainarekin itzuli beharko lukeen.

Adibidea aztertu ondoren, garbi ikusten da korreferentzia-ebazpena lagungarria izan daitekeela ezaugarri desberdinak dituzten hizkuntzen arteko itzulpen automatikoa gauzatzeko. Antzeko ideia aurkezten da Miculicich Werlen and Popescu-Belis (2017) lanean ere. Lan horretan korreferentzia-erlazioak erabiltzen dituzte gazteleratik ingeleserako itzulpen automatikoa hobetzeko. Zehatzago esanda, izenordainen itzulpena hobetzeko erabiltzen dituzte korreferentzia-erlazioak, gaztelera *pro-drop* hizkuntza izanik eta ingelesa ez, ingeleseko itzulpenean agertu beharreko izenordain batzuk elidituta ager baitaitezke gaztelerazko testuetan. Elididutako izenordain horien aurrekari zuzena identifikatzeak garrantzitsua dela erakusten dute autoreek.

Euskarara etorrita, ezaguna da hizkuntza gutxituek urri dituztela baliabide linguistikoak, hizkuntza handi eta ahaltsuen aldean. Gainera, euskara hizkuntza isolatua da ez baitu inongo loturarik eta antzekotasunik bere inguruan dituen indoeuropear hizkuntzekin. Hizkuntza eranskari, buru-azken, *pro-drop* eta ordena librekoa da (Laka, 1996). Hori guztia dela eta, hizkuntzaren prozesamenduko atazetarako tresna eraginkorrak garatzea erronka handia da.

Tesi-lan hau Euskal Herriko Unibertsitateko IXA ikerketa-taldearen jardunean kokatzen da. IXA taldeak ia 30 urte daramatza hizkuntzalaritza konputazionalaren arloan lanean, eta urte horietan zehar euskararen analisi linguistiko konputazionalerako analisi-kate sendo bat garatu du (Aduriz *et al.*, 2004; Otegi *et al.*, 2016). Analisi-katea testuak analizatzeko erabiltzen diren oinarrizko tresnen multzoa da, eta tresna horiek geruzaka antolatuta daude. Geruza bakoitzak aurrekoak eskaintzen dion informazioa erabiltzen du sarrera moduan, eta jasotako analisia informazio linguistiko berriarekin

aberasten du. Bukaeran, maila desberdinetan analizatutako testua itzultzen du.

Korreferentzia-ebazpenari dagokionez, IXA taldean euskarazko izenordainen ebazpenerako ikasketa automatikoan oinarritutako Arregi *et al.* (2010) lana bakarrik zegoen argitaratuta tesi lan hau hasi aurretik. Ondorioz, tesilan honen helburu nagusia hauxe da: euskarazko korreferentzia-ebazpen automatikoa gauzatzeko dagoen baliabide eta tresnen hutsunea betetzea eta garatutako baliabide horiek IXAko analisi-katean integratzea.

1.2 Tesi-txostenaren eskema

Honako tesi hau bost artikuluz osatutako bilduma da. Bilduma honetako artikuluak hiru atal nagusitan bana daitezke.

Lehenengo atalean aipamen-detekzioan egindako lana aurkertuzko da. Bigarrenean, erregelatan oinarritutako korreferentzia-ebazpenaren inguruan egindako lanak aurkeztuko dira. Hirugarren atalean, berriz, ikasketa automatikoan oinarritutako korrereferentzia-ebazpenean gauzatutako lanen nondik norakoak azalduko dira.

Artikulu-bildumaren aurretik, sarrera eta arloaren egoera azalduko dira. Amaieran berriz, ondorioak eta etorkizuneko lanak. Gainera, artikulu bakoitzaren aurretik egindakoaren euskarazko laburpena emango da.

Honakoak dira artikulu-bildumaren hiru atalak eta atal bakoitzean aurkezten diren artikuluak:

1. Atala: Aipamen-detekzioa.

1. Soraluze A., Arregi O., Arregi X., and Díaz De Ilarraza A. 2016. **Improving mention detection for Basque based on a deep error analysis**. *Natural Language Engineering*, 23(3), 351–384.

2. Atala: Korreferentzia-ebazpena.
Erregelatan oinarritutako sistema.

2. Soraluze A., Arregi O., Arregi X., and Díaz De Ilarraza A. In revision. **Adaptation of the Stanford Deterministic Coreference Resolution System to a morphologically rich language**. Submitted to *Language Resources and Evaluation*.

3. Soraluze A., Arregi O., Arregi X., and Díaz De Ilarraza A. 2017. **Enriching Basque Coreference Resolution System using Semantic Knowledge sources**. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, 8–16, Valencia, Spain.

3. Atala: Korreferentzia-ebazpena.
Ikasketa automatikoan oinarritutako sistema.

4. Soraluze A., Arregi O., Arregi X., and Díaz de Ilarraza A., Kabadjov M., and Poesio M. 2016. **Coreference Resolution for the Basque Language with BART**. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, 67–73, San Diego, California.

5. Soraluze A., Arregi O., Arregi X., and Díaz de Ilarraza A. In revision. **Improving BART coreference resolution system for Basque with semantic knowledge**. Submitted to *Natural Language Engineering: Special Issue on Knowledge-Rich Coreference Resolution*.

1.3 Tesi honen garapenetik atera diren beste argitalpenak

Jarrain tesiaren garapenean zehar argitaratu diren beste artikuluak zerrendatzen dira. 1., 3. eta 4. artikuluak tesiarekin zuzenki lotutakoak dira, hala ere, ez dira tesi-txosteneko artikulu bilduman sartu, bilduman agertzen diren artikuluak hauen hedapenak direlako. 2. artikulua kolaborazio lan baten emaitza da.

1. Soraluze A., Arregi O., Arregi X., Ceberio K., and Díaz de Ilarraza A. 2012. **Mention Detection: First Steps in the Development of a Basque Coreference Resolution System**. In *KONVENS 2012, The 11th Conference on Natural Language Processing*, 128–136, Vienna, Austria.

2. Gonzalez-Dios I., Aranzabe M.J., Díaz de Ilarraza A., and Soraluze A. 2013. **Detecting Apposition for Text Simplification in Basque**. In *Proceedings of the 14th International Conference on Computational*

Linguistics and Intelligent Text Processing - Volume 2, 513–524, Samos, Greece.

3. Soraluze A., Arregi O., Arregi X., and Díaz de Ilarraza A. 2015. **Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque**. *Procesamiento del Lenguaje Natural.* 55:23–30.

4. Soraluze A., Arregi O., Arregi X., and Díaz de Ilarraza A. 2015. **Korreferentzia-ebazpena euskaraz idatzitako testuetan**. In *I. Ikergazte: Nazioarteko ikerketa euskaraz*, 676–684, Durango.

1.4 Tesiaren ekarpen nagusiak

Laburbilduz, honakoak dira tesi-lan honen ekarpen nagusiak:

1. Atala: Aipamen-detekzioa.

- Euskararako aipamen-detektatzaile automatikoa diseinatu eta inplementatu da. Oinarrian egoera finituko teknologia erabiltzen du eta garapena euskarako aipamenen azterketa linguistikoan oinarrituta egin da. Gainera, aipamen-detekzioak korreferentzia-ebazpenean duen eragina aztertu da.

- Hizkuntzalarien etiketatze-lana erraztu ahal izateko eta aipamenak testu hutsetik etiketatzen hasi beharrik ez izateko, garatu dugun aipamen-detektatzaile automatikoa erabiliz EPEC corpusaren zati bat etiketatu da. Horretarako, aipamen-detektatzailearen emaitza MMAX2 tresnan erabili ahal izateko prestatu da. MMAX2 tresnak eskaintzen duen, maila anitzeko etiketatzeari esker, automatikoki etiketatutako aipamenen zuzenketa eta korreferentzia-kateen eskuzko etiketatzea ahalbidetu da. Guztira 44383 hitzez osatutako corpusa landu da, 12792 aipamen eta 1934 korreferentzia-erlazioz osatutako urre-patroia sortuz. Urre-patroi hori eskuragarri[1] jarri da.

- Aipamen-detektatzailea Gonzalez-Dios *et al.* (2013) lanean erabili da, euskarazko testuen sinplifikaziorako erabiltzen den aposizio-detektatzailearekin konbinatuz aposizio egituren detekzioa hobetzeko.

[1] http://ixa2.si.ehu.es/epec-koref/epec-koref_v1.0.tgz

2. Atala: Korreferentzia-ebazpena. Erregelatan oinarritutako sistema.

- Stanfordeko unibertsitatean ingeleserako diseinatu den erregelatan oinarritutako korreferentzia-ebazpenerako sistema (Lee *et al.*, 2013) euskararen ezaugarrietara egokitu da. Hala nola, euskararen morfologia aberatsa edo ordena-librea bezalako ezaugarriak kontuan hartu dira egokitzapena egiteko orduan.

- Euskaltzaindiak eskaintzen dituen baliabideak oinarri hartuta, Stanfordeko sisteman integratu diren bi lexikoi egokitu dira. Lexikoi horiek beste atazaren batean erabilgarriak izan daitezke. Honakoak dira egokituko lexikoiak:

  1) 3109 sarrera dituen euskal, frantses eta gaztelerazko pertsona izenen generoari buruzko informazioa duen lexikoia
  2) 373 sarrera dituen nazio mailako gentilizioen euskarazko lexikoia

- Euskaraz idatzitako testuen prozesamendu automatikorako ixaKat analisikatean integratu eta eskuragarri[2] jarri da egokitutako sistema.

- QTLeap[3] (Quality Translation by Deep Language Engineering Approaches) proiektuan (FP7-ICT-2013.4.1-610516) euskararako 4 corpusetan korreferentzia-erlazioak automatikoki etiketatu dira gure sistema erabiliz. Hauek dira corpus horien datuak:

Corpusa	Tokenak	Korreferentzia-erlazioak
Elhuyar-QTLeap	10.639.863	1.551.340
GNOME-QTLeap	4.194.823	563.570
News-QTLeap	20.869	2.278
QTLeap	56.927	6.103

- Hizkuntza- eta hizketa-teknologiak ikertzea eta ikerketaren emaitzak aplikazio bihurtu eta jendearen eskura jartzea helburu zuen Ber2Tek[4] proiektuan (IE12-333) erabili da euskarara egokitu dugun Stanfordeko

[2] http://ixa2.si.ehu.es/ixakat/ixa-pipe-coref-eu.php
[3] http://qtleap.eu/
[4] http://ber2tek.eus/eu/hasiera

korreferentzia-ebazpenerako sistema, literatura-kritiketako korreferentzia-erlazioak automatikoki etiketatzeko.

- Egokitutako sistema hobetu da ezagutza-base semantikoak erabiliz, zehazki esanda WordNet eta Wikipedia. Sistemaren hobekuntza, erroreanalisi sakona egin eta sistemak munduaren ezagutza eta ezagutza semantikoa beharrezkoa den korreferentzia-erlazioak ebazteko zuen gabeziaz jabetuta egin da.

3. Atala: Korreferentzia-ebazpena. Ikasketa automatikoan oinarritutako sistema.

- Erregelatan oinarritutako sistema egokitzeaz gain, BART ikasketa automatikoan oinarritzen den korreferentzia-ebazpenerako sistema (Versley *et al.*, 2008b) ere euskarara egokitu da.

- BART sistema euskal Wikipediako 263.316 orrialdetatik erauzitako ezagutzarekin aberastu da. Horretarako, batetik berbideraketa orriek eskaintzen diguten informazioa erabili da. Bestetik, *piped link* egituretatik entitate izendunen izen-laburdurak, pseudonimoak, laburtzapenak eta erauzi dira. Azkenik, gai konkretu bati buruzko orrialdeen zerrendak biltzen dituzten orrialdeetatik entitateen ezaugarriak lortu dira.

- Wikipediatik erauzitako informazio hori gordetzeko hiru taulaz osatutako datu-basea sortu da, 263.316 artikulu erabiliz. Entitate izendunen izen-laburdurak, pseudonimoak eta laburtzapenak gordetzen dituen taulak 93.240 sarrera ditu. 118.131 sarrera dituen taula sortu da Wikipediako berbideraketak erabiliz. Azkenik, entitateen ezaugarriak biltzen dituen taulak 20.276 sarrera ditu. Datu-basea hizkuntzaren prozesamenduko beste atazetan erabiltzeko baliabide aberatsa da.

2 Aurrekariak

Kapitulu honetan tesi-lan honen aurrekariak aurkeztuko ditugu. Horretarako, lehenbizi korreferentzia-ebazpenak izan duen bilakaera aztertuko dugu, hastapenetatik gaur arte argitaratu diren lan garrantzitsuenak azalduz. Ondoren, korreferentzia-ebazpenaren parte kontsideratzen den aipamen-detekzio atazaren nondik norakoak aurkeztuko dira. Jarraian, bete-betean korreferentzia-ebazpenaren gaiari helduko diogu. Ataza honetarako erabili ohi diren teknikak, ereduak eta ezaugarriak azalduko dira lehenik. Eskuragarri dauden corpusak aipatuko dira gero. Ezagunenak diren korreferentzia-ebazpenerako sistemak aurkeztuko dira ondoren. Korreferentzia-ebazpenean semantikaren erabilerak duen garrantzia ikusteko azpiatal bat eskainiko zaio gai honi eta hizkuntza desberdinetarako garatu diren sistemak aipatuko dira gero. Kapitulua amaitzeko, korreferentzia-ebazpena ebaluatzeko erabiltzen diren ebaluazio-metrikak azalduko dira.

2.1 Bilakaera

Korreferentzia-ebazpenaren hastapena 60-90eko hamarkadetan kokatu ohi da. Urte haietan korreferentzia-ebazpena lengoaia naturalaren prozesamenduko azpiatazatzat hartzen zen eta itzulpen automatikoaren edo sistema adituen testuinguruan kokatzen zen, adibidez Bobrow-ren (1964) STUDENT izeneko sistema eta Woods *et al.*-en (1974) LUNAR izenekoa.

Garai hartako lanik garrantzitsuena Hobbs-en (1978) "Resolving Pronoun References" dela esan daiteke. Urteak aurrera egin ahala hainbat lan argitaratu baziren ere, Carter-ek (1987) proposatzen zuen "Shallow Processing

Anaphor Resolver (SPAR)" eta Lappin and Leass-en (1994) "An algorithm for pronominal anaphora resolution" artikuluak dira esanguratsuenak. Azken lan hori hurrengo urteetan argitaratu zirenen aitzindari har genezake eta mugarria izan zen izenordainen ebazpenerako, RAP (Resolution of Anaphora Procedure) algoritmo sendoa deskribatzen baitu xehetasun maila handiz. Hori dela eta, korreferentzia-ebazpenaren arloan gehien erreferentziatutako lanetako bat da.

Alderdi teorikoan ikerketa handia egin zen *focusing* eta diskurtsoaren egituraren azterketan, horien artean daude Sidner-ek argitaratutako lanak (Sidner, 1979, 1981, 1983) eta Grosz and Sidner-en (1986) lanak. Gainera, *centering* teoriaren oinarriak ere garai horretan finkatu ziren (Grosz *et al.*, 1983).

Azken bi hamarkadei dagokienez, garrantzi handia eman zaio korreferentzia-ebazpenari eta gai honen inguruan kongresu ugari antolatu dira. *Message Understanding Conference* (MUC-6, 1995; MUC-7, 1998) konferentzietan korreferentziaren inguruko ataza espezifikoak antolatu ziren. Kongresu hauei esker, lehen aldiz, korreferentzia-ebazpenerako sistemen konparaketa ahalbidetu zen. Horretaz gain, korreferentzia-ebazpenerako etiketatutako lehen corpusak sortu ziren eta horrek corpusetan oinarritutako tekniken sorrera ekarri zuen ataza honetara. Garai horretakoa da hain ezaguna eta esanguratsua egin den Soon *et al.* (2001) lana. Lan horretan, gainbegiratutako erabakitze-zuhaitzen ikasketa egiten da MUC-6 eta MUC-7 kongresuetako corpusak erabilita. Etiketatutako corpusak eskuragarri izateak Lappin eta Leas-en erregelatan oinarritutako ikuspuntuaren birfintzea ahalbidetu zuen, eta RAP algoritmoaren ondorengotzat har daitezkeen lanen argitaratzea ekarri zuen. Adibidez, Mitkov-ek (1998) bere MARS sistemaren oinarriak garatu zituen. Izenordainen inguruan egindako ikerketa ugariz aparte, izenordainak ez ziren beste korrefentzia-erlazio mota espezifikoen lanketa ere egin zen garai haietan. Horien artean ditugu, adibidez, izen-sintagma mugatuen inguruan egindako Poesio and Vieira (1998) eta Vieira and Poesio (2000) lanak.

Garai bertsuan ospatu ziren Discourse Anaphora and Anaphor Resolution Colloquium konferentziak (DAARC, 1996, 1998, 2000). Korreferentzia-ebazpenaren inguruan lanean zebiltzan ikerlari ugarirentzako topagune egokia izan ziren konferentzia horiek. *The Automatic Content Extraction (ACE)* kongresuan ere jorratu zen korreferentzia-ebazpena, bertan aurredefinitutako entitate multzo baten arteko erlazioak identifikatzen saiatu ziren (Doddington *et al.*, 2004), eta *Anaphora Resolution Exercise (ARE)* kongresuak

anafora ebazpena eta izen-sintagmen arteko korreferentzia-ebazpena hartu zituen ardatz (Orasan *et al.*, 2008).

Aurrerago ikusiko dugun moduan, urteak aurrera joan ahala ingelesa ez zen beste hizkuntzetan ere korreferentzia-ebazpena garrantzia hartzen joan zen. Azken urteetako kongresuei dagokienez, *SemEval-2010 Task 1: Coreference Resolution in Multiple Languages* atazan korreferentzia-ebazpena gauzatu behar zen hizkuntza desberdinetan (Recasens *et al.*, 2010). Hurrengo urtean, *CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes* atazan (Pradhan *et al.*, 2011), parte-hartzaileek Ontonotes corpusean (Pradhan *et al.*, 2007) ebatzi behar izan zuten korreferentzia ingeleserako. 2011ko ataza ingeleserako soilik izan bazen ere, *CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes* atazak (Pradhan *et al.*, 2012) ingelesaren, txineraren eta arabieraren gaineko korreferentzia-ebazpena egitea eskatzen zuen.

Azken urteetan geldirik zegoela zirudien korreferentzia-ebazpenaren inguruko ikerketa, hala ere, 2016 eta 2017an ospatu diren *Coreference Resolution Beyond Ontonotes* (CORBON) kongresuetan (Ogrodniczuk and Ng, 2016, 2017) oraindik indarrean dagoela ikusi da eta hizkuntza gutxituen inguruan egindako lanen presentzia nabarmen hazi da.

2.2 Aipamen-detekzioa

Ikerketa ugaritan frogatu den moduan (Broscheit *et al.*, 2010a; Stoyanov *et al.*, 2009; Hacioglu *et al.*, 2005; Zhekova and Kübler, 2010) aipamen-detekzioak berebiziko garrantzia du korreferentzia-ebazpenerako sistemetan. Ataza honetan egindako erroreek hurrengo pausoetara hedatu eta hauetan lortzen den eraginkortasuna murrizten dute. Hori dela eta, korreferentzia-ebazpenerako sistemetan aipamen-detekzioa hobetzeak arloaren egoera nabarmen hobetuko luke. Ideia hori frogatu da korreferentzia-ebazpenean aipamen-detekzioaren eragina kuantifikatu den hainbat lanetan. Adibidez, Uryupina (2008) lanean aurkezten den korreferentzia-ebazpenerako sistemaren estaldura-erroren % 35 identifikatu gabeko aipamenen ondorioz dela adierazten da eta Uryupina (2010) lanean doitasun erroreen % 20 gaizki identifikatutako aipamenen ondorioz dela gehitzen dute. Chang *et al.* (2011) lanean bi sistema konparatzen dituzte, batetik, aipamenak automatikoki detektatzen dituen sistema, eta bestetik, aipamenen urre-patroia (*gold standard*) erabiltzen duen sistema. Azken horrek emaitzak % 15 eta % 18 bitartean hobetzen

ditu.

Domeinu espezifikoetan ere aipamen-detekzio egokia egitea funtsezkoa dela defendatu da. Kim *et al.*-ek (2011) diotenez, biomedikuntza arlo espezifikorako prestatutako korreferentzia-ebazpenerako sistemetan ere, aipamen-detektatzaile eraginkorra izatea oso garrantzitsua da. Autoreek behatu dutenen arabera, automatikoki detektatutako aipamenak edo urre-patroiak erabiliz, MUC ebaluazio-irizpideari jarraituz haien sistemak lortzen dituen emaitzak 49,69 izatetik 87,32 izatera pasatzen dira. Hobekuntza nabarmena da, zalantzarik gabe.

Aipamen-detekzioan erabilitako teknologiari dagokionez, bi multzo nagusi bereiz genitzake: erregelatan oinarritutako teknikak eta ikasketa automatikoan oinarritutakoak. Orokorrean, erregelatan oinarritutako aipamen-detektatzaileek doitasun balio altuagoak lortzen dituzte ikasketa automatikoan oinarritutakoek baino (Pradhan *et al.*, 2011). Aipamen-detekzioan estaldura balio baxuak lortzeak eragin zuzena du ondoren korreferentzia-ebazpenean, detektatu gabe utzitako aipamen horiek berreskuratzeko aukerarik ez baitago.

Semeval-2010 Task 1 atazan korreferentzia-ebazpenerako sistema gehienek erregelatan oinarritutako aipamen-detektatzaileak erabili zituzten (Zhekova and Kübler, 2010; Uryupina, 2010; Attardi *et al.*, 2010; Broscheit *et al.*, 2010a). Uryupina and Moschitti (2013) lanean aipatzen denez, *CoNLL 2011 Shared Task* atazan ere parte-hartzaile gehienek aipamenen mugak detektatzeko erregelatan oinarritutako teknikak erabili zituzten. Hala ere, urtebete beranduagoko *CoNLL 2012ko Shared Task* atazan, erregelatan oinarritutako aipamen-detektzaileek emaitza kaskarrak lortu zituzten txinerara eta arabierara egokitu behar izan zirenean. Argi ikusi zen erregelatan oinarritutako aipamen-detektatzaileak beste hizkuntza batera egokitzea lan konplexua zela.

2.3 Korreferentzia-ebazpena

2.3.1 Teknikak

Korreferentzia-ebazpenerako erabiltzen diren algoritmoen artean aipamen-detekzioan bezala, bi mota nagusi bereiz genitzake: erregelatan oinarritutakoak eta ikasketa automatikoa erabiltzen dutenak. Azken horiek ere bi multzo nagusitan banatzen dira: ikasketa gainbegiratuko tekniketan oinarri-

tutakoak eta ikasketa ez-gainbegiratukoak.

Ikasketa gainbegiratuko teknikek korreferentzia-kateekin etiketatutako corpusak erabiltzen dituzte ikasketa prozesua gauzatzeko, ez-gainbegiratukoek, aldiz, ez. Ikasketa prozesurako etiketatutako corpusik behar ez izateak ikasketa ez-gainbegiratuko teknikak erakargarri egiten baditu ere, gainbegiratutako teknikak arrakastatsuagoak dira.

Erregelatan oinarritutako teknikak

Gaur egun ezagutzen ditugun korreferentzia-ebazpenerako sistemen aurrekariak erregelatan oinarritzen ziren izenordainen ebazpenerako sistemak ziren (Hobbs, 1978; Rich and LuperFoy, 1988; Carbonell and Brown, 1988; Alshawi, 1990; Kameyama, 1997a; Tetreault, 2001; Palomar *et al.*, 2001) bereziki murriztapen eta lehentasun erregelak aplikatzen zituztenak.

MUC-6 eta MUC-7 konferentzietan ere erregelatan oinarritutako hainbat sistema (Appelt *et al.*, 1995; Gaizauskas *et al.*, 1995; Kameyama, 1997b; Humphreys *et al.*, 1998) aurkeztu ziren. Kongresu hauetan sortutako anotazio kopuru esanguratsua dela eta, ikasketa automatikoko tekniketan oinarritutako sistemen hazkunde nabarmena eman zen hurrengo urteetan, eta erregelatan oinarritutako tekniken erabilera nabarmen murriztu zen.

Hala ere, badirudi azken urteetan berriro erregelatan oinarritutako tekniken loraldia eman dela. Sistema ugari garatu dira, horien artean *CoNLL 2011 Shared Task*-ean lehenengo postua lortu zuen Stanfordeko korreferentzia-ebazpenerako sistema determinista (Lee *et al.*, 2013).

Ikasketa gainbegiratuko teknikak

Ikasketa gainbegiratuan oinarritutako korreferentzia-ebazpenerako sistema bat sortzeko garaian honako hauek hartu behar dira kontuan: i) entrenamendurako instantziak sortzeko metodoa edo eredua, ii) instantziak errepresentatzeko erabiliko diren ezaugarri linguistikoak, eta iii) sailkatzailea entrenatzeko erabiliko den ikasketa algoritmoa.

i) **Instantziak sortzeko metodoa**

- **Aipamen-bikote eredua** (*mention-pair model*)
  Aipamen-bikote ereduak korreferentzia-ebazpena sailkapen ataza moduan planteatzen du non sailkatzaile bat entrenatzen den bi

aipamen korrerefenteak diren edo ez erabakitzeko. Bi pauso nagusi ditu ereduak:

1) Aipamen bikoteak korreferente edo ez-korreferente gisa sailkatzen dira (*pairwise classification*).
2) Korreferente gisa sailkatu diren aipamen-bikoteetatik abiatuta korreferentzia-kateak sortzen dira, multzokatze-algoritmoak erabiliz.

Entrenamendurako instantziak sortzeko modu naturalena aipamen-bikote bakoitzeko instantzia berri bat sortzea bada ere, metodo hori ez da normalki erabiltzen. Testu bateko aipamen-bikote gehienak ez dira korreferenteak izaten eta ondorioz ikasketarako sortzen den multzoa ez da orekatua, hau da, instantzia negatibo askoz gehiago daude positiboak baino. Klaseen arteko desoreka hori konpondu nahian, metodo ugari proposatu dira, horien artean Soon *et al.* (1999, 2001) lanetan aurkezten dena. Metodo horren arabera, m_k aipamen bat emanik, instantzia positiboa sortzen da m_k eta justu aurretik gertuena duen m_j aurrekariarekin, eta instantzia negatiboak m_k aipamena eta m_j artean dauden beste $m_{j+1}, ..., m_{k-1}$ aipamenekin. Klaseen desoreka gehiago murrizteko asmotan, ikerlariek hainbat iragazketa-teknika proposatu dituzte instantziak sortzeko metodoaren gainean, adibidez, numero eta genero desberdina dituzten aipamen bikoteekin instantziarik ez sortzea, korreferente izateko aukerak txikiak baitira (Strube *et al.*, 2002; Yang *et al.*, 2003). Berrikiago, Sapena *et al.* (2011) lanean clustering algoritmo bat proposatzen dute instantzia negatiboen kopurua murrizteko. Clustering algoritmo horrek instantzia positiboak erabiltzen ditu zentroide moduan. Instantzia positibo bakoitzeko, d atalasea baino distantzia txikiago edo berdinera dauden instantzia negatiboak soilik erabiltzen dira ikasketa garaian. Bi instantziaren arteko distantzia kalkulatzeko balio desberdinak dituzten ezaugarriak begiratzen dira. Hau da, edozein instantzia positiborekiko d atalasea baino ezaugarri balio desberdin gehiago dituzten instantzia negatiborik ez da sortzen.

Test garaian, multzokatze algoritmo bat beharrezkoa da bikote horiek korreferentzia-kateetan biltzeko. Multzokatze-algoritmoei dagokienez, hiru bereiz genitzake. Gertuena-lehenengo (*closest-first*), hoberena-lehenengo (*best-first*) eta elkarketa agresiboa (*ag-*

gresive merge). Gertuena-lehenengo estrategiak, m_k aipamena testuan bere aurretik duen eta positiboki sailkatua izan den lehenengo aurrekari posiblearekin elkartzen du, edo elkartu gabe uzten du positiboki sailkatutako instantziarik topatzen ez bada (Soon *et al.*, 2001). Hoberena-lehenengo multzokatze-estrategia Ng and Cardie-k (2002c) proposatu zuten. Estrategia horren arabera, m_k aipamenaren aurrekaritzat, korreferente izateko baliorik hoberena duen aipamena hartzen da, beti ere balio horrek ezarri den atalasea gainditzen badu. Elkarketa agresiboan (McCarthy and Lehnert, 1995; Ng, 2005) aldiz, aipamen bakoitza aurretik dituen eta korreferente izateko positiboki sailkatu diren beste aipamen guztiekin elkartzen da estaldura hobetzea lortzeko asmoz. Hala ere, aipamen-bikoteen sailkapen fasean egiten diren eta multzokatze garaian egiten diren erroreek korreferentzia-kate desegokiak eta kontraesankorrak sor ditzakete. Korreferentzia-kateetan sor daitezkeen kontraesan horiek konpontzeko Hoste and van den Bosch-ek (2007) Levenshteinen distantzian oinarritutako korreferentzia-kateen post-zuzenketa metodo bat proposatzen dute, doitasuna zertxobait hobetzea lortuz.

(Soon *et al.*, 2001) eta (Ng and Cardie, 2002a) lanei esker ikasketa automatikoan oinarritutako korreferentzia-ebazpenerako eredurik esanguratsuenetakoa bilakatu da aipamen-bikote eredua, eta gaur egun korreferentzia-ebazpenerako oinarri-lerro estandar gisa erabili ohi bada ere, ereduak baditu hainbat ahultasun:

(a) Aipamen batentzako aurrekari hautagai posibleak independenteki begiratzen dira eta, beraz, ezinezkoa da hautagai bat beste hautagaiek baino hobea den edo ez jakitea.

(b) Ereduari espresibotasuna falta zaio, hau da, gerta daiteke bi aipamenetatik lortutako informazioa nahikoa ez izatea korreferentzia-erabaki egoki bat hartzeko.

- **Entitate-aipamen eredua** (*entity-mention model*)

  Aipamen-bikote ereduari kritikatzen zaion espresibotasun falta konpondu nahian sortu zen entitate-aipamen eredua (Luo *et al.*, 2004; Yangy *et al.*, 2004). Eredu mota honetan, aipamen bat, aurretik sortutako multzo batekin korreferentea den edo ez zehazten da.

  Ikasketarako instantzia bakoitzak, hiru osagai ditu; m_k aipamena, C_j aurreko clusterra edo multzoa, eta, klasea, m_k aipamena C_j

clusterraren partaide den edo ez adierazten duen etiketa. Ikasketarako instantziak sortzeko m_k aipamena eta bere aurreko cluster edo multzoak erabil daitezke. Hala ere, instantziak sortzeko modu honek klase negatibo eta positiboen arteko desoreka handia sor dezake, m_k aipamena aurreko multzo bakarreko partaide baita soilik. Klaseen arteko desoreka hau murrizteko, Yangy *et al.*-ek (2004) instantzia negatiboen kopurua txikitzen duen metodoa erabiltzen dute. Metodo horren arabera, m_k aipamenak ez badu aurrekaririk, ez da inolako ikasketarako instantziarik sortuko m_k aipamena erabiliz, ez positiborik noski, baina ez eta negatiborik ere. Bestela, instantzia negatibo bat sortuko da m_k eta aurreko C_j clusterra erabiliz baldin eta:

(a) m_k ez bada C_j clusterreko partaide

(b) m_k eta m_j artean dagoen aipamen bat C_j clusterreko partaide bada, non m_j $m_k$ren aurrekari gertuenekoa den

Ezaugarriei dagokionez, instantzia bakoitza cluster mailako ezaugarriez adieraz daiteke. Cluster mailako ezaugarri hauek lortzeko, aipamen-bikote ereduan erabilitako ezaugarriei predikatu logikoak (*guztiak*, *gehienak*, *edozein*) aplikatzen zaizkie. Adibidez, numeroaren bat-etortzea ezaugarriari, bi aipamenek numero berdina duten adierazten duenari, *guztiak* predikatu logikoa aplikatuz, egiazko balioa izango du m_k aipamenak C_j clusterreko aipamen guztien numero berdina badu, eta faltsua kontrako kasuan. Sarritan erabiltzen diren beste predikatu logikoak *guztiak* predikatuaren erlaxazioak dira, hala nola, *gehienak* predikatua, egiazkoa da baldin m_k aipamenak C_j clusterreko aipamenen erdien baino gehiagoren numero berdina badu eta *edozein* predikatua, egia da baldin m_k aipamenak C_j clusterreko aipamenen baten numero berdina badu.

Kontuan izan behar da test garaian, clusterrak inkrementalki sortzen direla, hau da, m_k aipamenaren aurrekari izan daitezkeen cluster hautagaiak partzialak direla. Cluster hauek $k-1$ aipamenentzat sortuak izan direnak dira.

Entitate-aipamen ereduak multzo mailako ezaugarriak erabiltzeko duen gaitasuna dela eta, aipamen-bikote ereduak baino espresibotasun handiagoa du.

- **Aipamen-mailakatze eredua** (*mention-ranking model*)
  Entitate-aipamen ereduak aipamen-bikote ereduak duen espresibotasun eza konpontzen badu ere, ez du aurrekaririk probableena identifikatzearen arazoa konpontzen. Aipamen-mailakatze ereduek, aldiz, aipamen bat emanik bere aurrekari probableena zein den zehazteko aukera eskaintzen dute, aurrekari hautagai guztiak aldi berean kontuan hartzeko aukera eskaintzen baitute.

  Mailakatzearen ideia, korreferentzia-ebazpenerako lehen aldiz (Connolly *et al.*, 1994, 1997) lanetan erabili zen. Ikasketarako instantzia bakoitza, m_k aipamena eta bi aurrekari hautagaiez osatzen da, m_i eta m_j, non bat m_k aipamenaren aurrekaria den eta bestea ez. Instantziaren klaseak bi hautagaietatik zein den hobea adierazten du. Eredu hau txapelketa (*tournament*) eredu gisa definitzen da (Iida *et al.*, 2003) lanean eta bikote-hautagai (*twin-candidate*) eredu moduan (Yang *et al.*, 2003, 2008) lanetan.

  Test garaian, aipamen bat ebazteko bi modu posible daude. Connolly *et al.* (1994) lanean, kanporatze-sinpleko txapelketa (*single-elimination tournament*) deritzon modua erabiltzen dute. Ikasketa garaian sortutako aipamen-mailakatze eredua, aurrekari hautagai posible guztiekin osatutako bikoteei aplikatze zaie aldi bakoitzean. Txapelketa galtzen duen hautagaia baztertu egiten da, hau da, aurrekaria izateko hobea den hautagaia erabiltzen da testeko instantzia berria sortzeko. Yang *et al.* (2003) lanean ere eredua aurrekari hautagai bikote bakoitzari aplikatzen zaio, baina aukeratuko den aurrekaria gehienetan hobekien sailkatu den hautagaia izango da.

  Ikasketa automatikoan egindako aurrerapenek ahalbidetu egin dute aurrekari hautagai guztiak batera mailakatzea eta ez bikoteka.

  Aipamen-multzokatze ereduek, aipamen-bikote ereduek lortzen dituzten emaitzak hobetzea lortu badute ere, ez dute aipamen-bikote ereduek baino espresibotasun handiagorik lortzen, ez baitira multzo mailako ezaugarriak erabiltzeko gai, entitate-aipamen ereduek erabiltzen dituzten moduan.

- **Multzo-mailakatze eredua** (*cluster-ranking model*)
  Multzo-mailakatze ereduek, entitate-aipamen eta aipamen-mailakatze ereduek dituzten ezaugarri hoberenak konbinatzen dituzte. Rahman and Ng (2009) autoreek proposatu zuten multzo-

mailakatze eredua, mailakatzaileetan multzo-mailako ezaugarriak erabili ahal izateko. Eredu hauek, m_k aipamen batentzako aurrekari hautagai posibleak mailakatu ordez, multzo edo clusterrak mailakatzen dituzte, ondorioz, aipamen-bikote ereduak dituen bi ahultasunak saihesten dituzte eta aipamen-mailakatze ereduak hobetzen.

ii) **Ezaugarri linguistikoak**

Ikasketarako erabiliko diren ezaugarri linguistikoak mota desberdinetakoak izan daitezke, hala nola, lexikalak, sintaktikoak, gramatikalak, distantzian oinarritutakoak, string-parekaketan oinarritutakoak, semantikoak eta cluster mailakoak. Gaur egungo korreferentzia-ebazpenerako sistema gehienek Soon *et al.* (2001) lanean aurkezten diren eta Ng and Cardie-k (2002c) hobetutako ezaugarriak erabiltzen dituzte. Ezaugarri horiek erabiliz korreferentzia-erlazio gehienak ebatz badaitezke ere, ez dira nahikoa hitz guztiz desberdinez osatutako aipamenak ebazteko, esate baterako, 4. adibideko *Osasuna*, *Taldea* eta *gorritxoek* aipamenak lotzeko.

(4) [**Osasunak**] lehenengo mailara igotzeko lehian azken astean bizi duen giroa oso polita da. [**Taldea**] lasaitzeko asmoz Oronozera eraman zituen Lotinak atzo guztiak. Oronozko kontzentrazioa beharrezkoa dute [**gorritxoek**].

Kasu horietan, ezaugarri lexikal, sintaktiko eta gramatikalez gain ezagutza semantikoa edo munduaren ezagutza beharrezkoa da. Hori dela eta, korreferentzia-ebazpenerako sistema berriagoetan WordNet eta Wikipedia bezalako ezagutza-baseetatik erauzitako ezaugarriak erabiltzen hasi zen (Ponzetto and Strube, 2006; Uryupina, 2006; Ng, 2007). Korreferentzia-sisteman ezagutza gehiago txertatzeko beste modu bat cluster mailako ezaugarriak erabiltzea izan da (Luo *et al.*, 2004; Poon and Domingos, 2008). Historian zehar erabili diren ezaugarri linguistikoen laburpena 2.1 taulan ikus daiteke.

iii) **Ikasketa-algoritmoak**

Korreferentzia-ebazpenerako sailkatzaileak entrenatzeko algoritmo ugari ibili izan dira. Horien artean gehien erabili izan direnak honakoak

Ezaugarri-mota	Deskribapena
Lexikalak	m_i eta m_j aipamenen string parekaketa
	m_i eta m_j aipamenen buruen string parekaketa
	m_i eta m_j aipamenen lemen string parekaketa
	m_i edo m_j bestearen azpi-stringa?
	...
Sintaktikoak	m_i eta m_j aposizio egituran daude?
	...
Gramatikalak	m_i eta m_j-k numero berdina?
	m_i eta m_j-k genero berdina?
	m_i (m_j) izenordaina?
	m_i (m_j) zehaztua?
	m_i (m_j) demostratiboa?
	m_i eta m_j izen bereziak?
	m_i eta m_j bizidunak/ez-bizidunak?
	...
Distantzian oinarritutakoak	m_i eta m_j-ren arteko distantzia esalditan
	m_i eta m_j-ren arteko distantzia aipamenetan
	m_i eta m_j-ren arteko distantzia hitzetan
	m_i eta m_j editatzeko distantzia minimoa
	m_i eta m_j-ren WordNeteko distantzia
	...
Semantikoak	m_i eta m_j-k izen-entitate berdina?
	m_i eta m_j-k rol semantiko berdina?
	m_j, m_i-ren aliasa da?
	m_i eta m_j-k WordNeteko klase semantiko berdina?
	m_i eta m_j-ren WordNeten oinarritutako antzekotasun semantikoa
	m_i, m_j-ren sinonimo/antonimo/hiperonimoa?
	...
Cluster mailakoak	X ezaugarria egiazkoa edozein bikoterentzat?
	Bikote guztiek X ezaugarria partekatzen dute?
	Bikote gehienek X ezaugarria partekatzen dute?
	...

2.1 taula – Korreferentzia-ebazpenerako ezaugarri linguistikoak.

dira: C4.5 erabaki-zuhaitzak (Quinlan, 1993), memorian oinarritutakoak (Daelemans and van den Bosch, 2005), entropia maximoan (Berger *et al.*, 1996; Zhang, 2004) oinarritutakoak, RIPPER algoritmoa (Cohen, 1995), bozketa bidezko pertzeptroiak (Freund and Schapire, 1999), bektore-euskarridun makinak (Support Vector Machine, SVM) (Vapnik, 1995), osoko programazio lineala (Integer Linear Programming, ILP) (Schrijver, 1986), hipergrafoen banaketarako teknikak eta sare neuronalak.

Ondorengo zerrendan aipatu berri ditugun algoritmo horiek erabili diren hainbat lanen adibideak ikus ditzakegu:

- C4.5 erabaki-zuhaitzak honako lanetan erabili dira (Connolly *et al.*, 1994; Aone and Bennett, 1995; McCarthy, 1996; Soon *et al.*, 2001; Yangy *et al.*, 2004; Yang *et al.*, 2008).
- Memorian oinarritutakoak lan hauetan Hoste (2005); Recasens and Hovy (2009) izan dira erabiliak.
- (Yang *et al.*, 2003; Luo *et al.*, 2004; Kehler *et al.*, 2004; Hendrickx *et al.*, 2007) lanetan entropia maximoan oinarritutakoak erabili dira.
- RIPPER algoritmoa Ng and Cardie (2002a, c, b); Hoste (2005); Uryupina (2006) lanetan erabili da.
- Bozketa bidezko pertzeptroiak Bengtson and Roth (2008) lanean erabili dira.
- Bektore-euskarridun makinak (SVM) lan hauetan Ng (2005); Uryupina (2006); Versley *et al.* (2008a); Rahman and Ng (2009) erabili dira.
- Osoko programazio lineala honako lanetan erabili da (Denis and Baldridge, 2007; Klenner, 2007; Finkel and Manning, 2008).
- Hipergrafoen banaketarako teknikak Cai and Strube (2010); Cai *et al.* (2011); Sapena *et al.* (2013) lanetan erabili dira.
- Saren neuronalak Wiseman *et al.* (2016) lanean erabili dira.

Ikasketa ez-gainbegiratuko teknikak

Korreferentzia-ebazpena gauzatzeko garatutako sistemen gehiengoak ikasketa gainbegiratuan oinarritutakoak badira ere, ikasketa ez-gainbegiratua erabil-

tzen duten hainbat lan ere badira. Gainbegiratutako ikasketan oinarritutako teknikek ikasteko behar dituzten eskuz etiketatutako corpusak sortzeak duen kostua dela eta, teknika ez-gainbegiratuak erabiltzea aukera egokia izan daiteke, horiek ez baitute ikasketarako etiketatutako corpusik behar. Hala ere, sistema hauek lortzen dituzten emaitzak orokorrean ikasketa gainbegiratuan oinarritutakoek lortzen dituztenak baino baxuagoak izan ohi dira.

Korreferentzia-ebazpenerako ikasketa ez-gainbegiratua erabiltzen lehenak (Haghighi and Klein, 2007) izan ziren. Horiek eredu Bayesiar sortzaile bat erabili zuten atazarako. Lan horretan oinarrituta, Ng-k (2008) Entropia Maximoan oinarritutako multzokatze prozesua aurkeztu zuen, Haghighi and Klein (2007) lanaren ahultasunak konpontzeko aldaketak proposatuz. Poon and Domingos (2008) autoreek Markov kateetan oinarritutako entitate-aipamen eredua azaldu zuten eta Ma *et al.* (2016) lanean ikasketa ez-gainbegiratuan oinarritutako multzo-mailakatze eredu sortzailea proposatzen zen.

2.3.2 Corpusak

Lehen aipatu bezela, *Message Understanding Conference* (MUC-6, 1995; MUC-7, 1998) kongresuetan korreferentzia-ebazpenerako etiketatutako lehen corpusak sortu ziren eta horrek corpusetan oinarritutako tekniken hedatzea ekarri zuen. Urteak aurrera joan ahala, korreferentzia-kateak etiketatuta dituzten corpusak asko ugaritu dira, bai ingeleserako baita beste hainbat hizkuntzatarako ere.

Poesio *et al.* (2016) liburuko "Annotated Corpora and Annotation Tools" kapituluan oso ongi azaltzen dira dauden corpusak eta dituzten tamainak. Kapitulu horretan aipatzen diren hizkuntza desberdinetarako gaur egun dauden corpusak eta guk gehitu ditugunak, adibidez, euskararako edo galizierako corpusak 2.2 taulan ikus daitezke. Nabarmena da hizkuntza gutxituetan eskuragarri dauden corpusak ez direla hain ugariak eta gehienetan tamainaz txikiagoak dira hizkuntza nagusiekin konparatzen baditugu.

2.3.3 Korreferentzia-ebazpenerako sistema ezagunenak

Korreferentzia-ebazpena gauzatzen duten eta erabiltzaileentzat eskuragarri dauden sistemen artean honako bi hauek dira azken urteetan ezagunenak bihurtu direnak:

Hizkuntza	Izena	Erreferentzia	Tamaina (hitzetan)
Alemana	Postdam commentary corpusa	(Stede, 2004)	50k
	TüBa-D/Z	(Hinrichs *et al.*, 2005)	600k
Arabiera	ACE-2005	(Walker *et al.*, 2006)	100k
	ONTONOTES 5.0	(Weischedel *et al.*, 2103)	300k
Bengaliera	ICON	(Sobha *et al.*, 2011a)	
Errusiera	RU-EVAL		
Euskera	EPEC	(Ceberio *et al.*, 2016)	45k
Frantsesa	CRISTAL-GRESEC/XRCE corpusa	(Tutin *et al.*, 2000)	1000k
	DEDE	(Gardent and Manuélian, 2005)	50k
Galiziera	-	(Garcia and Gamallo, 2014b)	45-51k
Gaztelera	ANCORA-CO-Es	(Recasens and Martí, 2010)	400k
	-	(Garcia and Gamallo, 2014b)	45-51k
Hindi	ICON	(Sobha *et al.*, 2011a)	
Ingelesa	MUC-6	(Grishman and Sundheim, 1996)	30k
	MUC-7	(Chinchor, 1998)	30k
	GNOME	(Poesio, 2004)	40k
	ACE-2		180k
	ACE-2005	(Walker *et al.*, 2006)	400k
	NP4Events	(Hasler *et al.*, 2006)	50k
	ARRAU 2.0	(Poesio and Artstein, 2008)	300k
	ICSI meeting corpusa	(Müller, 2008)	
	GENIA-MEDCO (izenordainak)	(Nguyen *et al.*, 2008)	800 dokumentu
	ONTONOTES 5.0	(Weischedel *et al.*, 2103)	1450k
	Phrase detectives	(Poesio *et al.*, 2013)	320k
Italiera	VENEX	(Poesio *et al.*, 2004)	40k
	i-Cab	(Magnini *et al.*, 2008)	250k
	LIVEMEMORIES 1.0	(Rodríguez *et al.*, 2010)	250k
Japoniera	NAIST testu corpusa	(Iida *et al.*, 2007)	38k esaldi
Katalana	ANCORA-CO-Ca	(Recasens and Martí, 2010)	400k
Nederlandera	COREA	(Hendrickx *et al.*, 2008)	325k
Portugesa	Summ-It	(Collovini *et al.*, 2007)	50 dokumentu
	-	(Garcia and Gamallo, 2014b)	45-51k
Tamil	ICON	(Sobha *et al.*, 2011a)	
Tibetera	Tusnelda (B11)	(Wagner and Zeisler, 2004)	<15k
Txekiera	Prague dependency Treebank	(Hajič *et al.*, 2000)	$\approx 800k$
Txinera	ACE-2005	(Walker *et al.*, 2006)	$\approx 200k$
	ONTONOTES 5.0	(Weischedel *et al.*, 2103)	1200k

2.2 taula – Korreferentzia-kateekin etiketatutako corpusak.

Stanford Deterministic Coreference Resolution System

Stanfordeko unibertsitatean garatutako korreferentzia-ebazpenerako sistema (Lee *et al.*, 2013) erregelatan oinarritutakoa da. Hiru multzo nagusitan bana daitezkeen 10 bahez (korreferentzia-ebazpenerako modulu espezifiko) osatua dago. Honako hauek dira hiru multzo horiek:

1. String-parekatzean oinarritzen direnak
2. Egitura bereziak tratatzen dituztenak, hala nola, aposizioak eta predikazioak
3. Izenordainen ebazpenaz arduratzen dena

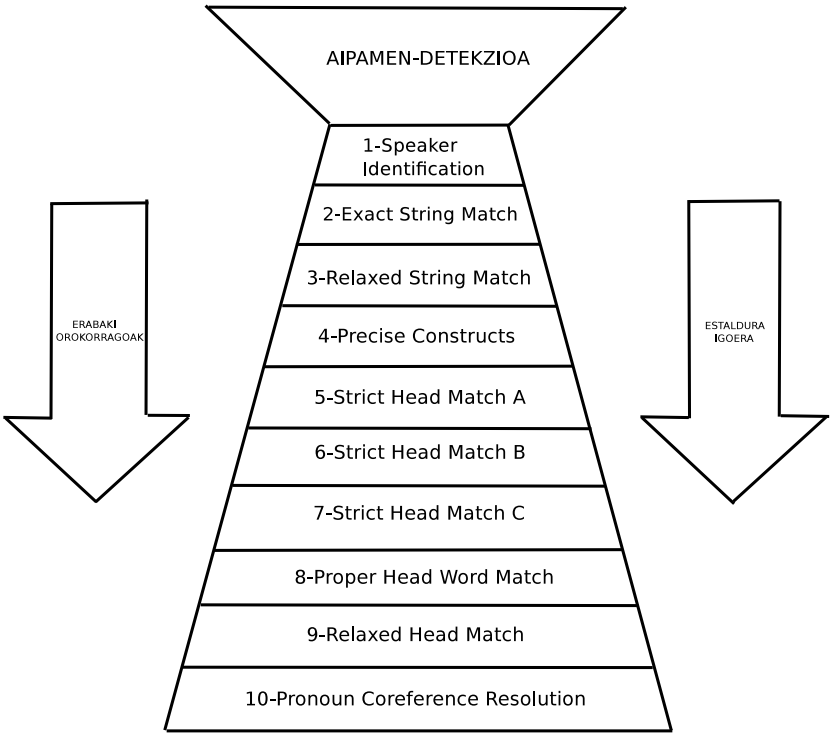
10 bahe horiek banan-banan aplikatzen dira, doitasun handiena lortzen dutenak aurrenik eta doitasun baxuagokoak ondoren. Planteamendu honi esker, lehenengo baheetan erabaki ziurrak hartzen dira (doitasun handia) eta ondorengoetan ez hain ziurrak, estaldura hobetuz baina batzuetan doitasuna kaltetuz. Bahe horiek sakonki azaltzen dira 4. kapituluko 4.4 azpiatalean.

Sistemaren arkitektura guztiz modularra izanik, erraz integra daitezke korreferentzia-ebazpenerako bahe berriak. Hori dela eta, ingelesa ez den beste hizkuntzetara ere nahiko erraz egokitzen da. *CoNLL-2011 shared task* atazan (Pradhan *et al.*, 2011) emaitzarik onenak lortu zituen. Sistemaren arkitektura 2.1 irudian ikus daiteke.

Beautiful Anaphora Resolution Toolkit

BART ikasketa automatikoan oinarritzen den korreferentzia-ebazpenerako sistema (Versley *et al.*, 2008b) Johns Hopkins Summer Workshopean hasi zen garatzen eta korreferentziaren ebazpenean egindako aurrerapenak plataforma batean biltzea zuen helburu. Aurreprozesaketarako hainbat metodo, ikasketarako ezaugarri desberdinak eta MMAX2 tresna (Müller and Strube, 2006) erabiliz errore analisiak egiteko aukera eskaintzen du tresnak, betiere modulartasun izaera mantenduz. Ondorioz, BART sistema corpus eta ezaugarri multzo desberdinetara arrakastaz egokitu da, baita hizkuntza desberdinetara ere.

Oinarrian bost modulu nagusi ditu BART sistemak: 1) aurreprozesaketarako pipelinea, 2) aipamen detekziorako erabiltzen den *Mention Factory* izeneko modulua, 3) ezaugarriak erauzteko modulua (*Feature extractor*), 4) dekoderra eta 5) enkoderra. Enkoderra ikasketarako instantziak sortzeaz arduratzen da, dekoderra aldiz, test garaian hartu diren korreferentzia-erabakiak



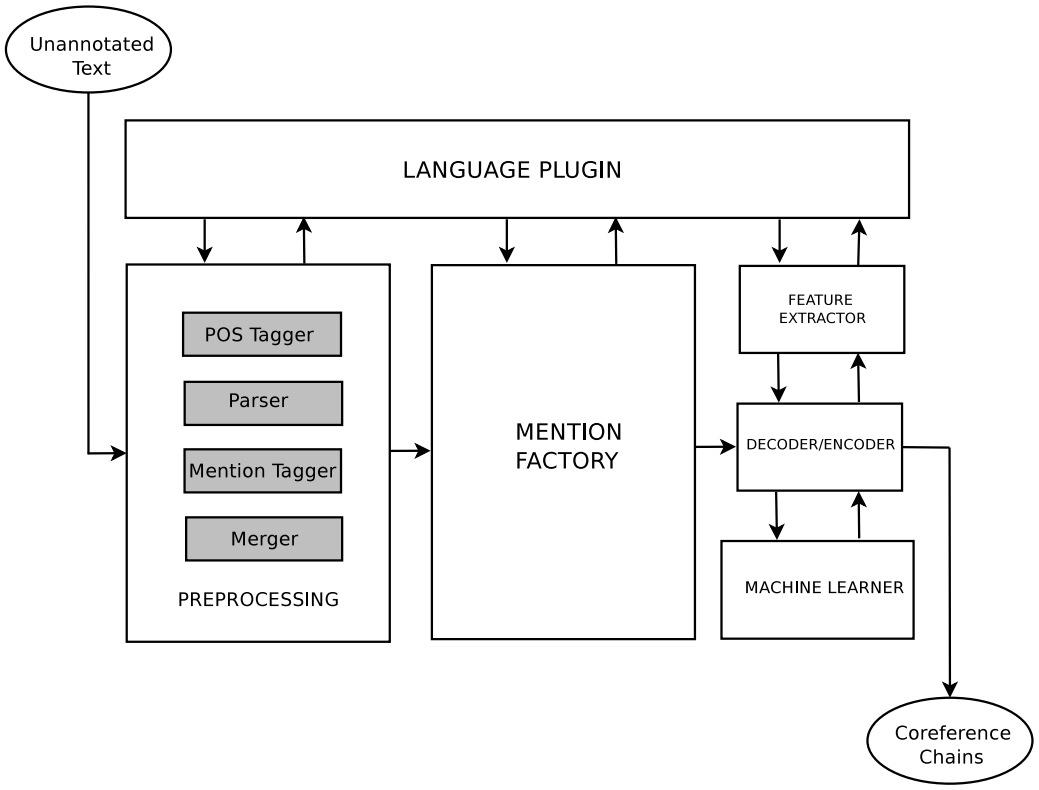
2.1 irudia – Stanfordeko korreferentzia-ebazpenerako sistemaren arkitektura.

multzo edo clusterretan biltzeaz. Horiez gain, *Language Plugin* deritzon moduluak hizkuntzari dagokion informazio espezifikoa tratatzen du. Bost modulu nagusiez aparte, ikasketarako erabili nahi den sailkatzailea hautatzeko aukera ere eskaintzen du. BART sistemaren arkitektura 2.2 irudian ikus dezakegu.

2.3.4 Semantikaren erabilera korreferentzia-ebazpenean

Milaka dokumentuz osatutako korreferentzia-kateez etiketatutako corpusetatik lortzen den informazio semantikoa datu berrietan erabiltzeko ez dela nahikoa ohartu ziren Durrett and Klein (2013) autoreak. Ondorioz, korreferentzia-ebazpenerako sistemen garatzaileak kanpo-baliabideak erabili beharrean aurkitzen dira ezagutza semantikoa lortzeko.

Recasens *et al.* (2013) lanean azaltzen denez aipamen-bikote opakuak (hitz guztiz desberdinez osatutako aipamenak, adibidez, *Reala* eta *futbol-*



2.2 irudia – BART korreferentzia-ebazpenerako sistemaren arkitektura.

talde txuri-urdina) arloaren egoerako korreferentzia sistema batek egiten dituen errore kausen % 65 dira. Hori dela eta, korreferentzia emaitzak % 60-70etik haratago lortzeko ezinbestekoa da ezagutza semantikoa hobeto ustiatzea aipamen opakuez osatutako korreferentzia-erlazioak ebazteko. Autoreek dokumentu konparagarriez osatutako corpusa erabiliz, ezizenez (*alias*) aparte, sinonimia eta metonimia erlazioak erauzten dituzte. CoNLL metrika erabiliz F_1 neurria % 0,7 hobetzea lortzen dute urrezko-aipamenak erabiltzen dituzten kasuan.

Stanfordeko korreferentzia-ebazpeneko sistemaren errore iturri handia, erroreen % 42, sistema honek ezagutza semantikoa erabiltzeko duen ezgaitasunaren ondorio dela aipatzen da Lee *et al.* (2013) lanean.

Informazio lexikal eta entziklopedikoa biltzen duten ezagutza-baseak, hala nola, WordNet, Wikipedia, Yago edo DBPedia sarri erabili izan dira korreferentzia-ebazpena hobetzeko asmoz.

WordNet (Fellbaum, 1998) ezagutza lexikala biltzen duen baliabiderik zaharrenetarikoa da. *Synset* deritzen egiturez osatua dago eta egitura hauek hitzen adiera sinonimikoak biltzen dituzte. WordNet sinonimoak eta erlazio hiperonimikoak lortzeko oso baliabide erabilgarria da.

Wikipedia eduki libreko entziklopedia, aportazio libre kolaboratiboen bidez osatzen dena da.

Yago (Suchanek *et al.*, 2007) Wikipediako sarrerak WordNetekin lotzen dituen ezagutza-basea da.

Eta azkenik, DBPediak (Mendes *et al.*, 2012) Wikipediako datuetatik erauzitako informazio ontologiko erabilgarria dauka.

Informazioa lexikal eta entziklopedikoa erabili duten lanei dagokienez, Ponzetto and Strube (2006) autoreak izan ziren WordNet eta Wikipedia erabiltzen lehenengotarikoak.

Uryupina *et al.* (2011) lanean bateragarritasun semantikoa eta ezizenen informazioa erauzi zuten autoreek eta korreferentzia-ebazpenerako sistema bati txertatu. Erauzitako ezagutza hori desanbiguazio eta iragazpenik gabe erabiltzeak oinarri-lerroarekiko inolako hobekuntzarik ez zekarrela frogatu zuten, aldiz, desanbiguazio eta iragazpen teknika sinpleak aplikatuz emaitzak hobetzeko aukera zegoela. Horrelako teknikak erabilita, haien sistema 2 eta 3 puntu bitartean hobetzea lortu zuten.

Rahman and Ng (2011) lanean Yago erabili zuten aipamenak ezagutza-atributuekin aberasteko, baina aberasketa hau zaratatsua izan daitekeela ohartu ziren.

Ratinov and Roth (2012) autoreek euren lanean Wikipediako orriak erabili zituzten ezaugarriak erauzteko eta honela beraien sistema hobetzeko.

WordNeten oinarritutako sinonimia eta hiponomia ezaugarriekin CoNLL metrikan 60,06tik 60,42rako hobekuntza lortu zuten aipamen automatikoak erabiliz Durrett and Klein (2013) autoreek, eta 75,08tik 76,68ra urrezko aipamenak erabiltzean.

Hajishirzi *et al.* (2013) lanean, korreferentzia-ebazpena eta entitate izendunen estekatzea aldi berean gauzatzen dituen NECo izeneko sistema aurkezten da. Bi atazak aldi berean egikarituz bakoitzean egiten diren erroreen kopurua jaistea lortzen dute autoreek. NECo-k, Stanfordeko korrefentzia-ebazpenerako sistema hedatzen du bahe berriak gehituz. Bahe horiek, entitate izendunen estekatzeari esker lortutako informazioa erabiltzen dute modu automatikoan aipamenak Wikipediara lotuz eta horrela hobekuntzak lortuz, zehazki 1,2 puntuko hobekuntza lortuz MUC neurrian eta 0,3koa B^3 neurrian.

Instantzia anbiguoen, hau da, laburtutako izenen gaineko entitate-izendunen detekzioak korreferentzia-ebazpena hobetu dezakeelako intuizioa dago Durrett and Klein (2014) lanaren oinarrian, baita Wikipediatik lortutako ezagutzak ere. Aldi berean, korreferentzia-ebazpenak entitate izendunen ezagutza hobetu dezake. Entitate izendunen detekzioa eta korreferentzia-ebazpena batera gauzatzen duten kasuan CoNLL neurrian 0,48 puntuko hobekuntza lortzen dute.

Versley *et al.* (2016) autoreen arabera, sistema batean informazio lexikala eta munduari buruzko ezagutza txertatzean ez da erraza hobekuntzak lortzea baina posiblea eta erabat beharrezkoa da.

2.3.5 Hizkuntzak

Korreferentzia-ebazpenerako sortutako lehen sistemak ingeleserako diseinatuak izan baziren ere azken urteetan ingelesa ez den beste hizkuntzetan korreferentzia-ebazpena gauzatzeak interesa sortu du, eta ondorioz, lan ugari argitaratu dira.

Lehen aipatu bezala, SemEval-2010 Task 1 atazan (Recasens *et al.*, 2010) katalana (Sapena *et al.*, 2011), nederlandera (Zhekova and Kübler, 2010; Kobdani and Schütze, 2010), alemana (Attardi *et al.*, 2010), italiera (Broscheit *et al.*, 2010a) eta gaztelera (Sapena *et al.*, 2011) hizkuntzetarako sistemak aurkeztu ziren. Hurrengo urtean ospatu zen *CoNLL 2012 Shared Task* atazan berriz, txinerarako (Martschat *et al.*, 2012; Chen and Ng, 2012; Björkelund and Farkas, 2012; Xu *et al.*, 2012) eta arabierarako (Fernandes *et al.*, 2012; Uryupina *et al.*, 2012; Stamborg *et al.*, 2012) sistemak agertu ziren.

Sarritan, hizkuntza berri baterako korreferentzia-sistema garatu nahi denean, sistema osoa hasieratik sortu beharrean, beste hizkuntza baterako sortua izan den sistema bat hartu eta hizkuntza berrira egokitzen da. Adibidez, *CoNLL 2012 Shared Task*-ean txinera eta arabiera tratatzeko Stanfordeko sistema egokitu zuten sei parte-hartzailek (Chen and Ng, 2012; Fernandes *et al.*, 2012; Shou and Zhao, 2012; Xiong and Liu, 2012; Yuan *et al.*, 2012; Zhang *et al.*, 2012).

BART sistema (Versley *et al.*, 2008b) ere hizkuntza askotara egokitua izan da. Ingeleserako sortu izan bazen ere, sistemak eskaintzen duen arkitektura modular eta malguak beste hizkuntzetarako egokitzeko erraztasunak eskaintzen ditu. Hori dela eta, lan ugari aurkeztu dira non BART sistema ingelesa ez den beste hizkuntzetara egokitua izan den. (Poesio *et al.*, 2010) lanean Wikipediako artikuluez osatutako Evalita corpusa erabiliz Italiera-

ra egokitua izan zen; TüBa-D/Z korreferentzia-corpusa erabiliz alemanerako egokitu zen (Broscheit *et al.*, 2010b) lanean; polonierarako ere egokitua dago (Kopeć and Ogrodniczuk, 2012) lanean azaltzen den moduan, eta azkenik Uryupina *et al.*-k (2012) arabiera eta txinerarako egokitu zuten. Orain dela gutxi, BART bengalerara (Sikdar *et al.*, 2013) eta euskarara (Soraluze *et al.*, 2016b) egokitua izan da.

Hizkuntza gutxituei dagokienez, ezaguna da horietan gertatu ohi den baliabide linguistikoen urritasuna. Hori dela eta, korreferentzia-ebazpena moduko atazetarako tresna eraginkorrak garatzea erronka izan ohi da. Hala ere, hasi dira hizkuntza gutxituetarako garatutako korreferentzia-ebazpenerako sistemak agertzen. Azken urteotako lanetan ikus dezakegu, hungarierarako (Mihăltz, 2008), polonierarako (Ogrodniczuk and Kopeć, 2011), txekierarako (Nguỵ *et al.*, 2009), hindi hizkuntzetarako (Sobha *et al.*, 2011b), galizierarako (Garcia and Gamallo, 2014a), persierarako (Nazaridoust *et al.*, 2014) edo lituanierarako (Žitkus and Nemuraitė, 2015) sistemak garatu direla.

Tesi lan hau hasi aurretik, euskarari dagokionez izenordainen ebazpenerako ikasketa automatikoan oinarritutako Arregi *et al.* (2010) lana bakarrik zegoen argitaratuta.

2.3.6 Ebaluazio-metrikak

Korreferentzia-ebazpenerako sistemen kalitatea nola neurtu zehaztea erabat beharrezkoa da. Ebaluazio-metrika egoki batek sistema baten benetako kalitatea erakusteaz gain, ikerketa lan ezberdinen arteko konparaketa ahalbidetu behar du. Azken hamarkadetan ebaluazio-metrika ugari proposatu dira. Proposatutako metrika guztiak korreferentzia-ebazpena ebaluatzeko sortuak izan dira edota aurretik proposatutako metriken gabeziak konpontzea dute helburu.

Korreferentzia-ebazpenean, urre-patroia sistema automatikoak itzuli duen erantzunarekin konparatzen da. Konparaketarako erabiltzen den urre-patroiko aipamen eta entitateen multzoari *key* (K) deitzen zaio, eta sistema automatikoak itzultzen dituen aipamen eta entitateen multzoari berriz *response* (R).

Gaur egun, korreferentzia-ebazpeneko sistemak ebaluatzeko orduan erabiltzen diren metrikak honakoak dira: MUC (Vilain *et al.*, 1995), B^3 (Bagga and Baldwin, 1998), CEAF_e (Luo, 2005), CEAF_m (Luo, 2005), BLANC (Recasens and Hovy, 2011) eta LEA (Moosavi and Strube, 2016).

MUC, B^3 eta CEAF_e neurrien batezbesteko aritmetikoa den CoNLL neurria ere, MELA izenez (Denis and Baldridge, 2009) ezaguna, erabiltzen da korreferentzia-ebazpenerako sistemen ebaluazioetan.

MUC

Korreferentzia-ebazpena ebaluatzeko erabili den metrika zaharrena eta erabiliena da. Loturetan oinarritutako (*link-based*) metrika da, hau da, *key* eta *response* artean dauden korreferentzia-loturak konparatzen dira ebaluazio balioak lortzeko.

Entitate konkretu baten estaldura kalkulatzeko, entitate horrek *key* multzoan dituen loturak ($|K|-1$), *response* multzoan falta diren lotura kopuruarekin konparatzen dira, azken hau *key* multzoko partizioen kopuru ($|p(K)|$) ken 1 bezala kalkulatzen da. $|p(K)|$, *key* multzoko entitatearen eta dagozkion *response* multzoko entitateen arteko ebakidura da. Beraz, entitate baten estaldura:

$$R = \frac{(|K-1|) - (|p(K)|-1)}{|K|-1} = \frac{|K| - |p(K)|}{|K|-1} \tag{2.1}$$

Dokumentu baten estaldura kalkulatzeko, entitate guztien estaldurak batzen dira.

$$R = \frac{\sum_{k_i \in K}(|k_i| - |p(k_i)|)}{\sum_{k_i \in K}(|k_i|-1)} \tag{2.2}$$

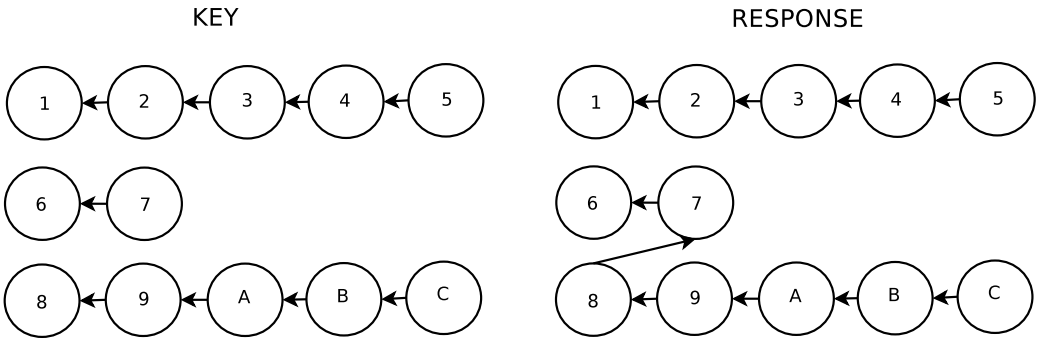
Doitasuna kalkulatzeko, aurreko planteamenduari buelta ematen zaio: Orain oinarri bezela *response* multzoa hartzen da eta egiten den galdera da zenbat lotura gehitu behar zaizkion *key* multzoari, *response* multzoa sortzeko.

$$P = \frac{\sum_{r_i \in R}(|r_i| - |p(r_i)|)}{\sum_{r_i \in R}(|R_i|-1)} \tag{2.3}$$

Azkenik, MUC F-balioa kalkulatzeko doitasunaren eta estalduaren batazbesteko harmonikoa kalkulatzen da.

$$F = \frac{2pr}{p+r} \tag{2.4}$$

(2.2) eta (2.3) ekuazioek konplexuak badirudite ere erraz kalkulatzen dira. MUC metrikaren balioak nola kalkulatzen diren modu errazean ikusteko zentra gaitezen 2.3 irudian agertzen diren *key* eta *response* multzoetan. Irudian 3 entitatetan banatzen den 12 *key* aipamenez osatutako dokumentu bat agertzen da, korreferentzia-ebazpenerako sistema automatiko batek aldiz dokumentua 2 entitatetan banatzen den 12 aipamenez osatuta dagoela itzultzen digu.



2.3 irudia – Korreferentzia-erlazioak *key* eta *response* multzoetan.

Estaldura kalkulatzeko *response* multzoko entitateen araberako *key* multzoko partizioak behar ditugu. Adibidez, {1,2,3,4,5} *key* entitatearen kasuan bere partizioa {{1,2,3,4,5}} da; {6,7} entitateari dagokion partizioa {{6,7}} eta {{8,9,A,B,C}} entitateari dagokiona berriz {{8,9,A,B,C}}. Ondorioz, estaldurak honako balioa du:

$$R = \frac{(5-1)+(2-1)+(5-1)}{(5-1)+(2-1)+(5-1)} = 1$$

Doitasuna kalkulatzeko, *response* eta *key* multzoen rolak alderanztea nahikoa da. Kasu honetan {1,2,3,4,5} *response* entitatearen partizioa {{1,2,3,4,5}} da eta {6,7,8,9,A,B,C} *response* dagokiona berriz {{6,7},{8,9,A,B,C}}. Ondorioz honakoa da doitasunaren balioa:

$$P = \frac{(5-1)+(7-2)}{(5-1)+(7-1)} = \frac{9}{10}$$

Behin doitasuna eta estaldura kalkulatuta, MUC F-balioa honakoa izango litzateke:

$$F = \frac{2 \times 1 \times \frac{9}{10}}{1 + \frac{9}{10}} = \frac{18}{19}$$

MUC metrikari kritikatu zaizkion puntu ahulen artean, korreferentzia-kate luzeak sortzen dituen sistemen emaitzak hobesten dituela da bat (Luo, 2005). Dokumentu bateko aipamen guztiak korreferentzia-kate berdinean kokatzen dituen sistemak % 100 estaldura lortzen du MUC metrika erabiltzean, eta doitasuna ez da asko kaltetzen. Kalkulatu berri ditugun 2.3. irudiaren estaldura eta doitasun balioak behatuz, garbi ikusten da oraintxe aipatu dugun ideia. Batu behar ez ziren bi entitate batuta lortzen den estaldura balioa % 100 da eta doitasuna oso gutxi kaltetzen da, 0,1 puntu hain zuzen.

Bestalde, metrikari kritikatzen zaion beste puntu ahul bat aipamen bakarrez osatutako entitateak (*singleton*) ez dituela kontuan hartzen da, horiek ez baitute inongo loturarik beste aipamen batekin. Ondorioz, aipamen hauetakoren bat ez dagokion entitatean edo clusterrean kokatzeak ez du inongo eraginik doitasunean (Bagga and Baldwin, 1998).

BCUBED

B^3 metrikak, estaldura kalkulatzeko, aipamenak konparatzen ditu ebaluazio balioak kalkulatzeko eta ez loturak MUC metrikak egiten duen moduan. Doitasun eta estaldura globalak, aipamen bakoitzaren doitasun eta estaldurak kalkulatuz lortzen dira eta ondorioz aipamenetan oinarritutako (*mention-based*) metrika dela esaten da.

B^3 metrikak m aipamenaren *response* entitatean dauden aipamen zuzenen frakzioa begiratzen du, *key* multzoko entitateetako m aipamen bakoitzeko.

$$R = \frac{\sum_{k_i \in K} \sum_{r_j \in R} \frac{|k_i \cap r_j|^2}{|k_i|}}{\sum_{k_i \in K} |k_i|} \tag{2.5}$$

MUC metrikan egiten den moduan, doitasuna *key* eta *response* multzoak aldatuz kalkulatzen da.

$$P = \frac{\sum_{k_i \in K} \sum_{r_j \in R} \frac{|k_i \cap r_j|^2}{|r_i|}}{\sum_{r_i \in R} |r_i|} \tag{2.6}$$

Azkenik, B^3 F-balioa kalkulatzeko doitasunaren eta estalduaren batazbesteko harmonikoa kalkulatzen da.

$$F = \frac{2pr}{p+r} \tag{2.7}$$

2.3. irudiko adibidera itzuliz, honela kalkulatuko lirateke B^3 metrikarekin estaldura eta doitasuna. Estaldura 1 da, *key* multzoko 3 entitateen eta *response* multzoko entitateen arteko ebakidurak kalkulatzen baditugu entitate multzo berdinak lortzen baititugu.

Doitasunari dagokionez, *response* multzoko {1,2,3,4,5} entitatearen ebakidura *key* multzoko entitateak izanda {1,2,3,4,5} da, beraz, entitate horrek doitasunean 5eko ekarpena egiten du. {6,7,8,9,A,B,C} entitatearen ebakidurak *key* multzoko entitateei dagokionez bi azpimultzo itzultzen ditu, {6,7} eta {8,9,A,B,C}, beraz, entitate horri dagokion doitasuna $\frac{2^2+5^2}{7} = \frac{29}{7}$.

Doitasun totala beraz:

$$P = \frac{5 + \frac{29}{7}}{12} = \frac{16}{21}$$

Eta F-measure balioa:

$$F = 2 \times \frac{1 \times \frac{16}{21}}{1 + \frac{16}{21}} = \frac{32}{37}$$

B^3 metrikak, MUC metrikak dituen ahuleziak konpontzen ditu, doitasun eta estaldura balioak m aipamen bakoitzarentzat kalkulatzen baititu, ondorioz, aipamen bakarreko entitateekin gertatzen den arazoa konpontzen da eta ez du emaitza hobea itzultzen korreferentzia-kate luzeak sortzen dituzten sistemetan. Hala ere, Luo (2005) autoreak aipatzen duen moduan, zenbait kasutan B^3 metrikak itzultzen dituen emaitzak nahasgarriak izan daitezke.

Adibidez, inongo loturarik sortzen ez duen eta *key* aipamen guztiak *singleton* moduan uzten dituen sistemak % 100-ko doitasuna lortzen du. Bestalde, *key* aipamen guztiak korreferentzia-kate batean biltzen dituen sistemak % 100-ko estaldura lortzen du. Azkenik, B^3 metriak aipamen errepikatuak kudeatzeko gaitasunik ez duela esaten da (Luo and Pradhan, 2016). Sistema batek, *key* multzoan agertzen den aipamen bat behin baino gehiagotan itzultzen badu bere erantzunean, errepikapen bakoitzeko B^3 balioa igo egiten da.

$\mathbf{CEAF}_m$ eta $\mathbf{CEAF}_e$

CEAF metrikak ϕ antzekotasun balioa erabiltzen du bi entitateren arteko antzekotasuna neurtzeko. Kuhn-Munkresen algoritmoa erabiltzen du, *key* entitateetatik *response* entitateetarako $g*$ mapaketa hoberena lortzeko. Suposatuz K^* mapaketa optimoan dauden *key* entitateen multzoa dela, estaldura honela kalkulatzen da:

$$R = \frac{\sum_{k_i \in K^*} \phi(k_i, g^*(k_i))}{\sum_{k_i \in K} \phi(k_i, k_i)} \tag{2.8}$$

CEAF metrika erabiliz doitasuna kalkulatzeko 2.8 ekuazioko zatitzailea, $\sum_{r_i \in R} \phi(r_i, r_i)$ zatitzailearekin ordezkatzen da:

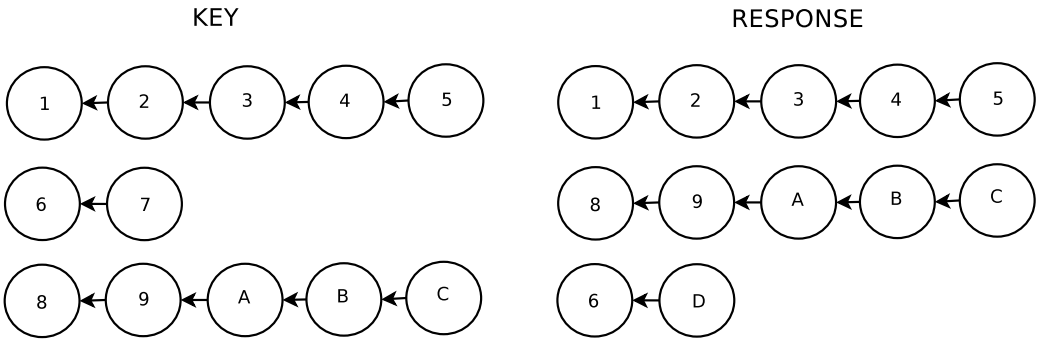
$$P = \frac{\sum_{k_i \in K^*} \phi(k_i, g^*(k_i))}{\sum_{r_i \in R} \phi(r_i, r_i)} \tag{2.9}$$

Azkenik, F-balioa honela kalkulatzen da:

$$F = \frac{2pr}{p + r} \tag{2.10}$$

ϕ mapaketan oinarrituta, CEAF metrikaren bi aldaera daude: CEAF_m eta CEAF_e. Lehenengoari, aipamenetan oinarritutako (*mention-based*) CEAF deitzen zaio eta bi entitateen arteko antzekotasuna bietan dauden aipamen berdinen kopuruarekin kalkulatzen du, hau da, $\phi(k_i, r_j) = |k_i \cap r_j|$. Bigarrenari, entitateetan oinarritutako (*entity-based*) CEAF deitzen zaio eta antzekotasuna kalkulatzeko bi entitate konparatzen ditu, non $\phi(k_i, r_j) = \frac{2 \times |k_i \cap r_j|}{|k_i| + |r_j|}$.

CEAF_m eta CEAF_e nola kalkulatzen diren garbiago har dezagun oinarritzat 2.4. irudiko adibidea.



2.4 irudia – Korreferentzia-erlazioak *key* eta *response* multzoetan.

Lehenik eta behin entitateen arteko mapaketa optimoa $g*$ lortu behar dugu. Honakoa izango litzateke *key* eta *response* multzoko entitateen arteko mapaketa optimoa:

$$\{1,2,3,4,5\} \leftrightarrow \{1,2,3,4,5\},$$
$$\{6,7\} \leftrightarrow \{6,D\},$$
$$\{8,9,A,B,C\} \leftrightarrow \{8,9,A,B,C\}.$$

Parekatutako 3 entitate pareek 5,1,5 aipamen komun dituzte, hurrez hurren. Beraz, CEAF_m $R = \frac{11}{12}$ da, $P = \frac{11}{12}$ eta ondorioz F-balioa ere $\frac{11}{12}$.

CEAF_m ordez CEAF_e aldaera erabiltzen badugu, emaitza aldatu egiten da. Parekatutako 3 entitateen F-balio lokalak 1, $\frac{1}{2}$ eta 1 dira, beraz CEAF_e estaldura $\frac{1+\frac{1}{2}+1}{3} = \frac{5}{6}$ da, doitasuna ere $\frac{5}{6}$. Ondorioz, F-balioa ere $\frac{5}{6}$ da.

CEAF metrikei ere kritikatu zaizkie hainbat ahulezi, (Denis and Baldridge, 2009) autoreek aipatzen duten moduan, CEAF metrikek mapeatu gabe gelditzen diren *response* multzoko entitateetan egindako lotura zuzenak ez dituzte kontuan hartzen. Horretaz gain CEAF_e metrikak, entitate guztiak berdin tratatzen ditu beraien tamaina kontuan hartu gabe (Stoyanov *et al.*, 2009).

BLANC

BLANC metrika, korreferentzia-ebazpena ebaluatzeko sortu zen Rand Indizearen (Rand, 1971) aldaera bat da. BLANC metrikak, korreferentzia-loturak eta ez-loturak erabiltzen ditu doitasuna, estaldura eta F balioa kalkulatzeko.

Izan bitez C_k eta C_r, *key* eta *response*ko korreferentzia-loturen multzoak hurrenez hurren eta N_k eta N_r *key* eta *response* multzoetan dauden ez-korreferentzia loturak. Korreferentzia-loturen estaldura, doitasunak eta F-balioak honela kalkulatzen dira:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|}, \quad P_c = \frac{|C_k \cap C_r|}{|C_r|}, \quad F_c = \frac{2R_cP_c}{R_c + P_c} \tag{2.11}$$

Ez-korreferentzia loturen estaldura, doitasun eta F-balioak berriz balioak honela kalkulatzen dira:

$$R_n = \frac{|N_k \cap N_r|}{|N_k|}, \quad P_n = \frac{|N_k \cap N_r|}{|N_r|}, \quad F_n = \frac{2R_nP_n}{R_n + P_n} \tag{2.12}$$

BLANC metrikaren estaldura eta doitasun balioak lortzeko, korreferentzia-loturen eta ez-korreferentzia loturen arteko batezbestekoa kalkulatzen da:

$$R = \frac{R_c + R_n}{2} \tag{2.13}$$

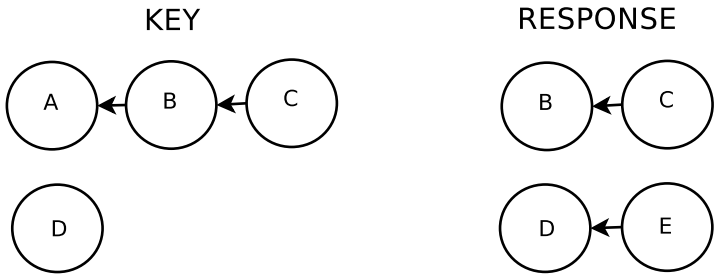
$$P = \frac{P_c + P_n}{2} \tag{2.14}$$

Eta F-balioa honela kalkulatzen da:

$$F = \frac{F_c + F_n}{2} \tag{2.15}$$

2.5. irudiko adibidea kontuan hartuz, honela kalkulatuko lirateke doitasuna, estaldura eta F-balioa BLANC metrika erabiltzen denean.

Lehenbizi, C_k, N_k, C_r eta N_r lortu behar ditugu. Honakoak izango lirateke:



2.5 irudia – Korreferentzia-erlazioak *key* eta *response* multzoetan.

$$C_k = \{(ab), (bc), (ac)\};$$
$$N_k = \{(ad), (bd), (cd)\};$$
$$C_r = \{(bc), (de)\};$$
$$N_r = \{(bd), (be), (cd), (ce)\};$$

Beraz, $C_k \cap C_r = \{(bc)\}$, $N_k \cap N_r = \{(bd), (cd)\}$.

Ondorioz, $R_c = \frac{1}{3}$, $P_c = \frac{1}{2}$, $F_c = \frac{2}{5}$, $R_n = \frac{2}{3}$, $P_n = \frac{2}{4}$, $F_n = \frac{4}{7}$.

Emaitza guzti horiek kalkulatu ostean, balio hauek hartuko lituzkete estaldura, doitasunak eta F-balioak:

$$R = \frac{\frac{1}{3} + \frac{2}{3}}{2} = \frac{1}{2}$$

$$P = \frac{\frac{1}{2} + \frac{2}{4}}{2} = \frac{1}{2}$$

$$F = \frac{\frac{2}{5} + \frac{4}{7}}{2} = \frac{17}{35}$$

LEA

LEA (Link-based Entity Aware) metrikak dokumentu bateko entitate bakoitzaren garrantzia eta ebazpenen prozesuan lortutako zuzentasuna kontuan hartzen ditu doitasun eta estaldura balioak kalkulatzerakoan.

Ondorioz, LEAk honela ebaluatzen ditu entitateak:

$$\frac{\sum_{e_i \in E}(garrantzia(e_i) \times ebazpen\text{-}balioa(e_i))}{\sum_{e_k \in Z} garrantzia(e_k)} \tag{2.16}$$

Non entitatearen tamaina (aipamen kopurua) erabiltzen den garrantzia kalkulatzeko, hau da, $garrantzia(e) = |e|$.

e entitate baten *ebazpen-balioa* berriz honela kalkulatzen da. n aipamen dituen e entitate batek $lotura(e) = n \times (n-1)/2$ ditu. Hori horrela izanik, *key* multzoko k_i entitate baten *ebazpen-balioa*, k_i entitate horretako zuzen ebatzitako loturen zatia da.

$$ebazpeneko\ balioa(k_i) = \sum_{r_j \in R} \frac{lotura(k_i \cap r_j)}{lotura(k_i)} \tag{2.17}$$

Behin *garrantzia* eta *ebazpen-balioa* definituta, LEA metrikan estaldura honela kalkulatzen da:

$$R = \frac{\sum_{k_i \in K}(|k_i| \times \sum_{r_j \in R} \frac{lotura(k_i \cap r_j)}{lotura(k_i)})}{\sum_{k_z \in K} |k_z|} \tag{2.18}$$

Doitasuna, *key* eta *response* multzoen rola aldatuz kalkulatzen da:

$$P = \frac{\sum_{r_i \in R}(|r_i| \times \sum_{k_j \in K} \frac{lotura(r_i \cap k_j)}{lotura(r_i)})}{\sum_{r_z \in R} |r_z|} \tag{2.19}$$

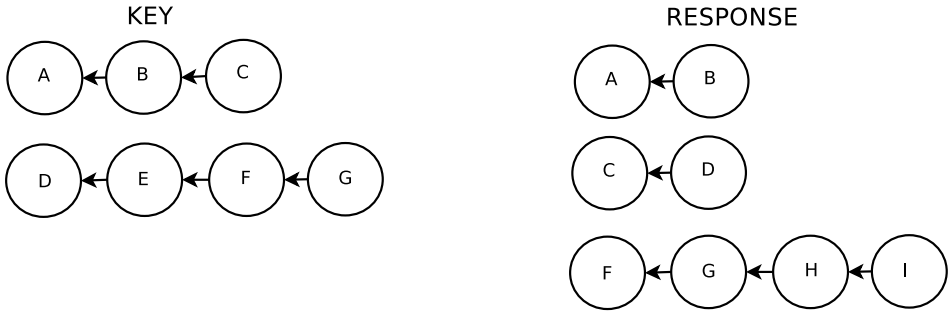
Eta F-balioa:

$$F = \frac{2pr}{p+r} \tag{2.20}$$

2.6. irudiko adibidean oinarrituz, kalkula ditzagun doitasun, estaldura eta F-balioak LEA metrika erabiliz.

key multzoko entitateen multzoa $K = \{k_1 = \{a, b, c\}, k_2 = \{d, e, f, g\}\}$ da eta *response* multzoko entitateen multzoa, berriz, $R = \{r_1 = \{a, b\}, r_2 = \{c, d\}, r_3 = \{f, g, h, i\}$.

garrantzia entitateen tamaina dela kontuan hartuz, $garrantzia(k_1) = 3$ eta $garrantzia(k_2) = 4$.



2.6 irudia – BLANC

k_1 eta $k_2$ren artean dauden korrefentzia-kateen multzoak {ab, ac, bc} eta {de, df, dg, ef, eg, fg} dira hurrenez hurren.

Ondorioz, $lotura(k_1) = 3$ eta $lotura(k_2) = 6$. *ab* lotura da k_1 eta r_1 multzoen artean dagoen lotura komun bakarra. Ez dago inolako lotura komunik k_1 eta gainerako *response* multzoko entitateen artean. Modu bertsuan, k_2-k lotura komun bakarra du r_3-rekin *fg*, eta bat ere ez r_1 edo r_2-rekin. Beraz, $ebazpen\text{-}balioa(k_1) = \frac{1+0+0}{3}$ eta $ebazpen\text{-}balioa(k_2) = \frac{0+0+1}{6}$.

Ondorioz estaldura:

$$R = \frac{3 \times \frac{1}{3} + 4 \times \frac{1}{6}}{3+4} \approx 0.24$$

Doitasuna:

$$P = \frac{2 \times \frac{1+0}{1} + 2 \times \frac{0+0}{1} + 4 \times \frac{0+1}{6}}{2+2+4} \approx 0.33$$

Azkenik, F-balioa:

$$F = \frac{2 \times 0.24 \times 0.33}{0.24 + 0.33} \approx 0.28$$

AIPAMEN-DETEKZIOA

3

Improving Mention Detection for Basque Based on a Deep Error Analysis

3.0 Laburpena

Artikulu honetan 2012an euskararako garatu genuen aipamen-detektatzailearen (Soraluze *et al.*, 2012) hobekuntza prozesua azaltzen da. Sistema erregelatan oinarritutakoa da eta euskarazko aipamenen egiturak kontuan hartzen ditu. Sistema hori corpus txiki batean ebaluatu zen eta azken urteotan eman den corpusaren handitzeak sistema hobetzeko bidea ireki digu.

Aipamen-detektatzaile bat garatzeko orduan garrantzitsua da aipamenen egiturak ondo aztertzea sistemak emaitza onak izango baditu. Hori dela eta, lehenbizi, euskarazko aipamen desberdinen egiturak analizatzen ditugu eta haien agerpen kopurua aztertzen dugu 12.792 aipamenez osatutako EPEC corpusean. 10 dira corpusean ageri diren aipamen motak, horietatik ugarienak izen-kate arruntak eta izen bereziak, % 80 baino gehiago. Horrek, indarra non jarri behar den jakiteko balio izan digu.

EPEC corpuseko aipamen moten banaketa beste hizkuntzetakoen antzekoa den ikusteko, ingeleseko OntoNotes eta gaztelerako eta katalaneko AnCora corpusetako aipamen moten agerpen kopuruarekin konparatu ditugu. Konparaketa horren emaitza 3.1 taulan ikus daiteke.

Izen-kate arruntak antzeko proportzioetan agertzen dira gainerako hizkuntzetako corpusetan, hala ere ez da berdina gertatzen izen berezien kasuan. OntoNotes corpusean % 2,72 soilik dira izen bereziak eta % 13,5 eta % 12,43 gaztelera eta katalaneko AnCora corpusetan. Beste desberdintasun

Aipamen-motak	Eu	En	Es	Cat
Izen bereziak	23,82	2,72	13,5	12,43
Izenordainak	2,97	25,85	14,65	14,24
Edutezkoak	1,36	8,78	3,55	3,12
Aditz-izenak	2,26	1,93	-	-
Postposizio-lokuzioetako aipamenak	4,68	-	-	-
Mendeko perpausa duten izen-kateak	3,15	-	-	-
Elipsia	0,61	-	-	-
Koordinazioa	2,87	-	3,41	3,97
Lekuzko adberbioak	0,53	-	-	-
Izen-kate arruntak	57,75	58,85	60,62	62,86

3.1 taula – Hizkuntza desberdinetako aipamen-moten konparaketa: Euskara (Eu), Ingelesa (En), Gaztelera (Es) and Katalana (Cat).

esanguratsu bat izenordainen kopuruan dago. Izenordainen agerpen kopuru handiena ingeleseko corpusean gertatzen da, % 25,85 hain zuzen ere, eta kopuru hori txikiagoa da gaztelerako eta katalaneko corpusetan, % 14,65 eta % 14,24 hurrenez hurren. Hala ere, diferentziarik handiena euskarazko izenordainen kopuruarekin gertatzen da, % 2,97 soilik baitira izenordainak EPEC corpusean. Recasens and Hovy (2010) lanean aipatzen denez, AnCora corpusean pertsona izenordainen kopurua ingeleseko corpusekoa baino txikiagoa da, katalana eta gaztelera *pro-drop* hizkuntzak direnez, subjektu posizioan dauden izenordainak eliditu egin baitaitezke. Euskaraz, subjektu posizioan dagoen izenordaina eliditzeko aukeraz gain, objektu zuzena eta zeharkako objektua ere eliditu daitezke, *three-way pro-drop* hizkuntza baita. Hori izan daiteke euskaraz izenordainen kopurua ingelesez, katalanez eta gazteleraz baino txikiagoa izatearena.

Corpus desberdinetan agertzen diren desberdintasunik handienak kontuan izanda, hau da, izenordainen eta izen berezien proportzioak, ikus daiteke bien baturak antzekoak direla hizkuntza guztietan. % 26,79 euskaraz, % 28,57 ingelesez, % 28,15 gazteleraz eta % 26,67 katalanez hurrenez hurren. Balio hauek hizkuntza bakoitzak izenordainen erabilera dela eta desberdin jokatzen dutela iradokitzera garamatza. Dirudienez, ingelesak izenordainak erabiltzeko joera handiagoa du testuan aurrerago agertu den zerbaiti erreferentzia egiteko, katalanak eta gaztelerak izen berezien eta izenordainen arteko oreka mantentzen dute, eta euskarak izenordainen ordez izen bereziak

erabiltzeko joera nabarmenagoa duela dirudi, ziurrenik, bestela sor daiteken anbiguotasuna saihesteko.

Aipamenak aztertu ondoren, aipamen detekziorako egindako lehen hurbilpena aurkezten eta ebaluatzen dugu. Ebaluaziorako, nazioartean ezagunak diren bi parekatze metodo erabili ditugu: *Lenient Matching* edo *Partial Matching* eta *Strict Matching*. Aipamen bat zuzena 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). Parekatze mota honi *Lenient Matching* edo *Partial Matching* esaten zaio. 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.

Oinarri-lerroa (B) finkatzeko lengoaia naturalaren prozesamendurako tresna generiko bat erabiltzen dugu, zatitzaile edo *chunkerra* hain zuzen ere. Aipamen egiturekin antzekotasun handiena duten egiturak zatitzaile edo *chunkerrak* itzultzen dituen izen-kateak izanik, egitura horiek aipamentzat hartzen ditugu hurbilpen honetan. % 64,28 puntuko F-measure balioa lortzen dugu ebaluaziorako *Exact Matching* protokoloarekin eta % 71,12 puntukoa *Lenient Matching* protokoloa erabiltzean. Zatitzaileak edo *chunkerrak* zehazki ez dira aipamenen detekzioa egiteko sortuak izan. Zenbait kasutan aipamen batzuk egitura sintaktiko konplexuagoak dituzte eta horrek, tresna hauen emaitzetatik lortzen diren aipamenen mugak batzutan zehazki zuzenak ez izatea dakar. Hori dela eta, tresna hauek doitu egin behar dira aipamenen detekzioa modu egokian egin ahal izateko.

Aipamenen mugak doitzeko, zatitzailearen irteera egoera finituko teknologia erabiliz egokitzen dugu eta horrela emaitzak hobetzea lortzen dugu. Eskuz definitutako erregelak konpilatuz Egoera Finituko Transduktoreak (Finite State Transducers, FST) lortu dira. Guztira 34 erregela definitu ditugu eta horiek konpilatu ostean 12 FST lortu dira. 3.2 taulan hurbilpen honekin, Basque Mention Detector (BMD), lortutako emaitzak eta oinarri-lerroarekin (B) lortutakoak ikus ditzakegu. Basque Mention Detector-arekin, % 73,36 puntuko F-measure balioa lortzen dugu *Exact Matching* protokolorekin eta % 79,68 puntukoa *Lenient Matching* protokoloarekin. Beraz, 9,08ko eta 8,56 puntuko hobekuntzak lortu dira, hurrenez hurren, oinarri-lerroarekiko.

Aipamen detekzioan hizkuntza baten ezaugarriak kontuan hartzen duen garrantziaz jabetzeko, gure aipamen-detektatzailearen emaitzak ingelesera-

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
B	62,24	66,46	64,28	67,02	75,75	71,12
BMD	73,86	72,87	73,36	78,69	80,71	79,68

3.2 taula – Oinarri-lerroa eta aipamen-detektatzailearen emaitzak. B=Oinarri-lerroa, BMD=Basque Mention Detector.

ko sortuak izan diren bi aipamen-detektzailek lortzen dituzten emaitzekin konparatu ditugu. Batetik, Stanfordeko sistemak erabiltzen duen aipamen-detektatzailea (SMD) eta bestetik BART sistemak erabiltzen duena (BRTMD) konparatzen dira. Lortutako emaitzak 3.3 taulan ikus daitezke. Gure aipamen detektatzaileak (BMD) 47,53 puntu hobetzen du konparatu den Stanfordeko aipamen-detektatzailearen (SMD) emaitza *Exact Matching* protokoloa erabiltzean denean eta 22.63 puntu *Lenient Matching* erabiltzean. BARTen aipamen-detektatzailearekiko (BRTMD) 35,01eko hobekuntza dago ebaluazioan *Exact Matching* protokoloa erabiltzen den kasuan eta 17.29koa *Lenient Matching* protokoloarekin. Lortutako emaitzek garbi erakusten dute aipamen-detektatzaileak sortzerako orduan hizkuntzaren ezaugarriak kontuan hartzeak emaitza hobea lortzea dakarrela.

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
SMD	23,15	29,22	25,83	50,25	65,98	57,05
BRTMD	34,69	42,87	38,35	54,89	72,27	62,39
BMD	73,86	72,87	73,36	78,69	80,71	79,68

3.3 taula – Aipamen-detektatzaile desberdinen konparaketa. SMD=Stanford Mention Detector, BRTMD= BART Mention Detector, BMD=Basque Mention Detector.

Hala ere, aipamen-detekzioan gertatzen diren erroreak sailkatu eta aipamen-detektzailea hobetzeko asmotan, aipamen-detektatzailearen ebaluazio kualitatibo bat aurkezten dugu ondoren. 1.904 aipamenez osatutako lagin bat erabiliz, aipamen-detekzioan gertatzen diren errore-moten eta errore-kausen sailkapen bana proposatzen dugu. Guztira 8 errore-mota eta 9 errore-kausa identifikatzen dugu.

Errore-moten eta errore-kausen sailkapenean oinarrituta sistemaren errore-analisia aurkezten da. Aipamenen % 32,09k erroreren bat dute. Horiek erabiliz errore-motak kuantifikatzen ditugu eta ondoren errore bakoitzaren kausa identifikatzen dugu. Errore-mota usuenak *Missing Mention (MM)*, hau da, urre-patroiko aipamen bat ez da detektatu kasuen % 11,50ean gertatzen da eta *Extra Mention (EM)*, urre-patroian ez dagoen aipamen bat itzultzen da, berriz, kasuen % 7,46an. Errore-kausa nagusienak, *Incorrect Chunk Tag (ICT)*, *chunkerrak* itzultzen dituen etiketaren bat okerra da % 27,93an eta *Missing Chunk Tag (MCT)*, *chunkerrak* ez du itzuli etiketaren bat % 18,92an.

Horrela, aipamen-detektatzailearen errore-mota eta eta errore-kausen banaketa lortzen dugu eta horren aurrean 5 hobekuntza proposatzen eta inplementatzen ditugu aipamen-detektatzailea hobetuz (BIMD). Proposatutako hobekuntza horiek inplementatu ostean, erroreen % 10 konpontzea lortzen dugu eta aipamen-detektatzailearen emaitzak hobetzen ditugu. Emaitza horiek 3.4 taulan ikus ditzakegu. Zehazki esanda % 74,57 puntuko F-measure balioa lortzen du sistemak *Exact Matching* protokoloarekin eta % 80,57 puntukoa *Lenient Matching* protokoloa erabiltzean.

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
BMD	73,86	72,87	73,36	78,69	80,71	79,68
BIMD	74,67	74,47	74,57	79,26	81,92	80,57

3.4 taula – Aipamen-detekzioan lortutako emaitzen hobekuntza. BMD=Basque Mention Detector, BIMD= Basque Improved Mention Detector.

Errore-analisiak erakutsi digu aurreprozesaketaren ondorioz errore ugari gertatzen direla. Hori dela eta, aurreprozesaketarako tresnetatik jasotzen diren erroreek aipamen-detekzioan izan dezaketen eragina kuantifikatzeko gure aipamen-detektatzailea partzialki urrezkoa den sarrerarekin ere ebaluatzen dugu. Sarrera partzialki urrezkoa dela diogu, sarrera horrek ez baititu hizkuntzaren maila guztiak eskuz etiketatuta. Eskuz etiketatutako lemak, kategoriak, azpikategoriak eta kate edo *chunkak* jasotzen ditugu. Funtzio sintaktikoak eta entitate izendunak automatikoki tratatuak izan dira. Egoera honetan, sistemak ($BIMD_g$) % 85,89ko F-measure balioa lortzen du *Strict Matching* protokoloarekin eta % 89.06koa *Lenient Matching* erabiltzean, hau da, 11,32 eta 8,49 puntuko diferentzia, hurrenez hurren, aipamen automati-

koak jasotzen dituen sistemaren ($BIMD_a$) emaitzarekin konparatzen baditugu. Emaitzak 3.5 taulan ageri dira. Garbi ikusten da tresna automatiko batek jasotzen duen sarrerak berebiziko garrantzia duela aipamen-detekzioan emaitza onak lortzeko orduan.

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
$BIMD_a$	74,67	74,47	74,57	79,26	81,92	80,57
$BIMD_g$	84,95	86,84	85,89	87,06	91,15	89,06

3.5 taula – Aipamen-detekzioan lortutako emaitzak sarrera automatikoarekin eta partzialik urrezkoa den sarrerarekin.

Azkenik, aipamen-detekzioan eginiko hobekuntzek korreferentzia-ebazpenean duten eragina aztertzen da. Horretarako Soon *et al.* (2001) lanean oinarritzen den korreferentzia-ebazpenerako sistema sinple batek lortzen dituen emaitzak ebaluatzen ditugu, aipamen-detektatzaile ezberdinek lortutako aipamenak erabiliz. Oinarri-lerroko aipamen-detektzailea (B), euskararako sortu dugun lehenengo aipamen-detektatzailea (BMD) eta hobekuntzak aplikatuta dituena (BIMD) izan dira aztertu direnak. Lortutako emaitzak ataza honetan ohikoak diren metrikak erabiliz, 3.6 taulan ikus daitezke.

	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
B	25,17	54,39	51,62	51,92	28,13	43,83
BMD	28,36	61,62	57,96	58,91	35,56	49,63
BIMD	29,28	62,65	58,69	59,65	36,77	50,53

3.6 taula – Aipamen-detekzioaren eragina korreferentzia-ebazpenean.

Aipamen-detekzioan oinarri-lerroa 9 puntu hobetzen dugun kasuan korreferentzia-ebazpeneko CoNLL neurria 5,8 puntu hobetzea lortu dugu, 43,83tik 49,63ra, eta errore-analisia egin ondoren proposatutako hobekuntzak inplementatu ostean aipamen-detekzioan hobetutako 1,21 puntuk CoNLL neurria 0,9 puntu hobetzea lortu dutela ikusi dugu, 49,63tik 50,63ra. Aipamen-detekzioan emaitza hobeak lortzeak korreferentzia-ebazpenerako emaitzetan eragin zuzena duela ondorioztatzen dugu.

Laburbilduz, euskarazko aipamen desberdinen egituren azterketan oinarritzen den aipamen-detektatzailea aurkeztu da. Aipamen-detektatzaile hori

hobetzeko hainbat estrategia aplikatzen dira, hala nola egoera finituko teknologiaren erabilera eta errore-analisi batetik ondorioztatuko hobekuntzen inplementazioa. Aipamen detekzioan hizkuntza baten ezaugarriak kontuak hartzeak lortzen diren emaitzak hobetzen dituela frogatzen da eta sistemak jasotzen duen sarrerak ere berebiziko garrantzia duela ondorioztatzen dugu. Emaitza onak lortzeak korreferentzia-ebazpenean duen eragina ere aztertzen da. Bukatzeko, aipamen-detektatzaileak % 74,57 puntuko F-measure balioa lortzen du ebaluaziorako *Exact Matching* protokoloa erabiltzen denean eta % 80,57 puntukoa *Lenient Matching* protokoloarekin.

Improving Mention Detection for Basque Based on a Deep Error Analysis

Ander Soraluze, Olatz Arregi, Xabier Arregi and Arantza Díaz de Ilarraza

Published in *Natural Language Engineering*, 23(3):351-384

Abstract

This paper presents the improvement process of a mention detector for Basque. The system is rule-based and takes into account the characteristics of mentions in Basque. A classification of error types is proposed based on the errors that occur during mention detection. A deep error analysis distinguishing error types and causes is presented and improvements are proposed. At the final stage, the system obtains an F-measure of 74.57% under the Exact Matching protocol and of 80.57% under Lenient Matching. We also show the performance of the mention detector with gold standard data as input, in order to omit errors caused by the previous stages of linguistic processing. In this scenario, we obtain an F-measure of 85.89% with Strict Matching and of 89.06% with Lenient Matching, i.e., a difference of 11.32 and 8.49 percentage points, respectively. Finally, how improvements in mention detection affect coreference resolution is analysed.

3.1 Introduction

Coreference resolution consists of identifying textual expressions (mentions) that refer to real-world objects (entities) and of determining whether these mentions refer to the same entity. It is well known that coreference resolution is essential in Natural Language Processing applications where a higher accuracy in discourse analysis leads to better performance. Some of the tasks that can benefit from coreference resolution include information extraction, question answering, machine translation, sentiment analysis, machine reading, text summarisation, and text simplification.

In Soraluze *et al.* (2012) we presented a preliminary mention detector for Basque, evaluated on a small corpus. In this paper we present an improved version

that has been evaluated on a considerably larger corpus to obtain more reliable results. A bigger corpus has also made it possible to carry out a deep error analysis, which in turn has allowed us to identify and address some of the main sources of error.

Basque is a non Indo-European language and differs considerably in grammar from the languages spoken in the surrounding regions. It is an agglutinative, head-final, pro-drop, free-word order language isolate (Laka, 1996), which presents a number of special problems regarding mention detection.

The paper is structured as follows. After reviewing related work in Section 3.2, we describe the types of mention structures in Basque in Section 3.3. Section 3.4 compares mention types in Basque, English, Catalan and Spanish. Section 3.5 describes the experiments we carried out in order to create a mention detector, and in Section 3.6 a classification of errors types and causes made during mention detection is presented. Moreover, after a deep error analysis is carried out, further improvements are designed. Section 3.7 outlines an experiment with running the detector on gold input, with the aim of observing the effect of general natural language preprocessing tools. In order to observe how improvements in mention detection affect coreference resolution, in Section 3.8 we compare the scores obtained by a coreference model before and after the improvements in mention detection are applied. Finally, conclusions and future work are discussed.

3.2 Related Work

Much attention has been paid to the problem of coreference resolution in the past two decades. Conferences specifically focusing on coreference resolution have been organised since 1995. The sixth and seventh Message Understanding Conferences (MUC-6, 1995; MUC-7, 1998) included a specific task on coreference resolution. The Automatic Context Extraction (ACE) Program focused on identifying certain types of relations between a predefined set of entities (Doddington *et al.*, 2004) while the Anaphora Resolution Exercise (ARE) involved anaphora resolution and NP coreference resolution (Orasan *et al.*, 2008).

More recently, SemEval-2010 Task 1 was dedicated to coreference resolution in multiple languages (Recasens *et al.*, 2010). A year later, in the CoNLL-2011 shared task (Pradhan *et al.*, 2011), participants had to model unrestricted coreference in the English-language OntoNotes corpora (Pradhan *et al.*, 2007) and CoNLL 2012 Shared Task (Pradhan *et al.*, 2012) involved predicting coreference in three languages: English, Chinese and Arabic.

The coreference resolution task is commonly divided into two main subtasks: mention detection and resolution of references (Pradhan *et al.*, 2011). Mention

detection is considered a crucial task for the accuracy of end-to-end coreference resolution systems and the mention detection step has a significant impact on coreference resolution systems: Uryupina (2008) reports that 35% of recall errors in their coreference resolution system are caused by missing mentions and in Uryupina (2010) adds that 20% of precision errors are due to inaccurate mention detection, while Stoyanov *et al.* (2009) conclude that improving the ability of coreference resolvers to identify coreference elements or mentions would likely improve the state-of-the-art greatly. Chang *et al.* (2011) argue that a robust mention detector is crucial, as detection errors propagate to the coreference stage; as they show, their system, which uses gold mentions, outperforms the system that uses predicted or system mentions by a large margin (from 15% to 18% absolute difference in F_1 score). Hacioglu *et al.* (2005) also show that errors in the first stages, such as mention detection, propagate and reduce the performance of subsequent stages, and Zhekova and Kübler (2010) demonstrated experimentally that during the development phase the detection of mentions is one of the most important steps to achieve a high accuracy in coreference resolution. Also in Broscheit *et al.* (2010a) the importance of mention detection is recognised.

There are two main approaches to the technology used to create mention detectors: rule-based systems and machine learning models. Machine learning models seem able to better balance precision and recall, and thus to achieve higher F-score values in the mention detection task, their recall tends to be quiet a bit lower than obtained by rule-based systems designed to favour recall (Pradhan *et al.*, 2011). In end-to-end coreference resolution systems, low recall in the mention detection step has negative effects on the whole process, since missed mentions are not available for coreference resolution in the later stages. As is pointed out in Uryupina and Moschitti (2013), in the CoNLL 2011 shared task the majority of the participants relied on rule-based modules to obtain mention boundaries for English. However, in the CoNLL 2012 shared task, participants with rule-based systems fell back to very simple baselines in mention detection, which demonstrated that rule-based systems are not easily adaptable to languages other than the one for which they were created.

The first systems that resolved coreferences were mainly designed for English. During the past few years, however, efforts in resolving coreference in languages other than English have been published. For example, in the SemEval-2010 Task 1, Catalan, Dutch, English, German and Italian coreference resolvers were evaluated. The CoNLL 2012 shared task added Arabic and Chinese. Less-resourced languages are also gaining presence in coreference resolution, as exemplified by recent work on Hungarian (Mihâltz, 2008), Polish (Ogrodniczuk and Kopeć, 2011), Czech (Nguy *et al.*, 2009), and also Indian Languages (Lalitha Devi *et al.*, 2014).

3.3 Mentions in Basque

Language-specific patterns that vary according to the features of each language have to be taken into consideration during mention detection. In Basque written text, while not all mentions represent a particular detection challenge, some can have complex structures and require a more sophisticated identification strategy.

In general, we consider noun phrases (NP), focusing on the largest span of the NP. In the case of nouns complemented by subordinate clauses and coordination, NPs embedded in a larger NP are also extracted.

These are the structures that we consider mentions:

- **Proper Nouns (PN):** Structures that have a proper noun as head.

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

  Clinton itxaropentsu agertu_ zen kazetarien aurrean
  Clinton hopefully appeared the_reporters in_front_of

  "[Clinton] appeared hopeful in front of the reporters."

- **Pronouns (Pro):** All personal pronouns are considered mentions. However, in Basque, the demonstrative determiners also act as third person pronouns (Laka, 1996). To deal with this ambiguity we mark as mentions the demonstratives used as pronouns (example 6).

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

  LDPko buruek Mori hautatu_ zuten apirilean Keizo Obuchi
  of_LDP The_heads Mori chose in_April Keizo Obuchi
  lehen ministroa ordezkatzeko hark tronbosia izan
  Prime the_Minister to_replace who_(he) a_thrombosis suffered
  ostean
  after

  "The heads of LDP chose Mori in April to replace the Prime Minister Keizo Obuchi after [who (he)] suffered a thrombosis"

  Note that "hark" acts as a pronoun in this context, but the same word may be used as a determiner in other contexts. For instance, "hark" is a determiner in the sentence "Gizon [hark] tronbosia izan zuen;" ("[That] man suffered a thrombosis").

- **Possessives (Poss):** We consider two types of possessives: possessive determiners, even if they are not the head of the NP as in example 7; and possessive pronouns, as in example 8.

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

  Ekisoainek Granollersera joatea nahiago_ du bere emazteak
  Ekiosain to_Granollers to_go prefers his wife
  lagunak baititu bertan
  friends has_because there

  "Ekisoain prefers to go to Granollers because [his] wife has friends there."

  (8) Escuderok euskal musika tradizionala eraberritu eta indartu zuen. [Harenak] dira, esate baterako, Illeta, Pinceladas Vascas eta Eusko Salmoa obrak.

  Escuderok euskal musika tradizionala eraberritu eta
  Escudero Basque music traditional renewed and
  indartu_ zuen Harenak dira esate_ baterako Illeta Pinceladas
  gave_prominence his are for_example Illeta Pinceladas
  Vascas eta Eusko Salmoa obrak
  Vascas and Eusko Salmoa The_works

  "Escudero renewed and gave prominence to traditional Basque music. The works Illeta, Pinceladas Vascas and Eusko Salmoa, for example, are [his]."

- **Verbal nouns (VN):** Verbs that have been nominalised and function as the head of the mention with the corresponding case marking suffix. The whole clause governed by the verbal noun has to be annotated.

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

  Europar Batasunaren zabaltze_ honek arazo asko
  European of_the_Community the_growth problems many
  konpontzera behartuko_ ditu
  to_solve will_force_them

  "[The growth of the European Community] will force them to solve many problems."

- **NPs as part of complex postpositions (CPost):** As Basque has a postpositional system, we mark the independent NP that precedes the complex postpositions. In example 10, the postposition is *aurka* ("against"), and we annotate the noun Athleticen (in this case, a proper noun) that precedes it, which is in genitive case (-en).

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

  Moreno eta Vlatko Djolonga Athleticen aurka jokatzeko
  Moreno and Vlatko Djolonga Athletic against to_play
  moduan daude.
  in_shape are

  "Moreno and Vlatko Djolonga are in shape to play against [Athletic]."

- **NPs containing subordinate clauses (SubrCl):** The head of these mentions is always a noun complemented by a subordinate clause. In example 11, the head noun is complemented by a subordinate clause that is called a complementary clause. We take the whole span of the NP (both the subordinate clause and the head noun) as a mention. In addition, when a relative clause adds information to the noun, as in example 12, the boundaries of the mention are set from the beginning of the relative clause to the end of the NP.

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

  Oslon hasitako prozesua gaur bukatuko_da
  in_Oslo that_began The_process today is_going_to_end

  "[The process that began in Oslo] is going to end today."

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

  Antimisilen_inguruan Pentagonoa atontzen ari_den sistema
  The_missile_defense the_Pentagon preparing that_is system
  aurkeztu_dute
  has_been_presented

  "[The missile defense system that the Pentagon is preparing] has been presented."

- **Ellipsis (Ellip):** In Basque, ellipsis is a common phenomenon.

  At morphosyntactic level, a noun-ellipsis occurs when the suffixes attached to the word correspond to a noun, even when the noun is not explicit in the word. We consider this type of ellipsis in the case of verbs that take suffixes indicating noun-ellipsis, as in example 13. The POS given by the analyser indicates the presence of an ellipsis, which is deduced by the presence of both the verb (*jaitsi ginen-* "we were relegated") and the empty mark (*-Ø-ekoa* "that in which"). All the information corresponding to both units is stored and treated as a noun.

  (13) Niretzat oso urte txarra izan zen [bigarren mailara jaitsi ginenekoa].

  Niretzat *oso* *urte txarra* *izan_zen* *bigarren*
  In_my_opinion really year a_bad it_was second
  mailara *jaitsi_ginenekoa*
  to_the_division that_in_which_we_were_relegated

  "In my opinion it was a really bad year, [that in which we were relegated to the second division]."

- **Coordination (Coor):** In the case of coordination, nominal groups of a conjoined NP are extracted. We also regard as mentions the nested NPs (*Xabier Mikel Errekondo*, and *Alvaro Jauregi*) and the whole coordinate structure (*Xabier Mikel Errekondo eta Alvaro Jauregi*).

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

  Xabier Mikel Errekondo eta *Alvaro Jauregi euren kontratuak*
  Xabier Mikel Errekondo and Alvaro Jauregi their contracts
  berritzear *daude*
  about_to_renew are

  "[[Xabier Mikel Errekondo] and [Alvaro Jauregi]] are about to renew their contracts."

- **Location Adverbs (LocAdv):** In general, adverbs are not referential, yet location adverbs do have a referential function. Therefore, location adverbs are considered mentions.

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

Futbol jokalariak Biarritzera joatekoak_ziren han festa
football The_players to_Biarritz intented_to_go there a_party
antolatuta baitzuten
organised because_they_had

"The football players intended to go to Biarritz, because they had a party organised [there]."

- **Common Noun Phrases (CNP):** Phrases that have a noun as head word.

(16) [Langileak] haserre daude hartutako erabakiarekin.

Langileak haserre daude hartutako erabakiarekin
The_workers angry are taken with_the_decision

"[The workers] are angry with the decision taken."

In Table 3.7 we can observe the distribution of mention types in Basque according to the classification of mentions presented above. This distribution has been calculated on the EPEC corpus (the Reference Corpus for the Processing of Basque), which has been previously annotated manually. It is a collection of news published in *Euskaldunon Egunkaria*, a Basque newspaper. More details about this corpus are presented in Section 3.5.

Mention type	#	%
Proper nouns	3047	23.82
Pronouns	379	2.97
Possessives	174	1.36
Verbal nouns	289	2.26
NPs as part of complex postpositions	599	4.68
NPs containing subordinate clauses	403	3.15
Ellipsis	78	0.61
Coordination	367	2.87
Location adverbs	68	0.53
Common noun phrases	7388	57.75
Total	12,792	100.00

3.7 Table – Mention types in EPEC corpus.

As can be observed, the majority of mentions are common noun phrases, (57.75%), followed by proper nouns (23.82%). The other structures are much less frequent in the corpus, ranging from about 0.5% to just under 5% of the corpus.

3.4 Comparison of mention types between different languages

This section compares the distribution of mentions in Basque with their distribution in English, Catalan, and Spanish.

Pradhan *et al.* (2011) present a distribution of mentions in OntoNotes corpus by their syntactic category. This English corpus collects newswire articles, magazine articles, broadcast news, broadcast conversations and web data. Recasens and Martí (2010) present the statistics of mention types in the AnCora corpus for Catalan and Spanish. Each corpus consists of newspaper and newswire articles.

The distributions for English, Catalan, and Spanish are not directly comparable with the distribution for Basque. For instance, the type "NPs as part of complex postpositions" has no equivalent type in English, Spanish or Catalan. To make them partially comparable, we strove to map the mention types we consider in Basque to the same or similar structures in these other languages.

Although the comparison is not detailed, it enables us to reach certain conclusions about the differences between Basque and other languages.

Table 3.8 presents the distributions in different languages. Some differences and similarities between languages can be noted. As expected, in all the languages the majority of mention structures are common noun phrases, with quite similar percentages: 57.75% in Basque, 58.85% in English, 60.62% in Spanish and 62.86% in Catalan. In Basque proper nouns also appear frequently, 23.82%; this does not happen in English, only 2.72%. In Spanish and Catalan the presence of proper nouns is similar, 13.5% and 12.43%, respectively; however, they are fewer than in English and more numerous than in Basque. Another remarkable difference occurs regarding the presence of pronouns. Again, the values are similar for Catalan and Spanish, but there is a significant difference between English and Basque: in English, pronouns are far more frequent than in Basque. As pointed out in Recasens and Hovy (2010), AnCora has a smaller number of personal pronouns compared with English because Spanish and Catalan are pro-drop languages, which allow zero subject pronouns that can be inferred from the verb. In addition to the pronominal subject, Basque allows the omission of direct and indirect objects (three-way pro-drop) and, contrary to English, Catalan and Spanish, it does not generally use pronouns to construct relative clauses. These characteristics can be the reason for the low presence of pronouns in Basque compared with English, Catalan and Spanish.

The other differences are far less dramatic. The main outlier is possessive structures in English, which are clearly more frequent than in the other languages.

After reviewing the presence of mention types in different languages, some general conclusions can be sketched out. In particular, it seems that the sum of

Mention type	Eu	En	Es	Cat
Proper nouns	23.82	2.72	13.5	12.43
Pronouns	2.97	25.85	14.65	14.24
Possessives	1.36	8.78	3.55	3.12
Verbal nouns	2.26	1.93	-	-
NPs as part of complex postpositions	4.68	-	-	-
NPs containing subordinate clauses	3.15	-	-	-
Ellipsis	0.61	-	-	-
Coordination	2.87	-	3.41	3.97
Location adverbs	0.53	-	-	-
Common noun phrases	57.75	58.85	60.62	62.86

3.8 Table – Comparison of mention types among different languages: Basque (Eu), English (En), Spanish (Es) and Catalan (Cat).

proper nouns and pronouns is about the same in each language: 26.79% for Basque, 28.57% for English, 28.15% for Spanish, and 26.67% for Catalan. This may suggest that each language behaves differently with regard to the use of pronouns. While English tends to use pronouns to reference mentions that have appeared previously, Spanish and Catalan balance the use of pronouns and proper nouns, using both almost equally. Pronouns have no gender in Basque, which seems to prioritise the use of proper nouns instead of pronouns, presumably to avoid ambiguity.

3.5 Experimental setup

3.5.1 Corpus

The EPEC corpus (the Reference Corpus for the Processing of Basque) aims to function as a "reference" corpus for the development and improvement of several NLP tools for Basque (Aduriz *et al.*, 2006). It is a 300,000-word sample collection of news published in *Euskaldunon Egunkaria*, a Basque-language newspaper. This corpus has been manually tagged at different levels (morphology, syntax, phrases, etc.). Recently, mentions and coreference chains have also been tagged by two expert linguists in a subpart of EPEC corpus. The estimation of the inter-annotator agreement, including the chance-based factor for the task of tagging mentions has not been researched yet according to Artstein and Poesio (2008). In Kopeć and Ogrodniczuk (2014) only the observed agreement is presented, since it is difficult to estimate the probability of a random tagging of a mention. The same procedure

is used in Ohta *et al.* (2012). One annotation is considered as gold and the other as system. We evaluated the annotation consistency in mention detection in the same way, which was scored using the Strict Matching protocol, explained in Section 3.5.2. The F-measure obtained for agreement was 94.07%.

The mention detector has been developed and tested using a subpart of the EPEC corpus consisting of 46,383 words that correspond to 12,792 mentions. The relation between the number of words and the number of mentions is about 27%, meaning that approximately every four words a mention is detected. Similar statistics regarding the proportion of words/mentions can be found in Màrquez *et al.* (2013), where English, Spanish and Catalan values are presented.

3.5.2 Scoring Protocols

The most used measures to evaluate mention detection are precision, recall, and F-measure. To calculate these, mentions extracted by the mention detector and the set of manually tagged mentions are compared. Typically, two matching methods have been used to make the comparison between mentions.

The first one, known as *Lenient Matching* or *Partial Matching*, considers mentions to be correct if their span is within the span of the gold mention and contains the head word (Kummerfeld *et al.*, 2011). This protocol was used in Message Understanding Conferences (MUC-6, 1995; MUC-7, 1998). Slightly different measure was used in SemEval-2010 Task 1 (Recasens *et al.*, 2010) to score the mention detection task. Mention were rewarded with 1 point if their boundaries coincide with those of gold mentions and the mentions that whose boundaries are within the gold mention including its head with 0.5 points. And in ACE programs (NIST, 2008) only system mentions that overlap more than 30% in terms of character span with the gold mention and which have the same entity type were credited.

The second scoring protocol is clearly a stricter variation, because it only considers correct mentions that are exactly the same as the gold mentions. This matching method is known as *Strict Matching* and has been used in conferences such as CoNLL-2011 Shared Task (Pradhan *et al.*, 2011).

We present our mention detector scores using both Lenient Matching and Strict Matching.

3.5.3 Nominal chunks as mentions

For mention detection, as a starting point we decided to use a rule-based approach using a generic NLP tool, namely a chunker (Aduriz and Díaz de Ilarraza, 2003), an analyser that identifies verbal and nominal chunks based on rule-based grammars. This tool is integrated in a pipeline for Basque processing where other modules

are also applied: i) A morphological analyser that performs word segmentation and PoS tagging (Alegria *et al.*, 1996), ii) A lemmatiser that also disambiguates the PoS and the syntactic function (Alegria *et al.*, 2002a), iii) A multi-word item identifier that determines which groups of two or more words are to be considered multi-word expressions (Alegria *et al.*, 2004), iv) A named-entity recogniser that identifies and classifies named entities (person, organization, location) in the text (Alegria *et al.*, 2003).

The chunker uses Constraint Grammar (CG) formalism (Karlsson *et al.*, 1995) and obtains a precision value of 81.08%, a recall of 81.09% and F-measure of 81.08% in chunking.

The nominal-chunks can be considered the structures that are most similar to mentions, so we consider them and discard the verbal chunks for mention detection. The obtained results can be seen in Table 3.9. We obtained an F-measure of 64.28% using the Strict Matching protocol and 71.12% using Lenient Matching. These scores were considered as baseline.

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
B	62.24	66.46	64.28	67.02	75.75	71.12

3.9 Table – Mention detection using nominal chunks. B=Baseline.

Regarding the scores obtained using only a chunker for mention detection, we conclude that mention detection itself is a very challenging task since expressions can have complex syntactic and semantic structures. Preprocessing with generic NLP tools, while helpful, did not always succeed in identifying mention boundaries correctly. The same fact is pointed out by Nguyen *et al.* (2008), who suggest that using a base noun phrase chunker is insufficient for mention detection in the bio-domain, and that therefore some adaptations are needed. In less resourced languages such as Basque, a processing of the output obtained by generic NLP tools is needed to obtain better results in mention detection. In contrast, in richer resource languages where generic NLP tools such as parsers obtain good results, these tools can be used for accurate mention detection with little or no processing.

The example in Figure 3.1 illustrates the differences between the chunker output and mentions. In this example, we deal with the sentence *Mikelek erosi zituen etxeak...* "The houses that Mikel bought...". The first row shows that *Mikel* "Mikel" and *etxeak* "houses" are the nominal chunks obtained by the chunker, which are automatically considered mentions in this first approach. *Mikel* is a mention itself but *etxeak* cannot be considered a mention, because it is only the head of a larger mention *Mikelek erosi zituen etxeak* "The houses that Mikel bought". It

is clear that some processing of the output generated by a chunker is needed to correctly identify mentions.

	[Mikelek] erosi zituen [etxeak] . . .
Chunker:	NP BVP EVP NP
Correct:	[[Mikelek] erosi zituen etxeak] . . .

3.1 Figure – Wrongly identified mention. NP =Noun Phrase, BVP = Begin Verbal Phrase, EVP = End Verbal Phase

3.5.4 Mention Detection with Finite State Transducers

As explained in Section 3.5.3 using a preprocessing tool such a chunker does not really succeed in mention identification. Some mention structures are more complex than nominal chunks and mentions exceed the chunk boundaries. Thus, to identify even the complex type of mentions, we decided to combine the chunker with a clause tagger, that is, an analyser that identifies clauses, combining rule-based-grammars and machine learning techniques (Arrieta, 2010).

We combined the tags provided by these tools and created a mention detector system, defining a set of hand-crafted rules that have been compiled into Finite State Transducers (FST).

The FSTs are able to detect complex structures that should be identified as mentions and that chunkers do not recognise. In reference to the classification of mentions presented in Section 3.3, the FSTs detect verbal nouns, NPs as part of complex postpositions, NPs containing subordinate clauses, ellipses, coordination cases, and location adverbs.

To better understand how the FSTs work, an example of a mention composed by an NP containing a relative clause is presented, and then a rule that identifies these structures and marks them as mentions is defined.

Suppose that the mention detector receives as input the sentence *Armada britainiarrak Ipar Irlandan dituen bi kuartel eraitsi dituzte*.

Armada britainiarrak Ipar Irlandan dituen bi
army British Northern in_Ireland that_the_has two
kuartel eraitsi_dituzte
military_barracks has_been_demolished
"Two military barracks that the British army has in Northern Ireland have been demolished"

Figure 3.2 shows the above sentence with information obtained during the preprocessing step and the mentions that the system proposes.

Input	[Armada	britainiarrak]	[Ipar	Irlandan]	dituen	[bi	kuartel]	eraitsi	dituzte
Chunker	BNP	ENP	BNP	ENP	VREL	BNP	ENP	BVP	EVP
Clause tagger	{CB{CB				CB}				CB}
Output	[[Armada	britainiarrak]	[Ipar	Irlandan]	dituen	bi	[kuartel]]	eraitsi	dituzte

3.2 Figure – The input and the output of the mention detector. BNP = Begin-NP, ENP = End-NP, BVP = Begin-VP, EVP = End-VP, VREL = Relative Verb, CB = Clause Boundary.

Using the preprocessing information, the rule in Figure 3.3 identifies the verb containing a relative suffix (*dituen* "that has"), which is tagged with VREL. Next, the right side boundary is established in the NP or coordinated NPs that follow the relative verb (*bi kuartel* "Two military barracks"). Finally, the left side boundary of the mention is established at the closest clause boundary ({CB) to the left (*Armada* "army"). Following these steps, the system proposes the correct mention (*Armada britainiarrak Ipar Irlandan dituen bi kuartel* "Two military barracks that the British army has in Northern Ireland") and tags the whole structure to show where the mention begins (<MENTION>) and ends(</MENTION>).

```
define RV Verb & $['VREL'];
define RM [CB W+ RV NP [[and|or|...]  NP]*] @->
                                "<MENTION>" ... "</MENTION>";
```

3.3 Figure – A simplified rule to recognise NPs containing subordinate clauses. NP = Noun Phrase, CB = Clause Boundary, W=Word, RV = Relative Verb, RM = Relative Mention.

Further discussion about the behaviour of the FSTs can be found in Soraluze *et al.* (2012).

The use of Finite State Technology enables the processing of large datasets at a high processing speed and with low memory usage. Using *foma*[1] (Hulden, 2009), an open source platform for finite-state automata and transducers, we defined 12 FSTs, composed of 34 hand-crafted rules.

[1]The regular expression syntax used in *foma* can be consulted at http://code.google.com/p/foma/

We compared the results of the mention detector thus obtained with a baseline, established by considering all the nominal chunks that the chunker outputs to be mentions. The results are presented in Table 3.10.

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
B	62.24	66.46	64.28	67.02	75.75	71.12
BMD	73.86	72.87	73.36	78.69	80.71	79.68

3.10 Table – Baseline and mention detection scores. B=Baseline, BMD=Basque Mention Detector.

Our mention detector outperforms the Strict Matching protocol F-measure baseline by 9.08 percentage points, and that of Lenient Matching by 8.56 percentage points. The improvements obtained by our system are significant.[2]

With the aim of comparing our mention detector system with standard systems, we carried out experiments with the Stanford Mention Detector presented in Lee *et al.* (2013) and with BART (Versley *et al.*, 2008b) for mention detection. Using a high-recall algorithm, Stanford Mention Detector identifies nominal and pronominal mentions from all noun phrases (NPs), pronouns, and named entity mentions, and then non-mentions are filtered out. BART uses a built-in mention extraction module, computing boundaries heuristically from the output of a parser. It creates a list of candidate mentions by merging basic NP chunks with named entities.

The tokens and PoS tags provided to both mention detectors were analysed by Basque Linguistic processors, nevertheless, the PoS tags were mapped to equivalent Penn PoS tags (Marcus *et al.*, 1993). The parse trees used by the two systems were created using the Stanford PCFG parser (Klein and Manning, 2003).

Table 3.11 presents the scores obtained by the three compared systems. Stanford Mention Detector (SMD), BART Mention Detector (BRTMD) and Basque Mention Detector (BMD).

Compared to the Stanford mention detector, our mention detector outperforms Strict Matching protocol F-measure by 47.53 and that of Lenient Matching by 22.63 percentage points. Basque Mention Detector also improves the scores that BART Mention Detector obtains by 35.01 when Strict Matching protocol is used, and by 17.29 using Lenient Matching. The great difference in F-measures of Stanford and BART mention detector compared with those obtained by our mention detector

[2]Statistical significance is tested with Paired Student's t-test. p-value < 0.01

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
SMD	23.15	29.22	25.83	50.25	65.98	57.05
BRTMD	34.69	42.87	38.35	54.89	72.27	62.39
BMD	73.86	72.87	73.36	78.69	80.71	79.68

3.11 Table – Comparison of different mention detectors. SMD=Stanford Mention Detector, BRTMD= BART Mention Detector, BMD=Basque Mention Detector.

clearly shows that language specific characteristics (agglutinative, free-word order etc. for Basque) should be taken into account when mention detection is performed.

3.6 Error analysis

Evaluation scores can show how efficient a mention detector is; however, they neither identify the type of errors that the mention detector makes, nor give any indication of how those errors might be corrected.

In Subsection 3.6.1 we present the classification of the error types we found during the error analysis, and in 3.6.2 the causes of these errors.

3.6.1 Error Types

To help identify the different types of errors, we carried out a qualitative evaluation on a subcorpus of 1,904 mentions that were used to develop the system.

We classify the error types made in the mention detection step in 8 different categories. The error categorization idea is based on (Kummerfeld and Klein 2013) where errors of a coreference system are classified. We have adapted this categorization to classify errors produced in mention detection. To better understand this, for each category we present an example. The first item in each example is the incorrect case ($\times$) and the second the respective correct one ($\surd$). The categories are the following:

1. **Missing mention (MM):** The mention in gold standard has not been identified.

   $\times$ Guk bere garaian apustu handia egin genuen Sa Pinto fitxatu genuenean, jokalaria egoera berezi batean aurkitzen baitzen.

√ Guk bere garaian apustu handia egin genuen Sa Pinto fitxatu genuenean, [jokalaria] egoera berezi batean aurkitzen baitzen.

Guk bere_ garaian apustu handia egin_ genuen Sa Pinto
we at_the_time gamble a_great took Sa Pinto
fitxatu_ genuenean jokalaria egoera berezi batean
we_hired the_player situation special in_a
aurkitzen_ baitzen
was_because

"At the time we hired Sa Pinto we took a great gamble because [the player] was in a special situation."

2. **Extra mention (EM):** The system proposes a mention that is not present in the gold standard.

× [Bestetik], Zestoako Udalak mozio bat onartu zuen aho batez AEKren kontrako prozesua salatzeko.

√ Bestetik, Zestoako Udalak mozio bat onartu zuen aho batez AEKren kontrako prozesua salatzeko.

Bestetik Zestoako_ Udalak mozio_ bat onartu_ zuen
Moreover Zestoa_City_Council a_motion accepted
aho_ batez AEKren kontrako prozesua salatzeko
unanimously AEK against the_trial to_condem

"Moreover, Zestoa City Council accepted a motion unanimously to condemn the trial against AEK."

3. **Span errors (SE):** The system mention does not match exactly with the true or gold mention. Two different span error cases can happen.

3.1 **Head identified (HI):** The head of the mention to be identified is inside the span proposed by the system.

3.1.1 **Missing text (MT):** The system mention span misses one or more words that should be considered part of the gold mention.

× [Jiri Vanek] txekiarrak irabazi egin zion atzo.

√ [Jiri Vanek txekiarrak] irabazi egin zion atzo.

Jiri Vanek txekiarrak irabazi_ egin_ zion atzo
Jiri Van czech beat_him/her yesterday

"[Czech Jiri Van] beat him/her yesterday."

3.1.2 **Extra text (ET):** The system mention span has one or more words that are not part of the gold mention.

× [Luis Uranga harrituta] azaldu da Portugalgo klubaren jokaerarekin.

√ [Luis Uranga] harrituta azaldu da Portugalgo klubaren jokaerarekin.

"[Luis Uranga] is surprised with the attitude of the club from Portugal."

Luis Uranga harrituta_ azaldu_ da Portugalgo
Luis Uranga is_surprised from_Portugal
klubaren jokaerarekin
of_the_club with_the_attitude

"[Luis Uranga] is surprised with the attitude of the club from Portugal."

3.2 **Head not identified (HNI):** The head of the mention to be identified is not inside the system mention span.

× Europako Kontseiluetan herrialde handiek indar [gehiago] izango dute.

√ Europako Kontseiluetan herrialde handiek [indar gehiago] izango dute.

Europako Kontseiluetan herrialde handiek indar
European in_councils nations The_great strength
gehiago izango_ dute
more will_have

"The great nations will have [more strength] in European councils."

4. **Divided mention (DM):** Mention to be identified is divided in two or more mentions.

× Eztandak ez du inor zauritu, eta kalte material txikiak eragin ditu, Hammersmith zubian, [Londres] [mendebaldean].

√ Eztandak ez du inor zauritu, eta kalte material txikiak eragin ditu, Hammersmith zubian, [Londres mendebaldean].

Eztandak *ez_du* *inor* *zauritu* *eta* *kalte* *material*
The_explosion have_not anyone hurt and damage material
txikiak *eragin_ditu* *Hammersmith_zubian* *Londres*
small has_been in_Hammersmith_Bridge London
mendebaldean
west

"The explosion in Hammersmith Bridge, [west London], has not hurt anyone and material damage has been small."

5. **Conflated mentions (CM):** Two or more mentions are identified as they were only one.

   × Gogotsu ariko da [Euskaltel Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.

   √ Gogotsu ariko da [Euskaltel] [Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.

   Gogotsu *ariko_da* *Euskaltel*
   enthusiastically will_participate Euskaltel
   Alpeetako_Klasikoan *eta* *Dauphine_Liberen*
   in_the_Classique_des_Alpes and in_Dauphine_Libere
   ariko_dira *hurrena*
   they_will_take_part then

   "[Euskaltel] will participate enthusiastically in the Classique des Alpes and then they will take part in Dauphine Libere."

6. **Combination of errors (CN):** Two or more errors explained above can be identified in the mention proposed by the system. The following examples illustrate two cases of combined errors.

   - **Divided Mention + Missing Text**

     × [Siriako Atzerri ministro Farouk] Al [Xaraak] ukatu egin du Israelek Xebaako landetxeak okupatzea onartzen duela.

     √ [Siriako Atzerri ministro Farouk Al Xaraak] ukatu egin du Israelek Xebaako landetxeak okupatzea onartzen duela.

Siriako Atzerri ministro Farouk Al Xaraak
The_Syrian Foreign Minister Farouk al Sharaa
ukatu_egin_du Israelek Xebaako landetxeak okupatzea
has_denied Israel of_Shebaa the_farms to_occupy
onartzen duela
accepts that

"[The Syrian Foreign Minister Farouk al Sharaa] has denied that Israel accepts to occupy the farms of Shebaa."

- **Conflated Mentions + Extra text**

  × [Clintonen irudiko, Alderdi Errepublikarraren egitasmoa] ez da egokia europarren interesetarako.

  √ [Clintonen] irudiko, [Alderdi Errepublikarraren egitasmoa] ez da egokia europarren interesetarako.

  Clintonen irudiko Alderdi_Errepublikarraren egitasmoa
  Clinton believes of_Republican_Party the_project
  ez_da egokia europarren interesetarako
  is_not suitable of_Europeans for_the_interest

  "Clinton believes the project of the Republican Party is not suitable for the interest of Europeans."

Tables 3.12 summarises the examples explained above.

Error Type	System	Gold
Missing Mention	-	-jokalaria
Extra Mention	-Bestetik	-
Missing Text	-Jiri Vanek	-Jiri Vanek txekiarrak
Extra Text	-Luis Uranga harrituta	-Luis Uranga
Head not Identified	-gehiago	-indar gehiago
Divided Mention	-Londres	-Londres mendebaldean
	-mendebaldean	-
Conflated Mentions	-Euskaltel Alpeetako Klasikoan	-Euskaltel
	-	-Alpeetako Klasikoan
Combination of Errors	-Siriako Atzerri ministro Farouk -Xaraak	-Siriako Atzerri ministro Farouk Al Xaraak

3.12 Table – Examples of error types.

Table 3.13 shows the number of each error type when automatic mention detection is carried out and the percentages. The percentages are calculated taking

into account the total number of correct mentions (1,904) in the subcorpus used for the evaluation.

Error Type	#	%
Missing Mention	219	11.50
Extra Mention	142	7.46
Missing Text	103	5.41
Extra Text	65	3.41
Head not Identified	5	0.26
Divided Mention	25	1.31
Conflated Mentions	32	1.68
Combination of Errors	20	1.05
Total	611	32.10

3.13 Table – Error Types Automatic Mentions.

As can be observed, the majority of errors are Missing Mentions (11.50%) and Extra Mentions (7.46%). The last row shows that the errors are 32.10% of the total correct mentions.

Each error type differently affects the precision and recall values. The differences also depend on the scoring protocol that has been used for scoring. For instance, the conflated mentions error type provokes one precision error and k (the number of mentions in which they have been conflated) recall errors when Strict Matching protocol is used. To better understand this, we can observe the following example, which has previously been presented.

> × Frantziako ibilbide gogorretan gogotsu ariko da [Euskaltel Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.
>
> √ Frantziako ibilbide gogorretan gogotsu ariko da [Euskaltel] [Alpeetako Klasikoan] eta Dauphine Liberen ariko dira hurrena.

In this example, the system conflates two mentions in one ([Euskaltel Alpeetako Klasikoan]), so k= 2. This mention is considered incorrect when Strict Matching is applied as it does not exactly match with any mention in the gold, and also when Lenient Matching is applied. Although it has the head of the correct mention inside its span, it exceeds the span of the gold mention. Consequently, it is considered a precision error. Regarding recall, as k is 2 in this case, two recall errors are counted because [Euskaltel] and [Alpeetako Klasikoan] gold mentions are not present in the system output.

Table 3.14 summarises how error types affect precision and recall when different scoring protocols are used. In the case of combination of errors, the effect in precision and recall depends on the nature of the combined errors. The same happens when the scoring protocol is Lenient Matching.

	Strict Matching		Lenient Matching	
Error Type	P	R	P	R
Missing Mention	0	1	0	1
Extra Mention	1	0	1	0
Missing Text	1	1	0	0
Extra Text	1	1	1	1
Head not Identified	1	1	1	1
Divided Mention	k	1	k-1	0
Conflated Mentions	1	k	1	k

3.14 Table – Effects of error types in P and R.

3.6.2 Error Causes

To help identify the error causes, we carried out a qualitative evaluation on a corpus of 1,070 mentions, a subcorpus of the 1,904 mentions that were used to analyse the error types.

We divided the errors encountered in nine categories. For each category we present some examples. The first item in each example is the incorrect case (×) and the second the respective correct one (√).

1. **Erroneous morphosyntactic analysis (EMA):** POS tags and/or syntactic functions are incorrectly disambiguated by the lemmatiser. The same string of letters can be a noun in some contexts and an adjective in others; this error occurs when a non-noun instance is tagged as a noun and, consequently, incorrectly regarded as a chunk. In addition, some complex postpositional are not identified correctly. In the following example *arte* is tagged as a noun (*arte*, n.: art) and, consequently, considered a nominal chunk, although in this context it is a part of a complex postposition that has not been identified, meaning "until" which is a preposition in English.

   × Futbolean azken minutura [arte] edozer gauza gerta daiteke.

   √ Futbolean azken minutura arte edozer gauza gerta daiteke.

Futbolean azken minutura arte edozer_gauza gerta_daiteke
In_football last the_minute until anything can_happen

"In football, anything can happen until the last minute."

2. **Incorrect chunk tag (ICT):** the chunker does not provide correct chunk tags.

   × Estatu Batuetako presidente Bill [Clintonek] antimisilen inguruan Pentagonoa atontzen ari den sistema babesteko eskatu zion atzo Europar Batasunari Lisboan.

   √ [Estatu Batuetako presidente Bill Clintonek] antimisilen inguruan Pentagonoa atontzen ari den sistema babesteko eskatu zion atzo Europar Batasunari Lisboan.

   Estatu Batuetako presidente Bill Clintonek antimisilen_inguruan
   States United president Bill Clinton the_anti-missile
   Pentagonoa atontzen_ari_den sistema babesteko eskatu_zion
   the_Pentagon that_is_preparing system to_support asked
   atzo Europar Batasunari Lisboan
   Yesterday European the_Community in_Lisbon

   "Yesterday in Lisbon, [United States president Bill Clinton] asked the European Community to support the anti-missile system that the Pentagon is preparing."

3. **Missing chunk tag (MCT):** the chunker does not tag a chunk, or it only tags the beginning or the end of the chunk. Not opened or not closed tags can occur when Constraint Grammar (CG) formalism (Karlsson *et al.*, 1995) is used.

   × [Luis Uranga harrituta azaldu da Portugalgo klubaren jokaerarekin.

   √ [Luis Uranga] harrituta azaldu da Portugalgo klubaren jokaerarekin.

   Luis Uranga harrituta_azaldu_da Portugalgo klubaren
   Luis Uranga is_susprised Portuguese of_the_club
   jokaerarekin
   at_the_attitude

   "[Luis Uranga] is surprised at the attitude of the Portuguese club."

The error could be partially solved in the case when the chunker missed the ending tag of a chunk. In this case we applied a heuristic that is explained in Section 3.6.4.

4. **Incorrect clause tag (ICLT):** the left clause boundary proposed by the clause tagger is not the same as the gold mention left boundary, so the mention is incorrect.

   × Behin betiko hitzarmen bat sinatu behar dute bi aldeek, [horrela Oslon hasitako prozesua] bukatzeko.

   √ Behin betiko hitzarmen bat sinatu behar dute bi aldeek, horrela [Oslon hasitako prozesua] bukatzeko.

   Behin_betiko hitzarmen bat sinatu behar_dute bi aldeek
   permanent agreement a sign are_going_to two the_parts
   horrela Oslon hasitako prozesua
   so_that in_Oslo that_began the_process
   bukatzeko
   will_be_brought_to_a_close

   "The two parts are going to sign a permanent agreement so that [the process that began in Oslo] will be brought to a close."

5. **Predication (PRE):** Two words are joined in one chunk even though there is a predication relation between them. In coreference resolution it is common to join two mentions in one cluster when they have a predicational relation. For this reason it is necessary to split the chunk into two parts when there is a predication. However, in some cases it is difficult to predict the predication relation and the chunk is not split, causing a mention detection error.

   × [Bush buru] duen Alderdi Errepublikarraren proiektua okerragoa da.

   √ [Bush] [buru] duen Alderdi Errepublikarraren proiektua okerragoa da.

   Bush buru duen Alderdi Errepublikarraren proiektua
   Bush leader whose_is Party of_the_Republican the_project
   okerragoa da
   worse is

   "The project of the Republican party whose [leader] is [Bush] is worse".

The opposite case also occurs, i.e., a chunk is split when it should not be, causing a mention detection error.

× [Madariagaren] [ametsa] da aspalditik.

√ [Madariagaren ametsa] da aspalditik.

Madariagaren ametsa da aspalditik
Madariaga's dream it_has_been long

"It has long been [Madariaga's dream]."

The error is caused by a rule that does not divide a mention when it is presented in a predication relation or divides it when it should not. The rule can be adapted to better identify whether the mentions should be divided.

6. **Coordination (COR):** In coordination cases the entire coordinate structure is considered a mention. However, in some cases the submentions to the left and right of the coordinating conjunction do not need to be marked. The detector should not split the coordinate structure when the submentions do not have to be included, otherwise we get incorrect mentions in the output.

× [[Lintxamendu] eta [jazarpenerako erabili nahi den aitzakia]] da.

√ [Lintxamendu eta jazarpenerako erabili nahi den aitzakia] da.

Lintxamendu eta jazarpenerako erabili_nahi_den aitzakia da
lynching and for_persecution to_use an_excuse is

"It is [an excuse to use for lynching and persecution]."

The same happens when the whole coordinate structure is not joined.

× [Bost urteko kontratua] eta [sekulako dirutza] lortu zituen.

√ [[Bost urteko kontratua] eta [sekulako dirutza]] lortu zituen.

Bost_urteko kontratua eta sekulako dirutza lortu_zituen
a_five-year contract and lots_of money He/she_obtained

"He/she obtained [[a five-year contract] and [lots of money]]."

In addition, sometimes coordinate structures are enumerations. In enumeration cases, apart from the entire coordinate mention, submentions divided by the comma must also be obtained as mentions. When these enumerations are not divided correctly mentions are missed.

> × [Van der Linden, Malafosse, Carboneau, Carrat, Venditti, Arbizu, Lamaison eta Smith] dira jokalariak.
>
> √ [[Van der Linden], [Malafosse], [Carboneau], [Carrat], [Venditti], [Arbizu], [Lamaison] eta [Smith]] dira jokalariak.
>
> *Van der Linden Malafosse Carboneau Carrat Venditti Arbizu*
> Van der Lindern Malafosse Carboneau Carrat Venditti Arbizu
> *Lamaison eta Smith dira jokalariak*
> Lamaison and Smith are the_players
>
> "The players are [[Van der Linden], [Malafosse], [Carboneau], [Carrat], [Venditti], [Arbizu], [Lamaison] and [Smith]]."

Adapting the rules that decide whether a coordinate structure needs to be divided into submentions and joined into bigger structures would necessarily improve mention detection scores. In addition, specific rules to treat enumerations would also improve mention detection.

7. **Comparatives (COMP):** In comparatives structures the related two elements are not joined and, consequently, the adverb (baino "than") is considered out of the mention.

> × [Bilerak] baino [oihartzun handiagoa] lortu zuen atentatuak.
>
> √[Bilerak baino oihartzun handiagoa] lortu zuen atentatuak.
>
> *Bilerak baino oihartzun handiagoa lortu_ zuen atentatuak.*
> the_meeting than coverage greater received the_attack
>
> "The attack received [greater coverage than the meeting]."

A specific rule in the mention detector to treat comparative structures would solve the problem.

8. **Referentiality (RF):** Some referential chunks are discarded as they do not have nouns inside them.

× Italiarrak onartezintzat hartu du erabakia.

√ [Italiarrak] onartezintzat hartu du erabakia.

Italiarrak *onartezintzat hartu_du erabakia*
The_Italian unacceptable considers the_decision

"[The Italian] considers the decision unacceptable."

On the contrary, non referential chunks that have nouns inside are considered as mentions, for example, linking words.

× [Era berean], lurralde suntsitua berreraikitzeko nazioarteko laguntza eskatu zuen.

√ Era berean, lurralde suntsitua berreraikitzeko nazioarteko laguntza eskatu zuen.

Era_berean *lurralde* *suntsitua* *berreraikitzeko*
In_the_same_way territory the_destroyed to_rebuild
nazioarteko *laguntza* *eskatu_zuen*
for_international help he/she_called

"In the same way, he/she called for international help to rebuild the destroyed territory."

9. **Miscellaneous (MISC):** errors with varied causes that cannot be classified into the categories discussed earlier. Some examples include:

   - **Parenthetical:** the chunks are divided by a parenthetical structure.

     × [David Trimble] ([UUP]) [lehen ministroak] adierazpenak egin zituen.

     √ [David Trimble ([UUP]) lehen ministroak] adierazpenak egin zituen.

     David Trimble (UUP) lehen *ministroak adierazpenak*
     David Trimble (UUP) Prime Minister statements
     egin_zituen
     made

     "[Prime Minister David Trimble ([UUP])] made statements."

- **Typing error:** Some texts have typing errors, missing spaces between words, unreadable characters... When processing these texts with automatic tools the typing errors are not always solved and cause a mention detection error.

  × [1995eanhartu] zuen parte.

  √ [1995ean] hartu zuen parte.

  [1995ean] hartu_zuen_parte
  in_1995 he/she_participated

  "He/she participated in [1995]."

  Solving these problems is complicated, and they would be better corrected in the preprocessing step.

- **Extra Words in Boundaries:** the chunks contain words that fall outside the mention boundaries.

  × Kolpistek elkarrizketak hautsi zituzten [erabaki horren ondorioz].

  √ Kolpistek elkarrizketak hautsi zituzten [erabaki horren] ondorioz.

  Kolpistek *elkarrizketak* *hautsi_zituzten*
  The_leaders_of_the_coup the_talks broke_off
  erabaki *horren_ondorioz*
  decision because_of_that

  "The leaders of the coup broke off the talks because of that decision."

  The error of mentions with extra words can be solved by defining a set of words that appear at the boundaries of mentions even though they are not part of the mention. This way such words can be removed from within the mention boundaries.

- **Unjoined chunks:** Even though the chunks are correct they do not match the mentions exactly. In most cases, the chunks should be joined to create a correct mention.

  × Gentzelek [atzo] [goizean] hitz egin zuen.

  √ Gentzelek [atzo goizean] hitz egin zuen.

  Gentzelek *atzo* *goizean* *hitz_egin_zuen*
  Gentzel yesterday morning talked

"Gentzel talked [yesterday morning]."

The error of unjoined chunks that should form a mention can be solved by identifying the exact cases in which this phenomenon occurs.

Although specific rules to treat miscellaneous cases would likely improve mention detection, the phenomenon is a broad one and the improvement may not be substantial.

We observed that the errors explained above can be classified in three main categories, some are caused by preprocessing tools (EMA, ICT, MCT), some others are because of using the output of general-purpose tools that must be adapted to our mentions specification (ICLT, PRE, COR, CONP), while the rest are caused by inaccurate mention detection rules (RF). The MISC errors fall in all the categories.

Table 3.15 shows the percentages of the main error causes resulting from the mention detection step.

	Error Cause	#	%
1	EMA	47	14.11
2	ICT	93	27.93
3	MCT	63	18.92
4	ICLT	25	7.51
5	PRE	5	1.50
6	COR	35	10.51
7	COMP	2	0.60
8	RF	32	9.61
9	MISC	31	9.31
	Total	333	100.00

3.15 Table – Main causes of errors.

As Table 3.15 shows, the automatic preprocessing tools are the main source of errors: 27.93% are caused by the chunker providing incorrect chunk tags, and 18.92% by it missing the chunk tag. The morphosyntactic analysis that the mention detector receives as input also causes a number of errors in mention detection, accounting for 14.11% of the errors. Another considerable cause of error is that the left clause boundaries identified by the clause tagger do not match the left boundaries of gold mentions; this accounts for 7.51% of the errors.

Given the above, we can conclude that the input that our mention detector receives has a substantial impact on mention detection errors. This is also pointed

out by Uryupina and Moschitti (2013), who suggest that better performance can be obtained if a robust preprocessing is achieved. The performance of the mention detector is highly dependent on the preprocessing tools, and errors caused by these tools are difficult to tackle. We can, however, attempt to address other types of errors.

We found that the greatest improvements can be achieved by treating the errors classified as COR and RF, which account for 10.51% and 9.61% of errors, respectively.

3.6.3 Distribution of errors among mention types

Apart from identifying the main causes of errors, observing how these errors are distributed among mention types is essential to implementing improvements effectively. Figure 3.4 shows this distribution in terms of percentages of the total number of errors.

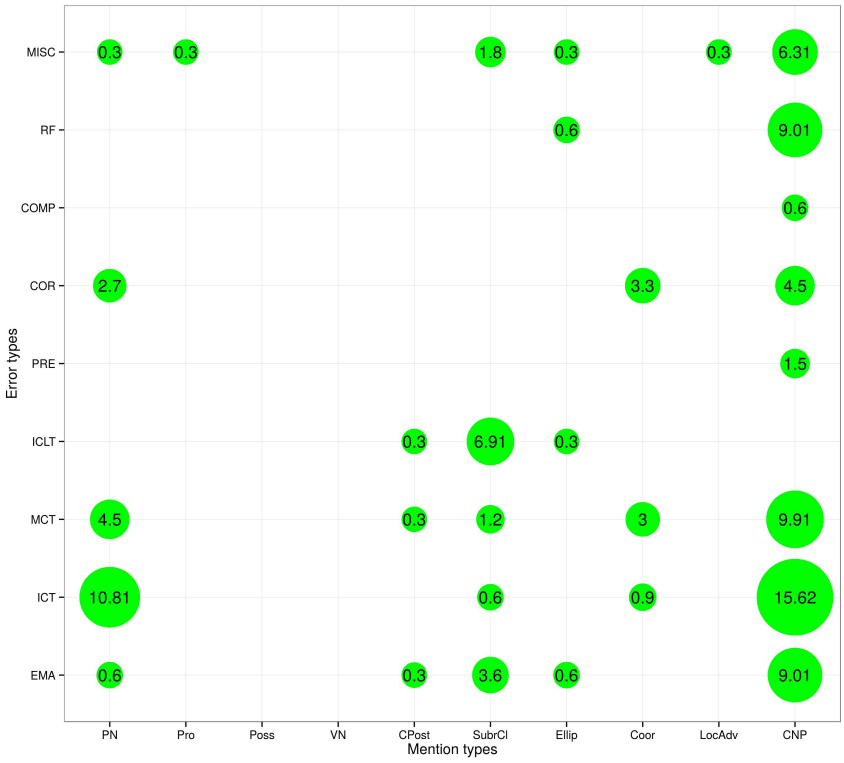
As Figure 3.4 shows, Incorrect Chunk Tag (ICT) error cause is mainly divided into two mention types, proper nouns (10.81%) and common noun phrases (15.62%). A similar pattern can be found in the Missing Chunk Tag (MCT) errors. They cause 18.92% of the total errors and mainly affect common noun phrases (9.91%) and proper nouns (4.50%). Incorrect Clause Tag (ICLT) errors, in turn, are mainly errors in NPs containing subordinate clauses (6.91%). Other considerable errors are errors caused by Erroneous Morphosyntactic Analysis (EMA) (9.01%), Referentiality (RF) (9.01%) and Miscellaneous (MISC) (6.31%) in Common Noun Phrases.

In the next Section, we explain how we improved our mention detection to overcome the problems identified by the error analysis.

3.6.4 Improvements in mention detection

To attempt to resolve some of the errors presented in Table 3.15, we implemented the following five adjustments and applied them in the mention detection process:

1. To solve the errors caused by the chunker only tagging the beginning of the chunk and not the end, we applied a heuristic that closes the nominal chunk when a verbal chunk starts. In other words, if a nominal chunk is not closed when a verbal one starts, we close it just before the verbal chunk. With this improvement we tried to solve some of the *MCT* errors (3nd row in Table 3.15). In example (a), the nominal chunk *borroka armatua* "armed struggle" is opened properly but the close tag is missed. Since a verbal chunk *itzultzeko* "to return" starts after the word *armatua* "armed", we close the nominal chunk, obtaining a correct nominal chunk as in example (b).



3.4 Figure – Distribution of error causes into mention types. Automatic processing.

(a) IRA [borroka armatura itzultzeko asmotan da.

IRA borroka armatura itzultzeko asmotan_da
The_IRA struggle to_armed to_return intends

"The IRA intends to return to [armed struggle."

(b) IRA [borroka armatura] itzultzeko asmotan da.

IRA borroka armatura itzultzeko asmotan_da
The_IRA struggle to_armed to_return intends

"The IRA intends to return to [armed struggle]."

2. We observed in the error analysis that some mentions in a predication relation are divided when they should not be, or not divided when they should be. We adjusted the rule in the FSTs that was responsible for deciding whether a mention should be divided or not, depending on whether it was in a predication relation. With this improvement, we aimed to solve *PRE* errors (5th row in Table 3.15). In the following examples, (a) is a case when a mention should be divided and (b) an example where the mention should not be divided.

   (a) [Ibarretxe] [lehendakaria] da.

   Ibarretxe lehendakaria da
   Ibarretxe the_president is

   "[Ibarretxe] is the [president]."

   (b) [Bushen proposamena] da.

   Bushen proposamen da
   Bush's proposal it_is

   "It is [Bush's proposal]."

3. We observed that chunks composed only of determiners were considered mentions. However, not all of these chunks are mentions. For instance, examples such as *biok* "we both" in (a) should be considered mentions, but examples such as *bestetik* "on the other hand" (b) should be discarded. We defined a new rule that discards chunks of undefined determiners because they are rarely referential. With this improvement, we aimed to detect some mentions that were missed and quantified as *RF* error (8th row in Table 3.15).

   (a) [Biok] egin genuen lan hura.

   Biok egin_ genuen lan_ hura
   We_both did the_work

   "[We both] did the work."

   (b) Bestetik, orain arte jasotako elkartasuna eskertu zuen.

Bestetik *orain_ arte* *jasotako*
On_the_other_hand so_far he_had_received
elkartasuna *eskertu_ zuen*
for_the_support he_expressed_his_thanks

"On the other hand, he expressed his thanks for the support he had received so far."

4. Chunks composed of only adjectives were all discarded. Upon closer examination, however, we observed that some of them could in fact be referential, as clarified by the examples below. In example (a), *handia* "big" is not referential so it must not be considered as a mention, but in example (b) the same word, *handia*, is referential and so is a mention. Therefore, we adjusted the rule that discarded chunks composed of nothing but adjectives so that it only discarded undefined adjectives and adjectives that were in a predication relation, as in example (a). With this improvement, our aim was to detect mentions that were missed and quantified as *RF* error (8th row in Table 3.15).

   (a) Etxea oso handia da.

   Etxea *oso* *handia* *da*
   The_house very big is

   " The house is very big."

   (b) [Handia] garestiagoa da.

   Handia *garestiagoa* *da*
   The_big_one more_expensive is

   "[The big one] is more expensive."

5. Whether chunks to the left and the right of the coordinating conjunction are considered a single chunk or separate chunks is determined by the nature of the chunk, in particular whether the element(s) in the chunk are declined. In example (a), both Pello and Ane are declined (-ri "to") and are therefore considered two separate chunks, but in (b) Pello is not declined (-Ø) and therefore the elements are considered a single chunk. We defined new rules that identify coordinate structures that are not joined but that should be considered mentions, e.g. (a). With this improvement, our goal was to detect coordinate structures that were missed, thereby decreasing *COR* error (6th row in Table 3.15).

(a) [Pellori] eta [Aneri] oparia eman zioten.

Pellori eta Aneri oparia eman_zioten
to_Pello and to_Ane the_present they_gave

"They gave the present to [Pello] and [Ane]."

(b) [Pello eta Aneri] oparia eman zioten.

Pello eta Aneri oparia eman_zioten
Pello and to_Ane the_present they_gave

"They gave the present to [Pello] and [Ane]."

After implementing these five improvements, we obtained a gain of 1.21% in the F-measure using Strict Matching and of 0.89% with Lenient Matching. These gains are significant.[3] The results can be seen in Table 3.16, and show that the improvements have succeeded in resolving some errors.

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
BMD	73.86	72.87	73.36	78.69	80.71	79.68
BIMD	74.67	74.47	74.57	79.26	81.92	80.57

3.16 Table – Improved scores in mention detection. BMD=Basque Mention Detector, BIMD= Basque Improved Mention Detector.

Table 3.17 shows the main error types with improved mention detection (BIMD) and without it (BMD). In absolute values, 61 errors have been resolved, 611 without improvements and 548 after applying them. The biggest improvement has been achieved in Missing Mention error type, 44 mentions that before were not identified are now in the system output and are correct mentions. In percentage values, an improvement of 3.3% is obtained, 28.80% being errors of total mentions.

3.7 Experiment with gold input

Given that the input received by the mention detector is a significant source of errors, we decided to experiment with what would happen with a perfect input.

[3]Statistical significance is tested with Paired Student's t-test. p-value < 0.01

	Automatic BMD		Automatic BIMD	
Error Type	#	%	#	%
Missing Mention	219	11.50	175	9.19
Extra Mention	142	7.46	138	7.25
Missing Text	103	5.41	102	5.36
Extra Text	65	3.41	66	3.47
Head not Identified	5	0.26	3	0.16
Divided Mention	25	1.31	20	1.05
Conflated Mentions	32	1.68	30	1.58
Combination of Errors	20	1.05	14	0.74
Total	611	32.10	548	28.80

3.17 Table – Comparison of main error types when improvements in mention detection are applied.

In "real-world" cases, the mention detector is integrated into a pipeline in which errors made in each step affect the following phases. Providing a partially gold input to our mention detector enables us to evaluate the accuracy of our system in an isolated way. Due to the lack of a corpus where all linguistic levels are manually tagged, we experimented with gold lemmas, gold PoS tags and gold chunks. The syntactic function, the named entity tags and clause tags are automatic, so the input is only partially gold. Table 3.18 shows the results obtained using partially gold information as input.

	Strict Matching			Lenient Matching		
	P	R	F_1	P	R	F_1
$BIMD_a$	74.67	74.47	74.57	79.26	81.92	80.57
$BIMD_g$	84.95	86.84	85.89	87.06	91.15	89.06

3.18 Table – Mention detection scores with partially gold annotation.

Using gold lemmas, PoS tags, and chunks, the scores clearly increase—in fact, they are highly promising. The F-measure value is 85.89% with Strict Matching and 89.06% with Lenient Matching. Comparing this to the results obtained by applying mention detection to automatic annotation, as shown in the first row, the results increase by 11.42% percentage points using Strict Matching and by 8.40% with Lenient Matching.

In view of these results, it is clear that the type of input provided is crucial. As Table 3.19 shows in last row, when partially gold input is used instead of automatic

preprocessed input, the error types drop from 548 (Automatic BIMD) to 285 (Gold BIMD). This is a significant drop and clearly shows that the output of preprocessing tools has a substantial effect on mention detection. In percentages, the difference is 13.80%, so about 14% of errors are produced by preprocessing tools.

	Automatic BMD		Automatic BIMD		Gold BIMD	
Error Type	#	%	#	%	#	%
Missing Mention	219	11.50	175	9.19	88	4.62
Extra Mention	142	7.46	138	7.25	87	4.57
Missing Text	103	5.41	102	5.36	33	1.73
Extra text	65	3.41	66	3.47	57	3.00
Head not Identified	5	0.26	3	0.16	0	0.00
Divided Mention	25	1.31	20	1.05	14	0.74
Conflated Mentions	32	1.68	30	1.58	2	0.11
Combination of Errors	20	1.05	14	0.74	4	0.21
Total	611	32.10	548	28.80	285	15.00

3.19 Table – Comparison of main error types when partially gold input is provided.

3.8 Effects of Mention Detection results in Coreference resolution

In order to evaluate how the improvements we obtained in mention detection affect coreference resolution, we used BART, a highly modular toolkit for developing coreference applications (Versley *et al.*, 2008b).

In our evaluation experiments, a simple coreference model has been used, namely, the mention-pair approach presented in Soon *et al.* (2001). The metrics used to evaluate the system performance are MUC (Vilain *et al.*, 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), $CEAF_m$ (Luo, 2005) and BLANC (Recasens and Hovy, 2011). The CoNLL metric is the arithmetic mean of MUC, B^3 and $CEAF_e$ metrics. The scores have been calculated using the reference implementation of the CoNLL scorer (Pradhan *et al.*, 2014).

Table 3.20 shows the results obtained in coreference resolution when different mention detection systems are used, Baseline (B), Basque Mention Detector (BMD) and Basque Improved Mention Detector (BIMD).

	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
B	25.17	54.39	51.62	51.92	28.13	43.83
BMD	28.36	61.62	57.96	58.91	35.56	49.63
BIMD	29.28	62.65	58.69	59.65	36.77	50.53

3.20 Table – F1 scores of Coreference resolution with different metrics

Basque Mention Detector (BMD) outperfoms the Baseline (B) scores in all the metrics. In CoNLL metric, Basque Mention Detector obtains a score of 49.63, which is 5.8 points higher than Baseline, which scores 43.83. The same happens when we compare Basque Mention Detector results with those obtained by Basque Improved Mention Detector (BIMD). BIMD outperforms in all the metrics and CoNLL score is outperformed by 0.9. All the improvements obtained when different mention detectors are used are significant.[4]

3.9 Conclusions

We have presented a mention detector system for Basque. The system has been created based on a linguistic analysis of mentions in Basque. We explain these structures and analyse their presence in the EPEC corpus; we also compare them with other languages.

Good results at the mention detection stage have a major impact on the success of coreference resolution systems, and therefore improving mention detection is crucial. We applied several different strategies in an attempt to obtain more accurate mentions.

First, we concluded that using only generic preprocessing tools for mention detection does not produce good results (F-measure of 64.28% using Strict Matching and 71.12% using Lenient Matching). As a consequence, we created several Finite State Transducers (FSTs) that make use of generic tools. These FSTs obtain better results: F-measure of 73.36% using Strict Matching and 79.68% when Lenient Matching is used. In other words, this approach outperforms the baseline by 9 points for Strict Matching and by 8 points for Lenient Matching.

Secondly, we carried out a deep error analysis to better understand and thereby perhaps mitigate the main causes of errors in mention detection. This error analysis enabled us to define five improvements, which increased the F-measure by 1.21 points using Strict Matching (74.57%), and by 0.89 points using Lenient Matching (80.57%).

[4]Statistical significance is tested with Paired Student's t-test. p-value < 0.01

We also considered it important to quantify the impact that automatic preprocessing tools have on mention detection. Thus, we carried out an experiment with partially gold input. The results revealed that errors decrease substantially when gold input is provided: the better the preprocessing tools, the better the mention detection scores. The F-measure value obtained with gold input ($BIMD_g$) is 85.89% with Strict Matching and 89.06% with Lenient Matching—a difference of 11.42 points for Strict Matching and of 8.48 points for Lenient Matching in comparison with results obtained when automatic input is used ($BIMD_a$).

We integrated the mention detector system into an end-to-end coreference resolution system for written Basque and analysed the influence that mention detection has in coreference resolution results.

KORREFERENTZIA-EBAZPENA

4

Adaptation of the Stanford Deterministic Coreference Resolution System to a morphologically rich language

4.0 Laburpena

Artikulu honetan Stanfordeko korreferentzia-ebazpenerako sistema (Lee *et al.*, 2013) euskararako egokitzeko egindako prozesua azaltzen da.

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 da. Gainera, hizkuntzarekiko guztiz independentea den sistema bat garatzea lan konplexua izan daiteke eta hizkuntzaren ezaugarriak kontuak hartzeak onurak ekartzen ditu. Egoera honetan, soluzio posible bat arloaren egoerako sistema bat hartu eta tratatu nahi den hizkuntzara egokitzea da. Gure kasuan *CoNLL-2011 shared task* atazan (Pradhan *et al.*, 2011) emaitzarik onenak lortu zituen Stanfordeko korreferentzia-ebazpenerako sistema egokitu dugu. Sistema hori erregelatan oinarritutakoa da, eta oinarrian 10 *bahe* edo korreferentzia-ebazpenerako modulu espezifiko ditu. Doitasun handiena lortzen duten baheetatik hasiz eta doitasun baxuagoa dutenekin amaituz, banan banan aplikatzen dira sistemaren 10 baheak. Filosofia honi esker, lehenengo baheetan erabaki ziurrak hartzen dira (doitasun handia) eta ondorengoetan ez hain ziurrak, batzutan doitasuna kaltetuz baina estaldura hobetuz. Sistemaren arkitektura guztiz modularra izanik erraz integra daitezke korreferentzia-ebazpenerako bahe berriak, hori dela eta, garatua izan

ez den beste hizkuntza baterako egokitzapena errazten da.

Dakigunez ingelesa eta euskara ezaugarri linguistiko desberdineko hizkuntzak dira. Adibidez, euskara eranskaria, buru-azkena, ordena librekoa eta *pro-drop* hizkuntza da, ingelesa aldiz ez. Ezaugarri horietaz gain, euskararen sistema nominalak ez du generorik eta izenordainek ez dute bizidun/ezbizidun propietaterik. Hori dela eta lehenik eta behin euskararen ezaugarriek korreferentzia-ebazpenean eduki dezaketen eragina aztertzen da, sistemaren egokitzapenean kontuan hartu beharrekoak, hain zuzen.

Kontuan hartu behar diren ezaugarriak finkatuta, Stanfordeko unibertsitatean garatutako sistema aurkezten dugu bahez-bahe. Bahe bakoitza arretaz aztertzen dugu, zuzenean bahe hori aplikatzeak euskarazko corpusean sortzen dituen arazoak azalduz eta arazo horiei aurre egiteko soluzioak planteatuz. Kasu gehienetan jatorrizko bahea egokitzea nahikoa izan zaigu, beste batzuetan, aldiz, jatorrizko bahea kendu eta berri batez ordezkatu beharrean aurkitu gara, adibidez, morfologiaren erabilera zorrotzagoa egin behar izan den kasuetan. Euskaraz gertatzen den elipsiaren fenomenoa ingelesez gertatzen ez denez, aipamen eliptikoak tratatzeko bahe espezializatu berri bat ere txertatu dugu sisteman, jatorrizko sistemako 10 baheak 11 izatera pasatuz. 4.1 irudian ikus daitezke egokitutako sistemak dituen baheak.

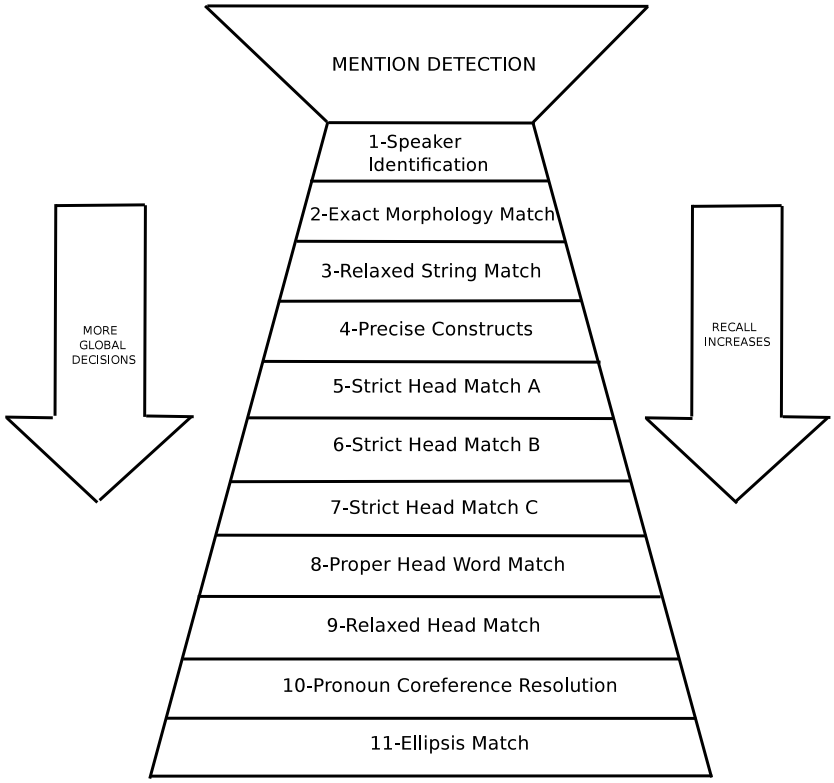
Egokitzapen-prozesuan erabilitako metodologia antzeko hizkuntzetarako baliagarria izan daiteke, adibidez, morfologia aberatsa duten hizkuntzetarako edota ordena libreko hizkuntzetarako.

Euskarako egokitutako sistemak IXA taldean garatutako analisi-katearekin prozesatutako testuak eta euskararako sortutako aipamen-detektzailearen irteera jasotzen ditu sarrera moduan. Horiek erabiliz, gai da euskarazko testuetako korreferentzia-erlazioak identifikatzeko.

Egokitutako sistema eta oinarri-lerrotzat hartu dugun jatorrizko Stanfordeko sistema ebaluatzen dira esperimentazio fasean. Ebaluaziorako EPEC corpusa erabili da, guztira 4.360 aipamenez osatutako zatia hain zuzen ere. Bi sistemek aipamen automatikoekin lortutako emaitzak 4.1 taulan ikus daitezke eta urrezko aipamenak erabiltzean lortzen dituztenak berriz 4.2 taulan.

Emaitzak konparatuz ikusi da egokitutako sistemak 7,07 puntuko hobekuntza lortzen duela CoNLL ebaluzio-metrikan aipamen-detekzioa automatikoki egiten den kasuan, eta 11,5eko hobekuntza urrezko aipamenak erabiltzen direnean. Hobekuntza horiek korreferentzia-ebazpenean hizkuntzaren ezaugarriak kontuan hartzeak berebiziko eragina duela erakusten dute.

Hizkuntzaren ezaugarriez gain, aurreprozesaketarako tresnek korreferentzia-ebazpenean duten eragina ere aztertu dugu. Horretarako egokitutako



4.1 irudia – Euskara egokitutako Stanfordeko sistemaren arkitektura.

sistemak lortzen dituen emaitzak konparatu dira automatikoki detektatutako aipamenak erabiltzean eta urrezko aipamenak pasatzen zaizkionean. 4.1 eta 4.2 tauletako bigarren errenkadak aztertzen baditugu, urrezko aipamenak erabiltzean CoNLL metrika 20,38 puntu hobetzen dela ikus dezakegu, 55,74 izatetik 76,12 izatera pasatzen baita. Diferentzia nabarmena da eta garbi ikusten da aipamen-detekzioa zuzen egiteak berezibiziko garrantzia duela korreferentzia-ebazpenean.

Ebaluazioa amaitzeko, bahe bakoitzak korreferentzia-ebazpenerako sisteman duen eragina aztertu da. Baheak modu inkrementalean gehitu dira eta bahe bakoitzaren ekarpena kuantifikatu da. Ekarpenik handiena jatorrizko sistemako *Exact String Match* bahea ordezkatzeko erabili den, eta jatorrizkoak baino morfologiaren erabilera sakona egiten duen, *Exact Morphology Match* izeneko baheak egiten duela ikusi da.

Egokitzapen-prozesuan jatorrizko Stanfordeko sistemaren baheen ordena-

Aipamen automatikoak							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
Oinarri-lerroa	74,57	27,46	59,96	56,86	58,61	36,75	48,67
Egokitua	74,57	42,32	62,94	61,54	61,98	43,18	55,74

4.1 taula – Oinarri-lerroko eta egokitutako sistemen emaitzak aipamen automatikoekin.

Urrezko aipamenak							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
Oinarri-lerroa	100	36,55	81,28	72,13	76,05	62,94	64,62
Egokitua	100	58,12	86,99	80,57	83,27	72,77	76,12

4.2 taula – Oinarri-lerroko eta egokitutako sistemen emaitzak urrezko aipamenekin.

keta errespetatu da, hala ere, litekeena da euskararako baheen beste aplikazio-orden bat egokiagoa izatea. Hori dela eta, euskararako baheen aplikazio-orden optimoena zein den ere lortu nahi izan da esperimentalki. Lortutako baheen orden berria jatorrizkoarekiko desberdina bada ere CoNLL neurrian ez da inolako hobekuntzarik lortzen.

Ebaluazio fasea amaituta, gure sistemak lortzen dituen emaitzak antzeko ezaugarriak dituzten beste hizkuntzetako sistemekin konparatu dira, hala nola, arabiera, alemana eta polonierarako sistemekin. Guk lortutako emaitzak, 4.3 taulan ikus daitekeen moduan, besteek lortzen dituztenen parekoak direla esan daiteke.

Sistema	**Hizkuntza**	**CoNLL F1**
Lan hau	Euskara	55,74
(Fernandes *et al.*, 2012)	Arabiera	45,2
(Chen and Ng, 2012)	Arabiera	32,4
(Kobdani and Schütze, 2010)	Alemana	55,03
(Zhekova and Kübler, 2010)	Alemana	33,93
(Ogrodniczuk and Kopeć, 2011)	Poloniera	61,95

4.3 taula – Hizkuntza desberdinetarako korreferentzia-ebazpenerako sistemek lortzen dituzten emaitzak.

Laburbilduz, Stanfordeko korreferentzia-ebazpenerako sistema euskararako egokitu da eta testuen prozesamendu automatikorako IXA taldeko analisikatean integratu. Egokitzapen prozesuan euskararen ezaugarriak kontuan hartu ditugu eta metodologia bahez-bahe azaldu da beste hizkuntza batera egokitzeko prozesuaren erreplikagarritasuna ziurtatuz.

Adaptation of the Stanford Deterministic Coreference Resolution System to a morphologically rich language

Ander Soraluze, Olatz Arregi, Xabier Arregi and Arantza Díaz de Ilarraza

Submitted to *Language Resources and Evaluation*

Abstract

This paper presents the adaptation process of the Stanford Coreference resolution module (Lee *et al.*, 2013) to the Basque language, taking into account the characteristics of the language. The module has been integrated in a linguistic analysis pipeline obtaining an end-to-end coreference resolution system for the Basque language. The adaptation process explained can benefit and facilitate other languages with similar characteristics in the implementation of their coreference resolution system. During the experimentation phase, we have demonstrated that language-specific characteristics have a noteworthy effect on coreference resolution, obtaining a gain in CoNLL score of 7.07 with respect to the baseline system. We have also analysed the effect that preprocessing has in coreference resolution, comparing the results obtained with automatic mentions versus gold mentions. When gold mentions are provided, the results increase 11.5 points in CoNLL score in comparison with results obtained when automatic mentions are used. Finally, the contribution of each sieve is analysed concluding that morphology is essential for agglutinative languages to obtain good performance in coreference resolution.

4.1 Introduction

Coreference resolution consists of identifying textual expressions (mentions) that refer to real-world objects (entities) and determining which of these mentions refer to the same entity. Coreference resolution is helpful in NLP applications where a higher level of comprehension of the discourse leads to better performance. Information Extraction, Question Answering, Machine Translation, Sentiment Analysis,

Machine Reading, Text Summarization, and Text Simplification, among others, can benefit from coreference resolution.

It is very common to divide the coreference resolution task into two main subtasks: mention detection and resolution of references (Pradhan *et al.*, 2011). Mention detection is concerned with identifying potential mentions of entities in the text and resolution of references involves determining which mentions refer to the same entity.

In less-resourced language it is particularly challenging to develop highly accurate tools for tasks like mention detection and coreference resolution. Besides, it could be complex to create completely language-independent systems and to take into account the characteristics of the language benefits the performance in these tasks. In this scenario a possible solution is to use a state-of-the art system with flexible modular architecture and adapt it to resolve coreference resolution in the new language to be treated. In our particular case, we have adapted the Stanford Coreference resolution system (Lee *et al.*, 2013) to Basque. The process we carried out demonstrates that using a modular architecture facilitates the development of robust coreference resolution systems for any other language with different characteristics to the language for what the system was originally created.

This paper is structured as follows. After reviewing Related work, we describe the most important characteristics of Basque and the challenges they present for coreference resolution. Then, the architecture of the end-to-end coreference resolution is presented and the adaption process explained. After that, the experiments we carried out to evaluate the coreference resolution system, comparing our results with other systems are described. Finally, conclusions and future work are discussed.

4.2 Related Work

Many coreference resolution conferences have focused on coreference resolution during the last decades. The sixth and seventh Message Understanding Conferences (MUC-6, 1995; MUC-7, 1998) were the first to include a specific task on coreference resolution. The Automatic Content Extraction (ACE) program (Doddington *et al.*, 2004) aimed at identifying certain types of relations between a predefined set of entities.

Nevertheless, all these conferences aimed to resolve coreference for English. In 2010, SemEval-2010 Task 1 (Recasens *et al.*, 2010) was the first conference where coreference in multiple languages (English, Dutch, German, Italian, Spanish and Catalan) had to be resolved. This conference aimed at answering interesting questions, such as, i) to what extent is it possible to implement a general coreference

resolution system portable to different languages?, ii) how much language-specific tuning is necessary to achieve this goal? and iii) how morphology, syntax and semantics can help to solve coreference in each language? were posed.

One year later, in the CoNLL 2011 Shared task (Pradhan *et al.*, 2011), participants had to model unrestricted coreference in the English-language Ontonotes corpora (Pradhan *et al.*, 2007). The system that obtained the best results in the 2011 edition was the Stanford system presented in Lee *et al.* (2013). CoNLL 2012 shared task (Pradhan *et al.*, 2012) focused on coreference resolution in a multilingual setting: English, Chinese and Arabic. The Stanford system was used by six participants (Chen and Ng, 2012; Fernandes *et al.*, 2012; Shou and Zhao, 2012; Xiong and Liu, 2012; Yuan *et al.*, 2012; Zhang *et al.*, 2012). Depending on the language to be resolved, participants used different strategies, e.g., Chen and Ng (2012) seek to improve the multi-pass sieve approach by incorporating lexical information using machine learning techniques. They employ different sieves depending on the language. (Fernandes *et al.*, 2012) took into consideration the special characteristics of each language and used the Stanford system to generate mention link candidates, which were then reranked by a supervised model.

Interest on coreference resolution in languages other than English has been increasing in the last few years, and because of this many works on adaptations of coreference resolution systems to other languages for which they were created have been published. The Beautiful Anaphora Resolution Toolkit (BART) (Versley *et al.*, 2008b) is the case where it has been adapted to many languages. BART was originally created for English, but its flexible modular architecture ensures its portability to other languages. That is why there has been a lot of work on extending the BART coreference toolkit to languages other than English. Recently, BART has been adapted to Basque (Soraluze *et al.*, 2016b).

4.3 Basque characteristics for coreference resolution

Basque is considered language isolate. It differs considerably in grammar from the languages spoken in its surrounding regions. It is an agglutinative, head-final, pro-drop, free-word order language. In addition there is no grammatical gender in the nominal system. Moreover, there are no distinct forms for third person pronouns and demonstratives are used instead (Laka, 1996). All these characteristics make coreference resolution for Basque more challenging in some aspects.

Regarding the agglutinative nature of the Basque language, a given lemma of nouns and adjectives takes many different word forms, depending on the case (genitive, locative, etc.) or the number (singular, plural, indefinite). This means

that looking only for the given exact word is not enough for Basque to resolve coreference when string matching techniques are applied.

Basque, unlike to English, is head-final, the head of a phrase follows its complements, and English is considered head-initial, the head of a phrase precedes its complements. The correct identification of mentions' head is really important in coreference resolution, therefore, the head directionality of the language must be considered.

In relation to word order typology, Basque is known to be a free word order language. Consequently, the same sentence can be written in different manners. For example, the sentence *Jonek liburua irakurri du* "John has read the book" is in neutral word order (SOV) but it has five more variations (SVO, OVS, OSV, VOS, VSO) due to free word order.

The free-word-order nature of Basque can the correct identification of syntactic function make more complex as it becomes more ambiguous. This has a direct effect in coreference resolution as the syntactic function feature is commonly used in the resolution phase.

In addition, Basque is a pro-drop language which allows zero subject pronouns that can be inferred from the verb. Moreover, Basque allows the omission of direct and indirect objects. It is said that Basque is three-way pro-drop language. As a consequence of this characteristic, many pronouns that should be resolved are omitted.

Finally, the lack of grammatical gender in the nominal system makes it impossible to use gender as a feature in the coreference resolution process, which has been proven particularly useful in the resolution of pronouns. The animacy feature cannot be used for pronoun resolution either because Basque pronouns are animacyless.

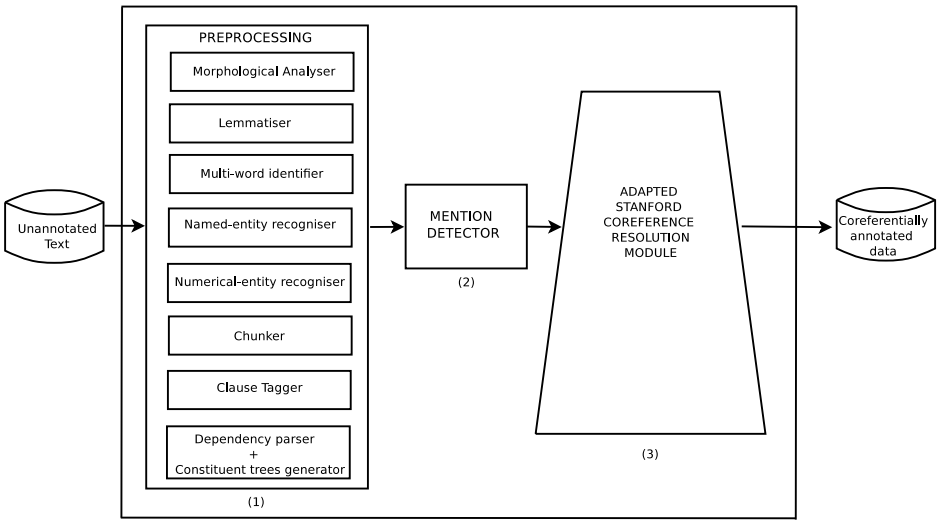
4.4 System architecture

In this section the adapted end-to-end coreference resolution system is presented. As shown in Figure 4.2, the system has three main components: i) preprocessing module, ii) mention detector and iii) coreference resolution module.

4.4.1 Preprocessing

The preprocessing step prepares the input that the coreference resolution system receives. In each step, Basque linguistic processors are applied to the text, thus obtaining linguistically annotated data.

The Basque linguistic processors used to create annotations are the following:



4.2 Figure – End-to-end coreference resolution for Basque.

i) A morphological analyser that performs word segmentation and PoS tagging (Alegria *et al.*, 1996), ii) A lemmatiser that resolves the ambiguity caused at the previous phase (Alegria *et al.*, 2002b), iii) A multi-word item identifier that determines which groups of two or more words are to be considered multi-word expressions (Alegria *et al.*, 2004), iv) A named-entity recogniser that identifies and classifies named entities (person, organization, location) in the text (Alegria *et al.*, 2003), v) A numerical-entity recognizer that identifies and classifies numerical entities (date, time, percent, number...) in the text (Soraluze *et al.*, 2011), vi) A Basque dependency parser (Bengoetxea and Gojenola, 2010); its output is then used to create constituent trees (Díaz de Ilarraza *et al.*, 2008).

4.4.2 Mention Detection

Language-specific patterns that vary according to the features of each language have to be taken into consideration during mention detection. In general, we consider noun phrases (NP), focusing on the largest span of the NP. In the case of nouns complemented by subordinate clauses and coordination, NPs embedded in a larger NP are also extracted.

We created a mention detector system (Soraluze *et al.*, 2016a), defining a set of hand-crafted rules that have been compiled into Finite State Transducers (FST). These FSTs are able to detect complex structures that should be identified as mentions. We defined 12 FSTs, composed of 34 hand-crafted rules using foma (Hulden, 2009).

The mention detector obtains an F-measure of 74.57 when automatic preprocessing is used and an F-measure of 85.89 with gold standard data.

4.4.3 Stanford Coreference resolution module

The Stanford coreference resolution module is a deterministic rule-based system which is based on ten independent coreference models or sieves that are precision-oriented, i.e., they are applied sequentially from highest to lowest precision. Each model selects a single best antecedent from a list of previous mentions or declines to propose a solution. Candidates in the same sentential clauses are sorted using left-to-right breadth-first traversal of syntactic trees to favour subjects (Hobbs, 1978). Nominal mentions in previous sentences are sorted using right-to-left to favour proximity. In the case of pronominal mentions, candidates in previous sentences are also treated left-to-right traversal in order to favour subjects that are more probable antecedents for pronouns. The sorting of candidates is important, as low quality negatively impacts the coreference links created.

The architecture is highly modular, which means that additional coreference models can be easily integrated. The system implements an entity-centric approach, allowing each coreference resolution decision to be globally informed by previously clustered mentions and their shared attributes. Finally, the lack of language-specific lexical features make the system easy to port to other languages (Lee *et al.*, 2013).

The system has been adapted to resolve coreference in Basque, modifying some sieves and adding new ones. It is an extended version of the system presented in Soraluze *et al.* (2015a). We will firstly present how the original sieves work, then the problems found are described and the adaptations proposed.

(S1) **Speaker Identification:** The sieve identifies speakers and links them with corresponding compatible pronouns. In conversational texts, the speakers are identified by searching the subjects of reporting verbs. In non-conversational texts, speaker information is provided in the dataset.

Adaptation: We have translated the list or reporting verbs and the list of pronouns.

(S2) **Exact String Match:** The exact string match sieve links two mentions with exactly same extent text.

In (17), the sieve matches the mentions [Milosevicek] and [Milosevicek] "[Milosevic]" and it considers them coreferent, however, [Milosevici] "on [Milosevic]" is not linked.

#	Mention	Translation	Lemmas	Number	Definiteness	Coreferent
1	txori politak	pretty bird	txori polit	plural	definite	-
2	txori politekin	with the pretty birds	txori polit	plural	definite	yes
3	txori politak	pretty bird	txori polit	singular	definite	no
4	txori politek	pretty birds	txori polit	plural	indefinite	no

4.4 Table – Examples to illustrate the suitability of the Exact Morphology Match sieve

(17) [Milosevicek] bere herriaren borondatea errespetatu beharko luke. Joe Lockhardek bozeramailearen ustez, [Milosevicek] lehenbailehen utzi beharko luke gobernua. Etxe Zuriko maizterrak esandakoak presio handiagoa egiten dio [Milosevici] boterea ahal den azkarren utz dezan.

"[Milosevic] should respect the will of its people. In the opinion of spokesman Joe Lockhard [Milosevic] should leave the government immediately. The words of the occupant of the White House put more pressure on [Milosevic] to leave power as soon as possible."

Problem: The constraint applied in the Exact String Match sieve is too restrictive in agglutinative languages, as the role of prepositions is played by suffixes added to word forms. Consequently, two mentions that refer to the same entity but differ in their word forms are not considered coreferent. That is the case of the mention [Milosevici] in example (17), it is not considered coreferent with [Milosevicek] as their extent text differs.

Adaptation: We created a specialisation of the Exact String Match sieve, named **Exact Morphology Match** sieve, and replaced it in the adapted version of the coreference resolution system. This sieve takes into account morphological features of mentions to consider if they corefer or not. Two mentions are linked if i) the lemmas of each word in both mentions are identical, and ii) if their number and definiteness are equal (or unknown in one of the mentions). Number attributes can take singular, plural or unknown values and definiteness values can be finite, indefinite or unknown (treated as wildcards, i.e., they can match any other value).

It is important that two mentions fulfil these three constraints at the same time because even if one or two are fulfilled, the mentions may not be coreferent as the examples in in Table 4.4 illustrate.

The first mention *txori politak*, and the second one, *txori politekin*, are coreferent because the conditions are fulfilled: the same lemmas, the same number

and the same definiteness. Nevertheless, although the first and third mentions are identical strings, they are not coreferent. The first mention *Txori politak* represents a plural mention in the absolutive case, and the same string in the third row corresponds with a mention in the singular ergative case (obviously this morphological information has been previously extracted by attending to the context). Finally, the first and fourth mentions have the same lemma and number but their definiteness differs (the first is definite while the second is indefinite), so they can not be considered coreferent.

After replacing the Exact String Match sieve with Exact Morphology Match the three mentions in example (17) are considered coreferent.

(S3) **Relaxed String Match:** This sieve considers two mentions as coreferent if the strings obtained by dropping the text following their head words, such as, relative clauses and participial postmodifiers (clauses headed by participial form of the verb) are identical.

Problem: In English, relative clauses follow the head word, however, in Basque they can follow or precede the head word. In the following example, the two possibilities in Basque for the relative clause "[Bill Clinton who accepted the new law] appeared hopeful in front of the reporters." are presented. Although, the following two examples in (18) are correct, the (a) case is the most common in Basque.

(18) a [Lege berria onartu duen Bill Clinton] itxaropentsu agertu zen kazetarien aurrean.

b [Bill Clinton zeinak lege berria onartu duen] itxaropentsu agertu zen kazetarien aurrean.

Similarly to relative clauses, participials in Basque can appear in two manners, however, they mostly precede the noun and, ocassionally, phrases can appear apposited to the right of the noun phrase.

Adaptation: The Relaxed String Match sieve has been modified in order to also consider relative clauses and participials that precede the noun. The adapted sieve is also able to drop the text of relative clauses and participials preceding the head word. To consider two mentions coreferent the compared mentions also have to fulfil the same morphological constraints applied in The Exact Morphology Match sieve.

After the adaptation, the sieve is able to link mentions the two mentions in example (19).

(19) [Gailurretako Rallyan] iragan urteko balentria handia errepikatu nahi izango dute pilotu zuberotarrek. [Igandean bukatuko den Gailurretako Rallyan] lehiakide zorrotzenekin topo egingo dute.

"In [Rally des Cimes] Soule pilots will want to repeat last year's great bravery. [In Rally des Cimes which ends on Sunday] they will meet the ablest opponents."

(S4) **Precise Constructs:** The Precise construct sieve links two mentions, if any of the conditions below is fulfilled:

- *Appositive*: The two mentions are in appositive relation.
- *Predicative nominative*: The two mentions are in copulative construction.
- *Acronym*: One of the mentions is an acronym of the other mention and both are tagged as a proper noun. The algorithm to detect acronyms marks as mention an acronym of the other if its text equals the sequence of upper case characters in the other mention.
- *Demonym*: One of the mentions is a demonym of the other. For demonym detection a static list of countries and their gentilic forms from Wikipedia is used.
- *Role appositive*: The candidate antecedent's head word is a noun and it appears as a modifier in a Noun Phrase whose head is this mention. For example, [[singer] Michael Jackson].
- *Relative pronoun*: The head of the antecedent Noun Phrase is modified by a relative pronoun mention, e.g., [the finance street [which] has already formed in the Waitan district].

Problem: The rules used to detect appositive and predicative nominative structures are not the most suitable for Basque. Furthermore, the algorithm to detect acronyms does not consider that acronyms can appear declined. In addition, the way in which our gold-standard corpus of mentions and coreference chains was annotated does not consider as mentions role appositives and relative pronoun structures. For example, in the case of singer Michael Jackson, Michael Jackson abeslaria in Basque, the whole structure is considered one mention [Michael Jackson abeslaria] and not two mentions [[singer] Michael Jackson] as in English.

Adaptation: A detector of appositive and predicative nominative for Basque (Gonzalez-Dios *et al.*, 2013) has been integrated in the preprocessing pipeline.

The information obtained by the detector then is provided to the coreference resolution system, and it links mentions that are considered appositive or predicative nominatives.

Referring to the algorithm to detect acronyms, it has been changed to treat declined acronyms. This way one mention is considered an acronym of the other if its lemma equals the sequence of upper case characters in the other mention. Example 20 illustrates the necessity to compare acronym lemmas instead of its text. If acronym text [AABek] "[CA]" were compared with [Amnistiaren Aldeko Batzordeak] "[Commission for Amnesty]", the two mentions would not be considered coreferent.

(20) Europako Giza Eskubideen Agiria ez dela betetzen salatuko dute [AABek] Nizan. Europako Batasuneko estatuburuen bilerara joango dira [Amnistiaren Aldeko Batzordeak] bertan onartuko den Giza Eskubideen Europako Agiria Euskal Herrian praktikara ez dela eramaten salatzera.

"[CA] will denounce in Nice that the European Convention on Human Rights is not met. [Commission for Amnesty] will go to the meeting of state heads of the European Union to denounce that European Convention on Human Rights to be accepted there is not put into practice in the Basque Country."

In addition, the original English static list of countries and their gentilic forms has been replaced by a Basque version to identify demonyms. The list has been created using the 38th rule of Euskaltzaindia (Royal Academy of the Basque Language).

Finally, the constraints of Role Appositive and Relative Pronoun have been deactivated. Role appositives are treated in sieve 9, (S9).

(S5-S7) **Strict Head Match A-C:** The Strict Head Match sieve links two mentions if all the following constraints are satisfied at the same time:

- *Entity head match*: The mention head word matches any head word in the antecedent entity.
- *Word inclusion*: All the non-stop words in the current entity to be solved are included in the set of non-stop words in the antecedent entity.
- *Compatible modifiers only*: The modifiers of the mention to be resolved are included in the modifiers of the antecedent candidate. As modifiers, nouns and adjectives are considered.

- *Not-i-within-i*: The two mentions are not in i-within-i construct, i.e., one cannot be a child NP in the other's NP constituent

The variants B and C of the Strict Head Match are relaxations of the constraints explained above. Strict Head Match B (S6) removes the constraint *compatible modifiers* and the Strict Head Match C (S7) removes the *word inclusion*.

Problem: This sieve is not suitable to apply directly in agglutinative languages, as in three of the four constraints that are applied word forms are compared.

Adaptation: We have modified the *Entity head match*, *Word inclusion* and *Compatible modifiers only* constraints. The adapted constraints compare lemmas instead of word forms. In addition, in *Entity Head Match* the compared head words must fulfil number and definiteness agreement.

After the adaptation of the three constraints, the sieve correctly links the mentions in example (21) as all the constraints are fulfilled. In addition, the mentions in example (22) are not linked. They are not coreferent, because even if the *Entity Head Match* constraint is fulfilled, the *Word inclusion* and *Compatible modifiers only* constraints are not satisfied.

(21) [Euskalteleko kirol zuzendari Julian Gorospek] emaitza onak lortzeko moduan daudela uste du. [Julian Gorosperentzat] Tourrean izateko aukera gutxi zuten, eta beraz, ondo hartu du berria.

"[Euskaltel sporting director Julian Gorospe] thinks they can get good results. For [Julian Gorospe] they had little chance to be on the Tour, and therefore, he took the news well."

(22) Eurokopako atezain onena bilakatu da Toldo, eta Italian heroi nazionala da. Merezita lortu du [Fiorentinako atezainak] gailurrera iristea. Peruzziri ere zor dio Toldok zerua ukitu izana. Buffonek titulartasuna kendu izana ez du ondo irentsi [Inter Milaneko atezainak] eta uko egin zion Eurokopari.

"Toldo has become the best goalkeeper in the European Championship, and he is a national hero in Italy. [Fiorentina goalkeeper] deserved it to reach the top. Toldo also owes Peruzzi for having touched the sky. [Inter Milan goalkeeper] did not take well that Buffon pulled off him the starting player position and he rejected the European Championship."

(S8) **Proper Head Word Match:** This sieve considers two mentions coreferent the following constrains are satisfied:

- Both mentions are headed by proper nouns and the head word is the last word of the mention.
- Not-i-within-i
- No location mismatches: The modifiers of two mentions do not have different location named entities, other proper nouns, or spatial modifiers.
- No numeric mismatches: The mention to be resolved does not have a number that does not appear in the antecedent mention.

Problem: Basque is a free word-order language, consequently, the head word of a mention should not neccesarily appear in the last position.

Adaptation: We changed the first constraint to permit the head word of the mention to appear in other positions apart from the last position. In addition, lemmas instead of word forms are used to compare head words and modifiers.

In example (23), the adapted sieve considers coreferent the mentions [Batistutarentzat] "For [Batistuta]" and [Gabriel Batistuta argentinarrak] "[Argentinian Gabriel Batistuta]" where the head word is Batistuta, and discards mentions in example (24) as they have a different spatial modifier.

(23) [Batistutarentzat] ez da normala Erromak beragatik ordaindutakoa. Italiako Erromarekin sinatu berri duen kontratuagatik harro dagoela esan zuen atzo [Gabriel Batistuta argentinarrak].

"For [Batistuta] is not normal what Rome has paid for him. [Argentinian Gabriel Batistuta] said yesterday he is proud of [the contract he has signed recently with Rome from Italy."

(24) Bonba bat lehertu da [Londresen]. Eztandak ez du inor zauritu, eta kalte material txikiak eragin ditu, Hammersmith zubian, [Londres mendebaldean].

"A bomb has exploded in [London]. The explosion has not hurt anyone and material damage have been small, in Hammersmith Bridge, in [west London]."

(S9) **Relaxed Head Match:** This sieve relaxes the entity head match constraint by allowing the mention head to match any word in the antecedent candidate

cluster. Apart from this constraint, both mentions have to be named entities and the types coincide. In addition, *word inclusion* and *i-within-i* constraints have to be fulfilled.

Problem: The same problem that occurs with the *Entity Head Match* constraint happens when comparing the head words because word forms are used and not lemmas. Moreover, as role appositives are not tagged in our corpus, they are not treated in the Precise Constructs sieve (S4). Nevertheless, they provide useful information that is needed to resolve some types of coreference relations that are treated in this sieve.

Adaptation: To compare head words, lemmas are used instead of word forms. Furthermore, we have relaxed the constraint that needs both mentions to be named entities. In our case, only the antecedent mention has to be a proper noun, the candidate mention could be nominal. The last change has been adopted to link mentions with their antecedents that are in role appositive constructions.

After the adaptation, the sieve considers the mentions [Portlandeko entrenatzaile Zupo Ekisoainekin] "with [Portland trainer Zupo Ekisoain]" and [Portlandeko entrenatzaileak] "[Portland trainer]" coreferent in example (25).

(25) Gentzelek atzo goizean hitz egin zuen [Portlandeko entrenatzaile Zupo Ekisoainekin], Granollersek egindako eskaintza onartzeko arrazoiak azaltzeko. Granollersen lagunak ditu bere emazteak, eta hara joatea nahiago izan dutela azaldu zuen atzo [Portlandeko entrenatzaileak].

"Gentzel talked with [Portland trainer Zupo Ekisoain] yesterday morning, to explain the reasons why he accepted the offer made by Granollers. His wife has friends in Granollers, and they have preferred to go there explained [Portland trainer] yesterday."

(S10) **Pronoun Resolution:** This sieve links pronominal mention with the antecededent mention based on these constraints:

- Number: the number attribute is assigned based on: i) a static list of pronouns; ii) mention marked as a named entity is considered singular with the exception of organizations, which can be both singular and plural; iii) NN*S part of speech tags are plural and other NN* tags are singular; iv) a static dictionary.
- Gender: Gender attributes are assigned from static lexicons.
- Person: person attributes are assigned to pronouns.

- Animacy: The animacy is set using: i) a static list of pronouns; ii) NER labels, e.g., PERSON is animate whereas LOCATION is not; iii) a dictionary boostrapped from the web.
- NER label.
- Pronoun distance: distance between a pronoun and its antecedent cannot be in sentences larger than 3.

Problem: Basque pronouns do not provide information about gender or animacy. In addition, the demonstrative determiners also act as third person pronouns.

Adaptation: We have translated the static list of pronouns to Basque. In addition, we have enriched the static lexicon for assigning gender with Basque, French and Spanish names. Furthermore, the pronoun distance constraint has been reduced to 2, i.e., the distance between a pronoun and its antecedent cannot be in sentences larger than 2. The modification of the pronoun distance value has been experimentally optimised. As pronouns in Basque do not provide information about their gender and animacy they are considered more ambiguous than in another languages that could exploit these features. Consequently, the distance between the antecedent and the pronoun may be shorter. Taking this into consideration, we have changed the mention candidate selection algorithm. Thereby, the antecedent candidates for pronouns are sorted right-to-left traversal in the same and previous sentences in order to favour proximity. Finally, to improve the precision of the Pronoun resolution sieve, we apply a new constraint apart from all those which are explained above. The definiteness of both mentions has to match in order to consider them coreferent. After all the modifications, the sieve is able to correctly link pronouns with their antecedents as in example (26).

(26) Partidaren hastapenean Miarritzeko gazte euskaldun batzuek [euskal presoen aldeko zapiak] zeramatzaten bizkarrean, eta [haiekin] sartzen saiatu ziren.

"At the beginning of the match some young Basques from Biarritz wore [scarves in favor of Basque prisoners] on their backs and they tried to enter with [them]."

(S11) **Ellipsis Match:** In our corpus elliptic mentions, the structures in which a noun ellipsis occur, i.e., the suffixes attached to the word correspond to a noun, even when the noun is not explicit in the word, are tagged and can be part of coreference chains. The Stanford system does not have any sieve to

treat elliptic mentions, therefore this sieve has been added to the coreference resolution system to treat ellipsis case.

The sieve links a mention with an elided noun with its antecedent. To link two mentions the candidate mention and the candidate antecedent have to agree in number and definiteness and they have to appear in the same sentence. For example, the mention with elided noun [kalitate handikoak] in example (27) is linked with its antecedent [Biak].

(27) Argentinako bi jokalari etorri ziren Gasteizera orain dela hamar urte: Nicola eta Guinazu. [Biak] oso gazteak ziren arren, [kalitate handikoak] ziren, eta etorkizun oparoa zuten.

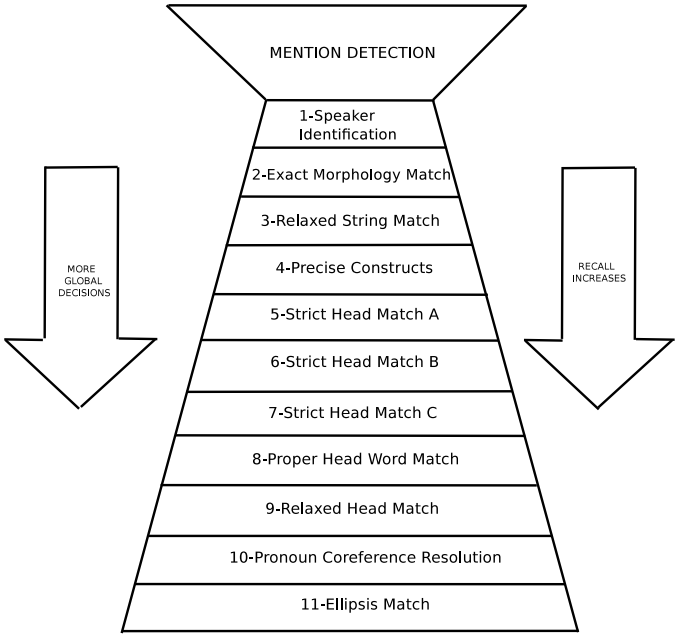
"Argentinian players came to Gasteiz ten years ago: Nicola and Guinazu. Although [both] were very young, they were [good quality players], and they had a promising future."

To summarize, in the adapted version of the Stanford system to Basque we make use of deeper morphological information to tackle the agglutinative nature of Basque. At the same time, changes related with the free-word order of Basque have also been implemented in the Basque system. We replaced an existing sieve with a new one, the Exact String Match sieve with the Exact Morphology Match sieve, modify 9 existing sieves, and introduce a new sieve, Ellipsis Match.

The architecture of the Basque coreference resolution system is shown in Figure 4.3, where the order of application of the sieves is illustrated.

4.5 System Evaluation

In this section, we evaluate the adapted Coreference resolution system. Several experiments have been carried out to measure different aspects of the system: i) we set the baseline with a copy of the original Stanford Corefence resolution system for English but it takes as input the output of the Basque linguistic processors and the Basque static lists of pronouns, demonyms and gender, ii) the adapted system is compared with the baseline system. Both systems are compared using automatic and gold mentions to distinguish the effect that preprocessing and mention detection have in the results, iii) in order to obtain an optimal sieve order, intuition guided *Hand-built ordering* and automatically obtained *Learned ordering* are considered, iv) an incremental evaluation of the adapted system has been performed to measure the contribution of individual sieves, v) our system is compared with other systems for different languages.



4.3 Figure – The architecture of the Basque coreference resolution system.

4.5.1 Corpus

EPEC (Reference Corpus for the Processing of Basque) (Aduriz *et al.*, 2006) is a 300,000 word sample collection of standard written Basque that has been manually annotated at different levels (morphology, surface syntax, phrases, etc.). The corpus is composed of news published in *Euskaldunon Egunkaria*, a Basque language newspaper. It is aimed to be a reference corpus for the development and improvement of several NLP tools for Basque.

Recently, mentions and coreference chains were also annotated by two linguists in a subset of the EPEC corpus which is composed of about 45,000 words (Ceberio *et al.*, 2016). First, automatically annotated mentions obtained by our mention detector were corrected; then, coreferent mentions were linked in clusters.

We divided the dataset into two main parts: one for developing the system and the other for testing. More detailed information about the two parts can be found in Table 4.5.

4.5.2 Metrics

The metrics used to evaluate the performance of the systems are MUC (Vilain *et al.*, 1995), B^3 (Bagga and Baldwin, 1998), CEAF_e (Luo, 2005), CEAF_m (Luo, 2005),

	Words	Mentions	Clusters	Singletons
Devel	30434	8432	1313	4383
Test	15949	4360	621	2445

4.5 Table – EPEC corpus division information

Automatic Mention Detection							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
Baseline	74.57	27.46	59.96	56.86	58.61	36.75	48.67
Adapted	74.57	42.32	62.94	61.54	61.98	43.18	55.74

4.6 Table – Performance of baseline and adapted systems with automatic mentions

and BLANC (Recasens and Hovy, 2011). The CoNLL metric is the arithmetic mean of MUC, B^3 and CEAF_e metrics. The scores have been calculated using the reference implementation of the CoNLL scorer (Pradhan *et al.*, 2014).

4.5.3 Automatic mentions vs. gold mentions

Table 4.6 shows the F1 scores obtained by the baseline (original Stanford system with Basque preprocessing and translate static lists) and the adapted system. In this case automatically identified mentions are used. The adapted system outperforms the baseline system according to F1 on all the metrics. In CoNLL metric, the adapted system has a score of 55.74, which is 7.07 points higher than the baseline system, which scores 48.67.

To isolate the behaviour of the resolution references from the mention detection, we have also compared the systems when gold mentions are provided. Scores obtained in this case are shown in Table 4.7. As we can observe, the adapted system also outperforms the baseline according to all the metrics. The official CoNLL metric is outperformed by 11.5 points.

It is interesting to observe the difference between the results obtained when automatic mentions and those obtained when gold mentions are provided. It is clear that having accurate preprocessing tools and a good mention detector are crucial to obtain good results in coreference resolution. The difference in CoNLL is about 15.95 points higher for the baseline system and 20.38 points for the adapted system when gold mentions are used.

Gold Mention Detection							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
Baseline	100	36.55	81.28	72.13	76.05	62.94	64.62
Adapted	100	58.12	86.99	80.57	83.27	72.77	76.12

4.7 Table – Performance of baseline and adapted systems with gold mentions

Hand-built ordering	**Learned ordering**
S1 Speaker Identification	S1 Speaker Identification
S2 Exact Morphology Match	S11 Ellipsis Match
S3 Relaxed String Match	S2 Exact Morphology Match
S4 Precise Constructs	S3 Relaxed String Match
S5 Strict Head Match A	S4 Precise Constructs
S6 Strict Head Match B	S8 Proper Head Word Match
S7 Strict Head Match C	S6 Strict Head Match B
S8 Proper Head Word Match	S5 Strict Head Match A
S9 Relaxed Head Match	S7 Strict Head Match C
S10 Pronoun Resolution	S10 Pronoun Resolution
S11 Ellipsis Match	S9 Relaxed Head Match

4.8 Table – Hand-built ordering and Learned ordering

4.5.4 Sieve ordering

The ordering of the sieves in the adapted system follows the intuition that is used in the original Stanford Coreference Resolution system, firstly the most precise sieves are applied and then those that are less precise.

Nevertheless, this order could not be the most optimal to be applied in Basque coreference resolution. Therefore, we performed and experiment to automatically obtain the best order of sieves. A greedy search was used, and the best precision sieve at each stage was chosen. The tuning of the sieve order was obtained using the development part of the EPEC corpus, and then evaluated in the test part.

Table 4.8 illustrates the new *Learned ordering* proposed by the optimisation algorithm in comparison with the *Hand-built ordering* represented in Figure 4.3.

The optimization resulted in some variations in all the metric scores although the CoNLL F1 remains at 55.74 points. It can be concluded that the two order configurations are optimal.

4.5.5 Incremental adding of sieves

In order to quantify the contribution of each individual, sieve we have evaluated our system by adding 11 sieves incrementally. The sieves have been added using the new Learned ordering proposes by the optimization algorithm. The results obtained are presented in Table 4.9.

The analysis of results reveals that the most significant improvements are due to the sieve Exact Morphology Match. This sieve accounts for an improvement of 14.08 CoNLL F1 points, which proves that replacing the original Exact String Match sieve with this, which takes into account morphological characteristics of the mentions, is necessary for morphologically rich languages.

The second biggest improvement in performance, around 1 point in CoNLL F1, is caused by Proper Head Word Match sieve followed by Strict Head Match B sieve, which improves by 0.83 points.

There is no gain in scores when Strict Head A sieve is applied. The reason for this is that Strict Head Match B is applied before Strict Head Match A sieve. As Strict Head Match B is a relaxation of the Strict Head Match A, all the mentions that A variation should link are resolved when B is applied.

We can observe that the CoNLL result drops slightly when Precise Constructs sieve is applied, exactly by 0.08 points. This drop is caused by the predication and apposition structure identifier, as it does not always correctly identify these type of structures.

The results show that linking elliptical mentions with their antecedent is a complex task. The improvement of CoNLL score of the Ellipsis Match sieve is slow. A deep analysis of ellipsis cases to improve the Ellipsis Match sieve is needed.

4.5.6 Comparison of results with other languages

To finish the evaluation of our systems it is interesting to compare the results obtained with similar systems for languages that share similar characteristics. The results of the compared systems are not directly comparable as the corpus used for evaluation is different but they give us an idea of the performance of our system.

In CoNLL 2012 Shared Task, two participants ((Fernandes *et al.*, 2012) and (Chen and Ng, 2012)) used an adaptation of the Stanford Systems to resolve coreference for Arabic, a morphologically rich language like Basque. (Fernandes *et al.*, 2012) obtained the best score for Arabic, 45.2 in CoNLL F1 and (Chen and Ng, 2012) obtained 32.4, the fifth position out of seven. Our system obtains 10.54 points higher than the best.

German is also a morphologically rich language, but while Basque is an agglutinative language German is a considered fusional language. The two systems presented in SemEval-2010 Task 1 (Recasens *et al.*, 2010) that resolved coreference

	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL
Speaker Identification	0	60.37	53.9	56.46	27.19	38.94
Ellipsis Match	0.3	60.37	53.95	56.53	27.25	39.06
Exact Morphology Match	35.74	63.13	60.06	60.58	39.95	53.15
Relaxed String Match	36.03	63.12	60.08	60.62	40.04	53.25
Precise Constructs	35.84	63.06	60.01	60.62	39.96	53.17
Proper Head Word Match	38.16	63.24	60.65	61.22	40.88	54.20
Strict Head Match B	40.15	63.36	61.16	61.61	41.74	55.04
Strict Head Match A	40.15	63.36	61.16	61.61	41.74	55.04
Strict Head Match C	40.79	63.32	61.14	61.51	42.18	55.20
Pronoun Resolution	41.55	63.02	61.28	61.76	42.73	55.44
Relaxed Head Match	42.32	62.95	61.51	61.97	43.14	55.74

4.9 Table – Performance of the system when sieves are added incrementally

System	**Language**	**CoNLL F1**
This work	Basque	55.74
(Fernandes *et al.*, 2012)	Arabic	45.2
(Chen and Ng, 2012)	Arabic	32.4
(Kobdani and Schütze, 2010)	German	55.03
(Zhekova and Kübler, 2010)	German	33.93
(Ogrodniczuk and Kopeć, 2011)	Polish	61.95

4.10 Table – Results obtained by coreference resolution system for different languages

for German are SUCRE (Kobdani and Schütze, 2010) and UBIU (Zhekova and Kübler, 2010). SUCRE obtained an score of 55.03 CoNLL F1 and UBIU 33.93. The best result is similar to our result.

Finally, it is interesting to compare the results with Polish, a similar language to Basque. Polish is also an inflectional and free-word order language. The system presented in Ogrodniczuk and Kopeć (2011) obtains a CoNLL F1 score of 61.95 points. The result is better than ours by 6.21 points.

Table 4.10 summarises the results obtained by the systems explained above.

4.6 Conclusions and Future Work

We have adapted the Stanford Coreference resolution system to the Basque language and integrated it into a global architecture of linguistic processors obtaining an end-to-end coreference resolution system. We have taken into account the characteristics of Basque and described the adaptation process in detail, sieve by sieve, facilitating the replicability of the process for other languages. Evaluation of the adapted system has been carried out, comparing the results with a baseline system (original Stanford system with Basque preprocessing and translated static lists) in two scenarios, automatic mentions versus gold mentions. The adapted system outperforms the baseline in all the metrics. In CoNLL F1 when automatic mentions are used the baseline is outperformed by 7.07 points and by 11.5 points when gold mentions are provided. We have also carried out an incremental experiment to quantify the contribution of each individual sieve.

The results obtained by our system have been compared with other language systems showing that our system obtains better results in some cases and worse in others.

As future work, we intend to make a deep error-analysis which can reveal our system's weak points and help to decide our future directions in the improvement of the system. Furthermore, we are interested in using semantic and word knowledge to link mention which has yet to be resolved. External resources such as Basque WordNet and Wikipedia are good resources that could help to resolve cases in which semantic information or word knowledge about mentions is needed. It would also be interesting to investigate other kinds of techniques such as machine-learning (to the extent that EPEC corpus is increased) and combine the strengths of rule-based methods and learning-based methods in a hybrid approach similar to the work presented in Chen and Ng (2012).

5

Enriching Basque Coreference Resolution System using Semantic Knowledge sources

5.0 Laburpena

Euskarara egokitutako Stanfordeko korreferentzia-ebazpenerako sistemaren errore-analisia egin ondoren ezagutza semantikoa eta munduaren ezagutzaren gabezia antzeman dugu. Beraz, sistema ezagutza horrekin aberasteko asmoz WordNet eta Wikipedia baliabideak nola erabili diren azaltzen da artikulu honetan.

Korreferentzia-ebazpena ebaluatzeko erabiltzen diren metrikek sistema bat korreferentzia-erlazioak ebazten zein ona den adierazten digute. Hala ere, metrika horiek ez dira gai sistema horren gabeziak detektatzeko, ez eta gabezia horiek nola konpon daitezkeen argitzeko ere. Errore-analisiak aukera egokia dira gabezia horiek identifikatu eta soluzio bideragarriak planteatzeko.

Errore-analisien onurak kontuan izanik, euskarara egokitutako Stanfordeko korreferentzia-ebazpenerako sistemari egindako errore-analisia aurkezten dugu lehenbizi artikulu honetan. Kummerfeld and Klein (2013) lanean korreferentzia-ebazpenean gertatzen diren errore-motak zazpi multzotan sailkatzen dituzte. Sailkapen hori oinarritzat hartu eta errore-mota horiek sortzen dituzten errore-kausak identifikatu eta definitzen ditugu guk. Guztira 7 errore-kausa identifikatu ditugu. Errore-kausa horien artean, batzuk aipatzearren, aurreprozesaketa okerra, aipamen-detekzioko akatsak, eta semantikaren eta munduaren ezagutzaren beharra identifikatu dira.

Behin errore-motak eta errore-kausak zehaztuta, gure korreferentzia-ebaz-

penerako sisteman errore-kausek errore-motetan zein proportziotan eragiten duten kuantifikatzen dugu. Errore-analisiaren ondorio nagusitzat gure sistemak hainbat korreferentzia-erlazio ebazteko duen gabezia aipa genezake, zehatzago esanda, ebazpenerako ezagutza semantikoa eta munduaren ezagutza edo ezagutza entziklopedikoa beharrezkoa den kasuetakoa alegia. Adibide gisa, korreferentzia-ebazpenerako sistema ez da gai jakiteko *Osasuna futbol taldea* eta *gorritxoak* aipamenak korreferenteak direla, gorritxoak Osasunari deitzeko erabiltzen den goitizena dela ez dakielako. Munduaren ezagutza beharrezkoa den kasuak errore guztien % 9,86 dira, semantika beharrezkoa den kasuetan sortzen diren erroreak, berriz, % 6,42.

Versley *et al.* (2016) autoreen arabera, sistema batean informazio lexikala eta munduari buruzko ezagutza txertatzean ez da erraza hobekuntzak lortzea baina posiblea eta erabat beharrezkoa da, hori dela eta, gure sistemak dituen gabeziek sortutako erroreak konpontzeko asmoz bi bahe berri, gehitu ditugu.

Lehenengo baheak, *Wikipedia sieve* deiturikoak, Wikipediako orrialdeetako informazioarekin aberastutako aipamenak erabiltzen ditu. Aipamenak aberasteko, entity-linkingeko teknikak erabiltzen dira, entitate izendunak diren aipamenak dagokien Wikipediako orrialdearekin lotzeko. Bahearen helburu nagusia ezagutza entziklopedikoa beharrezkoa den korreferentzia-erlazioak ebaztea da. Adibidez, bahea gai da *Osasuna* eta *gorritxoak* moduko aipamenak korreferenteak direla esateko.

Bigarrengo baheak, *Synonymy sieve* deiturikoak, euskarazko WordNetetik erauzitako sinonimoen zerrenda erabiltzen du oinarrian, aipamenaren burua sinominoa duten eta bateragarriak diren aipamenak lotzeko. Bahe honek sistemaren gabezi semantikoaren ondorioz sortutako erroreak konpontzea du helburu, adibidez, gai da *Libanoko parlamentua* eta *Libanoko Legebiltzarra* bezalako aipamenak korreferentzia-kate berean biltzeko.

Bi baheen nondik norakoak azalduta, horiek sisteman eragiten dituzten hobekuntzak aztertzeko, lau sistema ebaluatzen ditugu. Oinarri-lerrotzat bahe berriak gehitu gabe dituen sistema (1) hartzen dugu eta emaitza horiek *Wikipedia sieve* bahea bakarrik gehitzean lortzen direnekin (2) eta *Synonymy sieve* bahea soilik gehitzen direnekin (3) konparatzen dira lehenbizi. Azkenik, bi baheak batera gehitzean lortzen diren emaitzak (4) aurkezten dira. Aipatutako lau kasuetan aipamen automatikoak erabiltzean lortzen diren emaitzak 5.1 taulan ikus daitezke eta urrezko aipamenak erabiltzean lortzen direnak berriz, 5.2 taulan aurkezten dira.

Wikipedia sieve eta *Synonymy sieve* baheak aplikatu ostean oinarri-lerroarekiko 0,24 hobekuntza lortzen dugu CoNLL metrikan aipamen auto-

	Aipamen automatikoak																		
	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
1	34,1	55,76	42,32	57,98	68,83	62,94	60,78	62,31	61,54	66,02	58,41	61,98	38,41	53,57	43,18	46,71	51,78	49,12	55,74
2	34,41	55,70	42,54	58,09	68,64	62,93	60,73	62,26	61,49	65,94	58,49	61,99	38,65	53,27	43,35	46,82	51,64	49,11	55,82
3	34,57	56,03	42,76	58,08	68,80	62,98	60,85	62,38	61,61	65,99	58,51	62,03	38,53	53,65	43,31	46,83	51,97	49,27	55,92
4	34,88	55,90	**42,95***	58,19	68,60	**62,97**	60,80	62,33	**61,56**	65,92	58,60	**62,04**	38,77	53,33	**43,48***	46,94	51,83	**49,26**	**55,98***

5.1 taula – Aipamen automatikoekin lortutako emaitzak. 1=Baseline, 2=1+Wiki sieve, 3=2+Synonymy sieve, 4=1+Wiki sieve+Synonymy sieve.

	Urrezko aipamenak																		
	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
1	48,76	71,94	58,12	81,35	93,47	86,99	80,57	80,57	80,57	89,00	78,24	83,27	67,09	84,65	72,77	66,36	71,11	68,66	76,12
2	49,84	70,81	58,50	81,71	92,83	86,92	80,57	80,57	80,57	88,69	78,77	83,44	67,51	83,27	72,84	66,60	71,01	68,73	76,28
3	50,00	71,50	58,85	81,69	93,19	87,06	80,80	80,80	80,80	88,90	78,82	83,56	67,39	84,23	72,95	66,68	71,52	69,02	76,49
4	50,46	70,99	**58,99***	81,86	92,81	86,99	80,71	80,71	**80,71**	88,71	79,00	**83,57***	67,68	83,34	**73,00**	66,79	71,29	**68,97**	**76.51***

5.2 taula – Urrezko aipamenekin lortutako emaitzak. 1=Oinarri-lerroa, 2=1+Wiki sieve, 3=2+Synonymy sieve, 4=1+Wiki sieve+Synonymy sieve.

matikoak erabiltzen diren kasuan eta 0,39 puntutakoa urrezko aipamenekin. Bi hobekuntzak estatistikoki esanguratsuak dira Paired Student's t-testean oinarrituta. Orokorrean, nahiz eta oso handiak ez izan, metrika guztietan lortzen ditugu hobekuntzak, bai aipamen automatikoekin baita sistemari urrezko aipamenak pasatzen zaizkionean ere.

Amaitzeko, lortutako bahe berrien ekarpena nabarmenagoa ez izatearen kausak identifikatzen eta aurkezten ditugu. Guztira bost kausa nagusi zerrendatzen dira. Hala nola, hizkuntza gutxituetako baliabideek izan dezaketen estaldura falta.

Laburbilduz, euskarako korreferentzia-ebazpenerako sistema hobetu dugu bi bahe espezializatu berri gehituz. Bahe hauek ezagutza entziklopedikoa eta ezagutza semantikoa beharrezkoa duten hainbat korreferentzia-erlazio ebazteko gai dira.

Enriching Basque Coreference Resolution System using Semantic Knowledge sources

Ander Soraluze, Olatz Arregi, Xabier Arregi and Arantza Díaz de Ilarraza

Published in *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, 8–16, Valencia, Spain

Abstract

In this paper we present a Basque coreference resolution system enriched with semantic knowledge. An error analysis carried out revealed the deficiencies that the system had in resolving coreference cases in which semantic or world knowledge is needed. We attempt to improve the deficiencies using two semantic knowledge sources, specifically Wikipedia and WordNet.

5.1 Introduction

Coreference resolution consists of identifying textual expressions (mentions) that refer to real-world objects (entities) and determining which of these mentions refer to the same entity. While different string-matching techniques are useful to determine which of these mentions refer to the same entity, there are cases in which more knowledge is needed, that is the case of the Example in 28.

(28) [**Osasunak**] lehenengo mailara igotzeko lehian azken astean bizi duen giroa oso polita da. [**Taldea**] lasaitzeko asmoz Oronozera eraman zituen Lotinak atzo guztiak. Oronozko kontzentrazioa beharrezkoa dute [**gorritxoek**].

*"**[Osasuna]** is going through a beautiful moment in the last week in the race to ascend to the Premier League. In order to reassure **[the team]** Lotina has decided to give all of them to Oronoz. **[The reds]** need to concentrate in Oronoz."*

Having the world knowledge that *Osasuna* is a *football team* and its nickname is *the reds* would be helpful for establishing the coreference relations between the mentions [**Osasuna**], [**Taldea**] and [**gorritxoek**] in the example presented above.

Evaluation scores used in coreference resolution tasks can show how effective a system is; however, they neither identify deficiencies of the system, nor give any indication of how those errors might be corrected. Error analyses are a good option that can help to clear the deficiencies of a coreference resolver. Bearing this in mind, we have carried out an error analysis of the extended version of the coreference resolution system presented in Soraluze *et al.* (2015a). In this paper we present an improvement of this Basque coreference resolution system by using semantic knowledge sources in order to correctly resolve cases like in Example 28.

This paper is structured as follows. After presenting an error analysis of the coreference resolution system in Section 5.2, we analyse similar works to ours in which semantic knowledge sources have been used to improve coreference resolution in Section 5.3. Section 5.4 presents how we integrated the semantic knowledge in our system. The main experimental results are outlined in Section 5.5 and discussed in Section 5.6. Finally, we review the main conclusions and preview future work.

5.2 Error Analysis

A deep error-analysis can reveal the weak points of the coreference resolution system and help to decide future directions in the improvement of the system. The system we have evaluated is an adaptation of the Stanford Coreference resolution system (Lee *et al.*, 2013) to the Basque language. The Stanford coreference resolution module is a deterministic rule-based system which is based on ten independent coreference models or sieves that are precision-oriented, i.e., they are applied sequentially from highest to lowest precision. All the sieves of the system have been modified taking into account the characteristics of the Basque language and, one new sieve has been added, obtaining an end-to-end coreference resolution system.

The corpus used to carry out the error analysis is a part of EPEC (the Reference Corpus for the Processing of Basque) (Aduriz *et al.*, 2006). EPEC is a 300,000 word sample collection of news published in *Euskaldunon Egunkaria*, a Basque language newspaper. The part of the corpus we have used has about 45,000 words and it has been manually tagged at coreference level by two linguists (Ceberio *et al.*, 2016). First of all, automatically tagged mentions obtained by a mention detector (Soraluze *et al.*, 2016a) have been corrected; then, coreferent mentions have been linked in clusters.

More detailed information about the EPEC corpus can be found in Table 5.3.

	Words	Mentions	Clusters	Singletons
Devel	30434	8432	1313	4383
Test	15949	4360	621	2445

5.3 Table – EPEC corpus division information

5.2.1 Error types

The errors have been classified following the categorization presented in Kummerfeld and Klein (2013). The tool[1] presented in the paper has been used to help in identifying and quantifying the errors produced by the coreference resolution system:

- **Span Error (SE):** A mention span has been identified incorrectly.
- **Conflated Entities (CE):** Two entities have been unified creating a new incorrect one.
- **Extra Mention (EM):** An entity includes an incorrectly identified mention.
- **Extra Entity (EE):** An entity which consists of incorrectly identified mentions is outputted by the system.
- **Divided Entity (DE):** An entity has been divided in two entities.
- **Missing Mention (MM):** A not identified mention is missing in an entity.
- **Missing Entity (ME):** The system misses an entity which is present in the gold standard.

The error types are summarised in Table 5.4.

5.2.2 Error causes

Apart from classifying the errors committed by the coreference resolution system, it is important to observe the causes of these error types. These are the causes of errors we found:

- **Preprocessing (PP):** Errors in the preprocessing step (lemmatization, PoS tagging, etc.) provoke incorrect or missing links in coreference resolution.

[1] code.google.com/p/berkeley-coreference-analyser/

Error Type	System	Gold
Span Error	s_1	s_1 s_2
Conflated Entities	$\{m_1,\ m_2\}_{e1}$	$\{m_1\}_{e1}$
	-	$\{m_2\}_{e2}$
Extra Mention	$\{m_1,\ m_2\}$	$\{m_1\}$
Extra Entity	$\{m_1,\ m_2\}$	-
Divided Entity	$\{m_1\}_{e1}$	$\{m_1,\ m_2\}_{e1}$
	$\{m_2\}_{e2}$	-
Missing Mention	$\{m_1\}$	$\{m_1,\ m_2\}$
Missing Entity	-	$\{m_1,\ m_2\}$

5.4 Table – Error types. s=string, m=mention, e=entity

- **Mention Detection (MD):** These errors are provoked due to incorrectly identified (not a mention, incorrect boundaries..) or missed mentions during mention detection step. Missed mentions directly affect the recall of the system, and incorrectly identified mentions affect precision.

- **Pronominal Resolution (PR):** The system often generates incorrect links between the pronoun and its antecedent.

- **Ellipsis Resolution (ER):** Elliptical mentions do not provide much information as they omit the noun, as a consequence it is difficult to correctly link these types of mentions with their correct antecedent.

  For example, it is complicated to link the elliptical mention [Yosi Beilin Israelgo Justizia ministroak Jeruralemi buruz esandako-Ø[2]-ak] "what Yosi Beilin Israel Justice Minister said" with its antecedent [Beilin Justizia ministroaren hitzak] "Beilin Justice minister's words".

- **Semantic Knowledge (SK):** Errors related to a semantic relation (synonymy, hyperonymy, metonymy) between the heads of two mentions.

  For example, in mentions [Libanoko Parlamentuak] "Lebanon parliament" and [Libanoko Legebiltzarrak] "Lebanon parliament", *parlamentua* is a synomyn of *legebiltzarra*.

- **World Knowledge (WK):** In some cases the system is not able to link mentions as a consequence of the lack of world knowledge required to resolve them correctly.

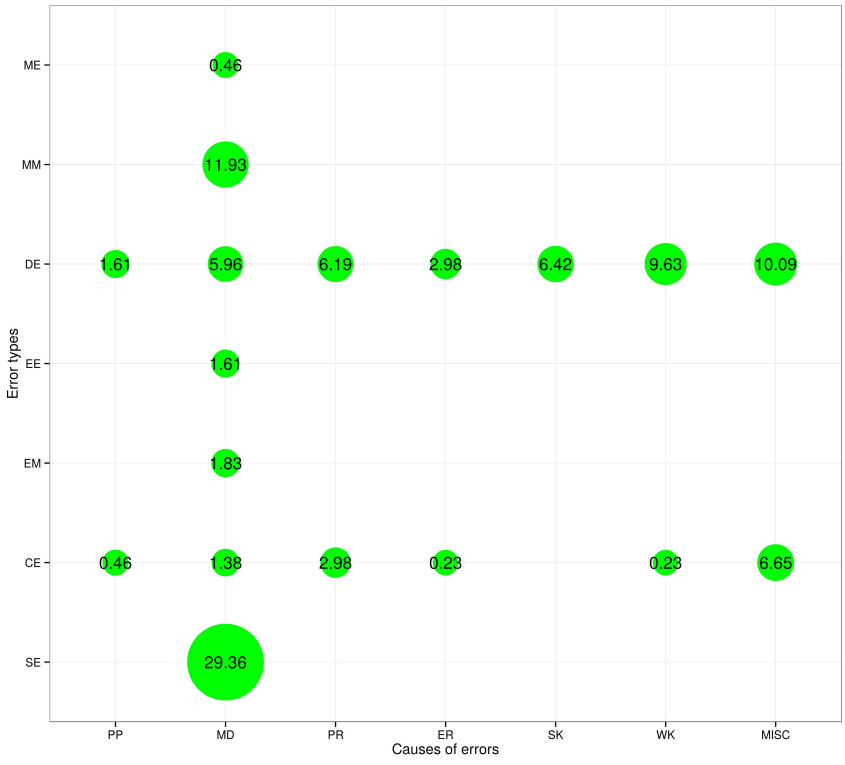
[2]In this case Ø refers to "what".

For example, to link the mention [Reala] "Reala" with the mention [talde txuri-urdinak] "white-blue team", it is necessary to know that *Reala* is a team and the nickname of the football team is *txuri-urdinak* "white-blue".

- **Miscellaneous (MISC):** In this category we classify the errors that are not contained in the above categories.

  An example of a miscellaneous error could be the following. The mention [Kelme, Euskaltel eta Lampre] should be linked with the mention [Hiru taldeak] "The three teams". In this specific example it is necessary to know that Kelme, Euskaltel and Lampre are teams and the enumerated mention has three elements.

After defining the error types and the error causes, we analysed how the error causes affect the error types in EPEC corpus. The distribution of errors is shown in Figure 5.1.



5.1 Figure – Distribution of error causes into error types.

As we observe in Figure 1, the most common errors types of the system fail in Span Error (29.36%), Conflated Entities (11.92%), Divided Entities (42.88%) and Missing Mention (11.92%) categories.

Observing the error causes, we can conclude that mention detection is crucial for coreference resolution, 52.52% of errors. Improving mention detection would likely improve the scores obtained in coreference resolution. Nevertheless, in order to identify deficiencies of a coreference resolution system, Pronominal Resolution (9.17%), Ellipsis Resolution (3.21%), Semantics (6.42%) and World Knowledge (9.86%) categories can reveal how the errors might be corrected. Due to the variety of errors classified in miscellaneous category, little improvement would be achieved despite making a big effort to solve them.

Among all the error causes, in this paper we are going to focus on errors provoked by the lack of semantic and world knowledge.

5.3 Related Work

Lexical and encyclopedic information sources, such as WordNet, Wikipedia, Yago or DBPedia have been widely used to improve coreference resolution.

WordNet (Fellbaum, 1998) is the one of oldest resources for lexical knowledge. It consists of *synsets*, which link synonymous word senses together. Using WordNet's structure, it is possible to find synonyms and hyperonymic relations. Wikipedia is a collaborative open source encyclopedia edited by volunteers and provides a very large domain-independent encyclopedic repository. Yago (Suchanek *et al.*, 2007) is a knowledge base, linking Wikipedia entries to the WordNet ontology. And finally, DBPedia (Mendes *et al.*, 2012) contains useful ontological information extracted from the data in Wikipedia.

Regarding works in which lexical and encyclopedic information sources have been exploited, Ponzetto and Strube (2006) were the earliest to use WordNet and Wikipedia.

Uryupina *et al.* (2011) extracted semantic compatibility and aliasing information from Wikipedia and Yago and incorporated it in coreference resolution system. They showed that using such knowledge with no disambiguation and filtering does not bring any improvement over the baseline, whereas a few very simple disambiguation and filtering techniques lead to better results. In the end, they improve their system's performance by 2-3 percentage points.

Rahman and Ng (2011) used Yago to inject knowledge attributes in mentions, but noticed that knowledge injection could be noisy.

Durrett and Klein (2013) observed that the semantic information contained even in a coreference corpus of thousands of documents is insufficient to generalize

to unseen data, so system designers have turned to external resources. Using specialised features, as well as WordNet-based hypernymy and synonymy and other resources, they obtained a gain from 60.06 in CoNLL score to 61.58 using automatic mentions, and from 75.08 to 76.68 with gold mentions.

Ratinov and Roth (2012) extract attributes from Wikipedia pages which they used to improve the recall in their system, based on a hybrid (Lee *et al.*, 2013).

In Hajishirzi *et al.* (2013) NECo, a new model for named entity linking and coreference resolution, which solves both problems jointly, reducing the errors of each is introduced. NECo extends the Stanford deterministic coreference resolution system by automatically linking mentions to Wikipedia and introducing new sieves which profit from information obtained by named entity linking.

As pointed out in Recasens *et al.* (2013), opaque mentions (mentions with very different words like *Google* and *the search giant*) account for 65% of the errors made by state-of-the-art systems, so to improve coreference scores beyond 60-70% it is necessary to make better use of semantic and world knowledge to deal with non-identical-string coreference. They use a corpus of comparable documents to extract aliases and they report that their method not only finds synonymy and instance relations, but also metonymic cases. They obtain a gain of 0.7% F1 score for the CoNLL metric using gold mentions.

Lee *et al.* (2013) mention that the biggest challenge in coreference resolution, accounting for 42% of errors in the state-of-the art Stanford system, is the inability to reason effectively about background semantic knowledge.

The intuition behind the work presented in Durrett and Klein (2014) is that named entity recognition on ambiguous instances can obtain benefit using coreference resolution, and similarly can benefit from Wikipedia knowledge. At the same time, coreference can profit from better named entity information.

5.4 Improving Coreference Resolution with Semantic Knowledge sources

This section explains the improvement process of the coreference resolution system with semantic knowledge sources. In order to treat cases where knowledge is needed, two new specialised sieves have been added to the coreference resolution system: One to extract knowledge from Wikipedia and the other to obtain semantic information from WordNet.

5.4.1 Enriching mentions with Named Entity Linking

Named Entity Linking is the task of matching mentions to corresponding entities in a knowledge base, such as Wikipedia.

As pointed out in Versley *et al.* (2016), named entity linking, or disambiguation of entity mentions, is beneficial to make full use of the information in Wikipedia.

The Basque version of Wikipedia, contained about 258,000 articles in September 2016, which is much smaller in size when compared with English Wikipedia, which contained about 5,250,837 pages on the same date.

In order to disambiguate and link mentions to Basque Wikipedia pages, the following formula has been applied to all the named entity mentions in a document:

$$P(s, c, e) = P\left(e \mid s\right) P\left(e \mid c\right)$$

$P\left(e \mid s\right)$ is the probability of being entity e given s string, i.e., the normalised probability of being entity e linked with string s in Wikipedia. $P\left(e \mid c\right)$ is the probability of being entity e given the context c. The context c is a window of size $[-50, +50]$ of the string s. To calculate $P\left(e \mid c\right)$ probability, UKB[3] software has been utilised. UKB software uses *Personalized Page Rank* algorithm presented in (Agirre and Soroa, 2009) and (Agirre *et al.*, 2014) to estimate the probabilities.

If a named-entity mention is linked with any page from Wikipedia, the page that UKB says it is the most probable is used to enrich the mention. From the Wikipedia page the following information is obtained:

- The title of the page. The title sometimes gives useful information. For example, for the named-entity mention *AEK*, the title of its Wikipedia page is *Alfabetatze Euskalduntze Koordinakundea* "Literacy and Euskaldunization Coordinator", where the extent of the acronym is obtained. Furthermore it gives the information that AEK is a coordinator, *koordinakundea*.

- The first sentence. The first paragraph of each Wikipedia article provides a very brief summary of the entity. Usually the most useful information is in the first sentence, this is where the entity is defined.

- If the Wikipedia page has an Infobox, we extract information from it. Infoboxes contain structured information in which the attributes of many entities are listed in a standardized way.

After the information is obtained from the Wikipedia page, this information is processed and the NPs are extracted.

[3] http://ixa2.si.ehu.es/ukb/

These NPs and their sub-phrases are used to enrich the mentions with world knowledge. To further reduce the noise, the NPs that are location named-entities in a Wikipedia page about a location are discarded.

Taking Example 28, the mention *Osasuna* is enriched as follows: The most probable Wikipedia page proposed by UKB for the mention *Osasuna* is *Osasuna futbol kluba* "Osasuna football club". Therefore, we obtain from this page the title, the first sentence and Infobox information. The NPs obtained after the information is processed are *gorritxoak* "the reds", *Osasuna futbol kluba* "Osasuna football club" and *Nafarroako futbol taldea* "football team from Navarre". So the mention *Osasuna* is enriched with the set of lemmas of the NPs and the lemmas of their sub-phrases: {gorritxo, Osasuna futbol klub, futbol klub, klub, Nafarroa futbol talde, futbol talde, talde} "{the reds, Osasuna football club, football club, club, football team from Navarre, football team, team}".

5.4.2 Wiki-alias sieve

The new Wiki-alias sieve uses the mentions enriched by information obtained from Wikipedia pages.

Using this information, the Wiki-alias sieve assumes that two mentions are coreferent if one of the two following conditions is fulfilled:

i) the set of enriched word lemmas in the potential antecedent has all the mention candidate's span lemmas. To better understand this constraint, suppose that the mention *Realak* is enriched with {talde, futbol talde, txuri-urdin} "{team, football team, white and blue}", as the potential antecedent *Realak* has all the lemmas in the mention candidate's span, i.e., *talde* "{team}" and *txuri-urdin* "{white and blue}", the mention *talde txuri-urdinak* "{white and blue team}" is considered coreferent of *Realak*.

ii) the head word lemma of the mention candidate is equal to the head word lemma of the potential antecedent or equal to any lemma in the set of enriched lemmas of the potential antecedent, and all the enriched lemmas of the potential antecedent appear in the cluster lemmas of the mention candidate. For example, this constraint considers coreferent the potential antecedent *Jacques Chiracek* and the mention candidate *Jacques Chirac Frantziako errepublikako presidentea*. After *Jacques Chiracek* mention has been enriched with lemmas {presidente, Frantzia presidente} "{president, France president}", the head word lemma of the mention candidate *presidente* is equal to a lemma in the set of enriched lemmas of the potential antecedent *presidente* and all the enriched lemmas of the potential antecedent appear in the cluster lemmas of the mention candidate, so the second constraint is fulfilled. This constraint aims to link coreferent mentions where a mention with novel information appears later in text than the less informative one.

As pointed out in Fox (1993), it is not common to introduce novel information in later mentions but it sometimes happens.

5.4.3 Synonymy sieve

To create this new sieve, we have extracted from Basque WordNet (Pociello *et al.*, 2011) all the words that are considered synonyms in this ontology. The Basque WordNet contains 32,456 synsets and 26,565 lemmas, and is complemented by a hand-tagged corpus comprising 59,968 annotations (Pociello *et al.*, 2011).

From all synsets, a static list of 16,771 sets of synonyms has been created and integrated in the coreference resolution system. Using the synonyms' static list, the *Synonymy sieve* considers two mentions as coreferent if the following constraints are fulfilled: i) the head word of the potential antecedent and the head word of the mention candidate are synonyms and ii) all the lemmas in the mention candidate's span are in the potential antecedent cluster word lemmas or *vice versa*. For example, the mention candidate *Libanoko legebiltzarra* "Lebanon parliament" and the *Libanoko parlametua* "Lebanon parliament" are considered coreferent as the head words *legebiltzarra* and *parlamentua* are synonyms and the lemma *Libano* "Lebanon" of the word *Libanoko* is present in the cluster word lemmas of the potential antecedent.

5.5 System evaluation

In order to quantify the impact of using semantic knowledge sources in coreference resolution, we have tested the enriched coreference resolution system using the EPEC corpus and compared the results with the baseline system. The experimentation has been carried out using automatic mentions and gold mentions. In both cases named entity disambiguation and entity linking has been performed automatically.

5.5.1 Metrics

The metrics used to evaluate the systems' performances are MUC (Vilain *et al.*, 1995), B^3 (Bagga and Baldwin, 1998), CEAF_e (Luo, 2005), CEAF_m (Luo, 2005), BLANC (Recasens and Hovy, 2011) and LEA (Moosavi and Strube, 2016). The CoNLL metrics is the arithmetic mean of MUC, B^3 and CEAF_e metrics. The scores have been calculated using the reference implementation of the CoNLL scorer (Pradhan *et al.*, 2014).

5.5.2 Experimental results

As pointed out in Rahman and Ng (2011), while different knowledge sources have been shown to be useful when applied in isolation to a coreference system, it is also interesting to observe if they offer complementary benefits and can therefore further improve a resolver when applied in combination. In order to quantify the individual improvement of each new sieve, we compared the baseline system (1) with the system in which the wiki-alias sieve has been added (2), with the one where the synonymy sieve has been added (3), and with the final system combining both sieves (4).

Table 5.5 shows the results obtained by the baseline system compared with those obtained by the coreference resolution system, which uses semantic knowledge sources. These scores are obtained with automatically detected mentions (F_1 =77.57).

	Automatic Mentions																		
	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
1	34.1	55.76	42.32	57.98	68.83	62.94	60.78	62.31	61.54	66.02	58.41	61.98	38.41	53.57	43.18	46.71	51.78	49.12	55.74
2	34.41	55.70	42.54	58.09	68.64	62.93	60.73	62.26	61.49	65.94	58.49	61.99	38.65	53.27	43.35	46.82	51.64	49.11	55.82
3	34.57	56.03	42.76	58.08	68.80	62.98	60.85	62.38	61.61	65.99	58.51	62.03	38.53	53.65	43.31	46.83	51.97	49.27	55.92
4	34.88	55.90	**42.95***	58.19	68.60	**62.97**	60.80	62.33	**61.56**	65.92	58.60	**62.04**	38.77	53.33	**43.48***	46.94	51.83	**49.26**	**55.98***

5.5 Table – Results obtained when automatic mentions are used. 1=Baseline, 2=1+Wiki sieve, 3=2+Synonymy sieve, 4=1+Wiki sieve+Synonymy sieve.

The scores obtained by systems using the gold mentions (F_1 =100), i.e., when providing all the correct mentions to the coreference resolution systems, are shown in Table 5.6.

5.6 Discussion

Observing the results presented in Table 5.5, we can see that the baseline system's F_1 scores are outperformed in all the metrics by the semantically enriched system. In CoNLL metric, the improved system has a score of 55.81, which is slightly higher than the baseline system, to be precise, 0.24 higher.

As shown in Table 5.6, the baseline F_1 scores are also outperformed in all the metrics, except in B^3 when gold mentions are used. The official CoNLL metric is improved by 0.39 points.

	Gold Mentions																		
	MUC			B^3			$CEAF_m$			$CEAF_e$			BLANC			LEA			CoNLL
	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
1	48.76	71.94	58.12	81.35	93.47	86.99	80.57	80.57	80.57	89.00	78.24	83.27	67.09	84.65	72.77	66.36	71.11	68.66	76.12
2	49.84	70.81	58.50	81.71	92.83	86.92	80.57	80.57	80.57	88.69	78.77	83.44	67.51	83.27	72.84	66.60	71.01	68.73	76.28
3	50.00	71.50	58.85	81.69	93.19	87.06	80.80	80.80	80.80	88.90	78.82	83.56	67.39	84.23	72.95	66.68	71.52	69.02	76.49
4	50.46	70.99	**58.99***	81.86	92.81	86.99	80.71	80.71	**80.71**	88.71	79.00	**83.57***	67.68	83.34	**73.00**	66.79	71.29	**68.97**	**76.51***

5.6 Table – Results obtained when gold mentions are used. 1=Baseline, 2=1+Wiki sieve, 3=2+Synonymy sieve, 4=1+Wiki sieve+Synonymy sieve.

Regarding recall and precision scores when automatic and gold mentions are used, all the metrics except CEAF$_e$ show an improvement in recall and decrease in precision when two new sieves are applied. The reason why the CEAF$_e$ metric is behaving differently could be that, as mentioned by Denis and Baldridge (2009), CEAF ignores all correct decisions of unaligned response entities. Consequently, the CEAF metric may lead to unreliable results.

It is interesting to compare the improvements obtained by the system which uses semantic knowledge sources in CoNLL scores. The improvement when automatic mentions are used is lower than when gold mentions are provided, 0.24 and 0.39 respectively. In both cases, even the improvements obtained are modest, they are statistically significant using Paired Student's t-test with p-value < 0.05.

As pointed out in Versley *et al.* (2016), in realistic settings, where the loss in precision would be amplified by the additional non-gold mentions, it is substantially harder to achieve gains by incorporate lexical and encyclopedic knowledge, but possible and necessary. A similar idea is concluded by Durrett and Klein (2013). They mention that despite the fact that absolute performance numbers are much higher on gold mentions and there is less room for improvement, the semantic features help much more than they do in system mentions.

To conclude the analysis of the results, it is also interesting to observe the difference between the results obtained by both systems when automatic mentions and when gold mentions are used. It is clear that having accurate preprocessing tools and a good mention detector are crucial to obtain good results in coreference resolution. In both systems the difference in CoNLL score is about 20.00 points higher when gold mentions are used.

The results obtained have enabled us to carry out a new error analysis in the development set. After applying the new two sieves, the error analysis has revealed five major issues that directly affect not obtaining bigger improvement when knowledge resources are used:

1. Some mentions do not have Wikipedia entry, as the coverage of Basque Wikipedia (257,546 pages) has less coverage than other languages, for example English (5,250,837 pages), i.e., Basque version is 21 times smaller.

2. Due to incorrect mention disambiguation, some mentions are linked to incorrect Wikipedia pages. The precision obtained in disambiguation is 87,84%.

3. Precision errors, provoked by cases where many proper noun mentions were potential antecedent for a common noun. For example, *Oslo* is linked by *hiriburu* "capital", nevertheless the correct antecedent for *hiriburu* is another capital that appears in text, in this specific case, *Jerusalem*.

4. Some indefinite mentions which do not have antecedent are linked incorrectly. For example, *estaturik* "state" is linked with *Frantziak* "France".

5. In the synonyms' static list, some synonyms that appear in texts are missing. In addition, many synonyms are so generic, i.e., they are synonyms depending on the context in which they appear. As a consequence of missing synonyms, some mentions with synonymy relations between them are not linked. The presence of very generic synonyms provokes to incorrectly link mentions that are not coreferent, so that precision decreases. Identifying the particular sense that a word has in context would likely help to improve the precision.

Regarding the issues that affect improvement of the systems when knowledge bases are used, Uryupina *et al.* (2011) suggest that in their particular case the errors introduced are not caused by any deficiencies in web knowledge bases, but reflect the complex nature of the coreference resolution task.

5.7 Conclusions and Future work

We have enriched the Basque coreference resolution adding new two sieves, *Wikialias* and *Synonymy sieve*, respectively. The first sieve uses the enriched information of named-entity mentions after they have been linked to their correspondent Wikipedia page, using Entity Linking techniques. The second sieve uses a static list of synonyms extracted from Basque WordNet to consider whether two mentions are coreferent.

Applying the two new sieves, the system obtains an improvement of 0.24 points in CoNLL F_1 when automatic mentions are used and the CoNLL score is outperformed by 0.39 points when the gold mentions are provided. The error analysis of the enriched system has revealed that the knowledge bases used, Basque Wikipedia and Basque WordNet, have deficiencies in their coverage compared with knowledge

bases in major languages, for example, English. We suggest that there is margin of improvement, as Basque Wikipedia and Basque WordNet coverage increase, bearing in mind that coreference resolution is a complex task.

As future work, we intend to improve the Pronoun resolution and Ellipsis Resolution, as we observed in the error analysis presented in Section 5.2 they are the cause of considerable coreference resolution errors, around % 12 of total errors.

6

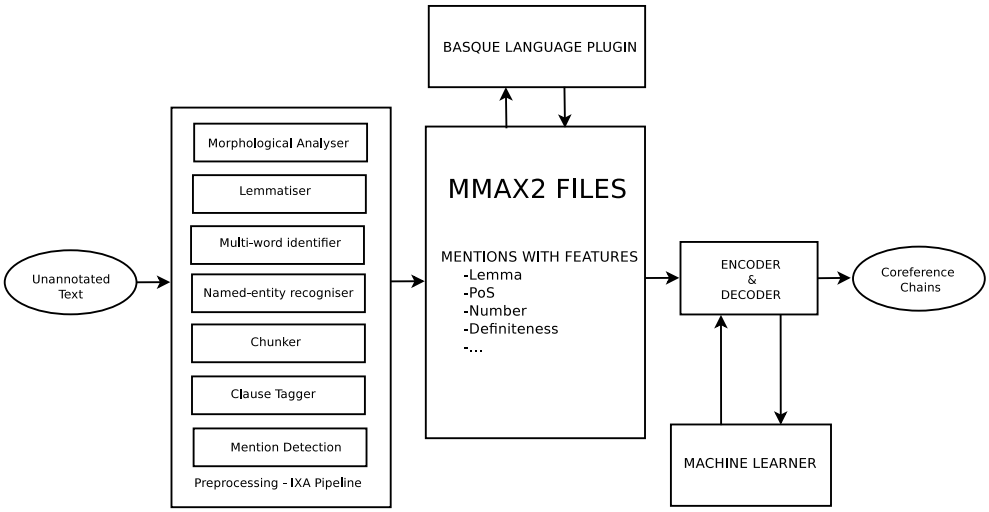
Coreference Resolution for the Basque Language with BART

6.0 Laburpena

Artikulu honetan ikasketa automatikoan oinarritzen den BART korreferentzia-ebazpenerako sistema euskarara egokitzeko egin dugun lehen hurbilpena aurkezten dugu.

Sistema ingeleserako sortua izan zenez, ez da gai zuzenean euskara bezalako hizkuntza batean korreferentzia-ebazpena gauzatzeko. Tesi-lan honetako 2. kapituluan azaldu dugun moduan, sistemak bost modulu nagusi ditu oinarrian eta horiez gain hizkuntzaren ezaugarriak definitzeko erabiltzen den *Language Plugin* deritzon modulua ere badu. Oinarrizko bost modulu horietako lau hizkuntzarekiko independenteak dira baina aurreprozesaketarako modulua ez dago ingelesa ez den beste hizkuntzak tratatzeko prestatua, hori dela eta, aipamenak kanpo-baliabideekin aurreprozesatuta eman behar zaizkio. Aurreprozesaketarako IXA taldean garatutako analisi-katea erabili da, analisi morfologikoa, analisi sintaktikoa, entitate izendunak eta *chunkak* lortuz. Gainera, aipamen-detekzioa egiteko euskararako garatu dugun aipamen-detektatzailea erabili da. Egokitutako sistemaren arkitektura 6.1 irudian ikus daiteke.

Aurreprozesaketako informazio guztia, MMAX2 anotazio tresnak erabiltzen duen formatuan gorde da, BART sistemak hortik lortzen baitu korreferentzia-ebazpena gauzatzeko behar duen informazio guztia tratatu beharreko hizkuntza ingelesa ez den kasuetan.



6.1 irudia – Euskarara egokitutako BART korreferentzia-ebazpenerako sistemaren arkitektura.

Aurreprozesaketa kanpo-baliabideekin egiteaz gain, *Language Plugin* modulua ere egokitu dugu. Modulu hori da hizkuntza baten ezaugarri bereziak definitzeko aukera eskaintzen duena. Hori dela eta, modulu horren barnean definitu ditugu euskaraz korreferentzia-ebazpena gauzatzeko garrantzitsuak diren ezaugarriak, hala nola, izenordainen zehaztapena.

Behin sistema euskara egokituta, EPEC corpusa erabiliz bi eredu ikasi ditugu. Lehenengoa oinarri-lerrotzat hartu dugu eta Soon *et al.* (2001) lanean oinarritutakoa da. Eredu hori aipamen-bikote eredua da, J48 sailkatzailea eta mota desberdinetako ezaugarriak (sintaktikoak, distantziak oinarritutakoak...) erabiltzen ditu korreferentzia-ebazpena gauzatzeko.

Bigarren eredua, euskararen ezaugarrietara hobekiago egokitzen den eredu hedatua da. Ikasketarako ezaugarri gehiago gehitu zaizkio, adibidez, *LemmaMatch* ezaugarri morfologikoa edota distantzian oinarritutako *DistanceSentence* eta *DistanceMarkable* ezaugarriak. Eredu bakoitzean erabiltzen diren ezaugarriak 6.1 taulan ikus daitezke.

Bi ereduen arteko konparaketa egiteko biak ebaluatu ditugu EPEC corpusaren testerako zatian. Ebaluazioak aipamen automatikoak eta urrezko aipamenak erabilita egin dira. Lortutako emaitzak 6.2 eta 6.3 tauletan ikus daitezke.

Ezaugarriak		Soon	Basque
Gender	m_i et m_j-k genero berdina	√	√
Number	m_i eta m_j-k numero berdina	√	√
Alias	m_j, m_i-ren aliasa	√	√
StringMatch	m_i eta m_j-ren arteko string parekaketa	√	√
SemClassAgree	m_i eta m_j-ren arteko bateragarritasun semantikoa	√	√
Appositive	m_i eta m_j aposizio egituran daude	√	√
DistanceSentence	m_i eta m_j-ren arteko distantzia esalditan	√	√
LemmaMatch	m_i eta m_j-k lema berdina	×	√
HeadMatch	m_i eta m_j-ren buruen string parekaketa	×	√
StringKernel	m_i eta m_j-ren stringen antzekotasuna	×	√
DistanceMarkable	m_i eta m_j-ren arteko distantzia aipamenetan	×	√
HeadPartofSpeech	m_i eta m_j-ren buruek kategoria berdina	×	√

6.1 taula – Esperimentuetan erabilitako bi ereduen ezaugarriak.

		R	P	F_1
Aipamen-detekzioa		72,91	74,69	73,79
MUC	Soon	18,37	67,23	28,86
	Basque	35,44	45,53	39,86
B^3	Soon	53,96	72,85	62,00
	Basque	58,10	65,27	61,48
$CEAF_m$	Soon	57,50	58,90	58,19
	Basque	58,67	60,10	59,38
$CEAF_e$	Soon	67,42	52,93	59,31
	Basque	61,63	58,15	59,84
BLANC	Soon	32,29	62,47	36,46
	Basque	38,70	48,81	42,41
CONLL	Soon	-	-	50,05
	Basque			53,72

6.2 taula – Aipamen automatikoekin lortutako emaitzak.

Emaitzek erakutsi digute euskararen ezaugarriak kontuan hartzen dituen ereduak nabarmenki hobetzen duela oinarri-lerroa metrika guztietan. Aipamen automatikoak erabiltzen diren kasuan oinarri-lerroa 3,67 hobetzen da CoNLL metrikan, 50,05 izatetik 53,72 izatera pasatzen baita. Urrezko aipamenak erabiltzean, aldiz, hobekuntza handiagoa da, 5,61 puntutakoa hain zuzen CoNLL metrikan, 66,81tik 72,42 igoz.

		R	P	F_1
Aipamen-detekzioa		100	100	100
MUC	Soon	23,62	78,66	36,34
	Basque	49,49	57,28	53,10
B^3	Soon	74,66	98,00	84,75
	Basque	81,21	87,78	84,37
$CEAF_m$	Soon	75,58	75,58	75,58
	Basque	76,59	76,59	76,59
$CEAF_e$	Soon	91,11	70,29	79,35
	Basque	82,10	77,64	79,81
BLANC	Soon	57,08	89,79	61,68
	Basque	66,78	75,99	70,34
CONLL	Soon	-	-	66,81
	Basque			72,42

6.3 taula – Urrezko aipamenekin lortutako emaitzak.

Aurreprozesaketak korreferentzia-ebazpenean duen eragina ere aztertu dugu, eredu bakoitzak aipamen automatikoekin eta urrezko aipamenekin lortzen dituen emaitzak konparatuz. Lehenenego ereduaren kasuan 16,76 puntu igotzen dira CoNLL neurrian urrezko aipamenak erabiltzen direnean. Bigarren ereduaren kasuan igoera 18,8 puntutakoa da. Igoera nabarmenak dira, aurreprozesaketa eta aipamen-detekzioa egoki egitearen onurak azaleratzen dituztenak.

Sistemaren gabeziez jabetzeko ebaluazio kualitatiboa egin da. Egindako errore-analisia azaltzen da, gaizki ebazten diren kasu nabarmenenak azalduz. Errore analisi horretatik ondorioztatuta, ikasketa automatikoa erabiltzeko euskara bezalako hizkuntza batean ager daitezkeen erronkak azaltzen ditugu, baliagarriak izan daitezkeen ezaugarrien inguruko hausnarketa planteatuz. Euskararen sistema nominalak generorik ez izatea, eta izenordainek bizidun/ez-bizidun propietateak ez izatea, izenordainen ebazpenerako distantzian oinarritutako ezaugarriak baliagarriak direla erakutsi digu hausnarketak, hala ere ezaugarri berrien azterketa sakonagoa egin beharra dago emaitza hobeak lortzeko.

Laburbilduz, BART ikasketa automatikoan oinarritutako korreferentzia-ebazpenerako sistema euskararako egokitu dugu. Horretarako aurreprozesaketa kanpo-baliabideekin egiteaz arduratu gara eta *Language Plugin* modulua egokitu dugu euskararen ezaugarriak erabili ahal izateko. Bi eredu ikasi eta

ebaluatu ditugu, oinarri-lerroa hobetzea lortuz eta ezaugarri morfologikoek eta distantzian oinarritutakoek euskara bezalako hizkuntza batean ikasketa automatikoa gauzatzeko orduan duten garrantziaz jabetuz.

Coreference Resolution for the Basque Language with BART

Ander Soraluze, Olatz Arregi, Xabier Arregi, Arantza Díaz de Ilarraza, Mijail Kabadjov and Massimo Poesio

Published in *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, 67–73, San Diego, California

Abstract

In this paper we present our work on Coreference Resolution in Basque, a unique language which poses interesting challenges for the problem of coreference. We explain how we extend the coreference resolution toolkit, BART, in order to enable it to process Basque. Then we run four different experiments showing both a significant improvement by extending a baseline feature set and the effect of calculating performance of hand-parsed mentions vs. automatically parsed mentions. Finally, we discuss some key characteristics of Basque which make it particularly challenging for coreference and draw a road map for future work.

6.1 Introduction

Basque is a language spoken by nearly three quarters of a million people, most of which live in the Basque country, a region spanning parts of northern Spain and southwestern France. One of the most surprising findings about the Basque language is that it cannot be linked with any of its Indo-European neighbours in Europe and, hence, has been classified as a language isolate. It differs considerably in grammar from the languages spoken in surrounding regions. It is an agglutinative, head-final, pro-drop, free-word order language (Laka, 1996).

Naturally, the Basque language has also inspired a lot of work in Computational Linguistics with tools for automatically processing it becoming increasingly available (Alegria *et al.*, 1996, 2002a, 2003; Aduriz *et al.*, 2003; Alegria *et al.*, 2008). However, as it is the case with most less-resourced languages, there are tools for the core processing levels, such as tokenisation, sentence splitting, morphological analysis, syntactic parsing/chunking, but much less so for higher semantic levels required in end goal applications such as Question Answering (Morton, 2000), Text

Summarisation (Steinberger *et al.*, 2007) or Information Extraction (MUC-6, 1995; MUC-7, 1998). One such intermediate problem which has been underresearched for Basque, and hence, no readily usable tools are publicly available yet, is that of Coreference Resolution (Poesio *et al.*, 2016).

However, preliminary work on Coreference for Basque is starting to emerge (Soraluze *et al.*, 2015a), and in this paper we describe our work on extending the coreference resolution toolkit, BART[1] Versley *et al.* (2008b) to the Basque language. BART benefits from an open architecture and provides a mechanism through language plugins which makes it particularly suitable for adaptations to new languages, and it attained good performance in the shared task on Multilingual Coreference at CoNLL 2012 (Uryupina *et al.*, 2012).

For our experiments we use the EPEC corpus annotated for coreference (Aduriz *et al.*, 2006) and we run experiments across two dimensions. First, we use a baseline model based on Soon *et al.* (2001) vs. a model that includes extra features reliably extracted for Basque with the tools at hand. Second, we measure performance on hand-parsed mentions vs. performance on automatically parsed mentions which illustrates the effect of pre-processing quality on the end results.

One of the key challenges that the Basque language introduces for Coreference is that it uses a genderless system for pronouns. In our experiments we look in more depth around this issue and show the challenges it presents as well as suggest viable solutions to model it with machine learning techniques.

The remainder of this paper is organised as follows: Section 6.2 briefly surveys related work, Section 6.3 gives details of EPEC, a coreference corpus, Section 6.4 describes the extension of BART to Basque, Section 6.5 presents results and provides a discussion on the challenges for coreference in Basque, and towards the end we draw conclusions and pointers to future work.

6.2 Related Work

Preliminary work on Coreference for Basque was done by Soraluze *et al.* (2015a) where they adapt the Stanford coreference resolution system (Lee *et al.*, 2013) to Basque. And there has been a lot of work on extending the BART coreference toolkit to languages other than English. Poesio *et al.* (2010) extend it to Italian using the Evalita corpus of Wikipedia articles (Broscheit *et al.*, 2010b) work on German using the TüBa-D/Z coreference corpus, Kopeć and Ogrodniczuk (2012) develop the Polish plug-in using a subset of the National Corpus of Polish, and finally Uryupina *et al.* (2012) run experiments on Arabic and Chinese.

[1] http://www.bart-coref.eu/

6.3 Annotated Corpus of Basque

EPEC (Reference Corpus for the Processing of Basque) (Aduriz *et al.*, 2006) is a 300,000 word sample collection of standard written Basque that has been manually annotated at different levels (morphology, surface syntax, phrases, etc.). The corpus is composed by news published in *Euskaldunon Egunkaria*, a Basque language newspaper. It is aimed to be a reference corpus for the development and improvement of several NLP tools for Basque.

Recently, mentions and coreference chains were also annotated by two linguists in a subset of the EPEC corpus which is composed of about 45,000 words. First, automatically annotated mentions obtained by our mention detector were corrected; then, coreferent mentions were linked in clusters. The mention detector is a set of hand-crafted rules that have been compiled into Finite State Transducers (FST). The FSTs match chunks and clauses provided by the preprocessing tools and identify the mentions and their boundaries. Further discussion about the FSTs' behaviour can be found in Soraluze *et al.* (2012).

All the annotation process has been carried out using the MMAX2 annotation tool Müller and Strube (2006). The coreference annotation of the EPEC corpus is explained more in detail in Ceberio *et al.* (2016).

To adapt BART to Basque, we divided the dataset into three main parts: one for training the system, the other for tuning, and the last for testing. More detailed information about the three parts can be found in Table 6.4.

	Words	Mentions	Clusters	Singletons
Train	23520	6525	1011	3401
Devel	6914	1907	302	982
Test	15949	4360	621	2445

6.4 Table – EPEC-coref corpus division information.

6.4 Extending BART to Basque

BART was originally created for English, but its flexible modular architecture ensures its portability to other languages.

BART consists of five main components: preprocessing pipeline, mention factory, feature extraction module, decoder and encoder. Furthermore, an additional independent *Language Plugin* module handles language specific information and is accessible from any component.

In the adaptation process of BART, we used a preprocessing pipeline of Basque linguistic processors, developed the *Basque Language Plugin* and added new features for coreference resolution specifically geared towards Basque.

6.4.1 Preprocessing and Mention Detection

The preprocessing pipeline takes raw texts and applies a series of Basque linguistic processors to analyse the texts: i) A morphological analyser that performs word segmentation and PoS tagging (Alegria *et al.*, 1996), ii) A lemmatiser that resolves the ambiguity caused at the previous phase (Alegria *et al.*, 2002a), iii) A multi-word item identifier that determines which groups of two or more words are to be considered multi-word expressions (Alegria *et al.*, 2004), iv) A named-entity recogniser that identifies and classifies named entities (person, organisation, location) in the text (Alegria *et al.*, 2003), v) A chunker, an analyser that identifies verbal and nominal chunks based on rule-based grammars (Aduriz *et al.*, 2003), vi) A clause tagger, that is, an analyser that identifies clauses, combining rule-based-grammars and machine learning techniques (Alegria *et al.*, 2008).

After the preprocessing step, mentions that are potential candidates to be part of coreference chains are identified using the mention detector explained in Section 6.3.

Finally, the linguistic information obtained by the preprocessing tools and the mentions identified by the mentions detector are stored in stand-off format of the MMAX2 annotation tool (Müller and Strube, 2006) that BART uses.

6.4.2 Basque Language Plugin

Developing a Basque language plugin for BART involved building on the system's already existing language plugins, and then translating closed-class words such as pronouns, mapping key part-of-speech tags and adapting lower-level heuristics for finding the head noun in noun phrases, person and number identification, as well as reading features made available by the preprocessing tools.

6.4.3 Feature engineering for Basque

Some kind of linguistic information from the mention is used by all the features implemented in BART. MentionFactory computes these properties when a language is supported by BART. In the case of a new language, such as Basque, they should be provided as part of the mention representation computed by external preprocessing facilities. So, we added in the MMAX2 files relevant features for coreference resolution in Basque, as are number and lemma.

Features		Baseline	Basque
Gender	M_i and M_j agree in gender	√	√
Number	M_i and M_j agree in number	√	√
Alias	Matches abbreviations and name variations	√	√
StringMatch	M_i and M_j have the same surface form	√	√
SemClassAgree	Assesses the semantic compatibility of M_i and M_j	√	√
Appositive	M_i and M_j are in apposition structure	√	√
DistanceSentence	Distance in sentences between M_i and M_j	√	√
LemmaMatch	M_i and M_j have the same surface lemma	×	√
HeadMatch	M_i and M_j have the same head	×	√
StringKernel	Computes the similarity M_i and M_j strings	×	√
DistanceMarkable	Distance in markables between M_i and M_j	×	√
HeadPartofSpeech	M_i and M_j head PoS are the same	×	√

6.5 Table – Features used for Coreference Resolution in our experiments. M_i is a candidate antecedent and M_j is a candidate anaphor.

For our experiments, we trained BART with two different models. The first one, is a simple model, presented by Soon *et al.* (2001).[2]

In the two models, gender agreement does not cause any improvement in the scores, as Basque is genderless. [3]

At this point the proposed new features to handle the specificity of Basque are not new and have also been used for other languages (see Poesio *et al.* (2016) for details).

6.5 Experimental Results

We have tested the two models presented in Subsection 6.4.3 in two different environments. In the first one automatically detected mentions are provided to the models and in the second one the mentions are gold.[4]

The metrics used in our evaluations are MUC (Vilain *et al.*, 1995), B^3 (Bagga and Baldwin, 1998), CEAF_e (Luo, 2005), CEAF_m (Luo, 2005), and BLANC (Recasens and Hovy, 2011). The scores have been calculated using the reference implementation of the CoNLL scorer (Pradhan *et al.*, 2014).

[2]Due to the way we integrated the preprocessing pipeline for Basque with BART, at this stage we were unable to incorporate all features in the original Soon *et al.* (2001) model. The second one, is an improved version of the first one where more Basque oriented features have been added. The features used in each model are presented in Table 6.5.

[3]We maintain this feature with the aim of not modifying the Soon *et al.* (2001) model.

[4]Since the official CoNLL scorer is used for the evaluation, it also takes care of the alignment between automatically detected mentions and gold ones.

Table 6.6 presents the results obtained by the two models when automatic mentions are used.

		R	P	F_1
Mention Detection		72.91	74.69	73.79
MUC	Soon	18.37	67.23	28.86
	Basque	35.44	45.53	39.86
B^3	Soon	53.96	72.85	62.00
	Basque	58.10	65.27	61.48
$CEAF_m$	Soon	57.50	58.90	58.19
	Basque	58.67	60.10	59.38
$CEAF_e$	Soon	67.42	52.93	59.31
	Basque	61.63	58.15	59.84
BLANC	Soon	32.29	62.47	36.46
	Basque	38.70	48.81	42.41
CONLL	Soon	-	-	50.05
	Basque			53.72

6.6 Table – Scores with automatic mentions.

In the case of automatically detected mentions, Basque model outperforms the Soon baseline model according to F_1 on all the metrics except B^3. In CoNLL metric, Basque model has a score of 53.72, which is 3.67 points higher than Soon Baseline, which scores 50.05.[5]

Scores obtained when gold mentions are provided are shown in Table 6.7.

When gold mentions are used the Basque model also outperforms the Soon baseline according to all the metrics, except B^3. The official CoNLL metric is outperformed by 5.61 points.

Comparing the results obtained when gold mentions are used with those obtained with the automatic mentions, there is a considerable difference. CoNLL F_1 of Soon baseline is 50.05 when automatic mentions are provided, while providing gold mentions this value raises to 66.81, an increase of 16.76. Similar increase in CoNLL F_1 happens with the Basque model. In this case, there is an increase of 18.7 points, from 53.72 with automatic mentions to 72.42 when gold mentions are used.

We also had a look at the pronoun resolution performance alone, but only MUC scores on automatic mentions as the CoNLL scorer does not provide a break-down of scores per anaphor type, and there was a small gain in performance from the Soon baseline to the Basque model from $F1 = 27.4$ to $F1 = 33.0$, respectively. The gain is due mostly to higher precision, suggesting the additional features in

[5]The CoNLL metric is the arithmetic mean of MUC, B^3 and $CEAF_e$ metrics.

		R	P	F_1
Mention Detection		100	100	100
MUC	Soon	23.62	78.66	36.34
	Basque	49.49	57.28	53.10
B^3	Soon	74.66	98.00	84.75
	Basque	81.21	87.78	84.37
$CEAF_m$	Soon	75.58	75.58	75.58
	Basque	76.59	76.59	76.59
$CEAF_e$	Soon	91.11	70.29	79.35
	Basque	82.10	77.64	79.81
BLANC	Soon	57.08	89.79	61.68
	Basque	66.78	75.99	70.34
CONLL	Soon	-	-	66.81
	Basque			72.42

6.7 Table – Scores with gold mentions.

the Basque model help discriminate better erroneously resolved pronouns in the baseline model, however, more work will need to be devoted to improving recall, which is particularly challenging in the case of Basque due to the lack of gender in the Basque pronoun system.

6.5.1 Error Analysis

In our error analysis we had a look at examples from our corpus covering the following four cases:

Case a. There were errors in the coreference resolution due to errors in the preprocessing which were propagated across the pipeline. Consider example 6.1, for instance:[6]

(6.1) **Gold mentions**: [Del Bosquek] prentsaurrekoa eman zuen atzo. [Vicente Del Bosque], [Real Madrileko entrenatzailea] , nahikoa kezkati azaldu zen.
Automatic mentions: [Del Bosquek] prentsaurrekoa eman zuen atzo. [Vicente Del Bosque , Real Madrileko entrenatzailea] , nahikoa kezkati azaldu zen.

Case b. Due to the challenges posed by the genderless pronoun system in Basque, there were pronouns easy to resolve in relative terms which were missed

[6]English translation: "[Del Bosque] gave a press conference yesterday. [Vicente Del Bosque], [Real Madrid coach], appeared quite concerned".

or incorrectly resolved. Example 6.2 illustrates this:[7]

(6.2) Lehendakari hautatu zutenetik, [Djukanovicek] aldaketa handia eman dio [bere] ildo politikoari.

Case c. Here with example 6.3 we illustrate an instance of a challenging cases of pronouns which are currently beyond the scope of our approach:[8]

(6.3) Gobernuaren bilera honen ondoren, oporretara joango da [Jospin], eta hauek baliatuko ditu, ziur aski, Chevenement kasuaz gogoetak egiteko eta konponbide batekin [bere] jarduerari eusteko.

In this example it is more challenging to resolve correctly the pronoun [bere] "[his]" as [bere] can refer to Jospin or to Chevenement.

Case d. Finally, with example 6.4 we show an instance of a correctly resolved pronoun by our system:[9]

(6.4) "[Guk] ez dugu inoiz penaltietan irabazi." Luzapena golik gabe amaitzean, itzal beltz batek estali zuen Arena estadioa . Rijkaard-ek esana zuen arreta bereziz prestatu zituztela penaltiak, "[gure] istoria ez errepikatzeko".

6.5.2 Discussion

Taking into consideration Basque most relevant grammatical characteristics, in some aspects it is more challenging to resolve coreferences in this language than in others.

Since Basque is an agglutinative language, a given lemma takes many different word forms, depending on the case (genitive, locative, etc.) or the number (singular, plural, indefinite) for nouns and adjectives. For example, the lemma lehendakari ("president") forms the inflections lehendakaria ("the president"), lehendakariak ("the president"), lehendakariari ("to the president"), lehendakariei ("to the presidents"), lehendakariaren ("of the president"), etc. This means that looking only for the given exact word, is not enough for Basque to resolve coreference when

[7] English translation: "Since he was elected as president, [Djukanovic] has greatly changed [his] policy lines".

[8] English translation: "After this government meeting, [Jospin] will go on holidays, and will surely use it to reflect on Chevenement case and to maintain [his] activity with a new solution".

[9] English translation: "[We] have never won on penalties." After the extension finished without goals, a large shadow turn off the stadium. Rijkaard said they prepared penalties with great attention,"so that [our] story would not occur again".

string matching techniques are applied and as we observed in our experiments the use of lemmas is more effective in morphologically rich languages.

Besides the agglutination, there is no grammatical gender in the nominal system. Nouns and adjectives have no distinct endings depending on gender. In addition, there are no distinct forms for third person pronouns in Basque, and demonstratives are used as third person pronominals (Laka, 1996).

This makes it impossible to use gender as a feature in the resolution process which has been proven particularly useful in the resolution of pronouns, for example. Furthermore, the animacy feature cannot be used for pronoun resolution either. In this scenario, distance-based features, like Sentence Distance and Markable distance could be the most effective features for pronoun resolution. Nevertheless, research will have to be devoted to finding other useful features to make up for the lack of gender and animacy.

6.6 Conclusions and future work

In this paper we presented our ongoing work on Coreference Resolution in Basque. We described the main resource we have been using which is the EPEC corpus annotated with coreferences and we explained how we have been adapting the coreference resolution toolkit, BART, to enable it to process Basque. We ran two levels of experiments one resolving coreferences using the gold mentions and one using automatically parsed mentions and we trained two different models for each, a baseline model based on Soon *et al.* (2001) and a Basque model with extended feature set. We showed that the Basque model significantly outperforms the baseline. We also discussed key characteristics of the Basque language which make it particularly challenging for coreference.

Next we plan to investigate more in depth suitable features that can both make up for the lack of gender and animacy and be extracted reliably from unrestricted text. We also plan to run an extrinsic evaluation guaging the effect of coreference on a higher level task.

7

Improving BART coreference resolution system for Basque with semantic knowledge

7.0 Laburpena

Artikulu honetan euskararako egokitutako BART korreferentzia-ebazpenerako sistemaren hobekuntza aurkezten dugu, horretarako Wikipediatik erauzitako ezagutza erabilita.

Oinarritzat 6. kapituluan aurkezten den euskarara egokitutako sistema, Soon *et al.* (2001) lanean oinarritutakoa, hartzen da. 16 ezaugarri erabiliz entropia maximoan oinarritutako sailkatzaile bat ikasten dugu EPEC corpusa erabiliz eta ikasketatik lortutako eredua oinarri-lerrotzat finkatzen da.

Ondoren, BART sistemari ezagutza entziklopedikoa txertatzeko jarraitu dugun metodologia azaltzen da. 2107ko euskarazko wikipediaren erauzketatik lortutako 263.316 artikuluak erabiliz, ikasketa garaian erabiltzeko hiru ezaugarri semantiko berri sortu dira eta horiek sisteman integratzeko beharrezko den inplementazioa gauzatu dugu. Erauzketa SQL formatuan egiten da, beraz, informazio hori biltegiratzeko datu-base erlazional bat sortu behar izan da. Sortutako ezaugarri bakoitzak erabiltzen duen informazioa datu-base erlazionaleko taula batean gordetzen da.

Lehenengo ezaugarriak, *WIKI_ALIAS* deiturikoak, wikipediako orrietan aurkitzen diren *piped link* izeneko egitura bereziak erabiliz lortutako entitateen izen-laburdurak, pseudonimoak eta laburtzapenak erabiltzen ditu. Informazio hori erabiliz gai da, adibidez, *Ibarretxe* izen laburtuaren jatorriko izena *Juan Jose Ibarretxe* dela jakiteko edota *PP* laburtzapenaren euskaraz-

ko hedapena *Alderdi Popularra* dela. Guztira 93.240 sarrera dituen taula sortu da datu-basean. Adibide moduan ikus 7.1 taula.

Aliasa	Orria
Ibarretxe	Juan Jose Ibarretxe
Joseba Irazu	Bernardo Atxaga
Alderdi Popularra	PP

7.1 taula – Piped link egiturak erabiliz Wikipediatik lortutako adibide batzuk.

Bigarren ezaugarriak *WIKI_REDIR* du izena eta Wikipediako orrialdeek erabiltzailearentzako modu gardenean egiten den orrialdeen arteko berbideraketatik lortutako informazioa erabiltzen du oinarrian. Wikipediako orrialde hauen helburua, erabiltzailea orrialde egokira bideratzea da. Horretarako, akronimoen hedapena eginez, idazketa erroreen zuzenketa proposatuz eta anbiguotasunik ez duten orrialdeak eskainiz. Berbideraketa horiek oso informazio erabilgarria eskaintzen digute zenbait korreferentzia-erlazio ebazteko. Adibidez, *Osasuna* bilaketa sarrerak *Osasuna futbol kluba* wikipediako orrialdera berbideratzen gaitu. Proposatutako berbideraketa berri horretatik, adibidez, *Osasuna* futbol klub bat dela lor dezakegu. Informazio hori gero, *Osasuna* eta *futbol kluba* moduko aipamenak korreferentzia-kate berean biltzeko oso baliagarria izango da. 118.131 sarrera dituen taula sortu da. Adibide moduan ikus 7.2 taula.

Aliasa	Orria
EEBB	Amerikako Estatu Batuak
Osasuna	Osasuna futbol kluba
Buda	Siddahata Gautama

7.2 taula – Berbideraketetatik lortutako adibideak.

Azken ezaugarriak, *WIKI_LISTS* deiturikoak alegia, Wikipediak eskaintzen dituen orrialde berezi batzuetatik lortutako informazioa erabiltzen du. Wikipedian badira gai konkretu bateko orrialdeen zerrenda biltzen dituzten orriak. Adibidez, *Euskal Herriko futbolarien zerrenda* orrialdeak Euskal Herriko futbolariak biltzen ditu eta futbolari bakoitzari dagokion orrialderako esteka gordetzen du. Horretarako, zerrenda-orrialde bakoitzean aurkitzen

den sarrerak hartu, eta sarrera bakoitzari zerrendak definitzen duen ezaugarria gehitu diogu. Adibidez, *Imanol Agirretxe* sarrera *Euskal Herriko futbolarien zerrenda* orrialdean aurkitzen denez, *Imanol Agirretxe futbolari* eta *euskal herriko futbolari* ezaugarriez aberastu da. 20.276 sarrera dituen taula sortu da. Adibide moduan ikus 7.3. taula.

Izena	Ezaugarriak
Imanol Agirretxe	futbolari#Euskal Herri futbolari
Txomin Agirre	idazle#Bizkaia idazle#euskaltzain
Izurde muturluze	espezie#balea espezie
Gregorio Ibarretxe	alkate#Bilbo alkate#ospetsu

7.3 taula – Zerrenda orrialdetatik lortutako adibideak.

Hiru ezaugarriak lortu eta inplementatu ondoren, oinarri-lerroko eta sistema aberastuaren ebaluazioa aurkezten da. Aipamen automatikoekin lortutako emaitzak 7.4 taulan ikus daitezke, eta urrezko aipamenekin lortutakoak, berriz, 7.5 taulan

	Aipamen automatikoak							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	LEA	CoNLL
B	73,79	40,45	61,49	59,54	59,80	43,17	45,40	53,91
B+Wiki	73,79	**40,98**	**61,66**	**59,82**	**60,00**	**43,48**	**45,97***	**54,21**

7.4 taula – Aipamen automatikoak erabiliz lortutako emaitzak. $*$ duten balioak esanguratsuak dira Paired Student's t-testarekin. p-value $< 0{,}05$

	Urrezko aipamenak							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	LEA	CoNLL
B	100	54,32	85,04	77,83	80,81	71,63	63,90	73,39
B+Wiki	100	**55,38***	**85,30***	**78,17**	**81,14**	**72,07***	**64,74***	**73,94***

7.5 taula – Urrezko aipamenak erabiltzean lortutako emaitzak. $*$ duten balioak esanguratsuak dira Paired Student's t-testarekin. p-value $< 0{,}05$

Ebaluazioan lortutako emaitzak aztertuz, metrika guztietan ezagutza entziklopedikoa erabiltzean oinarri-lerroko sistema hobetzea lortu dugula ikusten da. CoNLL metrika 0,3 puntu hobetu dugu aipamen automatikoak erabiltzean eta 0,55 puntu urrezko aipamenekin. Hizkuntza gutxitu batean ere

baliabide semantikoak erabiliz hobekuntzak lortzea posible dela erakutsi dugu, nahiz eta hizkuntza horietako baliabideen estaldura, hizkuntza nagusiagoetakoek dutena baino txikiagoa izan.

Laburbilduz, euskarako egokitutako BART sistema Wikipediatik erauzitako hiru ezaugarri semantiko berrirekin hobetzea lortu dugu. CoNLL metrika 0,3 puntu hobetu da aipamen automatikoak erabiltzean eta 0,55 puntu urrezko aipamenekin.

Improving BART coreference resolution system for Basque with semantic knowledge

Ander Soraluze, Olatz Arregi, Xabier Arregi and Arantza Díaz de Ilarraza

Submitted to *Natural Language Engineering: Special Issue on Knowledge-Rich Coreference Resolution*

Abstract

In this paper we present an improvement of the BART coreference resolution system for Basque using semantic knowledge extracted from Wikipedia. Precisely, three new semantic features have been extracted and integrated in the improved system and, using Maximum Entropy model for learning, we show that knowledge extracted from Wikipedia can improve coreference resolution system results, even when the sources used are for under-resourced language and have less coverage compared with knowledge bases in major languages, for example, English.

7.1 Introduction

Coreference resolution consists of identifying textual expressions (mentions) that refer to real-world objects (entities) and of determining whether these mentions refer to the same entity. It is well known that coreference resolution is essential in Natural Language Processing applications, where a higher accuracy in discourse analysis leads to better performance. Information extraction, question answering, machine translation, sentiment analysis, machine reading, text summarisation, and text simplification are some of the tasks that can benefit from coreference resolution.

In Soraluze *et al.* (2016b) we presented a preliminary work in the adaptation process of BART coreference resolution system to Basque.

Since Basque is an agglutinative language, a given lemma takes many different word forms, depending on the case (genitive, locative, etc.) or the number (singular, plural, indefinite) for nouns and adjectives. For example, the lemma *lehendakari* "president" forms the inflections *lehendakaria* "the president", *lehendakariak* "the president", *lehendakariari* "to the president", *lehendakariei* "to the

presidents", *lehendakariaren* "of the president", etc. This means that looking only for the exact given word is not enough for Basque to resolve coreference when string matching techniques are applied and, as we observed in our experiments, the use of lemmas is more effective in morphologically rich languages.

Besides the agglutination, there is no grammatical gender in the nominal system. Nouns and adjectives have no different endings depending on gender. In addition, there are no distinct forms for third person pronouns in Basque, and demonstratives are used as third person pronominals (Laka, 1996).

This makes it impossible to use gender as a feature in the resolution process, which has been proven particularly useful in the resolution of pronouns, for example. Furthermore, the animacy feature cannot be used for pronoun resolution either because Basque pronouns lack animacy.

In resolving coreference cases in which semantic or world knowledge is needed, semantic knowledge sources, specifically Wikipedia and WordNet, have less coverage in Basque language compared with knowledge bases in major languages, for example, English.

Taking everything into consideration, in some aspects it is more challenging to resolve coreferences in Basque than in other languages which have more resources.

The paper is structured as follows. After reviewing related work in Section 7.2, we describe BART coreference resolution system adapted for Basque in Section 7.3. Section 7.4 explains the methodology used to extract three new semantic features using Wikipedia. In order to observe how the effect of incorporating semantic knowledge affects the performance of the coreference resolution, we evaluate, in Section 7.5, the system before and after the new semantic features are applied in two scenarios: handparsed mentions vs. automatically detected mentions. The experimental results are presented in Section 7.6 and are discussed in Section 7.7. Finally, conclusions and future work are presented.

7.2 Related Work

Lexical and encyclopedic information sources, such as Wikipedia, have been widely used to improve coreference resolution.

Regarding works in which lexical and encyclopedic information sources have been exploited, Ponzetto and Strube (2006) were the earliest to use Wikipedia. By means of WikiRelate! method presented in Strube and Ponzetto (2006), they compute Wikipedia relatedness of mention pairs using the system of categories in Wikipedia as a semantic network. The authors demonstrate that including semantic relatedness scores in a coreference resolution system the results of the system improve.

Uryupina *et al.* (2011) extracted semantic compatibility and aliasing information from Wikipedia and Yago and incorporate it in coreference resolution system. They showed that using such knowledge with no disambiguation and filtering does not bring any improvement over the baseline, whereas a few very simple disambiguation and filtering techniques lead to better results. In the end, they improve their system's performance by 2-3 percentage points.

Ratinov and Roth (2012) extracted attributes from Wikipedia pages which they used to improve the recall in their system, based on a hybrid Lee *et al.* (2013) system.

In Hajishirzi *et al.* (2013), a new model for named entity linking and coreference resolution, which solves both problems jointly, reducing the errors of each is introduced. Their system automatically links mentions to Wikipedia and profits from information obtained by named entity linking.

Durrett and Klein (2013) observed that even the semantic information contained in a coreference corpus of thousands of documents is insufficient to generalize to unseen data, so system designers have turned to external resources. The intuition behind the work they presented in Durrett and Klein (2014) is that named entity recognition on ambiguous instances can obtain benefit for coreference resolution, and similarly can benefit from Wikipedia knowledge.

Referring to the use of BART coreference resolution toolkit in languages other that English, Poesio *et al.* (2010) should be mentioned, the authors extended it to Italian using the Evalita corpus of Wikipedia articles, Broscheit *et al.* (2010b) worked on German using the TüBa-D/Z coreference corpus. Besides, Kopeć and Ogrodniczuk (2012) developed the Polish language plug-in using a subset of the National Corpus of Polish, Uryupina *et al.* (2012) ran experiments on Arabic and Chinese and recently it has been adapted for Indian languages (Sikdar *et al.*, 2016).

7.3 Coreference resolution system

BART is a platform for integrating different state-of-the-art approaches to coreference resolution including the use of external knowledge resources such as Wikipedia and WordNet (Versley *et al.*, 2008b)

It consists of five main components: i) preprocessing pipeline, ii) mention factory, which creates mention objects based on the information of the MMAX files (Müller and Strube, 2006), iii) feature extraction module, iv) decoder and iv) encoder. Furthermore, an additional independent *Language Plugin* module handles language specific information and is accessible from any component. BART benefits from an open architecture and provides a mechanism through language plugins which makes it particularly suitable for adaptations to new languages.

In the adaptation process of BART, we developed the *Basque Language Plugin* and used a preprocessing pipeline of Basque linguistic processors explained in 7.3.2 to obtain the information needed for coreference resolution.

7.3.1 Corpus Used

The Reference Corpus for the Processing of Basque, EPEC (Aduriz, Aranzabe, Arriola, Atutxa, Díaz de Ilarraza, Ezeiza, Gojenola, Oronoz, Soroa and Urizar 2006) is a 300,000-word sample collection of standard written Basque that has been manually annotated at different levels (morphology, surface syntax, phrases, etc.). The corpus is composed of news published in *Euskaldunon Egunkaria*, a Basque language newspaper. It is aimed to be a reference corpus for the development and improvement of several NLP tools for Basque.

Recently, mentions and coreference chains were also annotated by two linguists in a subset of the EPEC corpus, which is composed of about 45,000 words. All the annotation process has been carried out using the MMAX2 annotation tool (Müller and Strube, 2006). The coreference annotation of the EPEC corpus is explained more in detail in Ceberio *et al.* (2016).

To adapt BART to Basque, we divided the dataset into three main parts: one for training the system, the other for tuning, and the last for testing. More detailed information about the three parts can be found in Table 7.6.

	Words	Mentions	Clusters	Singletons
Train	23520	6525	1011	3401
Devel	6914	1907	302	982
Test	15949	4360	621	2445

7.6 Table – EPEC-coref corpus division information.

7.3.2 Mention Detection Using External Preprocessing

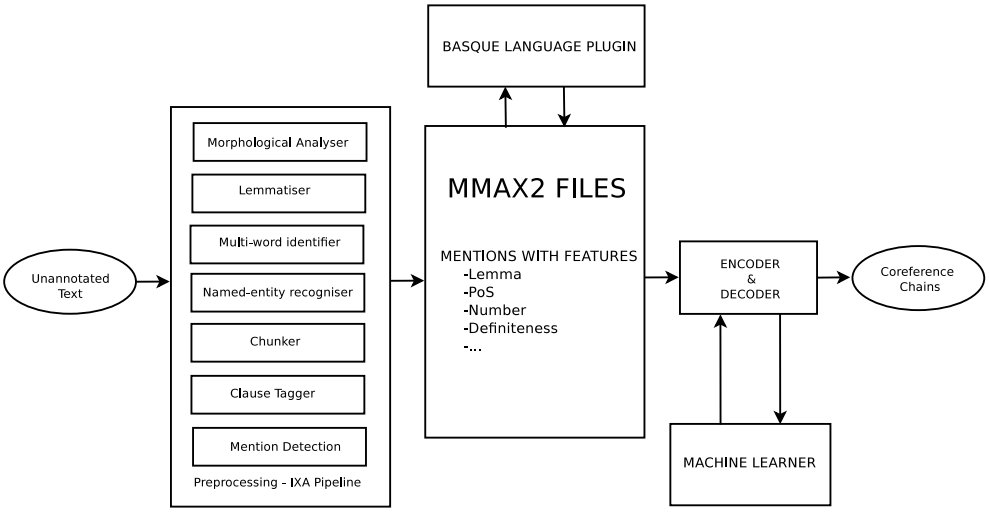
Some kind of linguistic information from the mention is used by all the features implemented in BART. MentionFactory computes these properties when a language is supported by BART. In the case of a new language, such as Basque, they should be provided as part of the mention representation computed by external preprocessing facilities.

We have used the following external preprocessing tools integrated in a pipeline for Basque language processing i) A morphological analyser that performs word segmentation and PoS tagging (Alegria *et al.*, 1996), ii) A lemmatiser that also disambiguates the PoS and the syntactic function (Alegria *et al.*, 2002a), iii) A

multi-word item identifier that determines which groups of two or more words are to be considered multi-word expressions (Alegria *et al.*, 2004), iv) A named-entity recogniser that identifies and classifies named entities (person, organization, location) in the text (Alegria *et al.*, 2003), v) A chunker (Aduriz and Díaz de Ilarraza, 2003), an analyser that identifies verbal and nominal chunks based on rule-based grammars, v) A clause tagger (Arrieta, 2010), that is, an analyser that identifies clauses, combining rule based-grammars and machine learning techniques (Alegria *et al.*, 2008) and vi) A mention detector that identifies mention candidates based on finite-state technology (Soraluze *et al.*, 2016a).

All the information obtained by the above preprocessing tools has been stored in MMAX2 files which are relevant features for coreference resolution in Basque, from which BART takes the information needed to perform coreference resolution.

Figure 7.1 shows the pipeline that takes unannotated Basque written texts and outputs coreference chains.



7.1 Figure – Coreference resolution system pipeline.

7.3.3 Features

The baseline system is a reimplementation of the Soon *et al.* (2001) system, presented in Soraluze *et al.* (2016b). In this version the Gender feature has been removed as it does not contribute to improving the results because Basque, as said, has a genderless nominal system.

The following are the features computed by the baseline system given a candidate antecedent m_i and candidate mention m_j.

(a) Lexical features

1. **STRING_MATCH** True if m_i and m_j have the same surface form; else False.
2. **ALIAS** True if m_j is an alias of m_i or vice versa; else False.
3. **LEMMA_MATCH** True if m_i and m_j have the same surface lemma; else False.
4. **HEAD_MATCH** True if m_i and m_j have the same head lemma; else False.
5. **STRING_KERNEL** The similarity value of m_i and m_j strings.

(b) Grammatical features

6. **MJ_IS_DEFINITE** True if m_j is definite. else False.
7. **MJ_IS_DEMONSTRATIVE** True if m_j is demonstrative. else False.
8. **MJ_IS_PRONOUN)** True if m_j is a pronoun. else False.
9. **MI_IS_PRONOUN** True if m_i is a pronoun. else False.
10. **NUMBER** True if m_i and m_j agree in number. else False.
11. **APPOSITIVE** True if m_i and m_j are in apposition structure; else False.
12. **PROPER_NAME** True if both m_i and m_j are proper names; else False.
13. **HEAD_POS** True if m_i's and m_j's head word have same PoS tag; else False.

(c) Distance features

14. **SENTENCE_DISTANCE** Distance in sentences betwen m_i and m_j.
15. **MARKABLE_DISTANCE** Distance in markables between m_i and m_j.

(d) Semantic features

16. **SEM_CLASS** True if m_i and m_j have same named-entity type; else False.

7.4 Incorporating Semantic Knowledge to the Baseline System

In order to incorporate semantic knowledge to BART system, we obtained Basque Wikipedia dumps from January 2017 in SQL format.[1] In total, this version of Basque Wikipedia contains 263,316 articles.

Following the strategies presented in Poesio *et al.* (2007), we have adapted three semantic features named **WIKI_ALIAS**, **WIKI_REDIR** and **WIKI_LIST** for Basque, and they have been added to the baseline feature set. The following subsection explains the procedure followed in the implementation process of these new features.

7.4.1 Semantic Features Extracted From Wikipedia

1. Wikipedia pages contain very useful information about entities, including shortened names, pseudonyms or abbreviations. Among different types of hyperlinks that the pages have, "piped links" contain the article that the text links to, followed by the visible text of the link, also known as alias, separated by the pipe character (|).

   For example, the following original text in Wikipedia, *Ibarretxe Eusko Jaurlaritzako 5. lehendakaria izan zen.* "Ibarretxe was the fith president of Basque Government" contains [[*Juan Jose Ibarretxe*|*Ibarretxe*]] piped link in its source file and the text "Ibarretxe" links to the article "Juan Jose Ibarretxe".

   ```
   Original text: Ibarretxe Eusko Jaurlaritzako 5. lehendakaria
       izan zen.
   Source text:[[Juan Jose Ibarretxe|Ibarretxe]] Eusko
      Jaurlaritzako 5. lehendakaria izan zen.
   ```

   In order to use the information that piped links provide, we have created a database in which each entry contains the text of the link and the article that the text links to. Table 7.7 shows some examples of information obtained from Basque Wikipedia piped links.

   In some cases, aliases links to more than one page and it is neccesary to take into account how often each string links to a given article. Consequently, a numeric weight associated with each article is calculated by the following formula:

[1] https://dumps.wikimedia.org/euwiki/20170101/

Alias	en	Pages
Ibarretxe	"Ibarretxe"	Juan Jose Ibarretxe
Joseba Irazu	"Joseba Irazu"	Bernardo Atxaga
Alderdi Popularra	"People's Party"	PP

7.7 Table – Some examples extracted from Wikipedia using piped links.

$$\frac{number\ of\ times\ the\ string\ X\ links\ to\ article\ Y}{number\ of\ times\ the\ string\ X\ links\ to\ any\ article}$$

Examples of aliases and the pages that they link to with the corresponding weights are presented in Table 7.8. They are used to compute **WIKI_ALIAS** feature.

Alias	Page	Weight	en
Hillary Rodham Clinton	Hillary Clinton	1	Hillary Clinton
UK	Erresuma Batua	0.819	Great Britain
	UK	0.091	UK
Sanse	Real Sociedad B	1	Sanse
Alderdi Popularra	Espainiako Alderdi Popularra	0.091	Spanish People Party
	PP	0.909	
SDLP	Alderdi Sozialdemokrata eta Laborista	1	Social Democratic and Labour Party

7.8 Table – Some examples of aliases extracted using piped links with their weights.

2. Many pages in Wikipedia transparently redirect the user to another article. The aim of these redirect pages is to expand acronyms ("EEBB" redirects to "Ameriketako Estatu Batuak"), correct spelling errors ("Bil Clinton" redirects to "Bill Clinton") and lead the user to the article that an unambigous name refers to. Examples of some redirect cases, which are used to compute **WIKI_REDIR** feature are shown in Table 7.9.

3. Wikipedia has several list pages, where articles associated with a certain topic are categorised. Consequently, if article X belongs to list Y, there is an indication that Y is a hypernym of X.

Alias	Pages	en
EEBB	Amerikako Estatu Batuak	"United States of America"
Osasuna	Osasuna futbol kluba	"Osasuna football team"
Buda	Siddahata Gautama	"Siddhartha Gautama"

7.9 Table – Some redirect examples.

For example, if the page *Euskal Herriko futbolarien zerrenda* "List of Basque football players" contains an entry of *Imanol Agirretxe*, we can suppose that Imanol Agirretxe is a football player, and, more specifically, a Basque football player.

To obtain this semantic knowledge from Wikipedia, some processing of information contained in Wikipedia dumps is needed. First of all, the pages that contain the word *zerrenda* "list" are extracted. For example, the page *Euskal Herriko futbolarien zerrenda* "List of Basque football players". After obtaining the set of the list pages, each list title (after dropping the word *zerrenda)* is analysed morphosintactically to obtain the NPs, *Euskal herriko futbolarien* "Basque footbal players" from the above example. Finally, the titles that the list page contains are enriched with the lemmas of the NPs and the lemmas of their sub-phrases lemmas. For example, "Imanol Agirretxe" is one of the titles that *Euskal Herriko futbolarien zerrenda* list page contains, so it is enriched with the lemmas *Euskal Herri futbolari* "Basque footbal player" and *futbolari* "footbal player". In addition, if one title is in more than one list page, the new information obtained from the new list title is added.

After the processing of the information provided by list pages, a table similar to the one presented in Table 7.10, which is used to compute **WIKI_LISTS** feature, is obtained. Each table entry contains a list item (*Imanol Agirretxe*, for example) followed by lemmas of the NPs and the lemmas of their sub-phrases of each list it belongs to (*futbolari* and *Euskal Herri futbolari*) separated by # character.

In order to extend the information obtained by using Wikipedia lists, the information obtained by the piped links has been used, i.e., if an alias links to a name from a wiki list, the alias has been enriched by the list information of the name that it links to.

For example, the alias *Agirretxe* links to *Imanol Agirretxe*, and *Imanol Agirretxe* has information of *futbolari* "football player" and *Euskal Herri futbolari* "Basque football player" obtained by Wikipedia lists, therefore, a new en-

Name	Lists
Imanol Agirretxe	futbolari#Euskal Herri futbolari
	"fotball player#Basque footbal player"
Txomin Agirre	idazle#Bizkaia idazle#euskaltzain
	"writer#Basque language keeper"
Izurde muturluze	espezie#balea espezie
	"species#whale species"
Gregorio Ibarretxe	alkate#Bilbo alkate#ospetsu
	"major#mayor from Bilbao#famous"

7.10 Table – Wiki list examples.

try is created with name *Agirretxe* and list information *futbolari* and *Euskal Herri futbolari*.

As a consequence of processing and extracting the information provided by Wikipedia explained above, three new semantic features have been added to the category of semantic used by the baseline system and explained in 7.3.3. The tables obtained using Wikipedia have been stored in a mysql database, from where the information needed to compute the features values is retrieved during training and testing.

17. **WIKI_ALIAS** Corresponding weight if m_j string links to m_i page in Wikipedia; else 0.0.
18. **WIKI_REDIR** True if m_j redirects to m_i page in Wikipedia; else False.
19. **WIKI_LISTS** True if m_i is an element of m_j list in Wikipedia; else False;

7.5 Evaluation

7.5.1 Learning algorithms

The learning algorithm used in our experiments has been the Maximum Entropy (Berger *et al.*, 1996) model. The MaxEnt model creates a probability for each category y (coreferent or not coreferent) of a candidate pair, conditioned by the context x in which the candidate occurs.

7.5.2 Evaluation Metrics

The metrics used to evaluate the performance of the systems are MUC (Vilain *et al.*, 1995), B^3 (Bagga and Baldwin, 1998), CEAF_e (Luo, 2005), CEAF_m (Luo, 2005), BLANC (Recasens and Hovy, 2011) and LEA (Moosavi and Strube, 2016). The CoNLL metrics is the arithmetic mean of MUC, B^3 and CEAF_e metrics. The scores have been calculated using the reference implementation of the CoNLL scorer (Pradhan *et al.*, 2014).

7.6 Experimental results

Table 7.11 shows the results obtained by the baseline system compared with those obtained by the coreference resolution system, in which the three semantic features (**WIKI_ALIAS**, **WIKI_REDIR** and **WIKI_LIST**) have been added. These scores are obtained with automatically detected mentions.

	Automatic Mentions							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	LEA	CoNLL
B	73.79	40.45	61.49	59.54	59.80	43.17	45.40	53.91
B+Wiki	73.79	**40.98**	**61.66**	**59.82**	**60.00**	**43.48**	**45.97***	**54.21**

7.11 Table – Results obtained when automatic mentions are used. Values with $*$ superscript are significant using Paired Student's t-test with p-value < 0.05

The scores obtained by the two systems using the gold mentions, i.e., when providing all the correct mentions to the coreference resolution systems, are shown in Table 7.12.

	Gold Mentions							
	MD	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	LEA	CoNLL
B	100	54.32	85.04	77.83	80.81	71.63	63.90	73.39
B+Wiki	100	**55.38***	**85.30***	**78.17**	**81.14**	**72.07***	**64.74***	**73.94***

7.12 Table – Results obtained when gold mentions are used. Values with $*$ superscript are significant using Paired Student's t-test with p-value < 0.05

7.7 Discussion

The results presented in Table 7.11 show that the scores of the baseline system are outperformed in all the metrics by the system that makes use of features extracted

using Wikipedia. In CoNLL metric, the improved system has a score of 55.81, which is slightly higher than the baseline system, to be precise, 0.3 higher.

As shown in Table 7.12, the baseline scores are also outperformed in all the metrics when gold mentions are used. The official CoNLL metric is improved by 0.55 points.

It is interesting to compare the improvements obtained by the system with semantic knowledge in CoNLL scores when automatic mentions are used with those obtained by gold mentions. The improvement when gold mentions are provided is higher than when automatic mentions are used, 0.55 and 0.3 respectively.

In this vein Durrett and Klein (2013) mention that, despite the fact that absolute performance numbers are much higher on gold mentions and there is less room for improvement, the semantic features help much more than they do in system mentions. Versley *et al.* (2016) point out that in realistic settings, where the loss in precision would be amplified by the additional non-gold mentions, it is substantially harder to achieve gains by incorporating lexical and encyclopedic knowledge, but it is possible and necessary.

Comparing the results obtained when gold mentions are used with those obtained with the automatic mentions, there is a considerable difference. CoNLL F_1 of the baseline is 53.91 when automatic mentions are provided, while providing gold mentions this value raises to 73.39, an increase of 19.48. A similar increase in CoNLL F_1 happens when semantic features are added. In this case, there is an increase of 19.73 points, from 54.21 with automatic mentions to 73.94 when gold mentions are used.

7.8 Conclusions and Future work

We have improved the BART coreference resolution system for Basque using semantic knowledge extracted from Wikipedia. Three new semantic features have been extracted and implemented in the adapted corefence resolver. We manage to improve the baseline system results in all the metrics when new semantic feaures are used and the CoNLL metric is also outperformed, in 0.3 when automatic mentions are used and in 0.55 when gold mentions are provided to the coreference resolution system. We demonstrate that, even when the sources used are from an under-resourced language and have less coverage compared with knowledge bases in major languages, for example, English, it is possible to obtain improvements.

Nowadays, we are adapting the Wordnet similarity feature used by BART to Basque. It would also be interesting to implement Wikipedia similarity feature. Moreover, we intend to make a deep error-analysis which can reveal the weak points of our system and help to decide our future directions in the improvement of the system.

ONDORIOAK ETA ETORKIZUNEKO LANAK

8 Ondorioak eta etorkizuneko lanak

Tesi-lan honetan euskararako korreferentzia-ebazpen automatikoan egin diren lanak aurkeztu dira. Azken kapitulu honetan, egindako lanaren analisitik ateratako ondorio nagusiak eta sortu diren etorkizunerako ikerlerroak zerrendatzen dira.

8.1 Ondorioak

1. Atala: Aipamen-detekzioa.

- Euskarazko aipamenen azterketa linguistikoan oinarritzen den euskararako aipamen-detektatzaile automatikoa garatu da. Aipamen desberdinen egiturak azaldu eta analizatu ditugu eta beraien agerpen kopurua aztertu dugu EPEC corpusean. Estrategia desberdinak aplikatu ditugu aipamen-detektatzaile sendo eta eraginkorra sortzeko eta aipamen hobeak lortzeko asmoz.

- Aipamen-detekziorarko besterik gabe aurreprozesaketarako erabiltzen diren tresna generikoak erabiltzeak ez dituela emaitza onak ematen ikusi dugu, hain zuzen, % 64,28ko F-measure balioa *Strict Matching* protokoloa erabiltzean eta % 71,12koa *Lenient Matching* erabiltzean. Ondorioz, aurreprozesaketarako tresna hauen irteera eta egoera finituko teknologia erabiliz sortutako hainbat transduktoreri esker aipamen-detekzioko emaitzak hobetzea lortu dugu, % 73,36ko F-measure balioa *Strict Matching* protokoloarekin ebaluatzean eta % 79,68koa *Lenient*

Matching erabiltzean. Beste modu batera esanda, transduktoreak gehitzearekin 9 puntuko hobekuntza lortzen dugu *Strict Matching* protokoloari dagokionez eta 8koa *Lenient Matching* aplikatzean.

- Errore-analisi sakona gauzatu dugu, aipamen-detekzioan gerta daitezkeen erroreen sailkapen berri bat proposatuz eta gure aipamen-detektatzaileak egiten dituen erroreen kausen azterketa eginez. Errore-analisi honetatik lortutako ondorioek bost hobekuntza proposatzera eraman gaituzte, 1,21 puntuko hobekuntza lortuz ebaluaziorako *String Matching* erabiltzean eta 0,89 puntukoa *Lenient Matching* ebaluazio-protokoloarekin.

- Aipamen-detekzioan aurreprozesaketarako tresna automatikoek izan dezaketen eragina aztertu dugu. Hori dela eta, partzialki urrezkoa den sarrera (eskuz etiketatutako lemak, kategoriak, azpikategoriak eta kate edo *chunkak*) erabiliz gure aipamen-detektatzailea ebaluatu dugu. Lortutako emaitzek erakutsi dute aipamen-detekzioan egindako erroreak nabarmenki murrizten direla urrezko sarrera erabiltzen den kasuan: zenbat eta aurreprozesaketarako tresna hobeak, are eta emaitza hobeak lortzen dira aipamen-detekzioan. Urrezko sarrera erabiltzean % 85,89ko F-measure balioa lortzen da *Strict Matching* protokoloa erabiltzean 11,42 puntu hobea sarrera automatikoa erabiltzean lortzen dena baino. Ebaluaziorako *Lenient Matching* protokoloa aplikatzean % 89,06ko F-measure balioa lortzen da, 8,48 puntu hobea sarrera automatikoarekin lortzen dena baino.

- Aipamen-detekzioak korreferentzia-ebazpeneko emaitzetan izan dezakeen eragina aztertzeko, garatu dugun aipamen-detektzailea korreferentzia-ebazpenerako sistema batean integratu dugu eta emaitzak aztertu aipamen-detektatzailea hobetu aurretik eta ondoren. Adibidez, aipamen-detekzioan oinarri-lerroa 9 puntu hobetzen dugun kasuan korreferentzia-ebazpeneko CoNLL neurria 5,8 puntu hobetzea lortu dugu eta errore-analisi ondoren proposatutako hobekuntzak inplementatu ostean hobetutako 1,21 puntuk CoNLL neurria 0,9 puntu hobetzea lortu dutela ikusi dugu. Esperimentu hauen emaitzak aztertuz garbi ikusi dugu aipamen-detekzioak berebiziko garrantzia duela korreferentzia-ebazpenean lortzen diren emaitzetan.

Baliabideak

- Euskarako aipamen-detektatzailea
- Aipamen eta korreferentzia-mailan etiketatutako urre-patroia[1]

2. Atala: Korreferentzia-ebazpena. Erregelatan oinarritutako sistema.

- Stanford unibertsitatean garatutako korreferentzia-ebazpenerako sistema automatikoa euskararako egokitu dugu bahe berriak sartuta eta besteak egokituta. 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 euskararen ezaugarriak kontuan hartu ditugu, adibidez, morfologiaren erabilera sakonagoa egiten dugu, hizkuntza eranskariaren ezaugarriak hobeto lantzeko. Egokitzapen prozesua modu zehatzean azaldu dugu, bahez bahe, sistema beste hizkuntzetara egokitzeko prozesuan lagungarri izan dadin.
- Egokitutako sistemaren ebaluazioak egin ditugu automatikoki lortutako aipamenak eta urrezko aipamenak erabiliz eta lortutako emaitzak oinarri-lerroko sistemarekin konparatu ditugu. Korreferentzia-ebazpenerako erabiltzen diren ebaluazio metrika guztietan egokitutako sistemak oinarri-lerroko sistemak lortzen dituen emaitzak gainditu ditu. CoNLL F_1 neurria 7,07 puntu hobetzen du egokitutako sistemak aipamen automatikoak erabiltzean eta 11,5 puntu sistemari urrezko aipamenak ematen zaizkionean. Sistemaren ebaluazioa amaitzeko, bahe bakoitzaren ekarpena aztertu da, horiek modu inkrementalean gehituz eta euskara bezalako hizkuntza eranskari batean korreferentzia-ebazpena egiteko morfologiak duen garrantziaz jabetu gara.
- Gure sistemak lortutako emaitzak beste hizkuntza batzuetarako garatu diren sistemek lortzen dituztenekin konparatu ditugu. Gure emaitzak parekoak dira.

[1] http://ixa2.si.ehu.es/epec-koref/epec-koref_v1.0.tgz

- Sistemaren errore-analiasi gauzatu dugu, errore-motak eta errore-kausak zehaztuz. Errore-analisi horren ondorioz gure sistemak munduaren ezagutza eta ezagutza semantikoa beharrezkoa den korreferentzia-erlazioak ebazteko duen gabeziaz jabetu gara.

- Euskararako egokitutako Stanfordeko sistema aberastu dugu bi bahe berri gehituz. Lehenengo baheak entitate izendunak diren aipamenak munduaren ezagutzarekin aberasten ditu dagokien Wikipediako orrialdearekin lotu ondoren. Bigarren baheak, euskal WordNetetik erauzitako sinonimoen zerrenda erabiltzen du bi aipamen korreferenteak diren edo ez erabakitzeko. Bi baheak aplikatu ostean, CoNLL F_1 neurria 0,24 hobetzea lortzen du sistemak aipamen automatikoak erabiltzean eta 0,39 puntu berriz urrezko aipamenekin.

- Bi bahe berri hauek erabiltzen dituen sistemaren errore-analisia egin ostean ikusi da, euskarako Wikipedia eta Wordnet ezagutza-baseen estaldura baxuagoa dela, adibidez, ingeleserako Wikipedia eta WordNetarekin konparatzen baditugu. Lortzen diren hobekuntzak, hala ere, beste hizkuntzetarako sistemei ezagutza semantikoa eta munduaren ezagutza txertatzean lortzen dituztenaren antzekoak dira. Bi baliabide hauen tamaina hobetzeak emaitzen hobekuntza handiagoa lortzen lagunduko ligukeela suposatzen dugu, nahiz eta garbi eduki sarritan ezagutza semantikoa eta munduaren ezagutza beharrezkoa duten korreferentzia-kasuak konplexuak izan ohi direla.

Baliabideak

- Erregelatan oinarritutako euskararako korreferentzia-ebazpenerako sistema[2]
- Bi lexikoi

[2] http://ixa2.si.ehu.es/ixakat/ixa-pipe-coref-eu.php

3. Atala: Korreferentzia-ebazpena. Ikasketa automatikoan oinarritutako sistema.

- BART ikasketa gainbegiratuan oinarritutako korreferentzia-ebazpenerako sistema egokitu dugu euskara tratatu ahal izateko. EPEC corpusa erabiliz bi eredu ikasi ditugu, lehenengoa Soon *et al.* (2001) lanean oinarritutako eredua oinarri-lerro bezala finkatu dugu, eta bigarrena, euskararen ezaugarrietara hobeki egokitzen den eredu hedatua. Bi ereduak ebaluatu ditugu, aipamen automatikoak eta urrezko aipamenak erabilita, eta emaitzek erakutsi digute euskararen ezaugarriak kontuan hartzen dituen ereduak nabarmenki hobetzen duela oinarri-lerroa metrika guztietan. Ikasketa automatikoa erabiltzeko euskara bezalako hizkuntza batean, adibidez, sistema nominalak generorik ez izateak eta izenordainek bizidun/ez-bizidun propietateak ez izateak sor ditzakeen erronkak azaldu ditugu.
- BART sistema hobetu dugu Wikipediatik erauzitako informazioa erabilita. Hiru ezaugarri semantiko berri lortu dira Wikipediatik erauzitako munduaren ezagutzatik eta BART sisteman inplementatu dira. Oinarri-lerroko sistema hobetzea lortu dugu metrika guztietan informazio semantikoa erabiltzean eta CoNLL metrika ere hobetu dugu, 0,3 puntu aipamen automatikoak erabiltzean eta 0,55 puntu urrezko aipamenekin.
- Hizkuntza gutxitu bateko baliabideak erabiliz ere hobekuntzak lortzea posible dela erakutsi dugu, nahiz eta hizkuntza horietako baliabideen estaldura ez izan, adibidez, ingelesa bezalako hizkuntza nagusi bateko baliabideen estaldura bezain zabala.

Baliabideak

- Ikasketa automatikoan oinarritutako euskararako korreferentzia-ebazpenerako sistema
- Wikipediatik erauzitako munduaren ezagutza duen datu-basea

8.2 Etorkizuneko lanak

Etorkizuneko lanei dagokionez honako ikerketa lerroak jorratzeko asmoa daukagu:

- Stanfordeko sistemari dagokionez, aipamen pronominalen eta eliptikoen korreferentzia-ebazpena hobetzeko asmoa dugu. 5.2 atalean aurkeztutako errore-analisiak erakutsi duen moduan errore guztien % 12 baitira gutxi gorabehera.

- Gaur egun, WordNet erabiliz bi hitzen arteko antzekotasun balioak ari gara ezaugarri moduan inplementatzen BART sisteman. Lan horrekin amaitu eta Strube and Ponzetto (2006) lanean aurkezten den metodologia jarraituz Wikipediako kategoria-sistema erabiliz erauzitako grafoarekin bi aipamenen arteko antzekotasuna lortzeko gai den ezaugarria ere inplementatu nahi dugu.

- WordNetetik eta Wikipediatik lortutako antzekotasun balioen ezaugarriak gehitu ondoren, beharrezkoa ikusten dugu, BART sistemaren errore-analisi sakonagoa egitea, gabeziez jabetzeko eta hobekuntza berriak planteatzeko.

- Euskararen sistema nominalak genero eta bizidun/ez-bizidun propietateak ez izateak sortzen dituen erroreak konpontzeko, ikasketarako ezaugarri berrien azterketa egin nahi dugu.

- Interesgarria izango litzateke erregelatan oinarritutako eta ikasketa automatikoko tekniken abantailak eta onurak konbinatzea, Chen and Ng (2012) edo Lee *et al.* (2017) lanetan aurkezten den antzera, sistema hibrido bat sortuz.

- EPEC corpusa osorik etiketatu nahi dugu aipamen eta korreferentzia mailan. Prozesua modu semi-automatikoan egin nahi dugu, horretarako lehenbizi garatu dugun aipamen-detektatzailea erabiliz aipamen guztiak etiketatuko ditugu. Ondoren, hizkuntzalariek sistemak egiten dituen akatsak eskuz zuzendu eta behin aipamen guztiak zuzen etiketatuta daudela korreferentzia-erlazioak etiketatuko ditugu modu automatikoan. Bukatzeko, erlazio okerrak eskuz zuzendu beharko lirateke.

- Korreferentzia-ebazpenak hizkuntzaren prozesamenduko aplikazioren batean, adibidez itzulpen automatikoan, izan dezakeen eragina sakon aztertu nahi dugu.

Bibliografia

Aduriz I., Aldezabal I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., and Gojenola K. Finite State Applications for Basque. *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing*, 3–11, 2003.

Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa M., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., and Urizar R. Methodology and Steps towards the Construction of EPEC, a Corpus of Written Basque Tagged at Morphological and Syntactic Levels for the Automatic Processing. *Language and Computers*, 56 lib., 1–15. Rodopi, Amsterdam, Netherlands, 2006.

Aduriz I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A., Gojenola K., Oronoz M., and Uria L. A Cascaded Syntactic Analyser for Basque. *Computational Linguistics and Intelligent Text Processing, 5th International Conference (CICLing 2004)*, 124–134, Seoul, Korea, 2004.

Aduriz I. and Díaz de Ilarraza A. Morphosyntactic Disambiguation and Shallow Parsing in Computational Processing of Basque. In Oyharçabal B., editor, *Inquiries into the lexicon-syntax relations in Basque*, 1–21. University of the Basque Country, 2003.

Agirre E., López de Lacalle O., and Soroa A. Random Walks for Knowledge-

based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84, 2014.

Agirre E. and Soroa A. Personalizing PageRank for Word Sense Disambiguation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, 33–41, Athens, Greece, 2009.

Alegria I., Ansa O., Artola X., Ezeiza N., Gojenola K., and Urizar R. Representation and Treatment of Multiword Expressions in Basque. *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, MWE '04, 48–55, Barcelona, Spain, 2004.

Alegria I., Aranzabe M.J., Ezeiza A., Ezeiza N., and Urizar R. Robustness and Customisation in an Analyser/Lemmatiser for Basque. *LREC-2002 Customizing knowledge in NLP applications workshop*, 1–6, 2002a.

Alegria I., Aranzabe M.J., Ezeiza N., Ezeiza A., and Urizar R. Using Finite State Technology in Natural Language Processing of Basque. In Watson B.W. and Wood D., editors, *Implementation and Application of Automata*, 2494 lib. of *Lecture Notes in Computer Science*, 1–12. Springer Berlin / Heidelberg, 2002b.

Alegria I., Arrieta B., Carreras X., Díaz de Ilarraza A., and Uria L. Chunk and Clause Identification for Basque by Filtering and Ranking with Perceptrons. *Procesamiento del Lenguaje Natural*, 41:5–12, 2008.

Alegria I., Artola X., Sarasola K., and Urkia M. Automatic Morphological Analysis of Basque. *Literary & Linguistic Computing*, 11(4):193–203, 1996.

Alegria I., Ezeiza N., Fernandez I., and Urizar R. Named Entity Recognition and Classification for texts in Basque. *II Jornadas de Tratamiento y Recuperación de Información*, (JOTRI 2003), 198–203, Madrid, Spain, 2003.

Alshawi H. Resolving quasi logical forms. *Computational Linguistics*, 16(3): 133–144, 1990.

Aone C. and Bennett S.W. Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. *Proceedings of the 33rd Annual*

Meeting on Association for Computational Linguistics, ACL '95, 122–129, Cambridge, Massachusetts, 1995.

Appelt D.E., Hobbs J.R., Bear J., Israel D., Kameyama M., Martin D., Myers K., and Tyson M. SRI International FASTUS System: MUC-6 Test Results and Analysis. *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, 237–248, Columbia, Maryland, 1995.

Arregi O., Ceberio K., Díaz de Ilarraza A., Goenaga I., Sierra B., and Zelaia A. A First Machine Learning Approach to Pronominal Anaphora Resolution in Basque. In Morales A.F.K. and Simari G.R., editors, *IBERAMIA 2010. LNAI 6433*, 6433 lib. of *Lecture Notes in Computer Science*, 234–243. 2010.

Arrieta B. *Azaleko sintaxiaren tratamendua ikasketa automatikoko teknikeen bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera koma-zuzentzaile batean*. Doktoretza-tesia, Computer Languages and Systems,University of the Basque Country, 2010.

Artstein R. and Poesio M. Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Attardi G., Simi M., and Dei Rossi S. Tanl-1: Coreference resolution by parse analysis and similarity clustering. *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), 108–111, Uppsala, Sweden, 2010.

Bagga A. and Baldwin B. Algorithms for Scoring Coreference Chains. *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, 563–566, Granada, Spain, 1998.

Bengoetxea K. and Gojenola K. Application of different techniques to dependency parsing of basque. *Proceedings of the NAACL HLT 2010 First Workshop on Statistical Parsing of Morphologically-Rich Languages*, SPMRL '10, 31–39, Los Angeles, California, 2010.

Bengtson E. and Roth D. Understanding the Value of Features for Coreference Resolution. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 294–303, Honolulu, Hawaii, 2008.

Berger A.L., Della Pietra V.J., and Della Pietra S. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22 (1):39–71, 1996.

Björkelund A. and Farkas R. Data-driven Multilingual Coreference Resolution using Resolver Stacking. *Joint Conference on EMNLP and CoNLL - Shared Task*, 49–55, Jeju Island, Korea, 2012.

Bobrow D.G. A Question-answering System for High School Algebra Word Problems. *Proceedings of the October 27-29, 1964, Fall Joint Computer Conference, Part I*, AFIPS '64 (Fall, part I), 591–614, San Francisco, California, 1964.

Broscheit S., Poesio M., Ponzetto S.P., Rodriguez K.J., Romano L., Uryupina O., Versley Y., and Zanoli R. BART: A Multilingual Anaphora Resolution System. *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), 104–107, Uppsala, Sweden, 2010a.

Broscheit S., Ponzetto S.P., Versley Y., and Poesio M. Extending BART to provide a coreference resolution system for German. *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 164–167, Valletta, Malta, 2010b.

Cai J., Mújdricza-Maydt E., and Strube M. Unrestricted Coreference Resolution via Global Hypergraph Partitioning. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, 56–60, Portland, Oregon, 2011.

Cai J. and Strube M. End-to-end Coreference Resolution via Hypergraph Partitioning. *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 143–151. 2010.

Carbonell J.G. and Brown R.D. Anaphora Resolution: A Multi-strategy Approach. *Proceedings of the 12th Conference on Computational Linguistics - Volume 1*, COLING '88, 96–101, Budapest, Hungry, 1988.

Carter D. *Interpreting Anaphors in Natural Language Texts*. Halsted Press, New York, NY, USA, 1987.

Ceberio K., Aduriz I., Díaz de Ilarraza A., and Garcia Azkoaga I. Erreferentziakidetasunaren azterketa eta anotazioa euskarazko corpus batean. *Gramatika Jaietan. P. Goenagaren 30 'Gramatika Bideetan' liburuaren omenez, X. Artiagoitia; J. A. Lakarra (Arg.)*, 153–172, 2008.

Ceberio K., Aduriz I., Díaz de Ilarraza A., and Garcia-Azkoaga I. Coreferential relations in Basque: the annotation process. *Theoretical Developments in Hispanic Linguistics*. The Ohio State University, 2016.

Chang K.W., Samdani R., Rozovskaya A., Rizzolo N., Sammons M., and Roth D. Inference Protocols for Coreference Resolution. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, 40–44, Portland, Oregon, 2011.

Chen C. and Ng V. Combining the best of two worlds: A hybrid approach to multilingual coreference resolution. *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, 56–63, Jeju Island, Korea, 2012.

Chinchor N.A. Overview of MUC-7/MET-2. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. 1998.

Cohen W.W. Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, 115–123, 1995.

Collovini S., Carbonel T., Fuchs Thielsen J., Coelho J.C., Rino L., and Vieira R. Summit: Um corpus anotado com informações discursivas visando à sumarização automática. *52nd Workshop em Tecnologia da Informação e da Linguagem Humana*, TIL'07, Rio de Janeiro, 2007.

Connolly D., Burger J.D., and Day D.S. A Machine Learning Approach to Anaphoric Reference. *Proceedings of the International Conference on New Methods in Language Processing*. ACL, 1994.

Connolly D., Burger J.D., and Day D.S. A machine learning approach to anaphoric reference. *New Methods in Language Processing*, 133–144, 1997.

DAARC. *Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquim (DAARC 1996)*. Lancaster, 1996.

DAARC. *Proceedings of the 2nd Discourse Anaphora and Anaphor Resolution Colloquim (DAARC 1998)*. Lancaster, 1998.

DAARC. *Proceedings of the 3rd Discourse Anaphora and Anaphor Resolution Colloquim (DAARC 2000)*. Lancaster, 2000.

Daelemans W. and van den Bosch A. *Memory-Based Language Processing*. Cambridge University Press, New York, NY, USA, 1st edition, 2005.

Denis P. and Baldridge J. Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 236–243, Rochester, New York, 2007.

Denis P. and Baldridge J. Global Joint Models for Coreference Resolution and Named Entity Classification. *Procesamiento del Lenguaje Natural*, 43: 87–96, 2009.

Díaz de Ilarraza A., Fernández-Terrones E., Aldezabal I., and Aranzabe M.J. From Dependencies to Constituents in the Reference Corpus for the Processing of Basque (EPEC). *Procesamiento del Lenguaje Natural*, 41, 2008.

Doddington G., Mitchell A., Przybocki M., Ramshaw L., Strassel S., and Weischedel R. The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. *Proceedings of Language Resources and Evaluation Conference*, (LREC 2004), 837–840, Lisbon, Portugal, 2004.

Durrett G. and Klein D. Easy Victories and Uphill Battles in Coreference Resolution. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1971–1982, Seattle, Washington, USA, 2013.

Durrett G. and Klein D. A Joint Model for Entity Analysis: Coreference, Typing, and Linking. *TACL*, 2:477–490, 2014.

Fellbaum C., editor. *WordNet: an electronic lexical database*. MIT Press, 1998.

Fernandes E., dos Santos C., and Milidiú R. Latent Structure Perceptron with Feature Induction for Unrestricted Coreference Resolution. *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, 41–48, Jeju Island, Korea, 2012.

Finkel J.R. and Manning C.D. Enforcing Transitivity in Coreference Resolution. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, HLT-Short '08, 45–48, Columbus, Ohio, 2008.

Fox B.A. *Discourse structure and anaphora: written and conversational English*. Cambridge Univeristy Press, 1993.

Freund Y. and Schapire R.E. Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, 37(3):277–296, 1999.

Gaizauskas R., Humphreys K., Cunningham H., and Wilks Y. University of Sheffield: Description of the LaSIE System As Used for MUC-6. *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, 207–220, Columbia, Maryland, 1995.

Garcia M. and Gamallo P. An Entity-Centric Coreference Resolution System for Person Entities with Rich Linguistic Information. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 741–752, Dublin, Ireland, 2014a.

Garcia M. and Gamallo P. Multilingual corpora with coreferential annotation of person entities. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, 2014b.

Garcia Azkoaga I. *Anafora eta erreferentziakidetasuna: izen-kohesiorako baliabideak*. Udako Euskal Unibertsitatea, Bilbo, 2016.

Gardent C. and Manuélian H. Création d'un corpus annoté pour le traitement des descriptions définies. *Traitement Automatique des Langues*, 1:115–140, 2005.

Gonzalez-Dios I., Aranzabe M.J., Díaz de Ilarraza A., and Soraluze A. Detecting Apposition for Text Simplification in Basque. *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2*, CICLing'13, 513–524, Berlin, Heidelberg, 2013.

Grishman R. and Sundheim B. Design of the MUC-6 Evaluation. *Proceedings of a Workshop on Held at Vienna, Virginia: May 6-8, 1996*, TIPSTER '96, 413–422, Vienna, Virginia, 1996.

Grosz B.J., Joshi A.K., and Weinstein S. Providing a Unified Account of Definite Noun Phrases in Discourse. *Proceedings of the 21st Annual Meeting on Association for Computational Linguistics*, ACL '83, 44–50, Cambridge, Massachusetts, 1983.

Grosz B.J. and Sidner C.L. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3):175–204, 1986.

Hacioglu K., Douglas B., and Chen Y. Detection of Entity Mentions Occurring in English and Chinese text. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, (HLT 2005), 379–386, Vancouver, British Columbia, Canada, 2005.

Haghighi A. and Klein D. Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 848–855, Prague, Czech Republic, 2007.

Hajič J., Böhmová A., Hajičová E., and Vidová-Hladká B. The prague dependency treebank: A three-level annotation scenario. In Abeillé A., editor, *Treebanks: Building and Using Parsed Corpora*, 103–127. Amsterdam:Kluwer, 2000.

Hajishirzi H., Zilles L., Weld D.S., and Zettlemoyer L. Joint Coreference Resolution and Named-Entity Linking with Multi-Pass Sieves. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 289–299, Seattle, Washington, USA, 2013.

Hasler L., Orasan C., and Naumann K. NPs for Events: Experiments in Coreference Annotation. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, 1167–1172, Genoa, Italy, 2006.

Hendrickx I., Bouma G., Coppens F., Daelemans W., Hoste V., Kloosterman G., Mineur A., Van Der Vloet J., and Verschelde J. A Coreference Corpus

and Resolution System for Dutch. *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*, 144–149. 2008.

Hendrickx I., Hoste V., and Daelemans W. Evaluating Hybrid Versus Data-driven Coreference Resolution. *Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Conference on Anaphora: Analysis, Algorithms and Applications*, DAARC'07, 137–150, Lagos, Portugal, 2007.

Hinrichs E.W., Kübler S., and Naumann K. A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations. *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, 13–20, Ann Arbor, Michigan, 2005.

Hobbs J. Resolving Pronoun References. *Lingua*, 44:311–338, 1978.

Hoste V. *Optimization Issues in Machine Learning of Coreference Resolution.* Doktoretza-tesia, Antwerp University, 2005.

Hoste V. and van den Bosch A. A modular approach to learning Dutch co-reference. *Proceedings from the First Bergen Workshop on Anaphora Resolution*, 51–75, Bergen, 2007.

Hulden M. Foma: A Finite-state Compiler and Library. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, EACL '09, 29–32, Athens, Greece, 2009.

Humphreys K., Gaizauskas R., Huyck C., Mitchell B., Cunningham H., and Wilks Y. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. *Proceedings of the Seventh Message Understanding Conference*, 84–89, 1998.

Iida R., Inui K., Takamura H., and Matsumoto Y. Incorporating contextual cues in trainable models for coreference resolution. *In Proceedings of the EACL Workshop on The Computational Treatment of Anaphora*, 23–30, 2003.

Iida R., Komachi M., Inui K., and Matsumoto Y. Annotating a Japanese Text Corpus with Predicate-argument and Coreference Relations. *Proceedings*

of the Linguistic Annotation Workshop, LAW '07, 132–139, Prague, Czech Republic, 2007.

Kameyama M. Intrasentential Centering: a Case Study. In Walker M., Prince E., and Joshi A., editors, *Centering Theory in Discourse*. Oxford University Press, 1997a.

Kameyama M. Recognizing referential links: An information extraction perspective. *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, ANARESOLUTION '97, 46–53, Madrid, Spain, 1997b.

Karlsson F., Voutilainen A., Heikkilä J., and Anttila A., editors. *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text*. Walter de Gruyter & Co., Hawthorne, NJ, USA, 1995.

Kehler A., Appelt D., Taylor L., and Simma A. The (Non)Utility of Predicate-Argument Frequencies for Pronoun Interpretation. In Susan Dumais D.M. and Roukos S., editors, *HLT-NAACL 2004: Main Proceedings*, 289–296, Boston, Massachusetts, USA, 2004.

Kim Y., Riloff E., and Gilbert N. The Taming of Reconcile as a Biomedical Coreference Resolver. *Proceedings of BioNLP Shared Task 2011 Workshop*, 89–93, Portland, Oregon, USA, 2011.

Klein D. and Manning C.D. Accurate Unlexicalized Parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 423–430, Sapporo, Japan, 2003.

Klenner M. Enforcing Consistency on Coreference Sets. *Recent Advances in Natural Language Processing (RANLP)*, 323–328, Borovets, Bulgary, 2007.

Kobdani H. and Schütze H. SUCRE: A Modular System for Coreference Resolution. *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), 92–95, Uppsala, Sweden, 2010.

Kopeć M. and Ogrodniczuk M. Creating a Coreference Resolution System for Polish. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, 192–195, Istanbul, Turkey, 2012.

Kopeć M. and Ogrodniczuk M. Inter-annotator Agreement in Coreference Annotation of Polish. In Sobecki J., Boonjing V., and Chittayasothorn S., editors, *Advanced Approaches to Intelligent Information and Database Systems*, 551 lib. of *Studies in Computational Intelligence*, 149–158. Springer International Publishing, Switzerland, 2014.

Kummerfeld J.K., Bansal M., Burkett D., and Klein D. Mention Detection: Heuristics for the OntoNotes Annotations. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 102–106, Portland, Oregon, USA, 2011.

Kummerfeld J.K. and Klein D. Error-Driven Analysis of Challenges in Coreference Resolution. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 265–277, Seattle, Washington, USA, 2013.

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

Lalitha Devi S., Sundar Ram V., and RK Rao P. A Generic Anaphora Resolution Engine for Indian Languages. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 1824–1833, Dublin, Ireland, 2014.

Lappin S. and Leass H.J. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., and Jurafsky D. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Computational Linguistics*, 39(4):885–916, 2013.

Lee H., Surdeanu M., and Jurafsky D. A scaffolding approach to coreference resolution integrating statistical and rule-based models. *Natural Language Engineering*, 1–30, 2017.

Luo X. On Coreference Resolution Performance Metrics. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, 25–32, Vancouver, British Columbia, Canada, 2005.

Luo X., Ittycheriah A., Jing H., Kambhatla N., and Roukos S. A Mention-synchronous Coreference Resolution Algorithm Based on the Bell Tree. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Barcelona, Spain, 2004.

Luo X. and Pradhan S. Evaluation Metrics. In Poesio M., Stuckardt R., and Versley Y., editors, *Anaphora Resolution: Algorithms, Resources, and Applications*, 141–163. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

Ma X., Liu Z., and Hovy E. Unsupervised Ranking Model for Entity Coreference Resolution. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1012–1018, San Diego, California, 2016.

Magnini B., Pianta E., Girardi C., Negri M., Romano L., Speranza M., Lenzi V.B., and Sprugnoli R. I-cab: the italian content annotation bank. *Proceegings of LREC*, 963–968, 2008.

Marcus M.P., Marcinkiewicz M.A., and Santorini B. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Màrquez L., Recasens M., and Sapena E. Coreference Resolution: an Empirical Study Based on SemEval-2010 Shared Task 1. *Language Resources and Evaluation*, 47(3):661–694, 2013.

Martschat S., Cai J., Broscheit S., Mújdricza-Maydt E., and Strube M. A Multigraph Model for Coreference Resolution. *Joint Conference on EMNLP and CoNLL - Shared Task*, 100–106, Jeju Island, Korea, 2012.

McCarthy J.F. A Trainable Approach to Conference Resolution for Information Extraction. Barne-txostena, University of Massachusetts, Amherst, MA, USA, 1996.

McCarthy J.F. and Lehnert W.G. Using Decision Trees for Coreference Resolution. *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, (IJCAI 1995), 1050–1055, San Francisco, CA, USA, 1995.

Mendes P., Jakob M., and Bizer C. DBpedia: A Multilingual Cross-domain Knowledge Base. In Chair) N.C.C., Choukri K., Declerck T., DoÄ§an

M.U., Maegaard B., Mariani J., Moreno A., Odijk J., and Piperidis S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012.

Miculicich Werlen L. and Popescu-Belis A. Using coreference links to improve spanish-to-english machine translation. *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, 30–40, Valencia, Spain, 2017.

Mihăltz M. Knowledge-based Coreference Resolution for Hungarian. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Mitkov R. Robust Pronoun Resolution with Limited Knowledge. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, 869–875, Montreal, Quebec, Canada, 1998.

Mitkov R. *Anaphora Resolution*. Longman, London, 2002.

Moosavi N.S. and Strube M. Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 632–642, Berlin, Germany, 2016.

Morton T.S. Coreference for NLP Applications. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, 173–180, Hong Kong, 2000.

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

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

Müller C. and Strube M. Multi-level Annotation of Linguistic Data with MMAX2. In Braun S., Kohn K., and Mukherjee J., editors, *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, 197–214. Peter Lang, Frankfurt a.M., Germany, 2006.

Müller M.C. *Fully automatic resolution of it, this and taht in unrestricted multy-party dialog*. Doktoretza-tesia, University of Tübingen, 2008.

Nazaridoust M., Bidgoli B.M., and Nazaridoust S. Co-reference Resolution in Farsi Corpora. In Jamshidi M., Kreinovich V., and Kacprzyk J., editors, *Advance Trends in Soft Computing: Proceedings of WCSC 2013, December 16-18, San Antonio, Texas, USA*, 155–162. Springer International Publishing, Cham, 2014.

Ng V. Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 157–164, Ann Arbor, Michigan, 2005.

Ng V. Shallow Semantics for Coreference Resolution. *Proceedings of IJCAI*, 1689–1694, 2007.

Ng V. Unsupervised Models for Coreference Resolution. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 640–649, Honolulu, Hawaii, 2008.

Ng V. and Cardie C. Combining Sample Selection and Error-driven Pruning for Machine Learning of Coreference Rules. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, EMNLP '02, 55–62, Philadelphia, Pennsylvania, 2002a.

Ng V. and Cardie C. Identifying Anaphoric and Non-anaphoric Noun Phrases to Improve Coreference Resolution. *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1*, COLING '02, 1–7, Taipei, Taiwan, 2002b.

Ng V. and Cardie C. Improving Machine Learning Approaches to Coreference Resolution. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 104–111, Philadelphia, Pennsylvania, 2002c.

Nguy G.L., Novák V., and Žabokrtský Z. Comparison of Classification and Ranking Approaches to Pronominal Anaphora Resolution in Czech. *Proceedings of the SIGDIAL 2009 Conference*, 276–285, London, UK, 2009.

Nguyen N.L.T., Kim J., and Tsujii J. Challenges in Pronoun Resolution System for Biomedical Text. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 2408–2412, Marrakech, Morocco, 2008.

Nicolov N., Salvetti F., and Ivanova S. Sentiment Analysis: Does Coreference Matter? *AISB 2008 Convention Communication, Interaction and Social Intelligence*, 37–40, 2008.

NIST. Automatic Content Extraction 2008 Evaluation Plan (ACE08), 2008.

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

Ogrodniczuk M. and Ng V., editors. *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*. Association for Computational Linguistics, San Diego, California, USA, 2016.

Ogrodniczuk M. and Ng V., editors. *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*. Association for Computational Linguistics, Valencia, Spain, 2017.

Ohta T., Pyysalo S., Tsujii J., and Ananiadou S. Open-domain Anatomical Entity Mention Detection. *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, ACL '12, 27–36, Jeju, Republic of Korea, 2012.

Orasan C., Cristea D., Mitkov R., and Branco A. Anaphora Resolution Exercise: an Overview. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Otegi A., Ezeiza N., Goenaga I., and Labaka G. A Modular Chain of NLP Tools for Basque. *Text, Speech, and Dialogue - 19th International Conference (TSD 2016)*, 93–100, Brno, Czech Republic, 2016.

Palomar M., Moreno L., Peral J., Muñoz R., Ferrández A., Martínez-Barco P., and Saiz-Noeda M. An Algorithm for Anaphora Resolution in Spanish Texts. *Computational Linguistics*, 27(4):545–567, 2001.

Peral J., Palomar M., and Ferrández A. Coreference-oriented Interlingual Slot Structure & Machine Translation. *Proceedings of the Workshop on Coreference and Its Applications*, CorefApp '99, 69–76, College Park, Maryland, 1999.

Pociello E., Agirre E., and Aldezabal I. Methodology and construction of the Basque WordNet. *Language Resources and Evaluation*, 45(2):121–142, 2011.

Poesio M. The MATE/GNOME proposals for anaphoric annotation, revisited. In Strube M. and Sidner C., editors, *Proceedings of the 5th SIGDIAL Workshop on Discourse and Dialogue*, 154–162, 2004.

Poesio M. and Artstein R. Anaphoric Annotation in the ARRAU Corpus. *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, 2008.

Poesio M., Chamberlain J., Kruschwitz U., Robaldo L., and Ducceschi L. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Trans. Interact. Intell. Syst.*, 3(1):3:1–3:44, 2013.

Poesio M., Day D., Artstein R., Duncan J., Eidelman V., Giuliano C., Hall R., Hitzeman J., Jern A., Kabadjov M., Wai Keong Yong S., Mann G., Moschitti A., Ponzetto S.P., Smith J., Steinberger J., Strube M., Su J., Versley Y., Yang X., and Wick M. ELERFED: Final Report, 2007.

Poesio M., Delmonte R., Bristot A., Chiran L., and Tonelli S. The VENEX corpus of anaphoric information in spoken and written Italian, 2004.

Poesio M., Stuckardt R., and Versley Y., editors. *Anaphora Resolution: Algorithms, Resources, and Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

Poesio M., Uryupina O., and Versley Y. Creating a Coreference Resolution System for Italian. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 713–716, Valletta, Malta, 2010.

Poesio M. and Vieira R. A Corpus-based Investigation of Definite Description Use. *Computatitonal Linguistics*, 24(2):183–216, 1998.

Ponzetto S.P. and Strube M. Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06, 192–199, New York, New York, 2006.

Poon H., Christensen J., Domingos P., Etzioni O., Hoffmann R., Kiddon C., Lin T., Ling X., Mausam, Ritter A., Schoenmackers S., Soderland S., Weld D., Wu F., and Zhang C. Machine Reading at the University of Washington. *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, FAM-LbR '10, 87–95, Los Angeles, California, 2010.

Poon H. and Domingos P. Joint Unsupervised Coreference Resolution with Markov Logic. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 650–659, Honolulu, Hawaii, 2008.

Pradhan S., Hovy E., Marcus M., Palmer M., Ramshaw L., and Weischedel R. OntoNotes: A Unified Relational Semantic Representation. *Proceedings of the International Conference on Semantic Computing*, (ICSC '07), 517–526, Washington, DC, USA, 2007.

Pradhan S., Luo X., Recasens M., Hovy E., Ng V., and Strube M. Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 30–35. 2014.

Pradhan S., Moschitti A., Xue N., Uryupina O., and Zhang Y. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. *Joint Conference on EMNLP and CoNLL - Shared Task*, 1–40, Jeju Island, Korea, 2012.

Pradhan S., Ramshaw L., Marcus M., Palmer M., Weischedel R., and Xue N. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, 1–27, Portland, Oregon, 2011.

Quinlan J.R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

Rahman A. and Ng V. Supervised Models for Coreference Resolution. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, 968–977, Singapore, 2009.

Rahman A. and Ng V. Coreference Resolution with World Knowledge. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 814–824, Portland, Oregon, 2011.

Rand W.M. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Ratinov L. and Roth D. Learning-based Multi-sieve Co-reference Resolution with Knowledge. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, 1234–1244, Jeju Island, Korea, 2012.

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

Recasens M., Can M., and Jurafsky D. Same Referent, Different Words: Unsupervised Mining of Opaque Coreferent Mentions. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 897–906, Atlanta, Georgia, 2013.

Recasens M. and Hovy E. A Deeper Look into Features for Coreference Resolution. *Anaphora Processing and Applications, 7th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2009, Goa, India*, 29–42, 2009.

Recasens M. and Hovy E. Coreference Resolution Across Corpora: Languages, Coding Schemes, and Preprocessing Information. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, 1423–1432, Uppsala, Sweden, 2010.

Recasens M. and Hovy E. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510, 2011.

Recasens M., Màrquez L., Sapena E., Martí M.A., Taulé M., Hoste V., Poesio M., and Versley Y. SemEval-2010 task 1: Coreference Resolution in Multiple Languages. *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), 1–8, Uppsala, Sweden, 2010.

Recasens M. and Martí M. AnCora-CO: Coreferentially Annotated Corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345, 2010.

Rich E. and LuperFoy S. An architecture for anaphora resolution. *Proceedings of the Second Conference on Applied Natural Language Processing*, ANLC '88, 18–24, Austin, Texas, 1988.

Rodríguez K.J., Delogu F., Versley Y., Stemle E.W., and Poesio M. Anaphoric Annotation of Wikipedia and Blogs in the Live Memories Corpus. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, 2010.

Sapena E., Padró L., and Turmo J. RelaxCor Participation in CoNLL Shared Task on Coreference Resolution. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, 35–39, Portland, Oregon, 2011.

Sapena E., Padró L., and Turmo J. A Constraint-Based Hypergraph Partitioning Approach to Coreference Resolution. *Computational Linguistics*, 39(4):847–884, 2013.

Schrijver A. *Theory of Linear and Integer Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1986.

Shou H. and Zhao H. System paper for CoNLL-2012 shared task: Hybrid Rule-based Algorithm for Coreference Resolution. *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, 118–121, Jeju Island, Korea, 2012.

Sidner C.L. Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse. Barne-txostena, Massachusetts Institute of Technology, Cambridge, MA, USA, 1979.

Sidner C.L. Focusing for Interpretation of Pronouns. *Computational Linguistics*, 7(4):217–231, 1981.

Sidner C.L. Focusing in the Comprehension of Definite Anaphora. In Brady M. and Berwick R.C., editors, *Computational Models of Discourse*, 267–330. MIT Press, Cambridge, MA, 1983.

Sikdar U.K., Ekbal A., and Saha S. A generalized framework for anaphora resolution in indian languages. *Knowl.-Based Syst.*, 109:147–159, 2016.

Sikdar U.K., Ekbal A., Saha S., Uryupina O., and Poesio M. Adapting a State-of-the-art Anaphora Resolution System for Resource-poor Language. *Sixth International Joint Conference on Natural Language Processing*, IJCNL 2013, 815–821, Nagoya, Japan, 2013.

Sobha L.D., Bandyopahdhyay S., Vijay Sundar Ram R., and Akilandeswari A. NLP tool contest @ICON2011 on anaphora resolution in Indian languages. *Proceedings of ICON*, Singapore, 2011a.

Sobha L.D., Pattabhi R.R., Ram R.V.S., Malarkodi C., and Akilandeswari A. Hybrid Approach for Coreference Resolution. *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, 93–96, Portland, Oregon, USA, 2011b.

Soon W.M., Ng H.T., and Lim C.Y. Corpus-Based Learning for Noun Phrase Coreference Resolution. *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 285–291, 1999.

Soon W.M., Ng H.T., and Lim D.C.Y. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27 (4):521–544, 2001.

Soraluze A., Alegria I., Ansa O., Arregi O., and Arregi X. Recognition and Classification of Numerical Entities in Basque. *Recent Advances in Natural Language Processing (RANLP)*, 764–769, Hissar, Bulgaria, 2011.

Soraluze A., Arregi O., Arregi X., Ceberio K., and Díaz de Ilarraza A. Mention detection: First steps in the development of a Basque coreference resolution system. *Proceedings of the 11th Conference on Natural Language Processing (KONVENS'12)*, 128–136, Vienna, Austria, 2012.

Soraluze A., Arregi O., Arregi X., and Díaz de Ilarraza A. Coreference Resolution for Morphologically Rich Languages. Adaptation of the Stanford System to Basque. *Procesamiento del Lenguaje Natural*, 55:23–30, 2015a.

Soraluze A., Arregi O., Arregi X., and Díaz de Ilarraza A. Korreferentzia-ebazpena euskaraz idatzitako testuetan. *I. Ikergazte: Nazioarteko ikerketa euskaraz*, 676–684, Durango, 2015b.

Soraluze A., Arregi O., Arregi X., and Díaz De Ilarraza A. Improving mention detection for Basque based on a deep error analysis. *Natural Language Engineering*, 23(3):351–384, 2016a.

Soraluze A., Arregi O., Arregi X., Díaz de Ilarraza A., Kabadjov M., and Poesio M. Coreference Resolution for the Basque Language with BART. *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, 67–73, San Diego, California, 2016b.

Stamborg M., Medved D., Exner P., and Nugues P. Using syntactic dependencies to solve coreferences. *Joint Conference on EMNLP and CoNLL - Shared Task*, 64–70, Jeju Island, Korea, 2012.

Stede M. The Potsdam Commentary Corpus. *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, DiscAnnotation '04, 96–102, Barcelona, Spain, 2004.

Steinberger J., Poesio M., Kabadjov M.A., and Jeek K. Two Uses of Anaphora Resolution in Summarization. *Information Processing and Management*, 43(6):1663–1680, 2007.

Stoyanov V., Gilbert N., Cardie C., and Riloff E. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-art. *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*, ACL '09, 656–664, Suntec, Singapore, 2009.

Strube M. and Ponzetto S.P. Wikirelate! computing semantic relatedness using wikipedia. *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, 1419–1424, Boston, Massachusetts, 2006.

Strube M., Rapp S., and Müller C. The Influence of Minimum Edit Distance on Reference Resolution. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, 312–319, Philadelphia, Pennsylvania, 2002.

Suchanek F.M., Kasneci G., and Weikum G. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. *Proceedings of the 16th International World Wide Web Conference (WWW 2007)*, 697–706, 2007.

Tetreault J.R. A Corpus-based Evaluation of Centering and Pronoun Resolution. *Computational Linguistics*, 27(4):507–520, 2001.

Tutin A., Trouilleux F., Clouzot C., Gaussier É., Zaenen A., Rayot S., and Antoniadis G. Annotating a large corpus with anaphoric links. *Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC 2000)*, page 2, United Kingdom, 2000.

Uryupina O. Coreference Resolution with and without Linguistic Knowledge. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, 2006. ACL Anthology Identifier: L06-1453.

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

Uryupina O. Corry: A System for Coreference Resolution. *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), 100–103, Uppsala, Sweden, 2010.

Uryupina O. and Moschitti A. Multilingual mention detection for coreference resolution. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 100–108, Nagoya, Japan, 2013.

Uryupina O., Moschitti A., and Poesio M. BART Goes Multilingual: The UniTN/Essex Submission to the CoNLL-2012 Shared Task. *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, 122–128, Jeju Island, Korea, 2012.

Uryupina O., Poesio M., Giuliano C., and Tymoshenko K. Disambiguation and Filtering Methods in Using Web Knowledge for Coreference Resolution. In Murray R.C. and McCarthy P.M., editors, *FLAIRS Conference*, 317–322. 2011.

van Deemter K. and Kibble R. On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26:629–637, 1995.

Vapnik V.N. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

Versley Y., Moschitti A., Poesio M., and Yang X. Coreference systems based on kernels methods. *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, 961–968, Stroudsburg, PA, USA, 2008a.

Versley Y., Poesio M., and Ponzetto S. Using Lexical and Encyclopedic Knowledge. In Poesio M., Stuckardt R., and Versley Y., editors, *Anaphora Resolution: Algorithms, Resources, and Applications*, 393–429. Springer Berlin Heidelberg, Berlin, Heidelberg, 2016.

Versley Y., Ponzetto S.P., Poesio M., Eidelman V., Jern A., Smith J., Yang X., and Moschitti A. BART: A Modular Toolkit for Coreference Resolution. *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, HLT-Demonstrations '08, 9–12, Columbus, Ohio, 2008b.

Vicedo J. and Ferrández A. Coreference In Q&A. *Advances in Open Domain Question Answering*, 32 lib. of *Text, Speech and Language Technology*, 71–96. Springer, 2006.

Vieira R. and Poesio M. Corpus-based Development and Evaluation of a System for Processing Definite Descriptions. *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00, 899–903, Stroudsburg, PA, USA, 2000.

Vilain M., Burger J., Aberdeen J., Connolly D., and Hirschman L. A Model-theoretic Coreference Scoring Scheme. *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, 45–52, Columbia, Maryland, 1995.

Wagner A. and Zeisler B. A Syntactically Annotated Corpus of Tibetan. *Proceedings of the Fourth International Conference on Language Resources and Evaluation, (LREC'04)*, Lisbon, Portugal, 2004.

Walker C., Strassel S., Medero J., and Maeda K. Ace 2005 Multilingual Training Corpus. LDC2006T06. *Linguistic Data Consortium, Philadelphia*, 2006.

Weischedel R., Palmer M., Marcus M., Hovy E., Pradhan S., Ramshaw L., Xue N., Taylor A., Kaufman J., Franchini M., El-Bachouti M., Belvin R., and Houston A. *Ontonotes Release 5.0.* Linguistic Data Consortium, Philadelphia, Pennsylvania, 2103.

Wiseman S., Rush A.M., and Shieber S.M. Learning Global Features for Coreference Resolution. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 994–1004, San Diego, California, 2016.

Woods W., Kaplan R., and Nash-Webber B. The Lunar Sciences Natural Language Information System: Final Report. Barne-txostena, Bolt, Beranek and Newman, Inc., Cambridge, MA, 1974.

Xiong H. and Liu Q. ICT: System Description for CoNLL-2012. *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, 71–75, Jeju Island, Korea, 2012.

Xu R., Xu J., Liu J., Liu C., Zou C., Gui L., Zheng Y., and Qu P. Incorporating Rule-based and Statistic-based Techniques for Coreference Resolution. *Joint Conference on EMNLP and CoNLL - Shared Task*, 107–112, Jeju Island, Korea, 2012.

Yang X., Su J., and Tan C.L. A Twin-Candidate Model for Learning-Based Anaphora Resolution. *Computational Linguistics*, 34(3):327–356, 2008.

Yang X., Zhou G., Su J., and Tan C.L. Coreference Resolution Using Competition Learning Approach. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, 176–183, Sapporo, Japan, 2003.

Yangy X., Su J., Zhou G., and Tan C.L. An NP-cluster Based Approach to Coreference Resolution. *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Geneva, Switzerland, 2004.

Yuan B., Chen Q., Xiang Y., Wang X., Ge L., Liu Z., Liao M., and Si X. A mixed deterministic model for coreference resolution. *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, 76–82, Jeju Island, Korea, 2012.

Zhang L. Maximum Entropy Modeling Toolkit for Python and C++ (version 20041229). *Natural Language Processing Lab, Northeastern*, 2004.

Zhang X., Wu C., and Zhao H. Chinese coreference resolution via ordered filtering. *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, 95–99, Jeju Island, Korea, 2012.

Zhekova D. and Kübler S. UBIU: A Language-Independent System for Coreference Resolution. *Proceedings of the 5th International Workshop on Semantic Evaluation*, (SemEval 2010), 96–99, Uppsala, Sweden, 2010.

Žitkus V. and Nemuraitė L. First Steps in Automatic Anaphora Resolution in Lithuanian Language Based on Morphological Annotations and Named Entity Recognition. In Dregvaite G. and Damasevicius R., editors, *Information and Software Technologies: 21st International Conference, ICIST 2015, Druskininkai, Lithuania, October 15-16, 2015, Proceedings*, 480–490. Springer International Publishing, Cham, 2015.