universidad del país vasco — euskal herriko unibertsitatea

Facultad de Informática — **Informatika Fakultatea**

Wikipedia eta Anbiguetate Lexikala

Ikaslea: Jn/An. Jokin Perez de Viñaspre Garralda
Zuzendariak: Jn/An. Eneko Agirre Bengoa
Jn/An. Aitor Soroa Echave

Karrera Bukaerako Proiektua, 2015eko uztaila

LABURPENA

Hemen aurkezten dugun txostena Ingeniaritza Informatikako Karrera Bukaerako Proiektu (KBP) batean sortu dugun memoria da. Bertan, izen-entitateak desanbiguatzeko bi tresna ebaluatu ditugu, zenbait datu multzo estandar erabilita.

Proiektua Eusko Jaurlaritzako lankidetza beka batekin egin dugu, EHUko IXA ikerketa taldearen baitan. *QTLeap* Europako proiektuan egin dugu lan, zeinek itzultzaile automatikoen mugak aztertu nahi dituzten. Bertan, hain zuzen, desanbiguazio edo izen-entitateen estekatze automatikoaren inguruan lan egin dugu.

Esan bezala 2 tresna ebaluatu ditugu, *Dbpedia Spotlight* eta UKB. Bestalde, hiru hizkuntzetan egin ditugu esperimentuak: ingelesean, gaztelanian eta euskaran. *Dbpedia Spotlight* tresna ingelesean eta gaztelanian erabili ditugu eta UKB aldiz, gaztelanian eta euskaran. Izen-entitateak desanbiguatzeko erabili dugun ezagutza-basea *Wikipedia* izan da.

Dbpedia Spotlight tresnaren erabileraren inguruan ikertu egin dugu, besteak beste bere desanbiguatzeko eta erantzuteko ahalmena, baita honen azkartasuna eta memoria kostua ere. Honetaz gain, tresna honen analisi sakona egin dugu, IXA taldearen baitan tresna honen ezagutza areagotu dadin.

Edukien aurkibidea

1. SARRERA....1

2. PROIEKTUAREN HELBURU-DOKUMENTUA....3
2.1. Irismena....3
2.2. Atazen zerrenda....5
2.3. Lanaren Deskonposaketa Egitura (LDE)....7
2.4. Lan metodologia....8
2.5. Denbora-estimazioa....8
2.5.1. Gantt diagrama....8
2.5.2. Estimazioak....10
2.6. Arriskuak....11
2.6.1. Arriskuen zerrenda....12
2.6.2. Kontingentzia-plana....14

3. AURREKARIAK....17
3.1. QTLeap....17
3.2. Wikipedia....17
3.2.1. Desanbiguazio orriak....18
3.2.2. Birbideratzeak....19
3.2.3. Aingurak....20
3.2.4. Hizkuntzen arteko loturak....21
3.3. Wikipediaren esportazioak....21
3.4. DBpedia Spotlight....22
3.5. UKB....22
3.6. Matx....24
3.7. Datu-multzoak (datasets)....25

3.7.1. TAC-KBP....26
3.7.2. AIDA / CoNLL....28
3.7.3. Euskarazko datu-multzoa....29
3.8. Ebaluazio-neurriak....31
3.9.I xati....32

4. SOFTWAREAREN ANALISIA....33
4.1. Dbpedia Spotlight....33
Web aplikazioa....33
Web Zerbitzaria....36
4.1.1. Desanbiguaziorako aukera ezberdinak....36
Spotlight Lucene....37
Spotlight Statistical....37
4.1.2. Parametroak....38
Testua / helbidea....39
Iragazkiak....39
Spotter....41
4.1.3. Ereduak....43
Eredu osoak(full)....44
Eredu txikituak(light)....44
4.1.4. Moduluak....45
Spot....45
Disambiguate....46
Annotate....46
Candidates....47
4.1.5. Sarrera / irteera formatuak....47
4.2. UKB....49

4.2.1. Parametroak....49
4.2.2. Baliabideak....50
Grafoak....50
Hiztegia....51
4.2.3. Sarrera / irteera formatuak....52

5. ESPERIMENTUAK....55
5.1. Ingelesezko esperimentuak....55
5.1.1. Lehenengo urratsa, aurreprozesamendua....56
5.1.2. Entitateen bihurketa....57
5.1.3. Ebaluaziorako datuak biltzen....58
5.1.4. Parametroen ikerketa....59
Iragazki parametroak....59
Atalasea....60
Aipamena....62
Testuingurua....64
Eredua....67
5.1.5. Denbora aurrezten....69
5.1.6. Errore analisiak....70
5.1.7. Ebaluazioaren emaitzak....72
5.2. Gaztelaniazko esperimentuak....73
5.2.1. Bi tresna testuinguru berdina....74
5.2.2. Matxen erabilera....75
5.2.3. Beharrezko aldaketak mapaketa....75
5.2.4. Spotlightekin desanbiguatzen....76
5.2.5. UKBrekin desanbiguatzen....80
5.2.6. Ezagutza-basea handitzen....82

5.2.7. Konparaketa....84
5.3. Euskarazko esperimentuak....85
5.3.1. I. Fernandezen datu-multzoarekin lanean....85
5.3.2. UKBren baliabideak....86
5.3.3. Datu-multzoa eta baliabideak ebaluatzeko algoritmo batzuk....87
MFS....87
Goi-bornea....88
5.3.4. Lematizazioa....89
5.3.5. UKBrekin desanbiguatzen....90
5.3.6. Aipamena....92
5.3.7. Ebaluazioa....95

6. ONDORIOAK ETA ETORKIZUNEKO LANAK....97
6.1. Ondorioak....97
6.1.1. Proiektuaren ondorioak....97
6.1.2. Estimazioak eta errealitatea....99
6.1.3. Ondorio pertsonalak....104
6.2. Etorkizuneko lana....105
6.2.1. Spotlighten inguruko etorkizuneko lanak....105
6.2.2. UKBren gaineko etorkizuneko lanak....106

BIBLIOGRAFIA....107
Liburuak....107
Artikuluak....108
Internet loturak....110

1. SARRERA

Entitateak Estekatzeko (EE) edo *Entity Linking* (EL) sistemak ebaluatuko ditugu proiektu honetan. Sistema hauen bidez testuetako izen-aipamenak ezagutza-baseetako entitateekin lotzen dira.

Guztioi inoiz gertatu zaigu, artikulu edo testu bat irakurtzen gauden bitartean, bertan ageri den izen bat interesgarria iruditzea. Jakin mina pizten zaigu eta nonbaitera jo behar dugu izen horretaz gehiago jakiteko. Orain dela urteak, liburuak miatzen genituen informazio gehiagoren bila. Gaur egun, ordea, Interneti esker segundo gutxiren buruan sarean dagoen edozer aurkitzeko gai gara.

Baina, gaur egunean, testuetako izen-entitateak ezagutza-baseetako entitateekin lotuta aurkitzea gero eta errazagoa da. Webguneetan ohikoa izaten da izen horiek *Wikipedia*ko artikuluetara lotuta egotea. Modu honetan, testuak ezagutza-baseetako informazioarekin aberasten dira, erabiltzailearen jakin mina asetzeko beharrezko denbora gutxituz.

1 Irudia: Webguneetan gero eta ohikoagoak diren Wikipedia orrietara egiten diren loturak ikus daitezke. Loturak esteken bidez egiten dira.

Adibidez, 1 irudian *John Adams*i buruzko testu bat ageri da. Bertan, "*John Adams*" esteka *Wikipedia*ko "*John_Adams*" artikulura zuzenduta dago. Hiperestekari esker irakurleak klik batean eskura dezake *John Adams*i buruzko informazioa.

Loturak ezartzeko, lehendabizi, testuan agertzen diren izen-aipamenak identifikatu behar dira. Dena dela, izen-aipamena ez da nahikoa lotura ezartzeko. Izan ere, izen-aipamen bakoitzari erreferentzia egiten dion entitate posible guztiak aztertu behar dira. Horiei desanbiguaziorako hautagaiak deritze. Azken urratsean, izen-aipamena hautagaien arteko egokienarekin lotuko da.

Beraz, EE sistema automatikoak hiru urrats edo ataza nagusi ditu: izen entitateak topatzea, hauen hautagai posibleak lortzea eta, azkenik, hautagaien artean onena erantzun moduan itzultzea.

Desanbiguatzerako orduan, ordea, zenbait arazo agertzen dira, anbiguotasuna eta aldakortasuna. Mundu zabal honetan zaila egiten da erakunde, toki edota pertsona baten izena bakarra izatea. Adibidez, *John Adams* izenarekin ingelesezko *Wikipedia*n, soilik politikoak kontuan izanik, 15 pertsona erreferentziatuak daude. 5 Irudian *John Adams*en desanbiguazio orria aurki daiteke (19 orrialdean). Gainera, entitate bat izendatzeko zenbait modu egon daitezke; *John Adams*ekin jarraituz, bere izenarekin erreferentziatua izateaz gain, AEBetako bigarren presidente bezala ere erreferentziatzen da.

Eskerrak, gaur egun, lan hori automatikoki egiten dituzten tresnak erabilgarri dauden. Proiektu honetan tresna hauek ebaluatuko dira ingelesa, gaztelania eta euskara hizkuntzetan.

2. PROIEKTUAREN HELBURU-DOKUMENTUA

Proiektuaren Helburu-Dokumentuan (PHD), proiektuan egikarituko denaren azalpen zehatza egingo dugu. Proiektuaren helburuak azaltzeaz gain, denbora--estimazioak, proiektuaren atazen zehaztapenak, gerta litezkeen arazoen kontingentzia-plana eta lan-metodologia azalduko ditugu.

2.1. Irismena

Proiektu honetan Hizkuntzaren Prozesamenduaren HParen) barne dagoen entitate-izenen desanbiguazio-munduan murgilduko gara. EHUko Donostiako Informatika Fakultateko IXA taldearekin elkar-lanean egin dugu, Eusko Jaurlaritzako lankidetza beka bati esker. Horregatik *QTLeap* proiektuan haientzat baliagarriak egin ahal zaizkien tresnak ebaluatuko ditugu hizkuntza ezberdinetako datu-multzo estandarretan. Tresna horiek *Dbpedia Spotlight* estatistikoa eta UKB tresnak dira, hurrenez hurren.

Hasteko, entitate-izenen desanbiguazioari buruzko informazioa bilatuko dugu, terminologiarekin trebatzeko asmoz. Ostean, proiektuan zehar erabiliko ditugun PERL eta BASH programazio lengoaiekin arituko gara, aurrerago arazo gutxiago edukitzeko.

Proiektuaren garapena hiru hizkuntzetarako egingo dugu: Ingelesa, gaztelania eta euskara. Lehenengo biak *Spotlight* estatistikoan ebaluatuko ditugu; euskara eta gaztelania, aldiz, UKBrekin. Hortaz, gaztelania bi tresnekin ebaluatuko dugu, *Spotlight*ekin eta UKBrekin.

Etorkizunean IXAk *Dbpedia Spotlight* estatistikoarekin lan egin ahal izateko beharrezko informazioa bildu beharko dugu ere. Horregatik, software analisi kapitulua sortu dugu eta honen baitan dagoen *Spotlight*i buruzko 4.1 azpi-ataza da horren mamitsua.

Ebaluazioa egin ahal izateko 6 datu-multzo erabiliko ditugu. Hauetatik hiru ingeleserako izango dira, TAC 2010 eta 2011 eta AIDA. Gaztelaniarako, ordea, TAC 2012. Euskararako (I. Fernandez, 2012) eskuz sortutako 2 datu-multzoak erabiliko ditugu.

Proiektua aurrera eramateko 7 iterazio identifikatu ditugu; ondorengo zerrendan azalduko dugu bakoitzean egingo duguna labur-labur:

1. Softwarearen analisia: *Dbpedia Spotlight* eta UKB softwareen analisia egingo dugu. *Spotlight*en kasuan, IXAk tresna erabili ahal izateko informazio osatua bildu beharko dugu. UKBrekin, ordea, IXAk garatutako tresna izanik, soilik proiektuan erabili ahal izateko informazioa bilduko dugu.
2. Garapena ingeleserako: Iterazio honetan *Spotlight* tresnarekin probak egingo ditugu parametro onenak hautatzeko asmoz. Horretarako TAC 2010 datu-multzoa erabiliko dugu.
3. Testa ingeleserako: Aurreko iterazioan lortutako parametroak aplikatuko ditugu TAC 2011 eta AIDA datu-multzoen gainean.
4. Gaztelania *Spotlight*en: Ingelesean lortutako parametro onenak erabiliko ditugu *Spotlight*en gaztelaniako eredua ebaluatzeko, horretarako TAC 2012 datu-multzoa erabiliko dugu. Aukera ikusten dugu proba ezberdinak egitea ikerketa osatuagoa izan dadin.

5. Gaztelania UKBn: UKB tresna ebaluatuko dugu zenbait baliabide ezberdin erabilita; aurreko iterazioko datu-multzo berdina erabiliko dugu. Aurrekoan bezala, ikerketa hobea egiteko proba gehiago egitea aurreikusten dugu; horretarako lehenago egindako emaitzen ondorioez baliatuko gara.
6. Garapena euskararako: Euskaran egingo diren probak UKB tresnarekin izango dira, I. Fernandezen garapen datu-multzoa erabiliko dugularik. Hau moldatu beharko dugu aurreko datu-multzoekin erabilitako softwarea berrerabili ahal izateko. Behin hori eginda, grafo ezberdinekin, lematizazioarekin eta lematizaziorik gabe eta desanbiguatu nahi den entitatea ezagututa edo ezagutu gabearekin proba ezberdinak egingo ditugu.
7. Testa euskararako: Aurreko iterazioan lortutako “parametro” onenak Fernandezen ebaluazio datu-multzoan testatuko ditugu.

2.2. Atazen zerrenda

Proiektua lau ataza nagusietan banatu dugu, prozesu taktikoak, prozesu operatiboak, formakuntza eta dokumentazioa.

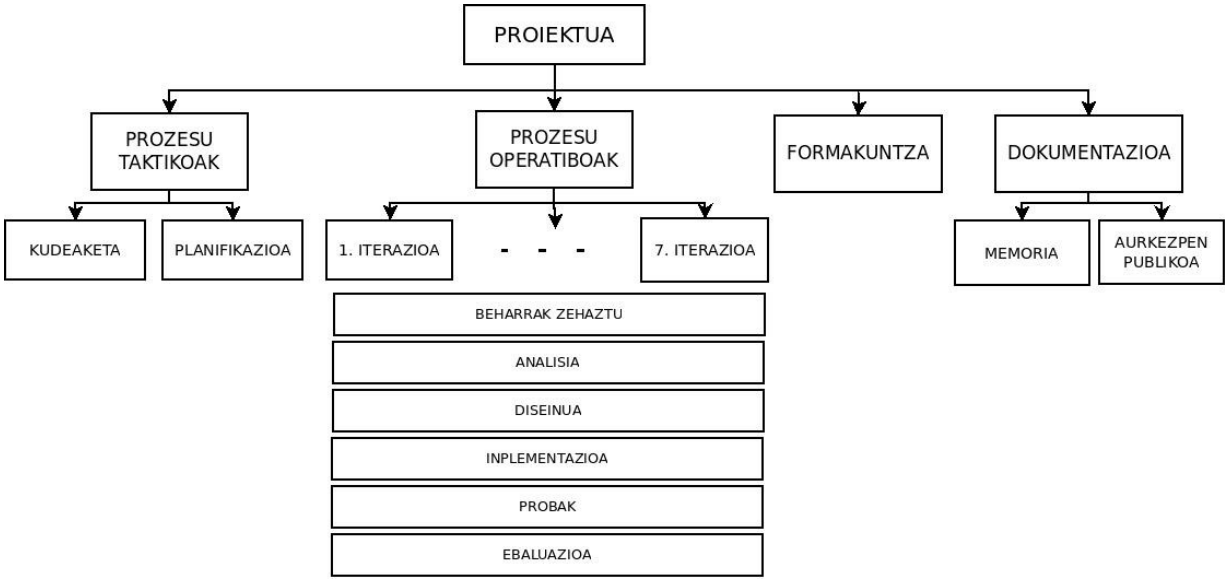
- Prozesu taktikoetan, kudeaketari eta planifikazioari dagozkien azpiataza guztiak daude, hala nola, bilerak egikaritzea, lan gaztiguaren kudeaketa, Proiektuaren Helburu Dokumentua (PHD), babes kopiak egitea eta abar.
- Prozesu operatiboetan, ordea, proiektuko helburuen zehaztapena, diseinua eta sorkuntzarekin loturiko azpiatazak daude. Aipatzekoa da jarraitutako metodologia dela eta, ur-jauzi metodologia, azpiataza hauek iterazio ezberdinetan zehar errepikatuko ditugula software edota datu multzo ezberdinak erabiliz. 7 iterazio egin ditugu orotara. Eta guztietan, ondorengo azpiatazak landu ditugu:

- “Beharrak zehaztu” atazan, iterazio bakoitzean egin beharrekoak zehaztuko ditugu. Horretarako, aurrez planifikatutakoa eta ordura arte egindako proben ondorioak kontuan hartuko ditugu.
- Analisian, aldiz, iterazioen beharrak eskatzen dituen ezinbesteko azterketak egingo ditugu. Hala nola, erabiliko ditugun baliabideenak, hots, tresnak, datu-multzoak, hiztegiak, etab.
- Diseinuan, analisitik ondorioztatutako beharrak asetuko dituen aplikazioa diseinatu beharko dugu.
- Inplementazio atazan, iterazio bakoitzean diseinatutako aplikazioa garatuko dugu, baita erabiliko diren baliabideak sortzeko edo moldatzeko beharrezkoak diren programa lagungarriak ere.
- Probetan, garaturiko programa nagusiaren gainean probak egingo ditugu, garapena burutzen den heinean eta amaitutakoan, eta baita aplikazio lagungarri guztiei, berauek bukatu ahala. Garatutako programa guztien funtzionamendu egokia ziurtatzeko asmoz.
- Ebaluazioari dagokionean, iterazio bakoitzean aukeratu diren datu-multzoak eta tresna ebaluatuko ditugu garaturiko programa nagusiaz lagunduta. Lortutako emaitzen azterketa egingo dugu ere.

- Dokumentazioaren atazak proiektuaren memoria, eta aurkezpena barnebiltzen ditu.
- Eta azkenik, proiektuaren gauzapenerako beharrezkoak diren gaitasunak lortzea barnebiltzen dituen formakuntza ataza dago. Hala nola, *DBpedia Spotlight* eta UKB tresnen funtzionamendua ulertzea eta beharrezkoak diren programazio lengoaiak ikastea.

2.3. Lanaren Deskonposaketa Egitura (LDE)

Lanaren Deskonposaketa Egitura diagrama proiektuan egingo den lanaren banaketa adierazten duen irudia da. Era honetan, proiektuan egin beharrekoak multzokatzen dira ataza ezberdinetan. Proiektuak lau ataza nagusi ditu; prozesu taktikoak, prozesu operatiboak, formakuntza eta dokumentazioa. 2 Irudian ikus daiteke egindako LDE diagrama.



2 Irudia: LDE diagrama. Bertan modu grafikoan proiektua garatzeko egin beharreko lana adierazten da. 4 ataza nagusiak eta bere baitako azpi-atazak ikus daitezke.

2.4. Lan metodologia

Proiektua aurrera eramateko iterazioetan oinarritutako metodologia[1] erabiliko dugu. Proiektuaren garapena iterazioetan zatitzen dugu. Honen bidez, proiektua konplexutasun txikiagoa duten proiektu txikiagoetan bilakatzen dugu. Hauetan, era berean, edozein proiektutan bezala, beharrak zehaztu, analisia, diseinua, inplementazioa, probak eta ebaluazio faseak egingo ditugu. Iterazio bakoitzari esker emaitzak hobetzen joango gara; hala nola, zehaztasun handiagoa lortuz edota funtzionalitate berriak gehituz.

Iterazioetan aipatzen ez badugu ere, iterazio bakoitza egin bitartean memoria osatuko dugu. Gainera, iterazioetatik kanpo geratzen den PHD egingo dugu iterazioekin hasi aurretik, hauek definitu ahal izateko.

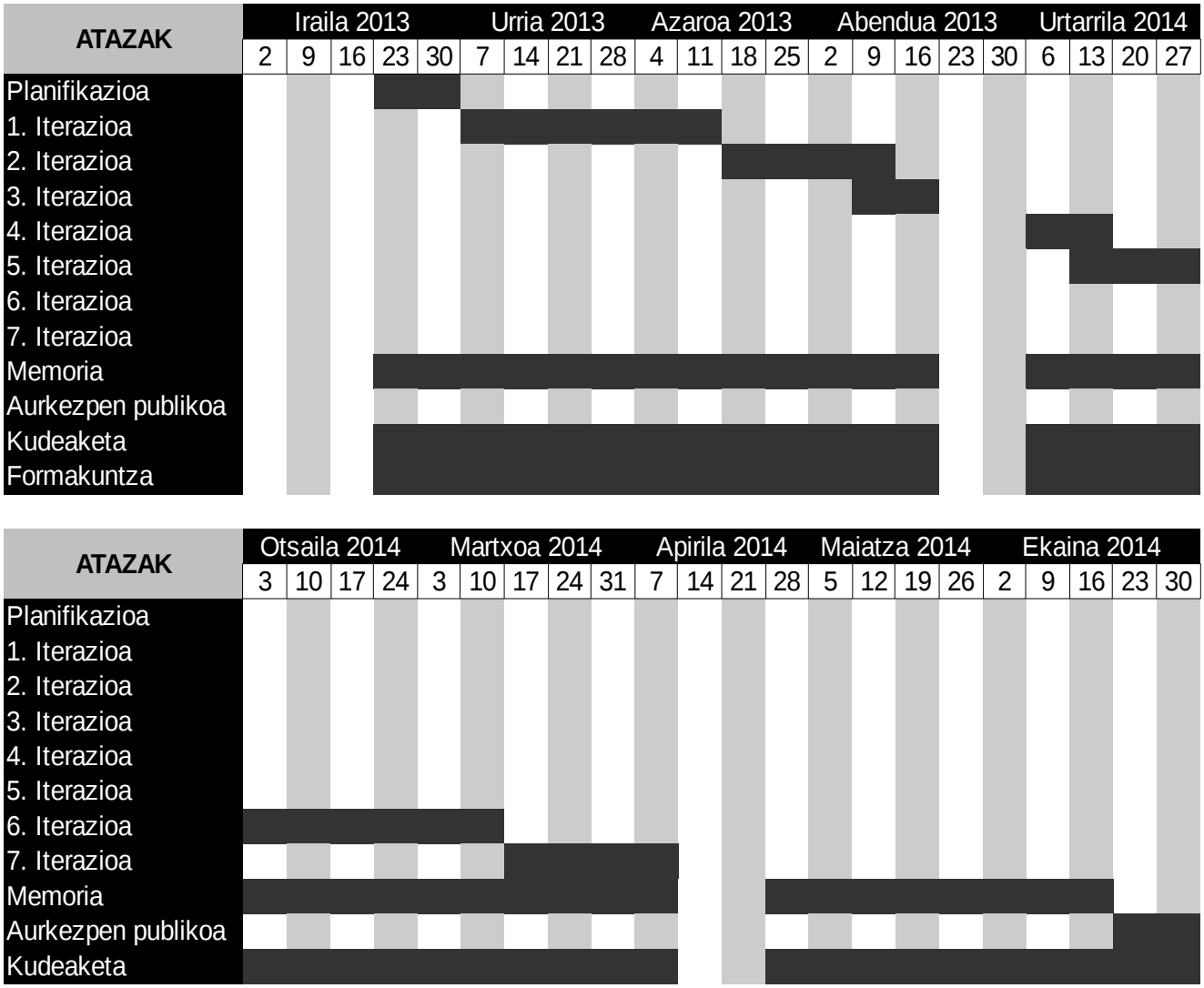
2.5. Denbora-estimazioa

Proiektua behar bezala egiteko, beharrezkoa da denboraren aurreikuspena egitea. Ondorengo ataletan adieraziko ditugu *Gantt* diagrama eta denbora-estimazioa.

2.5.1. *Gantt* diagrama

Atal honetan, planifikatutako ekintzak egutegian kokatuko ditugu adierazpide grafikoa erabiliz, *Gantt* diagrama. Era honetan, modu grafikoan ataza bakoitzerako aurreikusi dugun epea adierazi daiteke. Amaieran, 6.1.2 atalean, estimatutako denbora denbora errealarekin konparatuko dugu.

1 https://en.wikipedia.org/wiki/Iterative_and_incremental_development



3 Irudia: Gantt diagrama

3 Irudian ikus daitekeen *Gantt* diagrama, 2013/2014 ikasturtean zehar garatzeko diseinatuta dago. Irailaren amaieran hasi eta uztaileko deialdian bukatzen da. Abenduaren 23tik urtarrilaren 6ra arte eta apirilaren 14tik 27ra arte oporrak egongo dira eta ez da proiektuan aurrerapausorik emango. Proiektua ekainean amaitzea aurreikusten dugu.

2.5.2. Estimazioak

Ondoren, atazaka egindako denbora-estimazioak zehazten ditugu. Zehatz mehats, ataza bakoitzerako egindako estimazioak dira hauek, ordutan neurtuta. Txosten honen amaieran, egindako denbora-estimazioak eta errealak alderatuko ditugu.

Azaldu beharra dago, nahiz eta KBP batek 300 ordu izan behar dituen, nirea lankidetza beka baten barne dagoenez ordu kopurua handiagoa izan dela.

Atazak	Estimazioak
Guztira	384 ordu
Prozesu taktikoak	84 ordu
K- Kudeaketa	
K1- Bilerak	50 ordu
K2- Artxiboen kudeaketa	5 ordu
K3- Lan gaztiguen kudeaketa	4 ordu
P- Planifikazioa	
P1- Proiektuaren Helburu Dokumentua (PHD) egin	25 ordu
Prozesu operatiboak	160 ordu
I- Iterazioak (7 iterazio)	
I1- 1.Iterazioa	50 ordu
I2- 2.Iterazioa	30 ordu
I3- 3.Iterazioa	10 ordu
I4- 4.Iterazioa	15 ordu
I5- 5.Iterazioa	20 ordu
I6- 6.Iterazioa	25 ordu
I7- 7.Iterazioa	10 ordu
Dokumentazioa	90 ordu
M- Memoria	80 ordu
A- Aurkezpen publikoa	10 ordu
Formakuntza	50 ordu

4 Irudia: Estimazioak

4 Irudian ikus dezakegunez, ordu karga handiena prozesu operatiboei emango diegu. Formakuntza atalari ordu kopuru esanguratsua emango diogu, bertan tresna eta programazio lengoaia berrien ikasketari aurreikusten diogun denbora azaltzen baita. Dokumentazio atalak ere ordu kopuru handia dauka, karrera bukaerako proiektu bat izanik, egindako lan guztia oso ondo dokumentatu beharko dugulako. Aipatzekoa da bileren ordu kopurua, IXA ikerketa taldeak parte hartzen dagoen proiektuaren baitan lanean gaudenez eta talde lanean aurrera eramaten dugunez proiektua, koordinazio bilera kopurua handia da.

2.6. Arriskuak

Edozein proiektutan berebiziko garrantzia dauka gerta litezkeen arriskuak ezagutzea eta hauen eragina ahalik eta txikiena izateko hartu beharreko neurriak prest edukitzea. Horregatik, jarraian, arriskuen zerrenda eta hauen eragina murrizteko jarraituko dugun kontingentzia plana dago.

2.6.1. Arriskuen zerrenda

1 Taulan adierazten ditugu proiektuan zehar gerta litezkeen arriskuak, gertatzeko probabilitatearekin eta gertatzekotan honek dakarren larritasun-mailarekin.

Arriskua	Probabilitatea	Larritasun-maila
Formakuntzan arazoak	Ertaina	Txikia
Formakuntzako informazio galera	Txikia	Txikia
Bilera ezin egin	Ertaina	Txikia
Egitekoetan atzerapenak	Ertaina	Ertaina
Ordenagailua apurtzea	Oso txikia	Ertaina
Datu galera	Txikia	Ertaina
Softwarea erabiltzean arazoak	Ertaina	Ertaina
Emaitza arraroak lortzea	Ertaina	Ertaina
Beharrezko baliabideak garaiz ez edukitzea	Txikia	Handia
Softwarearen bertsio aldaketak	Ertaina	Handia

1 Taula: Arrisku-taula.

Jarraian, arriskuen taulan adierazitako arriskuak azalduko ditugu deskribapen txiki batekin:

- **Formakuntzan arazoak**

Proiektua egiteko beharrezkoa den informazioa aurkitzeko arazoak eduki daitezke; edota nahiz eta informazioa aurkitu, hau ulertzeko zailtasunak eduki.

- **Formakuntzako informazio galera**

Dagoeneko formatzeko aurkitutako informazioa gal daiteke.

- **Bilera ezin egin**

  Bileran parte hartu beharrekoak ezin bertaratzearen ondorioz, bilera ezin egitea.

- **Egitekoetan atzerapenak**

  Egitekoa egiteko estimatutako denbora baino gehiago behar izatea.

- **Ordenagailua apurtzea**

  Ordenagailua apurtzea eta, bertan proiektuarekin lanean ezin jarraitzea

- **Datu galera**

  Proiektuaren kudeaketa fitxategiak galtzea, proiektuaren garapeneko fitxategiak edota emaitzak galtzea.

- **Softwarea erabiltzean arazoak**

  Proiektua egiteko erabiltzen den softwarea maneiatzeko zailtasunak edukitzea.

- **Emaitza arraroak lortzea**

  Ebaluatzean lortutako emaitzak esperotakoen oso bestelakoak izatea.

- **Beharrezko baliabideak garaiz ez edukitzea**

  Proiektua garatzeko beharrezkoak diren baliabideak garaiz ez edukitzea.

- **Softwarearen bertsio aldaketak**

  Proiektu honetan zehar erabiliko diren softwarearen bertsio eguneratzea. Bi motatakoak aurreikusten dira, dagoeneko erabiltzen diren bertsioen hobekuntza txikia eta bertsioaren funtzionamenduan aldaketak dituztenak.

2.6.2. Kontingentzia-plana

Aurreko arriskuen aurrean, hurrengo kontingentzia-plana diseinatu dugu, arazoek ahalik eta eragin txikiena izan dezaten proiektuan:

- **Formakuntzan arazoak**

Zuzendariekin hitz egingo da informazio gehiago lortzeko edo ulertzeko beharrezkoak diren azalpenak jasotzeko.

- **Formakuntzako informazio galera**

Informazio hori ea babes kopian dagoen begiratuko da; hor aurkitzen ez bada, berriz ere interneten bilatuko da. Hala ere, arazoekin jarraituz gero, zuzendariei eskatuko zaie informazio bibliografiko osagarria.

- **Bilera ezin egin**

Egin beharreko bilerak etorkizuneko lana baldintzatzen baldin badu, orduan beste hitzordu bat jarriko da partaide guztien artean. Bestela, hurrengo asteko bileran berreskuratuko da. Gainera, korreoz konpontzen ahaleginduko da.

- **Egitekoetan atzerapenak**

Zuzendariei adierazi beharko zaie egitekoa burutzeko denbora gehiago beharrezkoa dela eta hori eragindakoaren zergatiak azalduko zaizkie; hiruren artean entregatzeko epe berria erabakiko da. Azkenik, atzerapenaren erregistroa eguneratuko da, planifikazioan aldaketak egiteko.

- **Ordenagailua apurtzea**

Proiektuaren informazioa berreskuratuko da babes kopiatik, beste ordenagailu batean lan egiten jarraitu ahal izateko.

- **Datu galera**

Babes kopiatik berreskuratuko dira galdutako datuak. Gainera, galera zergatik gertatu den analizatuko da berriz ere gerta ez dadin.

- **Softwarea erabiltzean arazoak**

Interneten softwarea erabiltzeko beharrezkoak diren gidak bilatuko dira. Bestela, foro espezializatuetan begiratuko da eta azalpenik ez baldin badaude bertan galdetuko da. Azken aukera moduan zuzendariei galdetuko zaie.

- **Emaitza arraroak lortzea**

Emaitza hori lortzeko jarraitu diren urratsak egiaztatuko dira. Zergatia aurkitzen ez baldin bada, emaitzen mapaketa egingo da emaitza arraroa duen adibide bat bilatzeko asmoz. Hala ere, zergatia aurkitzen ez baldin bada, zuzendariei galduko zaie.

- **Beharrezko baliabideak garaiz ez edukitzea**

Baliabidea lortu behar duenari korreoz gogoraraziko zaio baliabidea dagoeneko beharrezkoa dela. Erantzuna jaso bitartean beste egitekoekin jarraituko da denbora aprobetxatzeko asmoz. Bilera egunean erantzunik jaso ez baldin bada, bileran azalduko da konponbidea aurkitzeko asmoz.

- **Softwarearen bertsio aldaketak**

Softwarearen bertsio berriaren azterketa egingo da, eskura dagoen informazioa erabilita. Honekin aurretik definitu diren bertsio aldaketa mota identifikatuko da. Lehenengo motatakoa bada, ez du eragingo proiektuaren nondik norakoetan. Bigarren aldaketa mota gertatzen bada, ordea, analisi sakonagoa egingo da; proba txikiak eginez aldaketaren eragin erreala neurtzeko asmoz. Ondoren, beharrezkoa bada softwarearen bertsio berria erabiltzeko deiak moldatuko dira.

3. AURREKARIAK

Atal honetan proiektuko helburuak lortzeko erabilitako baliabideak azalduko ditugu. Baita proiektua kokatzeko beharrezko atalak.

3.1. QTLeap

QTLeap[2] Europako proiektua da, bere helburua gaur egungo itzultzaile automatikoen mugak gainditzea izanik. Horretarako Europako herrialde ezberdinetako zazpi unibertsitate elkarlanean aritzen dira; Pragako Charles Unibertsitatea, Berlingo Humboldt Unibertsitatea, Lisboako Unibertsitatea, Bulgariako Zientzien Akademia IICT-BAS, Alemaniako DKFI, Herbehereetako Groningengo Unibertsitatea eta bertako EHU, IXA ikerkuntza taldea hain zuzen ere.

IXA taldeak hiru ardatz nagusitan dihardu lanean. Batetik zuhaitz-bankuak (*tree-bank*), bestetik Interneteko baliabideak eta, azkenik, izen-entitateak. Gure proiektuan jorratu ditugunak azkeneko biak izan dira.

3.2. Wikipedia

Proiektu honetako ezagutza-basea *Wikipedia*ko entitateek osatzen dute. Entitate eta *Wikipedia* artikuluen arteko lotura zuzena da: *Wikipedia*n “http://en.wikipedia.org/wiki/John_Adams” artikulua bada, entitatea “*Jonh Adams*” izango da.

2 http://qtleap.eu/

Wikipedia, *Wikimedia Fundation*en entziklopedia eleanitza eta eduki askekoa da. Bertako artikuluak mundu osoko erabiltzaileek idazten dituzte eta bakoitzaren identifikadore unibokoa titulua da. Honen bidez artikuluaren kontzeptua deskribatzen da eta kontzeptuen aldaerak edo formak, birbideratze eta desanbiguazio orrien bidez lotzen dira artikulu nagusira.

Artikuluen barnean beste artikuluetara doazen estekak aingura bidez egiten dira eta artikuluan izen batez identifikatzen dira. Aingura izenak, sarrerak ez bezala, errepikatuak egon daitezke artikulu ezberdinetan. Honetaz gain, hizkuntzen arteko loturen bidez *Wikipedia*ko hizkuntz bertsio ezberdinen artikuluak estekatzen dira. Ezaugarri hauek, entitate-izenekin lan egiteko informazio-iturri gisa oso interesgarriak dira. Gainera, entitate-izenen desanbiguaziorako baliabide ezin hobea da *Wikipedia*.

3.2.1. Desanbiguazio orriak

*Wikipedia*ko desanbiguazio orriek bi adiera ezberdin edo gehiago dituzten kontzeptuen kasuan, adiera ezberdinen artean bereizteko loturak eskaintzen dituzte. Beraz orri hauek izen berdinaz erlazionatuak dauden artikuluen zerrendak eskaintzen dituzte.

John Adams (disambiguation)

From Wikipedia, the free encyclopedia

John Adams (1735–1826) was the second President of the United States (1797–1801).

John Adams may also refer to:

Contents [hide]
1 Politicians
2 Composers
3 Military
4 Scientists
5 Sports
6 Martyrs
7 Others
8 Other works
9 Places
10 Transport
11 See also

Politicians [edit]

- John Adams, 1st Baron Adams (1890–1960), British peer
- John Adams (Major General), former Chief of the Canadian Communications Security Establishment
- John Adams (MP) (by 1511–1571/75), Welsh MP for Pembroke Boroughs
- John Adams (New York) (1778–1854), United States Congressman from New York
- John Adams (Ohio politician) (born 1960), present-day member of Ohio House of Representatives
- John H. Addams (1822–1881), Illinois State Senator and father of Jane Addams
- John J. Adams (1848–1919), Congressman from New York during the 48th United States Congress
- John Q. Adams (Wisconsin) (1816–?), Wisconsin state legislator
- John Quincy Adams (1767–1848), 6th President of the United States and son of the 2nd President
- John Quincy Adams (1848–1919), general land and town-agent for the Chicago, Milwakee
- John Adams II (1803–1834), son of John Quincy Adams and grandson of John Adams
- John Quincy Adams II (1833–1894), American politician and grandson of the President of the same name
- John R. Adams (born 1955), United States federal judge
- John T. Adams (1862–1939), Republican National Committee chairman
- Tom Adams (politician) or Jon Adams (1931–1985), Prime Minister of Barbados

Composers [edit]

- John Adams (composer) (born 1947), American composer, came to prominence with *Shaker Loops* in 1978
- John Luther Adams (born 1953), American post-minimalist composer, 2014 Pulitzer Prize winner

Military [edit]

- John G. B. Adams (1841–1900), Civil War Medal of Honor recipient

5 Irudia: John Adams-en Wikipediak eskaintzen duen desanbiguazio orria.

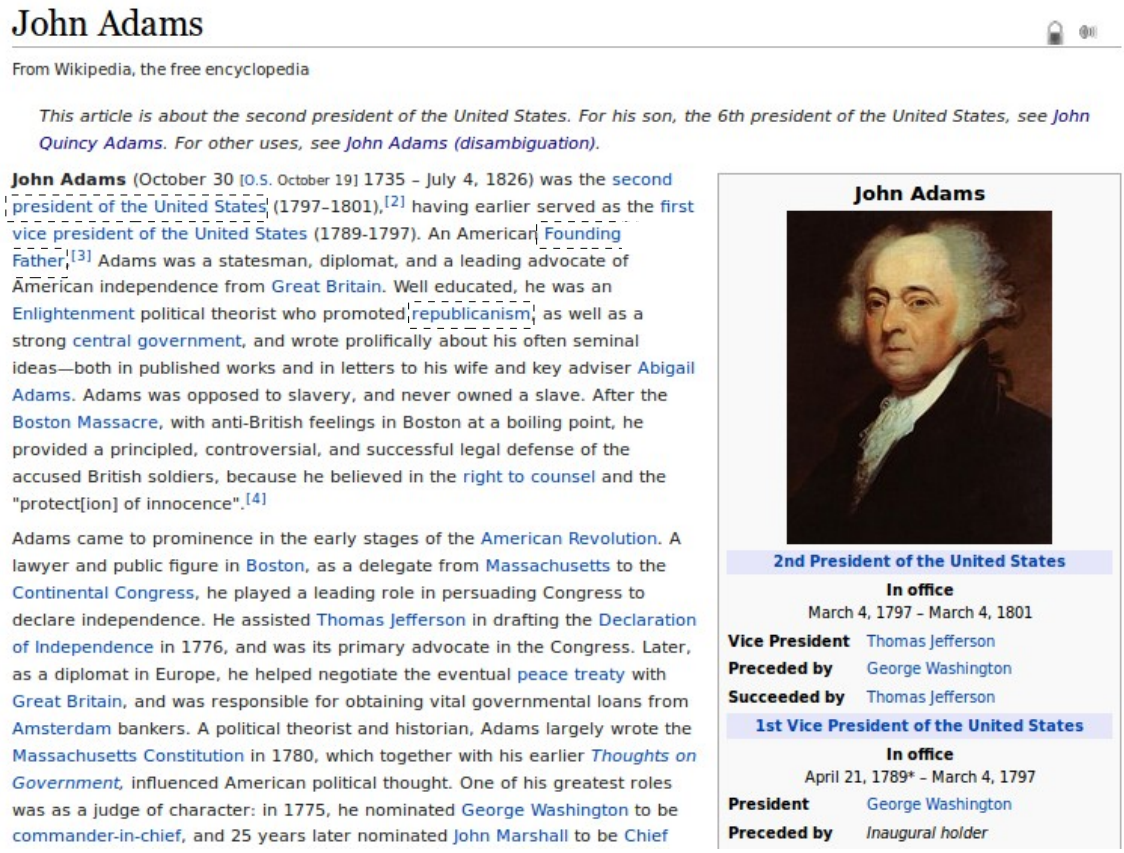
5 Irudian ikus daiteke "*John Adams*" izen-entitatearentzat *Wikipedia*ko desanbiguazio orriak eskaintzen dituen artikuluen zerrenda. Politikariaz gain, kirolari eta militarren artikuluetara loturak agertzen dira beste batzuen artean.

3.2.2. Birbideratzeak

*Wikipedia*ren orrialde asko birbideratze bidez atzitzen dira; adibidez, "*J. Q. Adams*" artikulua birbideratze artikulu bat da, hain zuzen, "*John Quincy Adams*" artikulura. Kasu hauetan, "*J. Q. Adams*" birbideratze orriak "*John Quincy Adams*" orrialdea ebazten duela esaten da. Orri hauen bidez, entitate berari izen ezberdinen bidez deitzeko arazoari aurre egiten zaio eta pluralizatze edo erlazionatutako hitzen erabilerari irtenbidea ematen zaie.

3.2.3. Aingurak

*Wikipedia*ren antolaketan, betebehar garrantzitsua daukate artikuluen arteko aingura testuek. Aingurak *Wikipedia* artikuluetan agertzen diren hiperestekak dira. Aingura testuan agertzen diren hitzek, gehienetan, erreferentziatu duten artikuluari buruzko informazio esanguratsua eskaintzen dute eta ez dute zertan artikuluaren izen berdina eduki behar. 6 Irudian ikus daiteke "*Jonh Adams*"en artikuluan, hiperesteka moduan dauden aingura ezberdinak: *president of the United States*, *Founding Father*, *republicanism*.



John Adams

From Wikipedia, the free encyclopedia

This article is about the second president of the United States. For his son, the 6th president of the United States, see John Quincy Adams. For other uses, see John Adams (disambiguation).

John Adams (October 30 [O.S. October 19] 1735 – July 4, 1826) was the second president of the United States (1797–1801),[2] having earlier served as the first vice president of the United States (1789-1797). An American Founding Father,[3] Adams was a statesman, diplomat, and a leading advocate of American independence from Great Britain. Well educated, he was an Enlightenment political theorist who promoted republicanism, as well as a strong central government, and wrote prolifically about his often seminal ideas—both in published works and in letters to his wife and key adviser Abigail Adams. Adams was opposed to slavery, and never owned a slave. After the Boston Massacre, with anti-British feelings in Boston at a boiling point, he provided a principled, controversial, and successful legal defense of the accused British soldiers, because he believed in the right to counsel and the "protect[ion] of innocence".[4]

Adams came to prominence in the early stages of the American Revolution. A lawyer and public figure in Boston, as a delegate from Massachusetts to the Continental Congress, he played a leading role in persuading Congress to declare independence. He assisted Thomas Jefferson in drafting the Declaration of Independence in 1776, and was its primary advocate in the Congress. Later, as a diplomat in Europe, he helped negotiate the eventual peace treaty with Great Britain, and was responsible for obtaining vital governmental loans from Amsterdam bankers. A political theorist and historian, Adams largely wrote the Massachusetts Constitution in 1780, which together with his earlier *Thoughts on Government*, influenced American political thought. One of his greatest roles was as a judge of character: in 1775, he nominated George Washington to be commander-in-chief, and 25 years later nominated John Marshall to be Chief

John Adams

6 Irudia: John Adams-en Wikipediako artikuluan ageri diren aingurak.

Aingura testuetan agertzen diren hitzak aztertuz, "*Founding Father*"k, adibidez, "*Founding Fathers of the United States*" artikulura erreferentzia egiten du. Propietate hau oso garrantzitsua izango da desanbiguazio orduan, izan ere, artikulu-izenak eta aingura-izenak ez dute zertan bat etorri behar.

3.2.4. Hizkuntzen arteko loturak

*Wikipedia*k hizkuntza askotako artikuluak ditu, eta artikuluak hizkuntzaren arabera dagokien *Wikipedia*ko bertsioan daude. Hizkuntza ezberdinetako artikuluak, era berean, ezagutza berdina defini dezakete. Horregatik *Wikipedia* hizkuntzen arteko (*interlingual*) estekez baliatzen da, artikulu batetik beste hizkuntza bateko artikulu berdinera estekatzeko.

Lotura hauek artikuluaren ezkerraldean agertzen den menuan daude; hizkuntza atalaren baitan. Hauen bidez ezagutza base handiagoa duen hizkuntza batean egin dezakegu bilaketa eta ondoren, ulergarriagoa egiten zaigun hizkuntzara igaro. Kontuan izan behar dugu, normalean, hizkuntzaren arabera artikuluaren informazio kantitatea aldatzen dela. Gomendagarriena artikuluaren ama hizkuntzan irakurtzea da; hori izaten baita informazio aberatsena duena.

3.3. Wikipediaren esportazioak

*Wikipediar*en esportazioak honek duen ezagutza basea *XML* erraldoi batean eskuratzeko aukera ematen du. Esportazio hauek egindako datarekin identifikatzen dira eta eduki ezberdina dute barnean; batzuk artikulu orriak dituzte, beste batzuk birbideratze orriak edota desanbiguazio orriak dituzte.

Gainera, *Wikipedia* atzigarri dagoen hizkuntza ezberdinen esportazioak daude. Hizkuntza desberdinetako artikulu berdinak erlazionatzeko 3.2.4 ataleko *interlingual* estekak erabilgarri daude. Hauei esker ingelesezko ezagutza handia erabili ahal izango dugu beste hizkuntza bateko ezagutza-basea aberasteko.

3.4. DBpedia Spotlight

2010eko ekainean *Freie University of Berlin*eko *Web Based Systems* taldeko ikerlariek hasitako proiektua da *DBpedia Spotlight*. Hala ere, gaur egun beste ikerlari batzuen laguntza dauka proiektuak. Sistema honek testuan agertzen diren entitateak identifikatu eta desanbiguatzen ditu, hots, IED (Izendun Entitateak Desanbiguatzeko) edo *NED* (*Named Entity Disambiguation*) problema ebazteko pentsatuta dago. Horretarako, hiru urrats egiten ditu; entitateak topatu, hautagaiak sortu eta desanbiguazioa.

*Dbpedia*ko ezagutza basea *Wikipedia*tik erauzten da. *Spotlight*en erantzunak *DBpedia*ko ezagutza baseko estekak dira, baina hauek era berean, *Wikipedia*ko identifikadore uniboko berdina dutenez, *Wikipedia*ko artikulu berdinera egiten dio erreferentzia. Hau da, *Dbpedia* ezagutza-baseko ondorengo artikuluak http://dbpedia.org/page/John_Adams *Wikipedia*ko artikulu berdinari egiten dio lotura http://en.wikipedia.org/wiki/John_Adams.

Spotlight bi modutan erabilgarri dago: Web aplikazioan, modu errazean testatzeko balio duena, eta norberak bere ordenagailuan instalatzeko web zerbitzarian. 4.1 Atalean tresna honen analisia egin dugu.

3.5. UKB

UKB grafoetan oinarritutako programa-bilduma da; hitz adieren desanbiguazioa burutzeko bi testuren antzekotasuna neurtzeko erabiltzen dena. Neurketa hori egiteko, hots, grafoko nodoak pisatzeko, zorizko ibilbideetan oinarritutako *PageRank* (*Brin and Page*, 1998) algoritmoaren aldaera erabiltzen da.

PageRank algoritmoa, jatorriz, webguneen garrantzia ebaluatzeko sistema da. Zer nolako garrantzia duen jakiteko, webguneen arteko estekak erabiltzen ditu. Esate baterako, A webguneak B webgunerako duen esteka Ak Bri emandako boto gisa hartzen du. Gainera, zenbat eta A orrialdea garrantzitsuagoa izan, orduan, bere pisua are handiagoa izango da.

Esan bezala, *PageRank* algoritmoak zorizko ibilbideen eredua erabiltzen du. Ibilbidea zoriz hautatutako grafoko nodo batetik hasten da eta urrats bakoitzean, momentuko nodoaren irteera-ertz bat aukeratzen du ausaz, ibilbideari jarraipena emanez. Baina, edozein momentutan, ertzak jarraitu beharrean, grafoko beste edozein nodotara alda daiteke ibilbideari jarraipena emateko. Nodo baten pisua ibilbidea eratzean nodo hori hautatua izateko probabilitatea izango da, zorizko ibilbideak mugarik gabe jarraituko duela suposatuz. Exekuzioa etengabea izan ez dadin atalase bat esleitzen zaio.

Hala ere, UKBk *PageRank* Pertsonalizatua (*Personalized PageRank*) izeneko aldaera erabiltzen du. Aldaera honi esker *PageRank* algoritmoaren arazoa gainditzen du; testuinguruarekiko erabat independentea izatea. Beraz, aldaera honi esker, grafoko zenbait nodoei pisu handiagoa emateko aukera ematen du, nodo horien garrantzia handituz eta pisu horiek grafoan zehar barreiatzea posible egiten du. Modu honetan, nodoen garrantzia ezberdinak izanik, algoritmoak jauzia egitean, garrantzitsuagoak diren nodoetara salto egitea lortzen da. Horrela adierazgarriagoak diren nodoetara jauzi egitea bideratzen da.

UKBk bi baliabide behar ditu erabili ahal izateko. Batetik, ezagutza-basea, grafo moduan adierazita. Eta, bestetik, hiztegia, zeinetan hitzen eta ezagutza-baseko nodoen arteko erlazioa definitzen den.

Beraz, UKBk aipamen anbiguo baten testuingurua emanik, aipamen horri dagokion ezagutza-baseko nodoen artean entitate egokiena aukeratzen du. Hori lortzeko, UKBk testuinguruan agertzen diren aipamenei dagozkien ezagutza-baseko nodoak, hiztegian definitutako nodoak, pisu berdinarekin hasieratzen ditu, eta zero balioarekin gainerakoak. Ondoren, *PageRank* Pertsonalizatua erabiltzen du. Hasierako aipamen anbiguoaren adiera diren nodoen artean *PageRank* balio altuena duen nodoa itzuliz ebazten du UKBk desanbiguazioa.

4.2 Atalean software honen analisia egingo dugu; bertan, parametroez, beharrezko baliabideez eta sarrera/irteera formatuez arituko gara.

3.6. Matx

Matx Ander Barrenak sortutako softwarea da, Hizkuntzaren Azterketa eta Prozesamendua Masterreko proiektuan zehar garatutakoa. Programa honek pasatutako testuan izendun entitateak topatzen ditu. Baina zehazki entitate luzeenak itzultzen ditu. Honetaz gain, guretzat beharrezkoa den entitate hauen testuingurua ere itzultzen digu. Testuinguru horri testuinguru leihoa deritzo; leiho honek izen-entitatearen aurrean dauden 50 hitzez eta ondoren dauden beste 50 hitzez osatuta dago.

Software honek “*Did you mean*” algoritmoa erabiltzen du, hiztegiaren ainguren eta izen-entitatearen arteko errore ortografiko edota tipografikoak gainditu ahal izateko. Modu honetan “*Jonh Adams*” agertuz gero “*John Adams*”i erreferentzia egiten diola lortzen da.

Matxen funtzionamendurako hiru heuristiko erabiltzen dira. Hasteko Matxek *multiword*ak, hitz-multzoak, topatzen ditu eta ondoren heuristikoak aplikatzen dizkio emaitza lortzeko. Lehenengo heuristikoak hitz-multzoari lagungarri duen *stopword* zerrendan agertzen diren hitzak kentzen dizkio. Bigarrengoak hizki larririk ez edukitzekotan *multiword*a baztertzen du. Hirugarrengo heuristikoak hitz-multzoaren hasiera eta amaieran hizki larririk ez dituzten hitzak kentzen ditu. Heuristikoak aplikatu osteko hitz-multzoa azken emaitza da.

Hala ere, proiektu honetan lehenengo eta bigarrengo heuristikoak erabili dira soilik, beste heuristikoak garapen fasean baitzegoen.

Multiword	of_the_USA	the_constraint_grammar	Ministry_of_Defence
1. Heuristikoa	USA	constraint_grammar	Ministry_Defence
2. Heuristikoa	USA		Ministry_Defence
3. Heuristikoa	USA		

2 Taula: Bertan Matxen hiru heuristikoen funtzionamendua ulertzeko hiru adibide daude. Ikus daiteke 1.heuristikoa aplikatu ondorengo emaitza nahiko ona dela eta ondorengo heuristikoak aplikatuz emaitza kopurua txikituz doala. Beraz, ahalik eta erantzun gehien lortzeko lehenengo heuristikoa aplikatzearekin nahikoa da; zehaztasuna hobetzea bilatuz gero, gainerako heuristikoak erabil daitezke.

3.7. Datu-multzoak (datasets)

Ikerkuntza arloan beharrezkoa da garatuko diren sistemak ebaluatzea, gainerako sistemekin modu erraz batean alderatu ahal izateko. Helburu horrekin zenbait erakundek euren datu-multzoak sortzen dituzte gainerakoek sortutako tresnak ebaluatu ahal izateko. Datu-multzoak eskuz daude anotatuak.

Proiektu honetan 3 hizkuntzetan ebaluatu dugunez, hizkuntza bezain beste datu-multzo behar izan ditugu. Hala ere, 5 erabili ditugu; batetik, ingelesezko TAC-KBP 2010, 2011 eta AIDA. Bestetik, gaztelerazko TAC-KBP 2012 eta, azkenik, euskarazkoa, Izaskun Fernandezek bere tesirako (Entitate-izenak euskaraz: identifikazioa, sailkapena, itzulpena eta desanbiguazioa) sortutakoa .

3.7.1. TAC-KBP

TAC-KBP edo *Text Analisys Conference - Knowledge Base Population* konferentziaren helburua, testu hutsetik entitate-izendunen ezagutza-baseak sortzeko eta aberasteko sistemak garatzea da. Horregatik, ingelesezko 2008ko *Wikipedia*ko entitateez osatutako ezagutza-base bat sortu zuten. Ezagutza-base hauen bidez lortutako datu-multzoak berebiziko garrantzia dute ikerkuntzan garatutako sistemak ebaluatu eta konparatu ahal izateko.

Konferentziako helburuetakoa den "*entity linking*" ataza testu ezberdinetako izen-aipamenak ezagutza-baseko entitateekin lotzean datza. Horregatik 2009 urteaz geroztik urtero datu-multzo berriak prestatzen dituzte. Datu-multzo hauek XML formatuan kodetuta daude, bere baitan "eskaerak" edo *query*ak (desanbiguatu beharreko adibideak) dituzte, hauek 7 irudian ikus daitezke. Eskaera bakoitzak identifikadorea (*query id*), lotu beharreko izena (*name*) eta izen-entitatea duen dokumentuaren identifikadorea (*docid*) dauka, hau da testuingurua duen dokumentuaren identifikadorea. Eskaera bakoitzak ezagutza-baseko entitate helburu bat izango du emaitza bezala.

```
...
<query id="EL1">
  <name>Abbas Moussawi</name>
  <docid>LTW_ENG_19960311.0047.LDC2007T07</docid>
</query>
<query id="EL2">
  <name>Abbas Moussawi</name>
  <docid>NYT_ENG_20000711.0026.LDC2007T07</docid>
</query>
...
```

Irudia 7: TAC datu-multzoko bi eskaera ikusten diren adibidea.

"*Gigaword*" kolekziotik lortzen dituzte dokumentuak, ingelesezko datu-multzoan "*English Gigaword*"eko 5. ediziotik eta gaztelaniazko datu-multzoaren kasuan "*Spanish Gigaword*"eko 3. bertsiotik. Bilduma honek, era berean, berriak, web orrietako laginak edota foroetako testuak ditu. Dokumentu hauek sasi-*XML* kodeketan daude; barneko testuak ez du inolako tratamendurik jaso, beraz, testu garbia lortzeko hauek erauztea ezin bestekoa da.

Izen-aipamenak entitateekin lotzerako garaian, aipamen batzuentzat ezagutza-basean entitate egokirik ez egotea gerta daiteke. Horrelakoetan, NIL entitate bereziari lotuko dira. Aukera honen bidez ezagutza basearekin lotu ezin diren aipamenak ebaluatzen dira. Beraz, eskaerak ezNIL eta NIL bezala ezberdintzen dituzte.

Hala ere, proiektu honetan ezNILen gainean lan egin dugu, hauen emaitzak ebaluatu nahi ditugulako, gainerako datu-multzoetan berdin jokatu dugu.

3 taulan erabili ditugun hiru TAC datu-multzoko eskaera kopuruak ikus ditzakegu.

TAC-KBP datu-multzoa	Eskaera guztiak	EzNIL eskaerak	NIL eskaerak
2010 EN	2250	1020	1230
2011 EN	2250	1124	1126
2012 ES	2066	923	1143

3 Taula: Urte ezberdinetan TAC-KBP datu-multzoen eskaera kopuruak ikus daitezke. Bertan EzNIL eta NIL motatako eskaera kopuruak adierazten dira.

Ebaluazioa egin ahal izateko, datu-multzo bakoitzak bere urre-patroia du; eskaera identifikatzailea eta helburu entitate bikotez osatuta. Helburu entitatea ezNIL bada ezagutza-baseko identifikatzailea agertuko da. NIL entitateentzat, ordea, "NILxxxx" egitura erabiltzen dute, non x horiek zenbakiak diren. Hauek zenbaki multzo zenbakia adierazten dute, eta hauen bidez entitateen multzokatzea edo *clustering*a egitea ahalbidetzen dute.

3.7.2. AIDA / CoNLL

Datu-multzo hau 2003ko *Conference on Computational Natural Language Learning* konferentzian sortu zuten, hain zuzen ere "*entity recognition*" atazan. Aurreko urteko konferentzia edizioan egin zuten haien lehen data-multzoa izen-entitateen ezagutzarako, gaztelaniarako eta nederlanderako. Konferentzia honek 2003an sortu zuen ingelesezko datu-multzoarekin batera Alemaniarrezko datu-multzoa ere sortu zuen. Izen entitateen lau mota ezberdin landu nahi izan dituzte; pertsonak, tokiak, erakundeak eta bestelako entitateak, non ez duten aurreko 3 taldeetan tokirik.

Gure proiektuan erabilitakoa ingelesezko datu-multzoa izan da. Ezagutza-basea sortzeko *Reuters Corpus*etik hartutako dokumentuez baliatu dira. Dokumentu hauek albisteak dira, 1996ko abuztuaren eta 1996ko abenduaren artean bildutakoak.

Datu-multzoa hiru azpi-multzotan banatuta dago: *train*, *test A* eta *test B*. Orokorrean, *train* ereduak entrenatzeko erabiltzen da, *test A* ereduen optimizazioa egiteko eta, azkenik, *test B* azpi-multzoa ebaluatzeko erabiltzen da. Proiektu honetan soilik *test B* erabiliko dugu, zeinek 231 dokumentuetatik sortua dagoen. Gainera, AIDA gisa izendatuko dugu datu-multzo hau hemendik aurrera.

Datu-multzoaren ezagutza-basea 2010eko *Wikipedia*ko esportazioa da.

Datu-multzoa	Eskaera guztiak	EzNIL eskaerak	NIL eskaerak
AIDA	5616	4485	1131

4 Taula: 2003an sortutako datu-multzoa honen "test B" azpi-multzoari dagozkion eskaera kopuruak ikus daitezke; eskaera totalak, EzNIL eta NIL kopuruak ezberdinduta.

3.7.3. Euskarazko datu-multzoa

Izen-entitatearen desanbiguazioa euskaraz ebaluatu ahal izateko, Izaskun Fernandezek bere doktoretza tesian erabilitako corpusa erabili dugu. Datu-multzo honek 2002 urteko Euskaldunon Egunkariako albisteez osatuta dago. Orotara 40.648 artikulu ditu, eta hauek bere baitan 130.505 izen-entitate etiketatuta dituzte.

Fernandezek, datu-multzoa sortzean, kontuan hartu du kazetariek berriak idazterako garaian pertsona, toki edo erakunde bat aipatzean, lehenbiziko aipamenean forma luzea edo osoa idatzi ohi dutela eta ondorengo aipamenetan, berriz, forma laburragoak. Baina hori ez da beti gertatzen, forma laburraren aurretik osoa ez duten artikuluak ere badaude. Ezaugarri hauek 8 irudiko adibidean ikus daitezke.

Ajeak dira biak, baina ezberdinak oso arrakastarena eta porrotarena. ***Aimar Olaizolak*** *eta Abel Barriolak bestondo erabat ezberdina izan zuten atzo...* ***Olaizola*** *kontrario eta lagunaren dohainak goratu zituen: "Kirolean burua da garrantzitsuena, eta* ***Aimarrek*** *bere tokian dauka".*

8 Irudia: (I. Fernandez, 2012) erabilitako adibidea paragrafoen ezaugarriak azaltzeko.

Honetaz gain, kontuan hartu behar da, *Wikipedia*ko entitateak datu-multzoaren urrezko-patroia sortzeko erabiltzen direla. *Wikipedia*n agertzen ez diren izen-entitateak NIL moduan adierazten dira eta agertzen direnak EzNIL.

Aurreko ezaugarriak kontuan hartuta, Fernandezek ondorengo paragrafo multzoak deskribatu zituen bere tesi lanean:

- “Izen-entitate laburrak, beraien forma desanbiguatua kontsidera daitekeen izen-entitate luzeagoa albistean agertzen da, eta azken hori bat dator *Wikipedia*ko sarrera batekin.
- Aurrekoan bezala izen-entitate anbiguorako izen-entitate desanbiguatua kontsidera daitekeen izen-entitate luzeagoa badago albistean, baina azken horrek ez du *Wikipedia*n sarrerarik.

- Izen-entitate anbiguoak ez du bera baino luzeagoa den izen-entitate agerpenik aurretik.”

Lehenengo multzoko paragrafoekin automatikoki lan egin zuten, helburu izen-entitatea *Wikipedia*ko ezagutza-basean agertzen baita, hots izen-entitatea *Wikipedia*ko artikulu izenburua baita. Baina ez da berdina gertatzen beste bi multzoekin. Bigarrengoan, forma luzeak *Wikipedia*n ez agertzeak ez du esan nahi espresio motzaren desanbiguazio formarentzako sarrerarik ez dagoenik. Hirugarren multzoan, albistean bera barne duen izen-entitate luzeagorik ez dagoenez, haiek ezarritako irizpideen arabera ezin zuten automatikoki desanbiguazio-forma identifikatu. Beraz, azken bi multzo hauen forma anbiguoaren desanbiguazio forma *Wikipedia*n eskuz egin zuten.

Multzo hauen ezaugarriak aztertu ostean, batetik modu automatikoan sortutako paragrafo sorta bat lortu zuten, non paragrafoko forma anbiguoari dagokion forma ez-anbiguoa *Wikipedia*ko sarrera ezaguna den (*corpusA*). Eta bestetik, eskuzko berrikuspenaren ondoren, beste paragrafo sorta bat definitu zuten, non izen-entitate anbiguoaren desanbiguazio adiera *Wikipedia*n existitzen zein existitzen ez diren paragrafoak nahasten diren (*corpusB*). Eskuzko lana minimizatzeko asmoz, lagin txikiak hartu ziren.

B corpusa, gainera, bitan banatu zuten; garapen eta test multzoetan. Etorkizunean sistema hobetzeko eta fintzeko banaketa interesgarria dela kontsideratu baitzuten.

Datu-multzoa	Eskaera guztiak	EzNIL eskaerak	NIL eskaerak
CorpusA	6500	6500	0
CorpusB-garapen	532	462	70
CorpusB-test	500	437	63

5 Taula: I. Fernandezek bere tesi lanerako egindako euskarazko datu-multzokoak ageri dira. Aurreko datu-multzoetan bezala eskaera moten araberako sailkapenaren kopuruak ageri dira.

Proiektu honetan *corpusB* datu-multzoko bi multzoak erabiliko ditugu. Izaskunen emaitzetan berak adierazitakoa kontuan hartu baitugu: “Beraz B corpusetako emaitzak A corpusekoak baino okerragoak badira ere, agertoki errealagoak irudikatzen duten B corpusetan sistemak egonkorragoak direla esan daiteke”.

3.8. Ebaluazio-neurriak

Proiektu honetan erabilitako desanbiguazio-sistemak ebaluatzeko 3.5 atalean zehazten diren datu-multzoak erabili ditugu. TAC-KBPren kasuan bi metrika erabiltzen dituzte; mikro zehaztasuna eta *Bcubed+*. Hala ere, proiektuan erabiliko ditugun neurriak beste sei hauek izango dira; doitasuna, estaldura (mikro zehaztasuna), f1, *coverage*, abiadura eta memoria.

- Doitasunak desanbiguazio adiera bat proposatzeko gai denean, zein asmatze-tasa duen adierazten du.

$$P = \frac{zuzen\ desanbiguatutako\ entitate\ kopurua}{desanbiguatutako\ entitate\ kopurua}$$

- Estaldura entitate-izendun anbiguoak ondo desanbiguatzeko gaitasuna da.

$$R = \frac{zuzen\ desanbiguatutako\ entitate\ kopurua}{entitate\ kopurua}$$

- F1 sistemaren ebaluazioaren ikuspegi orokorragoa izateko erabiltzen da; estaldura eta doitasuna konbinatuz.

$$f1 = \frac{2 * P * R}{P + R}$$

- *Coverage* entitate-izendun anbiguoak desanbiguatzeko gaitasuna da. Modu honetan sistemak erantzuteko gaitasuna neurtzen da.

$$C = \frac{desanbiguatutako\ entitate\ kopurua}{entitate\ kopurua}$$

- Abiadura (A/s) segundoko zenbat aipamen desanbiguatzeko gai den adierazteko erabili dugu. Horretarako 5.1.3 atalean zehaztutako tresnak erabili ditugu.

$$A/s = \frac{desanbiguatutako\ entitate\ kopurua}{desanbiguatzeko\ erabilitako\ denbora}$$

- Memoria ebaluazio-neurria sistemak erabiltzen duen *RAM* memoria neurtzeko erabili dugu.

Proiektu honetan datu-multzoen EzNIL eskaerak ebaluatuko dira soilik, NIL eskaerak alde batera utziz. Gainera, ebaluazioa urre-patroi batzuen kontra egingo dugu; hain zuzen ere 3.7 ataleko datu-multzoen kontra.

3.9. Ixati

EHUko IXA ikerketa taldeak garatutako euskararako analizatzaile sintaktiko sendoa da. Bere baitan duen modulu bakoitza zenbait geruza ezberdin ditu. Geruza hauek aurrekoaren informazioa jasotzen dute sarreratzat; modu honetan mailaz mailako analisi sintaktiko sakonagoa egiten du ur-jauzian. analisia egikaritzeko murrizketa gramatika (*Constraint Grammar*) formalismoa erabili da. Analizatzaile honi esker, euskarazko testuinguruak lematizatu ahal izango ditugu.

4. SOFTWAREAREN ANALISIA

Esperimentuekin hasi aurretik, lehendabizi softwarearen analisia egin dugu. Atal honetan UKB zein *Dbpedia Spotlight* analizatuko ditugu. Baina gehien analizatuko duguna *Spotlight* izango da, irismenean azaldu bezala, IXAk etorkizunean erabili ahal izateko beharrezko azalpenak eman behar ditugulako.

4.1. Dbpedia Spotlight

Tresna hau erabili ahal izateko bi aukera daude; egileek eskaintzen dizuten web aplikazioa erabiltzea edota guk geuk gure ordenagailuan web zerbitzari hori instalatzea.

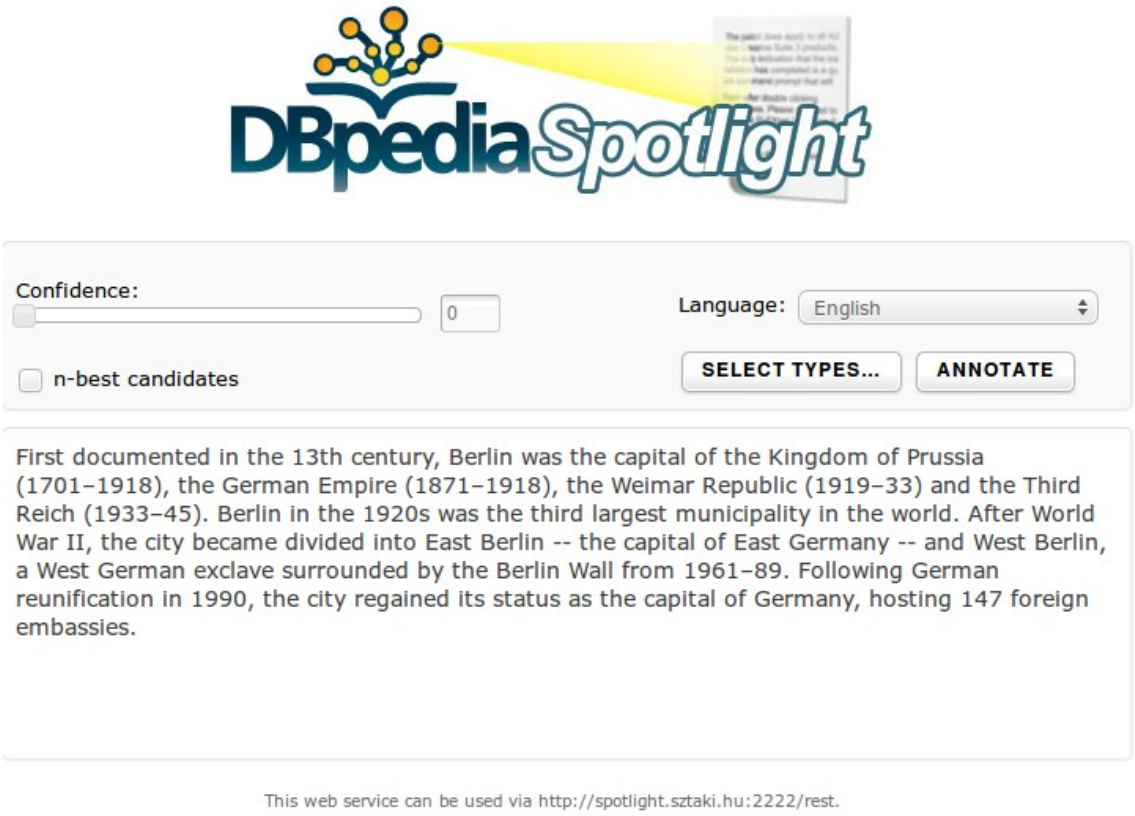
Web aplikazioa*

*Spotlight*eko web aplikazio honekin ezin dira *Spotlight*ek dituen parametro eta funtzio guztiak erabili, baina probatzeko eta ideia bat egiteko erabilgarria da oso.

Hasteko, hizkuntza adierazi beharko dugu; ondorengoak daude hautatzeko: Ingelesa, alemanera, nederlandera, frantsesa, portugesa, italiera, errusiera, hungariera, turkiera eta gaztelania.

* Esan beharra dago noizbehinka zerbitzaria ez dagoela atzigarri, eta web-aplikazioak ez du inolako mezurik adierazten.

Hizkuntzaz gain, *Confidence* balioa ere adierazten da; iragazki honen bidez zehazten dugun balioa baino handiago duten entitateak soilik adieraziko dira.



9 Irudia: DBpedia Spotlighten web aplikazioa. Bertan, lehenetsitako adibidea eta eskaintzen dizkigun aukerak ikus daitezke.

Gainera, *n-best candidates* aukerarekin aipamen baten adiera ezberdinak ikus daitezke; konfiantza langa gainditzen dituztenak. Sagua aipamenaren gainean jarriz gero aukera posibleak adieraziko ditu, bere konfiantzarekin.

Azkenik, *Select types* iragazkiarekin zein motatako entitateak desanbiguatu nahi den (edo zeintzuk ez) adieraz daiteke. Hiru motatako datu baseko motak erabil daitezke: *DBpedia*koa, *Freebase*koa eta *Schema.org*ekoa. Oso interesgarria da *Custom (SPARQL)* aukera. Modu honetan, *SPARQL* kontsulta eginez, soilik kontsulta hori betetzen duten entitateak desanbiguatuko dira.

Bestalde, menuaren azpian desanbiguatuko den testua idatzi behar da. Bertan, lehenetsita azaltzen den testua dago.

Annotate botoia sakatutakoan, zerbitzarira bidaliko dio eskaera eta emaitza testuan bertan azalduko da.

10 Irudia: DBpedia Spotlighten web aplikazioak emaitza emateko modua ikus daiteke. Aipatzekoa da, adibide honetan, "n-candidates" aukera sakatuta dagoenez, desanbiguatutako entitateen aukera posibleak adieraziko dizkigula.

Web Zerbitzaria

Dbpedia Spotlight web zerbitzaria gure makinan erabiltzeko zenbait modutara egin daiteke. Errazena eta erabilgarriena guretzako *.jar* fitxategi exekutagarria erabiltzea da. Modu honetan exekutagarri berdina erabil dezakegu zenbait zerbitzari martxan jartzeko. Honek modu errazean hizkuntza ezberdinetako ereduak dituzten zerbitzariak martxan jarri ahal izatea ahalbidetzen du.

Horretarako, *Dbpedia Spotlight* S*tatistical*en exekutagarria eta aukeratu dugun hizkuntzaren eredua jaitsi behar dugu webgunetik[3]. Behin eredua erauzita, zerbitzaria martxan jartzeko ondorengo aginduan zerbitzariak erabiliko duen portua eta eredua adieraziko dugu:

```
$ java -jar dbpedia-spotlight.jar EREDUA http://localhost:PORTUA/rest
```

Kontuan izan behar dugu zerbitzaria modu honetan martxan jartzeko beharrezkoa dela *JAVA* 1.6 baino berriagoa den bertsioa eta ereduaren araberakoa den memoria edukitzea; 8Gb *RAM* baino gehiago edukitzea gomendatzen dute. Ereduaren tamainaren arabera eta erabilgarri dugun memoriaren arabera zerbitzaria martxan jartzeko lau eta minutu baten arteko denbora behar izan dugu.

4.1.1. Desanbiguaziorako aukera ezberdinak

*Dbpedia Spotlight*ek lau urte hauetan zehar bi bertsio nagusi ditu: *Lucene* eta *Statistical*. Haien arteko ezberdintasunak erabilitako *back-end*a eta Lucenentzat soilik ingeles hizkuntzako eredua eskaintzea dira. Bigarrengoak, ordea, dagoeneko bederatzi hizkuntzetako ereduak eskaintzen ditu; horien artean alemana, ingelesa, gaztelania, italiera, frantsesa, portugesa, errusiera, daniera, hungariera, suediera eta turkiera. Hala ere, bi bertsioek internazionalizatzeko argibideak eskaintzen dituzte, honi esker *Wikipedia*n existitzen den edozein hizkuntzetako ereduak sortu eta erabili ahal dira.

3 http://spotlight.sztaki.hu/downloads/

Spotlight Lucene*

*DBpedia Spotlight*eko bertsio hau bietatik lehenago garatutako bertsioa izan da. Bertsio honek izen-aipamenak topatzeko (spoteatzeko) esaldietan kate zehatzak parekatzen ditu, zeinek *Aho-Corasick*en *LingPipe* inplementazioa erabiltzen duen. Entitate horiek desanbiguatzeko algoritmoa kosinu antzekotasunean eta *TF-IDF* pisuen moldaketan du oinarri, *Apache Lucene* erabiliz.

Nahiz eta, oraindik, deskargatzeko aukera egon, egileek estatistikoa erabiltzeko gomendatzen dute. Hori baita hobekuntzak jasoko dituena.

Spotlight Statistical

Bertsio hau sistema semantikoen 9. konferentzia internazionalean aurkeztu zuten. Haien helburu nagusiak bi izan ziren. Batetik, entitateak topatzeko eta desanbiguatzeko erabiltzen diren atazen errendimendua eta doitasuna hobetzea. Bestetik, *Dbpedia Spotlight* internazionalizatzea lortzea, hots beste zenbait hizkuntzetan erabilgarria izatea.

Spotlight estatistikoak erabiltzen duen estandarra entitateak topatzeko oso arina da hizkuntza prozesamenduaren eskakizunetan. Soilik hizkuntzaren menpeko urrats bat egin behar da: Tokenizazioa. Entitateak topatzeko aurrizkiz osatutako zuhaitza eta *Aho-Corasick* algoritmoa erabiltzen da, zeinek azpi-kateak parekatzea den bere helburua.

Back end hau izango da gure proiektuan erabiliko duguna. Beraz, hemendik aurrera hitz egingo duguna estatistikoari buruz izango da.

* Uztailean probatzeko zegoen demoa dagoeneko ez dago erabilgarri.

Spotlight estatistikoak bi bertsio ditu; 0.6 eta gaur egungo 0.7. Joachim Daiber egileetako batek esandakoaren arabera, 0.7 bertsioa *Spotlight* estatistikoaren lehenengo bertsio ofiziala da. Bertsio berri honen hobekuntza nabariak ondorengoak dira:

- Ereduak txikiagoak eta askoz azkarragoak izatea lortu da. Kontaketaren kuantifikazioari, bilaketaren optimizazioari eta zenbait inausketei esker.

- Zenbait zuzenketa *Spotlight*en eta *PigNLProc*en, *Wikipedia*tik datuak erauzteaz arduratzen dena.

- Laguntza *UIMA*n eta *confidence* balioan.

Baina, noski, ereduak berritu dituztenez ezin izango dira aurreko bertsioetako ereduak erabili 0.7 bertsioan. Gainera, entitateen kontaketak berriro ere egin dituzte *Dbpedia*ko 3.9 bertsioarekin, zeinek era berean 2013ko martxoaren eta apirilaren tarteko *Wikipedia*ko erauzketak erabili dituen.

4.1.2. Parametroak

Behin zerbitzaria martxan jarrita, eskaerak egiteari has gaitezke. Horretarako *cURL* aginduak erabiliko ditugu. CURL tresna komando-interpretatzailean erabiltzen da, honi esker URL sintaxia duten fitxategiak bidaltzea lortzen dugu. Agindu hauetan jarraian adierazten diren parametro guztiak erabil daitezke, *threshold*a izan ezik (hau zerbitzaria martxan jarri aurretik zehaztu behar baita).

11 irudian ikus daiteke *cURL* aginduaren adibide bat *POST* metodoaren bidez bidalita. Hala ere, *GET* metodoa erabiliz bidalketak egitea ere onartzen du, baina irakurgarriagoa eta ulergarriagoa *POST* metodoaren bidez egikaritzea da, batez ere desanbiguatu nahi den testua luzea denean.

```
$ curl http://localhost:2222/rest/annotate -H "Accept:application/json"
        --data-urlencode "text=You and I ought not to die before we have explained ourselves to each other
Adams wrote Jefferson in 1815 It is doubtful that today politicians will spend much time trying to explain
themselves to one another even after theu leave offive They are after all creatures of a culture in which it is
acceptable on the Senate floor for Vice President Divk Cheney to ."
        --data "confidence=0"
        --data "support=0"
```

11 Irudia: CURL aginduaren adibidea, deia annotate modulura egin da.

*DBpedia Spotlight*ek hiru motatako parametroak ditu; testua edo helbidea (*URL*a), iragazkiak eta *spotter*ra.

Gehienak zerbitzarira eskaera bidaltzerako garaian zehaztu behar zaizkio; baina badago bat zerbitzaria martxan jarri aurretik zehaztu behar dena. Jarraian denak azalduko ditugu.

Testua / helbidea

Parametro hau beharrezkoa da. *Text* parametroa erabiliz gero, bertan desanbiguatu nahi den testua ezarri behar da. Baina, onartzen duen testu tamaina 460KBekoa da, 460000 karaktere hain zuzen ere.

URL parametroaren bidez *Spotlight*i *HTML* fitxategia pasatu ahal izango zaio. Onartzen diren *HTML* fitxategien tamaina 490kBekoa da.

Iragazkiak

*Spotlight*ek itzultzen dituen emaitzak iragazteko erabiltzen dira. Modu honetan emaitza kaskarrak alde batera utz ditzakegu edota mota zehatz bateko entitateak lortu soilik.

- *Coreference resolution*

Parametro honek onartzen duen balioa boolearra da. Iragazki honen balioa egiazkoa denean, heuristiko bat aplikatzen du eta ez da beste inolako iragazkirik erabiltzen. *False* izatekotan, gainerako iragazkiak aplikatuko dira. Balio lehenetsia *True* da.

- *Support*

Iragazki honen bidez, guk ezarritako langatik gorako doitasuna duten entitateak itzuliko ditu soilik. Lehenetsitako balioa 10 da.

- *Confidence*

*Support*en antzekoa, entitateen *percentageOfSecondRank* delako balioa ezarritako balioaren karratua baino handiagoa duten entitate guztiak hautatzen ditu. Balio lehenetsia 0.1 da.

- *Types*

Nahiz eta proiektua egiteko erabili ez den, oso interesgarria da iragazki mota hau. *Types*ekin emaitza bezala itzuliko zukeen entitateak mota batekoak izatea lor dezakegu. Horrela desanbiguatu nahi dugun entitateari buruzko informazioa erabili ahal izango dugu emaitza hobetzeko. Defektuz ez da erabiltzen. Adibidez, pertsonak edota mendiak diren entitateak itzultzea lor dezakegu.

- *Sparql*

Beste hau ere ez dugu erabiltzen proiektuan, baina erabilgarria delakoan azalduko dugu. *Sparql*, *SPARQL Protocol and RDF Query Language*, RDF grafoen gainean kontsultak egiteko erabiltzen den estandarra da. Iragazki honekin kontsultak betetzen dituzten entitateak aukeratuko dira soilik. Defektuz desgaituta dago. Adibidez, Berlineko 1900. urteko testu historiko bat desanbiguatu nahi baldin badugu; *Spotlight*i adierazi ahal izango diogu “1900 baino lehenago eta Berlinen jaiotakoak” bilatzeko:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?name ?birth ?death ?person
WHERE {
        ?person dbo:birthPlace :Berlin .
        ?person dbo:birthDate ?birth .
        ?person foaf:name ?name .
        ?person dbo:deathDate ?death .
FILTER (?birth < "1900-01-01"^^xsd:date) . }
ORDER BY ?name
```

12 Irudia: SPARQL kontsulta baten adibidea.

- *Threshold*

Threshold iragazkia lehenengo urratsean erabiltzen da, hots, aipamenak topatzean. *Threshold*aren helburua aurkitutako aipamenen kalitatea ziurtatzea da. Horrela atalase balioa txikia ezarriz gero *Spotlight*ek aipamen gutxiago topatuko lituzke, zorrotzagoa bilakatuko genuke prozesua. Gure helburua ahalik eta aipamen gehien topatzea baldin bada, ordea, langa handiagotuz lor dezakegu. Baina honek aipamen desegokiak topatzea ekar lezake.

*Spotlighte*n zerbitzaria martxan jarri aurretik alda dezakegu balioa, 0tik 1erainoko balioak onartzen ditu. Horretarako, ereduaren karpetan dagoen *spotter_thresholds.txt* fitxategiaren azkeneko balioa aldatu behar dugu. Balio lehenetsia 0.7 bertsioan 0.3koa da.

Spotter

*Spotter*a aipamenak topatzeko tresna da, parametro honekin zein algoritmo erabiliko den zehaztu ahal dugu. Ondorengo algoritmoak onartzen ditu:

- *Default*: *Spotter* konfigurazio fitxategian zerrendatuta dagoen lehenengo *spotter*a hautatzen du. Defektuz lehenengo dagoen *LingPipeSpotter*a da.

- *LingPipeSpotter*: Ezagunak diren izenak dituen hiztegia erabiltzen du aipamenak topatzeko.

- *AtLeastOneNounSelector*: Aurrekoaren hiztegi berdina erabiltzen du, baina aipamenak izen bat eduki beharko du.

- *CoOccurrenceBasedSelector*: Aurrekoen hiztegi berdina erabiltzen du, honek, ordea, “entitateak ez diruditenak” kentzen ditu. Erabaki hori hartzeko *Co-ocurrence* estatistikak erabiltzen ditu.

- *NESpotter*: Entitateen izenen ezagutzerako, *Named Entity Recognition*erako (*NER*erako), *OpenNLP*eko defektuzko modeloak erabiltzen ditu.

- *KeyphraseSpotter*: *KEA*ko lehenetsitako modeloak erabiltzen ditu.

- *WikiMarkupSpotter*: Beste tresna batek egin duela eta aipamenak *WikiMarkup* formatuan kodetu dituela onartzen du.

- *SpotXmlParser*: Aurrekoaren berdina baina *SpotXml* formatuan kodetuta.

Horien artean proiektuan erabiliko direnak bi izango dira: lehenetsitakoa (*LingPipeSpotte*ra) eta *SpotXmlParse*ra.

*SpotXmlParse*ri esker, esan bezala, zuzenean adierazi ahal izango diogu *Spotlight*i zein entitate desanbiguatzea nahi dugun. Horretarako *text* parametroaren balioan *SpotXml* deritzon formatuan dagoen testu soila jarriko dugu, bertan adierazten dira testuingurua eta zein aipamen desanbiguatu nahi diren.

Beraz, *spotter* parametroari *SpotXmlParser* balioa emanda Spotlightek aipamenak topatzeko urratsa ez du egingo, zuzenean *SpotXml* formatuan adierazitako aipamenak desanbiguatuko ditu.

* Lerro saltoak eta tabulazioak formatua argiago ikusteko erabili dira. Parametroa pasatzerako garaian lerro bakarrean egin behar da. Kontuz *XML*aren atributuen kakotxekin, sinpleak erabiliz gero arazorik ez.

```
<annotation text="You and I ought not to die before we have explained ourselves to each
other Adams wrote Jefferson in 1815 It is doubtful that today politicians will spend
much time trying to explain themselves to one another even after they leave office They
are after all creatures of a culture in which it is acceptable on the Senate floor for
Vice President Dick Cheney to .">
  <surfaceForm name="You" offset="213"/>
  <surfaceForm name="Adams" offset="289"/>
  <surfaceForm name="Jefferson" offset="301"/>
  <surfaceForm name="It" offset="319"/>
  <surfaceForm name="politicians" offset="345"/>
  <surfaceForm name="They" offset="451"/>
  <surfaceForm name="creatures" offset="470"/>
  <surfaceForm name="culture" offset="485"/>
  <surfaceForm name="Senate" offset="526"/>
  <surfaceForm name="Vice President Dick Cheney" offset="543"/>
</annotation>
```

13 Irudia: SpotXml formatua duen adibidea.

13 Irudian, desanbiguatu nahi diren entitateak *annotation* elementuaren barnean dauden *surfaceForm*eko *name* atributuan azaltzen dira. Honetaz gain *offset*ean entitatea zenbatgarren karakterean hasten den adierazi behar da, kontuz, 0tik hasten baita (*You*ko Y hizkia 0a izango litzateke).

4.1.3. Ereduak

*Dbpedia Spotlight*ek zenbait hizkuntzetarako ereduak atzigarri ditu sarean edozein erabiltzaile erabiltzeko prest.

Guk erabilitako ereduak ingelesezkoa (en) eta gaztelerazkoa (es) dira. Baina, noski, *Wikipedia*ko ezagutza ingelesezko bertsioan eta gaztelerazkoan ez dira alderagarriak, lehenengoak 4,6 milioi artikulu eta 34,2 milioi orrialde baino gehiago ditu eta bigarrengoak, aldiz, 1,1 milioi artikulu eta 4,7 milioi orrialde baino gehiago. Honetaz ohartuta eta jakinda erabiltzaile askok azkartasuna nahiago izaten dutela ingelesezko eredu txikitua eskaintzen dute. Beraz, bi eredu mota eskaintzen dituzte; osoak eta txikituak.

Eredu osoak(full)

Eredu hauetan *Dbpedia*tik erauzitako informazio guztia dauka; hala nola, entitateak, entitateen kontaketak, testuinguruak, artikuluen arteko erlazioak, eta abar. Horretarako oraindik dokumentatuta ez dagoen prozedura jarraitzen dute.

Eredu txikituak(light)

Eredu txikituak, eredu osoekin alderatuz, arinagoak dira. Ez dute hainbeste denbora erabiltzen eta, gainera, memoria kontsumoa ere txikiagoa da. Horretarako, *Spotlight*eko 0.6 bertsioko ingeleseko eredu txikituak testuinguruak ez ditu kontuan hartzen. Eredu honek, beraz, aipamen berdinerako beti entitate berdina itzuliko luke nahiz eta testuingurua guztiz ezberdina izan.

Aldiz, *Spotlight* 0.7 bertsioarekin guk geuk sor ditzakegu gure beharretara gehien hurbiltzen den eredu txikitua. Hori lortzeko bi modu eskaintzen dizkigu; batetik, aurreko bertsioan egiten dena, hots, testuinguruak kontuan ez hartzea. Bestetik, eredua sortzean erabiltzen diren inausketa parametroak handiagoak zehaztea.

Eredua birsortzeko beste inausketa parametro batzuekin ondorengo urratsak jarraitu behar dira:

1. Webgunetik[4] nahi dugun hizkuntzako datuak jaitsi.
2. Eredua sortu komando lerrotik:

```
$ mvn clean install
$ export MAVEN_OPTS="-Xmx26G" //erabili nahi dugun memoria kopurua
$ export JAVA_HOME="/usr/lib/jvm/java-6-oracle/jre" //Java 6ren kokapena
$ mvn -pl index exec:java -Dexec.mainClass=org.dbpedia.spotlight.db.CreateSpotlightModel
-Dexec.args="en_US en_US en3+5 None en/stopwords.list EnglishStemmer prune=3,5"
```

4 http:/spotlight.sztaki.hu/downloads/raw/

Aurreko adibidean sortutako ereduan adibidez, "*prune=3, 5*" zehaztu dugunez, eredu berri honetan soilik 3 bider baino gehiago agertutako entitateak eta 5 alditan baino gehiagotan agertutako testuinguru aipamenak egongo dira.

4.1.4. Moduluak

*Spotlight*eko moduluei esker zein lan burutzea adierazi ahal izango diogu. Lau Aukera ezberdin eskaintzen dizkigu: *Spot*, *Disambiguate*, *Annotate* eta *Candidates*. Horretarako *CURL* aginduan zerbitzariaren helbidearen amaieran zehaztuko dugu zein modulu deitu nahi dugun, 4.1.5 Sarrera / irteera formatuak atalean adierazten dugu zehatzago.

Spot

Testu bat jasotzen sarreratzat eta bertan agertzen diren entitateak topatzen eta prestatzen ditu desanbiguatu ahal izateko. Urrats hau egiteko zenbait teknika daude erabilgarri; hala nola hiztegi begizta eta *Named Entity Recognition (NER)*. Esan bezala, erabiliko den teknika *spotter* parametroak zehazten du.

```
<annotation text="You and I ought not to die before we have explained ourselves to
each other Adams wrote Jefferson in 1815 It is doubtful that today politicians will
spend much time trying to explain themselves to one another even after they leave
office They are after all creatures of a culture in which it is acceptable on the
Senate floor for Vice President Dick Cheney to .">
  <surfaceForm name="You" offset="213"/>
  <surfaceForm name="Adams" offset="289"/>
  <surfaceForm name="Jefferson" offset="301"/>
  <surfaceForm name="It" offset="319"/>
  <surfaceForm name="politicians" offset="345"/>
  <surfaceForm name="They" offset="451"/>
  <surfaceForm name="creatures" offset="470"/>
  <surfaceForm name="culture" offset="485"/>
  <surfaceForm name="Senate" offset="526"/>
  <surfaceForm name="Vice President Dick Cheney" offset="543"/>
```

14 Irudia: Spot modulura egindako eskaera bati erantzundako emaitza da. Bertan, zein entitate-izen topatu dituen ikus daiteke

Disambiguate

Sarrera testua jasotzen du, baina honek dagoeneko entitateak topatuta eta markatuta ditu XML moduan. Testuan dagoen testuinguruaren arabera entitate bakoitza identifikatzen du, hots desanbiguatu egiten du, Dbpediako identifikadore uniboko bat emanez.

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<title>DBpedia Spotlight annotation</title>
<meta http-equiv="Content-type" content="text/html;charset=UTF-8">
</head>
<body>
<div>
<a href="http://dbpedia.org/resource/Youzhou" title="http://dbpedia.org/resource/Youzhou"
target="_blank">You</a> and I ought not to die before we have explained ourselves to each other <a
href="http://dbpedia.org/resource/John_Adams" title="http://dbpedia.org/resource/John_Adams"
target="_blank">Adams</a> wrote <a href="http://dbpedia.org/resource/Thomas_Jefferson" title="http://
dbpedia.org/resource/Thomas_Jefferson" target="_blank">Jefferson</a> in 1815 <a href="http://
dbpedia.org/resource/Wikipedia" title="http://dbpedia.org/resource/Wikipedia" target="_blank">It</a> is
doubtful that today <a href="http://dbpedia.org/resource/Politician" title="http://dbpedia.org/resource/
Politician" target="_blank">politicians</a> will spend much time trying to explain themselves to one
another even after they leave office <a href="http://dbpedia.org/resource/New_Gods" title="http://
dbpedia.org/resource/New_Gods" target="_blank">They</a> are after all <a href="http://dbpedia.org/
resource/Legendary_creature" title="http://dbpedia.org/resource/Legendary_creature"
target="_blank">creatures</a> of a <a href="http://dbpedia.org/resource/Culture" title="http://
dbpedia.org/resource/Culture" target="_blank">culture</a> in which it is acceptable on the <a
href="http://dbpedia.org/resource/United_States_Senate" title="http://dbpedia.org/resource/
United_States_Senate" target="_blank">Senate</a> floor for <a href="http://dbpedia.org/resource/
Dick_Cheney" title="http://dbpedia.org/resource/Dick_Cheney" target="_blank">Vice President Dick
Cheney</a> to .
</div>
</body>
</html>
```

15 Irudia: Disambiguate edo annotate modulura egindako eskaera bati erantzundako emaitza da. Bertan, HTML formatua duen emaitzan, testuinguruan zehar emaitza gisa desanbiguazio entitateen estekak ageri dira.

Annotate

Aurreko bi moduen kateaketa da. Testua sarrera da; entitateak topatzen ditu. Ondoren, testuingurua kontuan hartuta desanbiguatzen ditu; identifikadore bat itzuliz entitate bakoitzarentzat.

Candidates

*Annotate*ren oso antzekoa, baina entitate bakoitzari identifikadore bakarra eman beharrean identifikadore lista bat itzultzen du. Listan dauden identifikadoreak entitatea desanbiguatzeko hautagaiak dira, *finalScore*ren arabera ordenatuak.


```
<annotation text="You and I ought not to die before we have explained ourselves to each other Adams wrote Jefferson
in 1815 It is doubtful that today politicians will spend much time trying to explain themselves to one another even
after they leave office They are after all creatures of a culture in which it is acceptable on the Senate floor for
Vice President Dick Cheney to . ">
  <surfaceForm name="You" offset="213">
    <resource label="Youzhou" uri="Youzhou" contextualScore="0.9999630851864232"
percentageOfSecondRank="1.1336565924762143E-4" support="136" priorScore="1.1308067623608009E-6"
finalScore="0.9998641084775284" types=""/>
    <resource label="You" uri="You" contextualScore="8.094156436887603E-6"
percentageOfSecondRank="0.1727928207355762" support="873" priorScore="7.25878164368367E-6"
finalScore="1.1335025381559028E-4" types=""/>
    <resource label="You (George Harrison song)" uri="You_(George_Harrison_song)"
contextualScore="5.2666644483183995E-6" percentageOfSecondRank="0.07393925061163428" support="107"
priorScore="8.896788497985713E-7" finalScore="1.958611008788935E-5" types="Schema:CreativeWork, DBpedia:Work,
DBpedia:MusicalWork, DBpedia:Single"/>
    <resource label="You (Gong album)" uri="You_(Gong_album)" contextualScore="1.2896995145889802E-5"
percentageOfSecondRank="0.7518237204975854" support="21" priorScore="1.746098677174766E-7"
finalScore="1.448182302295509E-6" types="Schema:CreativeWork, DBpedia:Work, DBpedia:MusicalWork, Schema:MusicAlbum,
DBpedia:Album"/>
    <resource label="You (Robin Stjernberg song)" uri="You_(Robin_Stjernberg_song)"
contextualScore="8.144864174010632E-6" percentageOfSecondRank="0.2313236860083988" support="20"
priorScore="1.662951121118825E-7" finalScore="1.0887778064705685E-6" types="Schema:CreativeWork, DBpedia:Work,
DBpedia:MusicalWork, DBpedia:Single"/>
    <resource label="You (Lloyd song)" uri="You_(Lloyd_song)" contextualScore="1.2560666685132665E-6"
percentageOfSecondRank="0.0" support="25" priorScore="2.078688901398531E-7" finalScore="2.518600954369l1E-7"
types="Schema:CreativeWork, DBpedia:Work, DBpedia:MusicalWork, DBpedia:Single"/>
  </surfaceForm>
  <surfaceForm name="Senate" offset="526">
    <resource label="United States Senate" uri="United_States_Senate" contextualScore="0.20309016784471717"
percentageOfSecondRank="0.0249164742041648l2" support="28867" priorScore="2.4002205006668558E-4"
finalScore="0.9752219313382804" types="DBpedia:Agent, Schema:Organization, DBpedia:Organisation,
DBpedia:Legislature"/>
```


16 Irudia: Candidates modulura egindako eskaera bati erantzundako emaitza da, XML formatuan. Bertan, izen-entitate batek izan ditzakeen adiera ezberdinak ikus daiteke.

4.1.5. Sarrera / irteera formatuak

Eskaeren emaitzak jasotzeko zenbait formatu eskaintzen dizkigu *Spotlight*ek; *HTML*, *XML*, *XHTML+XML* eta *JSON* besteak beste. Zein formatu erabili nahi dugun adierazteko, *cURL* aginduan -H "*Accept:X/Y*" gehitu behar diogu; non X *text* edo *application* eta Y formatu ezberdinak izango diren[5]. Honen bidez *cURL* aginduak egindako *HTTP* deiari zehazten diogu zein den jasoko dugun erantzunaren formatua. Honako hauek dira formatu bakoitzaren berezitasunak:

5 *HTML* eta *XML* *text*ekin batera doaz eta beste biak, *XHTML+XML* eta *JSON*, *application*ekin

- *HTML* formatuak soilik desanbiguatu nahi den aipamenaren entitatea itzuliko digu testuinguruaren artean. 3.3.6 irudian formatu honen adibidea dago.
- *XHTML+XML* formatudun erantzunak *RDF*a formatua dute bere baitan. *RDF*a txertatzearen helburua dokumentuei semantika txertatzea da, horretarako *XHTML*ko *meta* elementuaren eta esteken atributuetaz baliatzen da. Modu honetan *W3C*k sortutako *Distiller and Parser tresnarekin* beste formatu batzuetara pasatu ahal izango dugu, *Turtle* eta *N Triples* besteak beste.
- *XML* formatuan jasotzen ditugun erantzunak, *JSON*ekin batera, informazio gehien eskaintzen dizkigutenak dira. Entitatearen motaz, entitate anbiguoaz eta kokapenaz gain, honek jaso dituen *similarityScore*a , *support*a eta *percentageOfSecondRank*a adierazten da. Hauei esker emaitzaren fidagarritasuna neurtu ahal izango dugu. 17 Irudian ikus daiteke formatu honen adibidea.


```
<Annotation text="You and I ought not to die before we have explained ourselves to each other Adams
wrote Jefferson in 1815 It is doubtful that today politicians will spend much time trying to explain
themselves to one another even after they leave office They are after all creatures of a culture in
which it is acceptable on the Senate floor for Vice President Dick Cheney to ." confidence="0.0"
support="0" types="" sparql="" policy="whitelist">
  <Resources>
    <Resource URI="http://dbpedia.org/resource/Youzhou" support="136" types="" surfaceForm="You"
offset="0" similarityScore="0.9998641084775284" percentageOfSecondRank="1.1336565924762143E-4"/>
    <Resource URI="http://dbpedia.org/resource/John_Adams" support="1754"
types="DBpedia:Agent,Schema:Person,Http://xmlns.com/foaf/0.1/
Person,DBpedia:Person,DBpedia:OfficeHolder" surfaceForm="Adams" offset="76"
similarityScore="0.9998541767068397" percentageOfSecondRank="1.4240104728890235E-4"/>
    <Resource URI="http://dbpedia.org/resource/Thomas_Jefferson" support="4432"
types="DBpedia:Agent,Schema:Person,Http://xmlns.com/foaf/0.1/
Person,DBpedia:Person,DBpedia:OfficeHolder" surfaceForm="Jefferson" offset="88"
similarityScore="0.9999943809397522" percentageOfSecondRank="2.052769124862812E-6"/>
    <Resource URI="http://dbpedia.org/resource/Wikipedia" support="21130"
types="Schema:CreativeWork,DBpedia:Work,Schema:WebPage,DBpedia:Website" surfaceForm="It" offset="106"
similarityScore="0.791561497487455" percentageOfSecondRank="0.18871888138134668"/>
    <Resource URI="http://dbpedia.org/resource/Politician" support="16500" types=""
surfaceForm="politicians" offset="132" similarityScore="0.6891733108232766"
percentageOfSecondRank="0.3328203772250774"/>
    <Resource URI="http://dbpedia.org/resource/New_Gods" support="276" types="" surfaceForm="They"
offset="238" similarityScore="0.9464038275243911" percentageOfSecondRank="0.036708854951869735"/>
    <Resource URI="http://dbpedia.org/resource/Legendary_creature" support="299" types=""
```


17 Irudia: XML formatua duen emaitzaren zatia ikus daiteke. Bertan, egindako deiaren parametroak eta desanbiguatutako entitateak ageri dira.

Kontuan hartu behar da, modulu guztiek ez dituztela irteera formatu guztiak onartzen eta aukeratzen den formatuaren arabera informazio ezberdina jasotzen dugula.

4.2. UKB

Ondorengo ataletan UKB programa-bildumako hitz adieren desanbiguazioa burutzeko erabiltzen diren parametroak, baliabideak eta sarrera-irteera formatuak analizatuko ditugu

4.2.1. Parametroak

UKBk parametro bidez jasotzen ditu ezagutza-baseko grafoa eta hiztegia. Hauetaz gain, *allranks, dict_weight, ppr_w2w, nopos, minput* eta *prank_iter* parametroak ere baditu. Ondorengo zerrendan parametro bakoitza azalduko dugu.

- *K*: Bertan ezagutza basea pasatu behar diogu.
- *D*: Erabiliko dugun hiztegia adierazi behar diogu.
- *minput*: Parametro honi esker UKB derrigortzen dugu nahiz eta sarrera datuak erroreak izan aurrera egitera. Sarrera fitxategiak desanbiguatzeko eskaera asko dituenean, nahiz eta UKBk eskaera bat prozesatzean errorea izan hurrengoko eskaera desanbiguatzen jarraituko du. Modu honetan denbora aurreztu ahal izango dugu erroreren bat gertatuz gero.
- *nopos*: Parametro honekin entitateen identifikadoreak edozein forma eduki dezake '#' eta zuriuneak izan ezik.
- *ppr_w2w*: Helburu hitzak desanbiguatzean *PageRank* Pertsonalizatua erabiltzen da hitzez hitz.
- *dict_weight*: Aipamenak entitateetara pisu bidez lotzeko.
- *prank_iter*: *PageRank* algoritmoan egingo diren iterazio kopurua; lehenetsitako balioa 30.

- *allranks*: Parametro honi esker emaitza posible guztiak zerrendatuta itzultzen ditu *confidence* (konfiantza) balioaren arabera ordenatuta.

```
run_64/ukb_wsd2.1 -K ~jirhizts/Corpus/Wikipedia_es/11Feb2012/wiki2012.Ab.csr64
-D ~jirhizts/Corpus/Wikipedia_es/11Feb2012/dict_full.txt
--nopos--minput --ppr_w2w --dict_weight --allranks input > output
```

18 Irudia: UKBri egindako dei baten adibidea.

4.2.2. Baliabideak

UKBk beharrezkoak dituen baliabideak hiztegia eta grafoa dira. Hiztegia desanbiguazio hautagaiak sortzeko erabiltzen da eta grafoa ezagutza-basea izango da. Hauen ezaugarrien arabera emaitza ezberdinak itzuliko dituelarik.

Grafoak

Grafoak sortzeko *Wikipedia*ko erauzketa[6] erabiltzen dugu, bertako artikuluen arteko loturak lortzeko. *Wikipedia*ko 2012ko otsailaren 11ko erauzketa erabili dugu, hain zuzen ere. Guretzako artikuluak *Wikipedia*ko orrialde guztiak izango dira; birbideratze, desanbiguazio eta kategoria orriak izan ezik. Artikulu bakoitza nodo bati dagokio. Gure grafoan artikulu batetik beste batera lotura zuzendua egongo da, lehenengo artikuluko testuan bigarren artikulura erreferentzia egiten dion esteka, aingura, baldin badu.

Gaztelaniazko esperimentuetan gaztelaniazko hiru grafo ezberdinez baliatu gara. Grafo hauen ezaugarri bereizgarria grafoaren nodoen arteko loturetan dago. Hiru grafo ezberdin sortu ditugu; zuzendua (AD) eta ez-zuzenduak (AU eta AB). 19 Irudian ikus ditzakegu grafo hauen azalpen grafikoa.

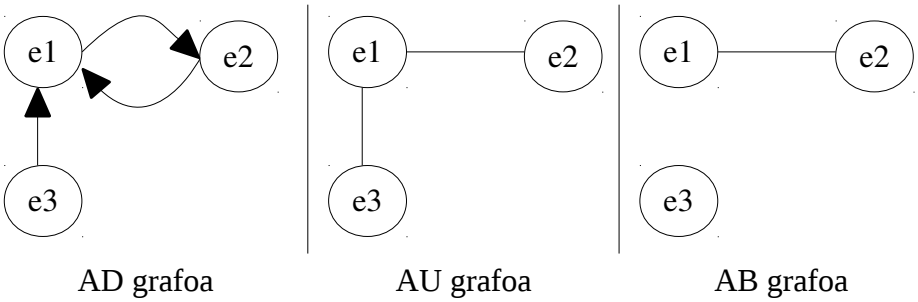
Grafo zuzendua tamainaz ertaina da, zentzu ezberdinetako loturak ditu bere baitan. Ezagutza-baseko erlazio guztiak ditu.

6 *Dump*a.

AU grafo ez-zuzendua ezagutza-baseko lotura guztiak bikoitzak bilakatzen ditu. Hots, behin bi nodoen arteko lotura izan, hauek lotuta geratuko dira bi noranzkoetan. Horrela, nodoen arteko erlazioak bi noranzkokoak direla esan dezakegu. Horregatik, hiru grafoen artean handiena da.

AB grafoa txikiena da. Nodoen arteko loturak bi noranzkokoak direnean jarriko ditu soilik. Hau da, noranzko bakarreko loturak alde batera uzten ditu. Horrela erlazio handia duten nodoak egongo dira lotuak, ziurtasun handiagoa espero daiteke.

Nahiz eta 19 Irudian AU eta AB grafoen nodoen arteko lotura ez-zuzendu bakarra azaldu, UKBek, benetan, bi lotura zuzendu ditu, bat norabide bakoitzarentzat.



19 Irudia: UKBrekin desanbiguatzeko sortutako 3 grafo moten azalpen grafikoa. Bertan kasu zehatza grafo mota ezberdinetan duen adierazpena ikus daiteke.

Hiztegia

Testua grafoko nodoekin erlazionatzeko hiztegia erabiltzen dugu, hots, aipamenaren eta aipamen horrek erreferentzia ditzakeen gainerako artikuluen artean erlazionatzeko erabiltzen dugu. Emandako aipamenaren artikulu hautagaiak sortzeko beharrezko baliabidea dugu.

Hiztegia sortzeko grafoa sortu dugun *Wikipedia*ko erauzketa berdina baliatu dugu eta honen artikulu izenburuak, birbideratzeak, desanbiguazio orriak eta aingurak erabili ditugu. Aipamenak hizki xehetara pasatzen dira eta parentesien arteko guztia ezabatzen da. Aingurak desanbiguazio orri bati erreferentzia egiten badio, testua desanbiguazio orriak dituen artikulu guztietara erlazionatuko da. Aipamen eta artikuluen arteko erlazioaren probabilitatea aldez aurretik kalkulatzen da, estimatzeko aipamena aingura horrekin erreferentziatutako kopurua zati aipamen hori erreferentziatua den aingura guztien kopurua egiten dugu.

Aingura	Artikuluak (hautagaiak)
determinantes	Determinante_(matemática):23 Artículo_(gramática):1 Lingüistica:4 Determinante_(lingüistica):22
dq	Dragon_Quest:1 DQ:2
miarritze	Biarritz:2

6 Taula: UKBk jasotzen duen hiztegiaren adibideak dituen taula da.

6 Taula erabilitako hiztegiaren zatia dugu, bertan ageri diren adibideetan ikus daiteke, hiztegian hautagaien ":" bi puntuen ondoren zenbaki bat dutela. Zenbaki horrek artikulu hori aingura horrekin zenbatetan izan den erreferentziatua adierazten du.

4.2.3. Sarrera / irteera formatuak

UKBk, desanbiguazioa egin ahal izateko, testu soila da sarreratzat jasotzen duena. Testu horrek, ordea, formatu zehatzean adierazita egon behar du, UKBk interpretatu ahal izateko. UKB prestatuta dago zenbait eskaerei aurrea egiteko, horregatik sarrera testuan desanbiguazio eskaera bat baino gehiago egon daiteke.

```
AFP_SPA_20071003.0158
lancaster##EL_SPA_01861#1 bush##id#0 corea_del_norte##id#0 irán##id#0 eeuu##id#0 3##id#0
oct##id#0 2007##id#0 george_w_bush##id#0 estados_unidos##id#0 2006##id#0
```

20 Irudia: UKBk jasotzen duen sarrera dokumentu baten adibidea.

Sarrera formatua 20 Irudian ikus daiteke. Bertan zenbait atal ezberdin aipatu behar ditugu. Lehenengo lerroan identifikadore bat jasotzen du, gure kasuan dokumentuaren identifikadorea da. Hurrengo lerroan desanbiguatu nahi den testua adierazten da hurrengo formatuarekin; aipamena##aipamenaren_ida#[0-1]. Azkenengo atalak desanbiguatu nahi den edo ez adierazten du; 1ekoarekin desanbiguatuko da. Aipamenak ezberdintzeko hutsuneak erabiltzen dira; beraz aipamen konposatuak azpi-gidoiaren bidez osatuko dira.

AFP_SPA_20071003.0158
EL_SPA_01861 Lancaster_(Pensilvania)/0.726765
Juan_de_Gante/0.0538932 Condado_de_Lancaster_(Pensilvania)/0.0516175
Lancaster_(Lancashire)/0.0456574 Avro_683_Lancaster/0.0352058
Condado_de_Lancaster_(Nebraska)/0.0286633
Estrecho_de_Lancaster/0.0238097 Lancashire/0.0223989
Condado_de_Lancaster_(Virginia)/0.00437475
Lancaster_(California)/0.0030923 Lancaster_(Ohio)/0.000858973
NH_Lancaster_Hotel/0.00083317
Condado_de_Lancaster_(Carolina_del_Sur)/0.000817888
Lancaster_(Nuevo_Hampshire)/0.000515385
Lancaster_(Texas)/0.000515384 Lancaster_(Kentucky)/0.00025741
Lancaster_(Carolina_del_Sur)/0.000220972
Lancaster_(Wisconsin)/0.00019814 Joseph_Lancaster/0.000171795
Lancaster_(Misuri)/0.000133152 !! lancaster

21 Irudia: UKBk itzulitako emaitza.

UKBk itzulitako emaitza 21 Irudian ikus daiteke eta honen formatua ondorengoa da. Hasteko, dokumentuaren identifikadorea zehazten du. Hurrengo lerroan desanbiguatu nahi zen aipamenaren identifikadorea eta jarraian entitate posibleak adierazten ditu haien konfiantza balioarekin batera ordenatuta, *allranks* parametroarekin egindako deia baldin bada, bestela konfiantza balio handiena duen entitatea adierazten da soilik. Amaitzeko desanbiguatu nahi den aipamena adierazten da.

5. ESPERIMENTUAK

Atal honetan *Dbpedia Spotlight* zein UKB erreminten gaineko esperimentuei buruz arituko gara. Esperimentuen helburu nagusiak hizkuntza ezberdinetan tresna hauen erantzuteko ahalmena, erantzun hauen zehaztasuna eta honek duen memoria eta denbora kostua neurtzea izan dira. Gainera, beste zenbait ezaugarri ikertu ditugu; hala nola, testuinguruak, ezagutza-baseak eta lematizazioak duten eragina.

Ondorengo ataletan hizkuntz bakoitzean egin dugun esperimentuak egikaritzeko jarraitutako ataza nagusiak azalduko ditugu.

5.1. Ingelesezko esperimentuak

Ingelesezko esperimentuak egiteko *Dbpedia Spotlight* tresna erabili dugu, 0.7 eta 0.6 bertsioak. Neurketa ezberdinak egiteko zenbait parametro erabili ditugu: *Spotlight*en iragazki parametroak parametroak, aipamena eta testuingurua. Esperimentuak parametro hauei balio ezberdinak emanez egin dira. Baina esperimentu nagusi hauetaz gain, 4.1.3 atalean azaltzen diren *Spotlight* 0.6 bertsioan eskaintzen diguten ingelesezko bi ereduen arteko aldea ezagutu nahi dugu; zenbait erabilpenerako interesgarria izan daitekeelako.

Irismenean azaldu bezala, garapenerako, hots parametroen ikerketarako, erabiliko dugun datu-multzoa TAC 2010ekoa izango da; ebaluaziorako, ordea, AIDA eta TAC 2011.

Horretarako jarraitu dugun metodologia parametroak banan-banan probatzea izan da. Hasteko, *Spotlight*ek eskaintzen dizkigun parametro onenak bilatu ditugu. Ondoren, *spot*a, testuingurua eta ereduak ikertu ditugu. Eta behin parametroen eraginak jakinda, hobekien datorkigun konbinaketarekin ebaluazioa egin dugu.

5.1.1. Lehenengo urratsa, aurreprozesamendua

Dbpedia Spotlight tresna modu automatikoan erabiltzeko *script*a garatu dugu. Modu honetan, sarreratzat datu-multzoa izanik, eskaera bakoitzeko *Spotlight*i egin beharreko deia prestatzen du, azken honek prozesatu ahal izateko.

Hasiera batean *Spotlight*ek jasotzen dituen deiak eskuz aldatzen genituen esperimentu ezberdinak egiteko, hau da, *confidence* parametroaren balio ezberdinak probatzeko *script*eko deiaren *confidence* balioa eskuz moldatzen genuen. Ez da modurik egokiena, baina bai emaitzak edukitzen hasteko modurik azkarrena.

Ikertu nahi dugun parametro batek, aipamenak, urrats bat gehiago ematea eskatzen du. 4.1.2 *Spotter* atalean azaldu bezala, desanbiguatu beharreko izen-aipamena zein den zehaztu ahal diogu *Spotlight*i eskaera barruan *spotXML* formatudun karaktere katea txertatuta. Honetaz baliatu gara Matx prozesuan lortutako izen-aipamena desanbiguatu ahal izateko. Horrela *Spotlight*en *spotter*ra eta Matx alderatu ahal izan ditugu.

Tresnak deiak prozesatu ostean, emaitzak itzuliko ditu banan-banan. *Spotlight*en bi irteera formatu ezberdin erabili ditugu emaitza horiek jasotzeko, *HTML* eta *JSON*. Modu honetan, softwarearen analisian ikusi dugun bezala, emaitzen informazio ezberdina lortu ahal izan dugu. Hala ere, *JSON* erabili dugu *HTML* baino azkarragoa baita.

Zein aipamen desanbiguatu nahi dugun zehaztu ezean, *Spotlight*ek identifikatzen dituen aipamen guztiak desanbiguatuko ditu. Beraz, erantzun guzti horien artean datu-multzoak zehazten digun aipamenarena topatu behar dugu. Horretarako, TACen kasuan, honek zehazten digun aipamena barne duen *Spotlight*ek identifikatutako aipamen anbiguoen artean lehenengoa aukeratuko dugu. AIDAn, ordea, zehazki adierazten digute zein aipamen desanbiguatu. Aipamen anbiguo horri dagokion izen-entitatea izango da emaitza.

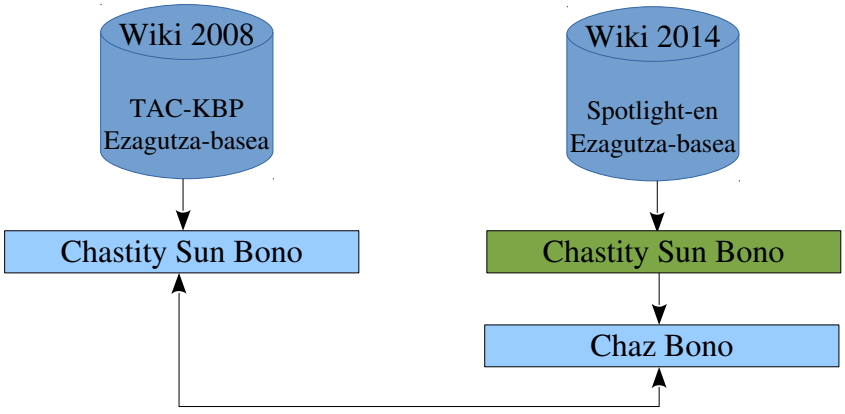
*Spotlight*ek emandako emaitzak identifikadore eta desanbiguatutako entitate bikoteetan bilakatuko ditugu, ondoren ebaluazioan erabiliko diren *script*en sarrera formatu berdina izan dezaten.

5.1.2. Entitateen bihurketa

Aurrekarietan eta software analisi ataletan azaldu dugu *Spotlight*ek duen ezagutza-basea eta datu-multzoena ez direla berdinak. Bi entitate alderatu ahal izateko ezagutza-base berdinekoak izan behar dira.

Hau kontuan izanik, datu-multzo ezberdinetan prozedura ezberdina erabili dugu. Batetik, *Spotlight*ek erantzundako izen-entitateak TACeko ezagutza-basera bihurtzea. Bestetik, AIDAren kasuan, honen ezagutza-basea *Wikipediako* 2013ko bertsiora bihurtu dugu, hain zuzen *Spotlight*en ezagutza-basera. AIDAn horrela jokatzearen arrazoia honen ezagutza-basearen *Wikipedia* erauzketa prozesaturik ez izatea izan da. TACen ezagutza-baseari dagokion *Wikipedia* erauzketa, ordea, badugu prozesatuta.

Baina, *Wikipedia*ko bertsioko izen-entitateak bihurtzeko *Wikipedia*k duen funtzionamendua azaldu beharra dago.



Irudia 22: Izenburu ezberdina duten arren, entitate berdinera erreferentzia egiten dioten Wikipediako bi bertsio ezberdinen entitateak ikus daitezke. Adibideko kasua famatu batek egindako izen aldaketa batek eragindakoa da.

*Wikipedia*ko artikulu ezberdinen identifikadore unibokoa izenburua bera da. Baina gerta daiteke izenburu horrek denboran zehar aldaketak jasatea. Egoera hauen aurrean *Wikipedia*k izenburu zaharra duen orria birbideratze orrialde bilakatzen du, honen adibidea 23 Irudian ikus dezakegu. Beraz, aurreko urteetako izenburua topatzeko birbideratze hauek ebatzi behar ditugu.

Gerta daiteke *Spotlight*ek erantzundako entitatea TACeko ezagutza-basera bihurtzean galtzea, hots, entitate hori ezagutza-basean ez existitzea. Kasu hauetan NIL bihurtu ditugu, 2013an existitzen diren artikulu batzuk *Wikipedia*ko 2008 bertsioan ez baitzeuden. AIDAren kasuan izen-entitateak 2010etik 2013ra bihurtzen direnez ez ditugu inolako arazorik izan.

5.1.3. Ebaluaziorako datuak biltzen

Behin *Spotlight*eko emaitzak datu-multzoko ezagutza-baseko entitate bilakatu ditugula, honen ebaluazioa egin dugu. Horretarako emaitzak urre patroiarekin alderatu ditugu, azken emaitzak lortu ahal izateko. Emaitza guzti horiek aurrekarietan adierazitako ebaluazio-neurriak dira.

```
sh 00-desanbiguatu_ebaluatu.sh ~/azken_emaitzak/7.00FT1MW-originala
datasetak/tac2010_id_string_MatchString_Windowtext.txt  en JSON
~/home_zaharra/EntityLinking/wikipedia-miner/04April2013/
~/home_zaharra/EntityLinking/entitylinking/erauzitako_datasetak/TAC-
KBP/gs/2010_test.htm 2223 2 false 0 0
desanbiguatzailea_JSON.pl exekutatzen

Total number of CPU-seconds that the process spent in user mode: 7.97
Total number of CPU-seconds that the process spent in kernel mode: 0.82
Elapsed real time: 0:41.70
Percentage of the CPU that this job got, computed as (7.97 + 0.82) /
0:41.70: 21%
Average total (data+stack+text) memory use of the process, in Kbytes: 0
Maximum resident set size of the process during its lifetime, in Kbytes:
70736
```

23 Irudia: Bertan, desanbiguatzeko scripta martxan jartzean, time komandoak eskaintzen dizkigun datuak daude. Denbora erreala hartu dugu erreferentziatzat gure esperimentuetako neurketentzat.

Abiadura eta memoria neurtzeko *Linux*eko *time* eta *top* komandoetaz lagundu gara. Lehenengoarekin *Spotlight*ek eskaerak prozesatzeko behar izan duen denbora lortu dugu. Bigarrengoarekin, ordea, tresnak erabilitako memoria neurtu ahal izan dugu. Denbora neurketa egiteko, esperimentuak errepikatu dira bataz besteko denbora lortzeko. Gainera, esperimentuak zerbitzariaren egoera berdintsuetan egin ditugu, denbora zerbitzariaren lan-kargaren araberakoa baita.

5.1.4. Parametroen ikerketa

*Dbpedia Spotlight*en parametro optimoak bilatu ditugu; hots iragazki parametroak eta *threshold*a. Baina, esan bezala, *spot*a, testuingurua eta eredua bezalako beste zenbait ezaugarri ere ikertu ditugu.

Esperimentu hauetarako testuinguru osoa eta *spoter*ra *Spotlight*ek egindakoa erabili dugu, noski dagozkien esperimentutan izan ezik. Esperimentu hauek TAC 2010 datu-multzoan egikaritu ditugu.

Iragazki parametroak

Hasteko, *Spotlight*ek eskaintzen dituen iragazki parametroekin aritu gara; *confidence*, *support* eta *coreference resolution*. Irismenean azaldu bezala, parametro hauen konbinaketa onena bilatu dugu.

Esperimentua egiteko bi konbinaketa mota erabili ditugu. Batetik, balio lehenetsiak, *coreference resolution* izan ezik eta, bestetik, guk emandako balioak. *Confidence*n eta *support*en kasuan 0ko balioa izan dute, *Spotlight*ek erantzun gehiago itzultzeko asmoz, eta *coreference resolution* parametroak, aldiz, *false* balioa hartu behar du, definizioz bestela ez baitira gainerako iragazkiak kontutan hartzen.

Ez ditugu 0.6 bertsioko emaitzak jartzen ez baitira adierazkorrak. Hau da, konbinaketa ezberdinen emaitzen ebaluazio-neurriak ia-ia berdinak dira.

	Tresna	P	R	f1	C	SYS	OK
1	SP-0.7 confi=0.1, sup=10 coref=F	86.63	67.35	75.79	77.75	793	687
2	SP-0.7 confi=0, sup=0 coref=F	86.79	67.65	76.03	77.94	795	690

7 Taula: Spotlighten argumentuekin aldatuz egindako esperimentuen emaitzak. Bertan, doitasuna (P), estalsura (R), f1, coverage (C), sistemak erantzundako eskaera kopurua (SYS) eta sistemak ondo erantzundakoak (OK) ageri dira. Kontuan izan behar dugu, TAC 2010 1020 EzNIL eskaera daudela.

7 Taulan 0.7 bertsioan egindako esperimentuen emaitzak daude. Bertan, *Spotlight*eko analisian ikusi genuen bezala, emaitzek baieztatzen dute: Parametroak 0ra hurbiltzean erantzun gehiago itzultzen ditu, nahiz eta hauek ziurtasun txikiagokoak izan. Beraz, ondorengo esperimentuetan guk ezarritako balioak erabiliko ditugu, alegia, *confidence* 0, *support* 0 eta *coreference-resolution false.*

Atalasea

Atalase iragazkiaren bidez, *Spotlight*ek berak aipamenak topatzeaz (*spotter*ra egiteaz) arduratzen denean, aipamen gehiago identifikatzea lor daiteke. Modu honetan, aukera gehiago ditugu erantzun egokia itzultzeko. Hori alderatu ahal izateko aurrekoan bezala balio lehenetsiak eta guk ezarritakoak erabili ditugu. Guk ezarritako balioa 1.0 izan da, ahalik eta erantzun gehien itzul ditzan.

	Tresna	P	R	f1	C	SYS	OK
1	SP-0.6 Atalasea Def.(0.25)	87.59	61.57	72.31	70.29	717	628
2	SP-0.6 Atalasea 1	87.94	65.78	75.27	74.80	763	671

8 Taula: Spotlighteko 0.6 bertsioan iragazki parametroen konbinaketa onenarekin thresholda moldatuz eginiko esperimentuen emaitzak ageri dira.

*Spotlight*eko 0.6 bertsioa erabilita lortutako emaitzak 8 Taulan daude. Atalasea igotzean 0.6 bertsioak jasaten duen aldaketa nabaria dela ikus dezakegu bertan; 46 eskaera gehiago erantzuten baititu eta 43 gehiago ondo. Hori argi ikusten dugu estaldura eta *coverage*a 4 puntu igo baitira. Gainera nahiz eta, berez, atalasea 1ean ezarrita lortutako emaitza estra hauen ziurtasuna eskasa izan[7], ikusten dugu doitasun handiagoa lortu dugula.

	Tresna	P	R	f1	C	SYS	OK
1	SP-0.7 Atalasea Def.(0.3)	86.79	67.65	76.03	77.94	795	690
2	SP-0.7 Atalasea 1	88.78	69.80	78.16	78.63	802	712

9 Taula: Aurreko emaitzak zerbitzariaren atalase parametroa aldatzean lortutako emaitzekin alderatzen ditugu.

9 Taulan ikus daiteke 0.7 bertsioan atalasea aldatzeak ez duela gehiegi eragiten emaitzetan. Honen arrazoia bertsio honetan *Spotlight*ek erabiltzen duen *spotter*ak aurrekoarena baino erantzuteko gaitasun handiagoa daukala da.

Hala ere, aipatzekoa da badaudela kasuak non lehenetsitako atalaseak eta guk zehaztutakoak itzultzen duten emaitzak ez datozela bat. Arrazoiak gure programaren inplementatzeko modua eta atalasea 1 zehaztean aipamen gehiago identifikatzea dira. Gogoratu gure programak erantzun guztietatik desanbiguatu nahi dugun aipamen anbiguoa barne duen lehenengo erantzuna aukeratzen duela egokitzat. Horregatik atalase txikiagoarekin erantzuten ez duen aipamena aukeratzen dugu egokitzat 1eko atalasearekin. 10 Taulan ikus daitezke hau gertatzen diren zenbait kasu

Gainera, NIL izateak ez du esan nahi *Spotlight*ek ez duenik erantzun, baizik eta entitate bihurketa egitean *Spotlight*ek erantzundako entitatea ezagutza-basean ez dela existitzen.

7 Gogoratu horregatik baztertzen direla 0.25eko atalase balioarekin.

Atalasea 1		Atalasea 0.3	
Spotlightek erantzundakoa	Ezagutza-basera bihurtutakoa	Spotlightek erantzundakoa	Ezagutza-basera bihurtutakoa
Davos	davos	World_Economic_Forum	world_economic_forum
Governmanet_of_Pakistan	nil	Pakistan	pakistan
Memphis,_Tennessee	memphis,_tennessee	Memphis_International_ Airport	memphis_international_ airport

10 Taula: Spotlightek eskaera berdinei emaitza ezberdina itzulitako kasuen adibideak.

Aipamena

Aipamena guretzako desanbiguatu beharreko aipamen anbiguoa da, eta hau, era berean, aipamen identifikatzaile edo *spotter* tresna baten ondoren lortzen dugu. *Spotlight*ek aipamena topatzeko ahalmena neurtzeko zenbait aukera erabili ditugu.

Batetik, berak defektuz erabiltzen duena; eta, bestetik, guk emandako aipamenak. Bigarren honetan desanbiguazioa egiteko 1. urratsa aurrezten diogu, hain zuzen, aipamen identifikatzaileak egin beharrekoa. Beste modu batean esanda, guk desanbiguatu beharreko aipamena zehaztu diogunez, *Spotlight*ek ez du aipamenik identifikatu beharko. *Spotlight*i pasatu dizkiogun aipamenak, era berean, bi modutan lortu ditugu. Datu-multzoak berak adierazten duen desanbiguatu beharreko aipamena[8] eta Matx prozesuaren ondoren originala barne duen aipamenik luzeena[9]. Matx prozesuan 1. heuristikoa erabili dugu. 11 Taulan ikus daiteke aipamen mota hauek.

Aipamen originala	Matx aipamena
AZ	Scottsdale AZ
Annapolis	Annapolis peace conference
Baltimore City	Baltimore City

11 Taula: Guk ezarritako aipamenen ezberdintasunen adibideak daude. Ikus daiteke batzuetan bat datozela biak.

8 Aipamen Originala bezala izendatuko duguna hemendik aurrera.

9 Matx aipamena moduan izendatuko duguna.

Horretarako *spotter* parametroaz baliatu gara; bertan zein aipamen identifikatzaile erabili dugun zehatz daiteke. *Spotlight*ek lehenetsitakoa erabiltzeko ez dugu inolako aldaketarik egin behar. Aldiz, guk zehaztutako aipamenak erabiltzeko *spotter* parametroak *SpotXmlParser* balioa eta *text* parametroari aurrez prestatutako *spotXml* formatudun testua ezarri dizkigu.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.6 *Spotlight* (Thr=1)	87.94	65.78	75.27	74.80	763	671	8
2	SP-0.6 Originala	88.32	72.65	79.72	82.25	839	741	20
3	SP-0.6 Matx	88.02	70.59	78.35	80.20	818	720	20

12 Taula: Spotlight 0.6 bertsioan spot ezberdinak hautatuta eginiko probak ageri dira, aurretik egindako threshold esperimentuen emaitzekin alderatuta. A/s zutabe berriak sistemak zenbat aipamen segundoko desanbiguatzeko gai den adierazten du.

12 Taulan 0.6 bertsioak izan dituen emaitzak adierazten dira. Bertan emaitzarik onenak aipamen originalak lortu dituela ikus daiteke. Aipamenak desanbiguatzeko abiaduran argi ikusten da aurretik azaldutakoa, urrats bat aurrezten dela, guk zehaztutako aipamenak erabiltzean bi bider azkarrago desanbiguatzea lortzen baita. Doitasunari erreparatuz gero, berdintasun handia dago; hiru esperimentuek 0.38 puntuko aldea baitute. Estaldura eta erantzuteko ahalmena neurtzen duen *coverage* ebaluazio neurriak analizatuta, guk zehaztutako aipamenak dira onenak, *Spotlight*en aipamenak lortutakoak 5 eta 7 puntura geratuz.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.7 *Spotlight* (Thr=1)	88.78	69.80	78.16	78.63	802	712	14
2	SP-0.7 Originala	85.54	69.61	76.76	81.37	830	710	33
3	SP-0.7 Matx	85.27	66.96	75.01	78.53	801	683	33

13 Taula: Aurreko esperimentuetako parametroak erabilita eta soilik spota topatzeko erabilitako algoritmoa aldatzean lorturiko emaitza.

13 Taulan ikus daiteke argi eta garbi *Spotlight*ek topatu beharreko aipamena zehaztean, askoz denbora gutxiago behar duela erantzuteko, aurreko bertsioan gertatzen den moduan. Honetaz gain, argi dago aipamen originala dela erantzuteko gaitasun handiena duena, *coverage* baliorik handiena baitu. *Spotlight*en *spotter*ak lortu dituen emaitzak erantzun egokiak kontuan harturik hobeak dira besteekin alderatuz gero. Matx izan da emaitza txarrenak lortutakoa, baina hurrengo atalean ikusiko dugunez testuinguruaren arabera honen emaitzak hobe daitezke. Aipatzekoa da, kasu honetan Matx aipamena erabiltzea okerragoa da, abiadura ez diren ebaluazio-neurri guztietan.

*Spotlight*en bertsio ezberdinak alderatuz gero, nabarmentzekoa da 0.7 bertsioak emaitza okerragoak dituela, beraz honek erabiltzen duen eredua okerragoa da. Hala ere, *Spotlight*ek erabiltzen duen aipamenak identifikatzeko tresna (*spotter*ra) hobetu egin dute, 40 eskaera gehiago erantzun izana honen erakusle da. 0.6 bertsioak 763 erantzuteko gai izan da hauetatik 671 ondo eta 0.7 bertsioak, aldiz, 802 erantzun ditu hauetatik 712 ondo.

Testuingurua

Testuinguruak *Spotlight*en zer nolako eragina duen jakiteko bi aldaera erabili ditugu; testuinguru osoa eta 50eko leihoa. Lehenengoa datu-multzoak eskaera bakoitzean ematen duen testuinguru guztia da. Bigarrena, ordea, Matx prozesuan sortzen da. Honek aurkitutako aipamen luzeenaren aurretik eta ondoren dauden hitzak hartzen ditu; beti ere 50eko maximoarekin, horregatik deritzogu 50eko leihoa.

Interesgarria egiten zaigu, testuingurua analizatzean, guk zehaztutako aipamena erabiltzea; Matxen analisi sakonagoa egin nahi baitugu.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.6 testuinguru osoa	87.94	65.78	75.27	74.80	763	671	8
2	SP-0.6 testuinguru osoa (Matx)	88.02	70.59	78.35	80.20	818	720	20
3	SP-0.6 testuinguru osoa (Orig.)	88.32	72.65	79.72	82.25	839	741	20
4	SP-0.6 testuingurua 50L	90.61	55.78	69.05	61.57	628	569	12
5	SP-0.6 testuingurua 50L (Matx)	89.37	72.55	80.09	81.18	828	740	32
6	SP-0.6 testuingurua 50L (Orig.)	89.23	73.92	80.86	82.84	845	754	32

14 Taula: Spotilght 0.6 bertsioan egindako esperimentuetan izandako emaitzak. Parentesien artean zein spot erabili den adierazten dugu, parentesirik gabekoak Spotlightek identifikatutakoa da, atalase balioa 1ekoa izanik.

14 Taula *Spotlight*en 0.6 bertsioa erabilita testuinguruarekin egindako esperimentu ezberdinen emaitzen taula da. Honetan doitasuna kontuan izanik, 50eko leihoko testuinguruarekin osoarekin baino emaitza hobeak lortzen dira. Hauen artean onena Spotlighten *spot*arekin egindakoa da, hala ere honek ez dauka erantzuteko gaitasun onik, *coverage* eta estaldura baliorik kaskarrenak lortu baitu (61.57 eta 55.78, hurrenez hurren). Gainerako ebaluazio neurriak begiratuz gero, 50eko leihoa erabilitakoak lortzen dute emaitzarik onenak. Argi ikusten da testuinguruaren tamainak duen eragina denboran, testuinguru txikiagoekin aipamen gehiago desanbiguatzeko gai baita.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.7 testuinguru osoa	88.78	69.80	78.16	78.63	802	712	14
2	SP-0.7 testuinguru osoa (Matx)	85.27	66.96	75.01	78.53	801	683	33
3	SP-0.7 testuinguru osoa (Orig.)	85.54	69.61	76.76	81.37	830	710	33
4	SP-0.7 testuingurua 50L	89.56	62.25	73.45	69.51	709	635	28
5	SP-0.7 testuingurua 50L (Matx)	88.48	68.53	77.24	77.45	790	699	52
6	SP-0.7 testuingurua 50L (Orig.)	87.30	69.41	77.33	79.51	811	708	52

15 Taula: Spotlightek jasotzen duen testuinguru ezberdinekin egindako taula da. M hizkiak spoteatzeko Matx erabilitako esperimentuak direla adierazten du.

15 Taulan 0.7 bertsioa erabilita testuinguruarekin egindako esperimentuen emaitzak adierazten dira. Bertan, agerikoa da 50eko leihoa erabilitako esperimentuetan lortu den doitasuna testuinguru osoarena baino hobea dela orokorrean. Gainera, denbora kontuan hartuz azkartasun nabaria dago 50eko leihoa erabilita, noski, zenbat eta testuinguru txikiagoa eman orduan prozesaketa denbora gutxiago behar baitu. Ondo erantzuteko gaitasuna neurtzen duen estaldura begiratuta, testuinguru osoa erabiltzen diren esperimentuetakoak dira onenak. Hala ere, aipamena zehaztuta 50eko leihoko testuingurua erabiltzean lortu ditugun emaitzak onak dira.

Erantzuteko gaitasunean, *coverage*an, erreparatuz gero, testuinguru osoa erabilita emaitza hobeak lortu dira. Bereziki aipagarria da originalaren emaitzak, noski desanbiguatu beharreko aipamen laburra izanik, anbiguoagoa denez, erantzuteko aukera gehiago baititu. Denbora aldetik, 50eko leihoa duen testuingurua azkarrago desanbiguatzeko gai da, bereziki aipamena zehaztu baldin badiogu.

Aipatzekoa da *Spotlight*en bi bertsioetan 50eko leihoa erabilita bera aipamenaz topatzeaz arduratzean huts egiten duela. Honen erakusgarri 0.6 bertsioan 140 eskaera gutxiago eta 0.7n ia 100 erantzuteko gai izatea da. Beraz, ondoriozta dezakegu testuinguru osoa dela onena *Spotlight*en *spotter*a erabiliko bada behintzat.

Eredua

Esperimentu honetan *Spotlight*ek eskaintzen dizkigun eredu ezberdinak alderatu ditugu. Orain arte egindako esperimentuetan *Spotlight*en 0.6ko eta 0.7 bertsioko eredua osoak erabili ditugu. Oraingoan bi eredu hauek 0.6 bertsioko eredu txikituarekin konparatu ditugu. Gogora dezagun eredu txikituak, edo light ereduak, testuingururik gabeko eredua dela. Esperimentu honi esker jakin ahal izango dugu zein neurriraino merezi duen eredu osoa erabiltzea edota nahikoa den testuinguru gabeko ereduarekin.

Horretarako bi parametro konbinaketa ezberdin erabilita probatu ditugu ereduak. Lehenengoan, *Spotlighte*n parametro optimoak eta Matx prozesuan lortutako aipamena eta testuingurua. Bigarrengoan, *Spotlight*en parametro optimoak, *Spotlighte*n aipamena eta testuinguru osoa erabili ditugu.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.7	88.48	68.53	77.24	77.45	790	699	52
2	SP-0.6 Osoa	89.37	72.55	80.09	81.18	828	740	32
3	SP-0.6 Txikitua	86.85	68.63	76.67	79.02	806	700	38

16 Taula: Spotlighten bertsio ezberdinek edukitako emaitzak. Parametro optimoak eta Matx prozesuan sortako testuingurua eta spota erabilita.

16 Taula lehenengo konbinaketa erabilita lortutako emaitzak ditu. Bertan adierazten diren emaitzetan argi ikusten da 0.6 bertsioko eredu osoak lortu dituela emaitzarik onenak, abiaduran izan ezik. Doitasuna analizatuz gero, onena, esan bezala, 0.6ko eredu osoa da, ondoren puntu bat baino gutxiagora 0.7ko eredua eta, azkenik, eredu txikitua bi puntu eta erdira. Estaldura erreparatuta, 0.6ko eredu osoak emaitzarik onena du; ondoren eredu txikitua eta 0.7ko eredua onenetik lau puntura; bi azken hauen aldea oso txikia da. *Coverage* ebaluazio-neurrian, aurrekoek izandako posizio berdinak errepikatzen dira, bakoitzaren arteko aldea bi puntukoa izanik.

Abiadura ikertuta, 0.7 bertsioak emaitzarik onena du, bertsio honek jasandako hobekuntzen erakusle; ondoren 0.6ko eredu murriztua eta, azkenik, eredu osoa. Hala ere, azken bi hauen arteko aldea ez da horren nabaria

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.7	88.78	69.80	78.16	78.63	802	712	14
2	SP-0.6 Osoa	87.94	65.78	75.27	74.80	763	671	8
3	SP-0.6 Txikitua	86.94	64.61	74.13	74.31	758	659	10

17 Taula: Spotlighten bertsio ezberdinetan parametro optimoekin, spota Spotlightek lortua eta testuinguru osoarekin egindako esperimentuen emaitzen taula.

Datu-multzoa bere gordintasunean Spotlighti pasatuz gero eta honek *spot*a lortzeaz arduratzean lortzen diren emaitzak ikus daitezke 17 Taulan. Bertan, 0.7 bertsio berrienak ditu emaitzarik onenak ebaluazio-neurri guztietan. Bigarren onena 0.6 bertsioko eredu osoa eta, azkenik, eredu txikitua; hau ebaluazio-neurri guztietan gertatzen da, abiaduran izan ezik.

Gainera, memoriaren erabilera azaltzen dugu ondorengo taulan, baita zerbitzaria martxan jartzeko beharrezkoa duten denborak.

	Tresna	Denbora martxan jarri	MEM
1	Spotlight 0.7	~ 2 minutu	~ 9 Gb
2	Spotlight 0.6 Osoa	~ 4 minutu	~ 13 Gb
3	Spotlight 0.6 Txikitua	~ 3 minutu	~ 8.5 Gb

18 Taula: Bertan, Spotlight Statisticalek martxan jartzeko beharrezko denbora eta beharrezkoa duen memoria adierazten da.

18 Taula ikusita argi dago 0.7 bertsiora egindako saltoaren eraginez softwarearen eraginkortasuna denbora eta memoria aldetik nabariak direla. Hala ere, kontuan izan behar dugu 0.7 bertsioan erabilitako eredua pixka bat murriztua dagoela.

5.1.5. Denbora aurrezten

5.1.1 Atalean azaldu bezala desanbiguatzeko erabilitako lehenengo *script*ean argumentuak eskuz aldatu behar genituen; gainera ebaluaziorako eman beharreko urratsak independenteak ziren, hots, *script* ezberdinek erabiltzen genituen bata bestearen atzean. Honek esperimentu eta esperimentu artean denbora galtzea dakar, parametroak eskuz aldatu eta ebaluazio *script*ak banan-banan martxan jarri behar baikenituen. Honetaz gain, egindako esperimentuak txukun sailkatzeko ez zitzaigun erraza egiten, kontsolatik scripta exekutatzean ez baitzuen adierazten erabilitako parametroak.

Horregatik desanbiguatzeko scripta moldatu dugu beharrezko baliabide eta adierazpide guztiak argumentu bidez jaso ditzan. Horien artean *Spotlight*en parametroak, sarrera fitxategia eta prozesuan sortuko diren fitxategien izena daude. Bertsio berri honi esker, probak egiterako garaian eskuz egindako lana aurrezten dugu, gainera eskuz txertatzeak ekar lituzkeen akatsak gutxitzen ditugu parametro bidez jasotzen baititu. Honetaz gain, *Spotlight*eko hizkuntza ezberdinak jasateko prestatu dugu, hurrengo esperimentuetan erabili ahal izateko. Bestalde, IXA ikerketa taldean, nitaz gain beste batzuek ere *Spotlight* estatistikoa erabiltzen hasi ziren. *Spotlight* zerbitzariak deiak jasotzeko erabiltzen duen portuan arazorik ez edukitzeko, erabiltzaile bakoitzak portu ezberdina erabili behar du. Horregatik *script* berri honen parametro gisa zerbitzaria zein portutan entzuten dagoen adierazi diogu ere.

Bestetik, ebaluazioa egiteko eman beharreko urratsak batu ditugu, horrela *script* bakarra exekutatzea nahikoa dugu ebaluazioa egiteko. Honek, era berean, argumentu moduan ebaluatzeko erabiltzen diren urre-patroia eta ezagutza-base bihurketarako erabiltzen dugun katalogoa jasotzen ditu. Behin bi *script* berriak izanik, interesgarria egiten zitzaigun desanbiguazioa eta ebaluazioa dei bakar batean egitea. 24 Irudian ikus daiteke honek jasotzen dituen parametroak.

```
$ sh 00-desanbiguatu_ebaluatu.sh IZENA SARRERA HIZKUNTZA SCRIPTa
KB_bihurketarako_KATALOGOA URRE_PATROIA PORTUA AIPAMENA COREF.RESO.
SUPPORT CONFIDENCE
```

24 Irudia: Kontsolatik sortutako scripta exekutatzeko behar dituen parametroak ikusten diren adibidea. Bertan irteera fitxategiaren izena, sarrera fitxategia, hizkuntza, desanbiguatzeko scripta, ezagutza-base bihurketarako katalogoa, urre-patroia, portua, zein aipamen eta Spotlightek jasotzen dituen parametroak daude.

5.1.6. Errore analisiak

Behin zenbait emaitza bagenituela, hauen analisia egin genuen akatsak topatzeko asmoz. Akatsak topatzeko 19 Taulan adierazitako erantzunen taula egin genuen. Taula hauen bidez *Spotlight*ek itzulitako emaitza, ezagutza-baseko bihurketa eta urrezko patroiko erantzun zuzena adierazten dugu. Modu honetan egin ditugun urratsak grafikoki adieraziko ditugularik.

19 Taulan ikusten denez, batzuetan desanbiguatutako entitatea ez da ezagutza-basean aurkitzen eta NIL bihurtzen da; beste batzuetan bihurketak gertatzen dira, baina ez da beti emaitza hona topatzen eta azkenik, badaude beste kasu batzuk zeinetan entitatea ez den aldatu ezagutza-base batetik bestera, hauetan ere asmatzen ez diren kasuak daude, hasieran desanbiguatutako terminoa okerra baita.

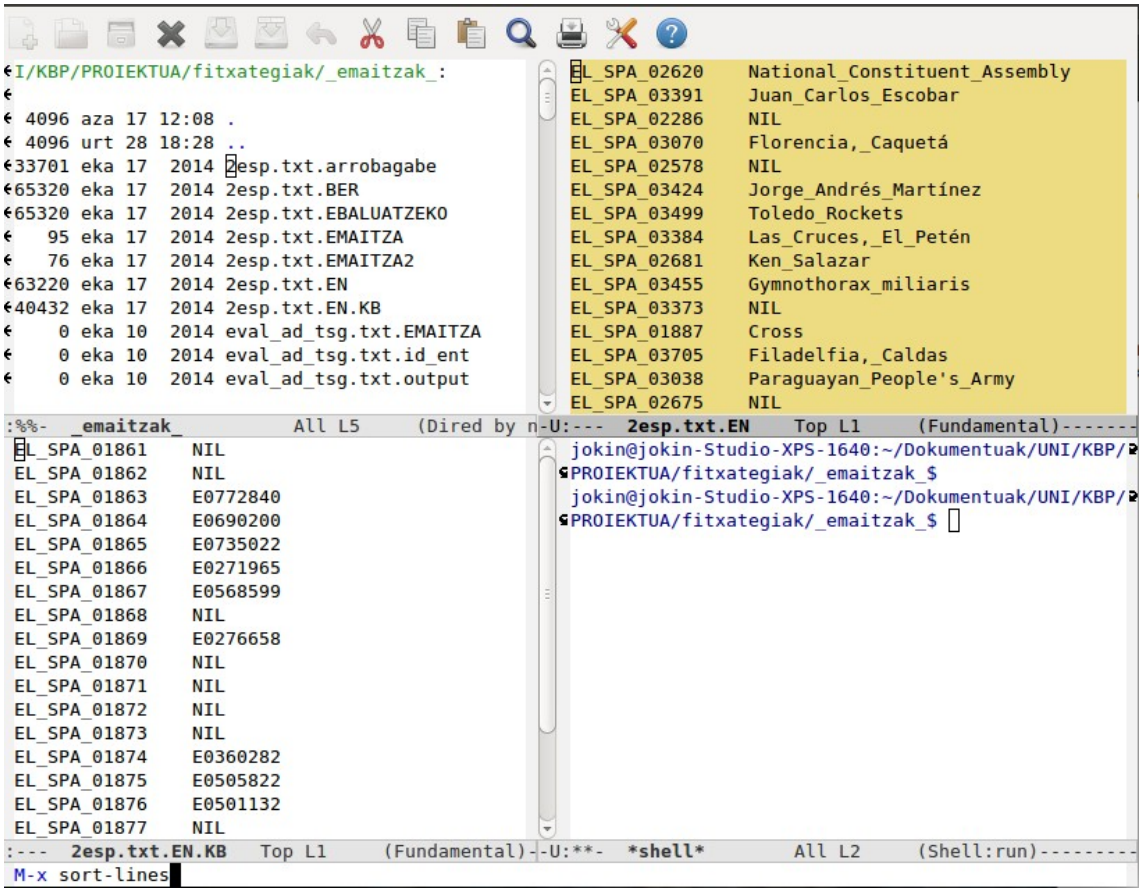
Spotlighten erantzuna	KB entitatea	Urrezko patroia
warsaw_ghetto	nil	warsaw
st._louis	st._lois,_missuri	st._lois,_missuri
detroit	detroit_city	detroit,_michigan
korean_war	korean_war	south_korea
arizona	arizona	arizona

19 Taula: Erantzunen taula. Modu honetan argi ikusten ditugu urratsez urrats egindako aldaketak, horrela akatsak identifikatu ahal izateko. "Spotlighten erantzuna" zutabean Spotlightek erantzundako izen-entitatea daude, "KB entitatea" aurreko izen-entitatea TACeko ezagutza-basera bihurtutakoan lorturiko izen-entitatea da eta, azkenik, "urrezko patroia" zutabean eskaera horri dagokion izen-entitate zuzena adierazten da.

Baina desanbiguazio script*e*an egindako urratsak ez daude aurreko tauletan adierazita. Hots, 5.1.1 atalean azaldutako *script*ean egindako urratsen errorerik ezin ditugu aurreko taulatik lortu. Beraz, honi aurre egiteko scriptean bertan kontrol aginduak txertatu genituen. Horrela eskuzko eginiko arazketari esker beste zenbait errore txikiren analisia egin dugu. Erroreen artean *spotXML* formatudun fitxategian txertatzen genuen kodeketa arazoa konpondu genuen. Hau konpontzeko *Perl*ek *XML* egiturak erabiltzeko moduluaz baliatu ginen.

Gainera, erabilitako datu-multzoak aurrez ez ditugunez prozesatu, zenbait datu galera izan genituen. Batetik, datu-multzo hauek duten kodetze formatuagatik eta, bestetik, testuinguruetan azaltzen diren karaktere bereziengatik. *Emacs* testu editoreari esker, galera horiek begi bistaz ikusi ahal izan genituen eta ondoren, kodeketa arazoa konpontzen duen *script*a erabiliz eta 5.1.1 atalekoari partxe batzuk jarriz konpondu genituen.

Nabarmentzekoa da, *Emacs*ek eskaintzen duen erabilgarritasuna. Honi esker zenbait leihoetatik egiten genuen nabigazioa oso azkarra gertatu zitzaigun. Bertan komando lerroa erabiltzeko aukera dago, datuak ordenatzeko ahalmena eta abar. Hots, beharrezkoak genituen tresna guztiak esku-eskura.



25 Irudia: Bertan Emacs testu editoreak zenbait leiho irekita ditu, horien artean komando lerroa. Gainera horiz hautatuta dagoen fitxategia lerroka ordenatzeko agindua idatzita dago.

5.1.7. Ebaluazioaren emaitzak

Behin garapen fasea amaituta, ebaluaziorako erabili ditugun datu-multzoetan testatu ditugu emaitzak, hau da, TAC 2011n eta AIDAn. Kontuan hartu beharra dago lan honetatik lortutako emaitzen ondorioak hurrengo esperimentuetan erabakiak hartzeko erabili ditugula.

Testatzeko erabili ditugun parametroen konbinaketa ondorengoa izan da: *Spotlight*eko iragazki parametro optimoak, hots, *confidence* 0, *support* 0 eta *coreference-resolution 0*, eta Matx prozesuan lortutako baliabideak, hau da, Matx aipamena eta 50eko leihoko testuingurua.

Ondorengo taulan ebaluaziorako erabilitako datu-multzoetan lortutako emaitzak artearen egoerarekin alderatuko dugu.

	Tresna	Datu-multzoa	R
1	Spotlight	TAC 2011	62.90
	Ixanpei	TAC 2011	81
2	Spotlight	AIDA	73.80
	WLDA-full	AIDA	84.89

20 Taula: Spotligth Statistical 0.7 bertsioak ebaluazio datu-multzoetan izan dituen emaitzak adierazten dituen taula.

Emaitzak adierazten dituen 20 Taularen estaldura ebaluazio-neurria erreparatuta, TAC 2011an gure sistema 19 puntu beherago aurkitzen dela Ixanpei sistematik (A. Barrena, 2013)[10] eta AIDAn, ordea, gertuago gaude, 11 puntura baikaude. WLDA-full (N. Houlsby & M. Ciaramita, 2014) *Wikipedia*ko eta *Freebase*ko esteken ezagutza eta inferentzia probabilistikoa erabiltzen duena.

5.2. Gaztelaniazko esperimentuak

Gaztelaniaz egindako esperimentuak bi tresnekin, *Spotlight*ekin eta UKBrekin, egin ditugu. Soilik datu-multzo bakarra erabili dugu garapenerako, parametroen azterketarako, eta ebaluaziorako; gaztelaniazko TAC 2012a. Garapenean aurrera eraman dugun metodologia parametroak banan-banan aztertzearena izan da. Esperimentuak tresnen arabera egin ditugu, lehenago *Spotlight*ekin aritu gara eta ondoren UKBrekin.

10 (X. Han & L. Sun, 2012) sistemaren berrinplementazioa.

Esan beharra dago gaztelerazko TAC 2012 datu-multzoa *cross-lingual* motatakoa dela. Honen eraginez, datu-multzoan ingelesezko zein gaztelaniazko dokumentuak aurkituko ditugu; zehazki 1991 gaztelaniazko eta 75 ingelesezko dokumentuak dira, lehenengo motatakoetan 868 EzNIL eta 1123 NIL eta ingelesezkoetan 55 EzNIL eta 20 NIL daude.

*Spotlight*en esperimentuetan 5.1 atalean egindako esperimentuetatik lortutako ondorioak erabili ditugu, denbora eraginkorki erabiltzeko asmoz. Beraz, iragazki parametroekin eta atalasearekin ez ditugu inolako esperimenturik egingo. Bai, ordea, aipamenekin eta ereduekin. *Spotlight*en bertsio ezberdinen ereduak ere aztertu ditugu, gaztelaniazkoak zein ingelesezkoak.

UKB tresna aztertzeko, honek jasotzen duen grafoaren aldaera ezberdinak erabili ditugu. Gainera, *Spotlight*ekin egin dugun bezala, ingelesezko ezagutza-basea erabili dugu beste zenbait grafo sortzeko. Ondorengo ataletan, esperimentuak egikaritzeko egindako lana azaldu dugu.

5.2.1. Bi tresna testuinguru berdina

Gaztelaniazko esperimentuak egiteko bi tresna erabili ditugunez, eta bien arteko emaitzak alderatu ahal izateko, tresnek jasotzen duten sarrerako datuak berdinak izan behar dira. Horrela UKBk sarreratzat duen entitate zerrenda *Spotligth*teko sarrera izatea lortu behar dugu. Baina *Spotlight*en ezaugarriak ahaztu gabe, hala nola, azken honek, UKBk ez bezala, hizki larri eta xeheen arteko ezberdintzea egiten duela. UKBren sarrera hizki xeheak izan behar ditu soilik.

Horren ondorioz, 5.2.2 atalean azaltzen den Matx tresnan zenbait aldaketa egin ditugu. Modu honetan, Matxek itzultzen dizkigun entitate guztiek hasierako formatua mantendu beharrean, soilik desanbiguatu beharreko entitatea izango ditu (edukitzekotan) hizki larriak. Horrela egitearen arrazoia gainerako aipamenen forma *Spotlight*entzat oso-oso eragin txikia duelako da.

5.2.2. Matxen erabilera

Ingelesezko emaitzetatik Matxek itzultzen digun aipamenak eta hauen inguruan dagoen 50 hitzeko testuingurua erabilita emaitza onak lortzen direla ondorioztatu dugu. TAC 2012ko datu-multzotik UKBk jasoko duen sarrera lortu dugu, baina noski Matxek prozesatu ondoren.

Alde batetik, Matxek desanbiguatu beharreko entitatea barne duen aipamenik luzeena itzultzen du. Bestetik, honen inguruan dagoen 50eko leihoa duen testuingurua. Behin hau izanda, UKBren sarreraren formatua kontuan izanik honen sarrera prestatu dugu esperimentuak egin ahal izateko.

5.2.3. Beharrezko aldaketak mapaketa

Aurreko esperimentuan bezala, honetan ere mapaketa, entitate bihurketa, egin dugu bertsio ezberdinen arteko arazoa konpontzeko. Baina honetaz gain, hizkuntzaren koska dugu. Nahiz eta gaztelaniazko TAC 2012 datu-multzoa izan, bere urrezko patroia ingelesezko *Wikipedia*ko 2008ko bertsioko entitateak dira, TACeko datu-multzo guztiak bezala. Beraz, gaztelaniazko entitateak, ebaluatu aurretik, ingelesezko entitate bihurtu behar ditugu.

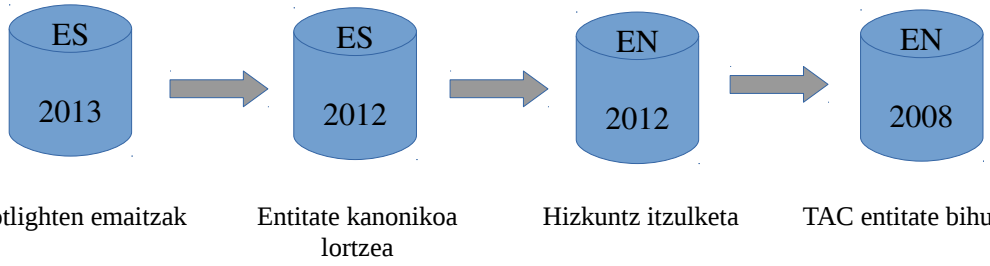
Hasteko, argi dago *Wikipedia*k eskaintzen dituen hizkuntzen arteko loturak (*interlingual links*) erabil ditzakegula entitateak gaztelaniatik ingelesera itzultzeko. Dagoeneko IXAk badu lotura hauek tratatuta; honako hirukoteaz osatuta: artikulu identifikadorea eta ingelesezko eta gaztelaniazko entitateak.

ID	Ingelesezko entitatea	Gaztelaniazko entitatea
839582	Goverment_ministers_of_Portugal	Ministros_de_Portugal
4775117	Sugar_Creek,_Wisconsin	Sugar_Creek_(Wisconsin)
3108177	Brito_(Guimarães)	Brito

21 Taula: Wikipediako ingelesezko eta gaztelaniazko artikuluen izenburuen arteko erlazioak ageri dira, izenburu hauek interlingual linken bidez erlazionatzen dituztenak.

Baliabide hau, ordea, orain arte erabili ez dugun *Wikipedia* bertsioarekin osatu zenez, erabili ahal izateko beste urrats bat gehitu behar diogu. *Spotlight*ek itzulitako gaztelaniazko izen-entitateak erabili nahi dugun baliabidearen *Wikipedia* bertsioko entitate kanonikoak bihurtuko ditugu.

Hau egin ostean, gaztelaniazko izen-entitateak ingelesezko izen-entitateak bihurtu ahal izango ditugu. Ondoren, aurreko esperimentuan eginiko mapaketa urratsak jarraituta emaitzak lortu ahal izateko. 26 Irudian eginiko prozesua modu grafikoan adierazten dugu.



26 Irudia: Spotlighteko emaitzak ebaluatzeko egin beharreko mapaketa. Bertan, Spotlighten itzulitako emaitzak entitate kanoniko bilakatzen ditugu, ondoren gaztelaniatik ingeleserako itzulketa egiteko. Azkenik, TACen ezagutza-baseko entitatea bilakatzen dugu.

5.2.4. *Spotlight*ekin desanbiguatzen

Tresna honekin egin diren esperimentuak ingelesezko esperimentuetako *Spotlight*en parametro onenekin egin dira, hots, *support* eta *confidence* zero, *coreference resolution false* eta *threshold*a bat. Behin parametro onenak zehaztuta, gainerako parametroak ikertuko ditu; testuingurua izan ezik, UKBrekin bat etor daitezen[11]. Ikertu dugun parametroa beraz, aipamena[12] da. Horretarako *Spotlight*ek eskaintzen duen gaztelaniazko eredua erabili dugu. Atal honetako esperimentuak *Spotlight*eko bi bertsioekin egin dira.

11 5.2.1 Atalean azaldutakoagatik.

12 Originala, Matxek lortutakoa edo *Spotlight*ena.

Baina, horretaz gain, ingelesezko ereduak duen eragina neurtu dugu. Arrazoia ezagutza-base honek jakituria gehiago izatea da. *Wikipedia*ko ezagutza basea oso handia da, bereziki ingelesezko bertsioa aurrekarietan esan bezala. Gainera, kontuan hartu behar dugu ingelesa eta gaztelania hizkuntzak munduan barrena oso hedatuak daudela.

Aurrekoa kontuan izanda, *Spotlight*ek eskaintzen dituen ingelesezko ereduak erabiliz TAC 2012ko datu-multzoa desanbiguatu dugu. Honen bidez ingelesezko ereduaren eragina ikuskatu nahi dugu.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.6 ES eredua	69.54	50.70	58.65	72.91	673	468	22
2	SP-0.6 ES eredua (Matx)	70.13	51.14	59.15	72.91	673	472	54
3	SP-0.6 ES eredua (Orig.)	59.97	42.04	49.43	70.10	647	388	54
4	SP-0.6 EN eredua	75.24	60.24	66.91	80.07	739	556	12
5	SP-0.6 EN eredua (Matx)	74.79	57.20	64.83	76.49	706	528	32
6	SP-0.6 EN eredua (Orig.)	66.48	51.79	58.22	77.90	719	478	32

22 Taula: Spotlighteko 0.6 bertsioan izandako emaitzak. Datu-multzo honetan 868 EzNIL entitate daude.

22 Taulan *Spotlight*eko 0.6 bertsioa erabilita lortutako emaitzak azaltzen dira, ingelesezko zein gaztelaniazko ereduak baliatuta. Bertako emaitzak kontuan izanik, ingelesezko eredua erabiltzen duten esperimentuek emaitzarik onenak lortu dituzte; aipamen originala erabiltzen den kasuetan izan ezik. Honen zergatia erabiltzen diren aipamen originalen anbiguotasun handia dutela izan daiteke.

Doitasun ebaluazio-neurria ikuskatzean 4 eta 5 esperimentuak emaitzarik onenak dituzte, ondoren 1 eta 2 esperimentuak, hots parametro berdinak baina eredu ezberdina erabilita. Estalduraren ikuspuntutik, 4 esperimentuak 60.24ko balioarekin lortu du onena, gainerakoetatik 4-18 puntuko ezberdintasunarekin.

Erantzuteko gaitasuna neurtzen duen *coverage* ebaluazio-neurriari erreparatuz gero, berriro ere emaitzarik onena 4 esperimentuak lortzen du 80.07ko balioarekin. Gainerako esperimentuak 2-10 puntuko ezberdintasuna dute onenarekiko, gertuenak ingelesezko eredua erabiltzen duten esperimentuak izanik. Aipatzekoa da estaldura eta *coverage* ebaluazio-neurrietan ingelesezko ereduarekin egindako esperimentu guztiek lortzen dituztela gaztelaniazko ereduarekin baino emaitza hobeak.

Denbora kontuan hartuz, azkarrenak gaztelaniazko eredua erabili dutenak dira, aipamena aurrez emandakoak hain zuzen. Arrazoi nagusia bi ereduen tamainaren aldea da, 24 Taulan ikus daiteke hauen memoriaren erabilera 6Gbeko ezberdintasuna dela.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	SP-0.7 ES eredua	76.12	57.31	65.39	75.30	695	529	37
2	SP-0.7 ES eredua (Matx)	75.62	56.45	64.64	74.65	689	521	88
3	SP-0.7 ES eredua (Orig.)	66.42	48.00	55.72	72.26	667	443	88
4	SP-0.7 EN eredua	75.56	61.97	68.10	82.02	757	572	28
5	SP-0.7 EN eredua (Matx)	75.77	58.29	65.89	76.92	710	538	52
6	SP-0.7 EN eredua (Orig.)	66.08	52.76	58.67	79.85	737	487	52

23 Taula: Spotlighteko 0.7 bertsioan izandako emaitzak.

23 Taulan *Spotlight*en 0.7 bertsioko gaztelaniazko eta ingelesezko ereduak erabilita lortutako emaitzak laburbiltzen ditugu. Taulako balio onenak[13] ikusita, 0.6 bertsioko emaitza antzekoak direla ohar gaitezke, 4 esperimentuarekin lortzen direla emaitzarik onenak ia-ia.

13 berdez koloreztatutako gelaxkak.

Ebaluazio-neurriak banan-banan aztertuta; doitasun neurriarekin hasita, onena 1 esperimentua da baina 2, 4 eta 5 esperimentuak puntu erdira daude. Estalduran onena 4 esperimentua da, hurbilen dauden esperimentuak 5, 1 eta 2 dira hurrenez hurren. Gainera eredu ezberdinak erabilitako esperimentu baliokideetan ingelesezko ereduak emaitza hobeak lortzen ditu.

Coverage ebaluazio-neurria analizatuta, berriz ere 4 esperimentuak emaitzarik onena du, ingelesezko eredua erabiltzen dituzten esperimentuak gaztelaniazkoak baino emaitza hobeak lortu ditu. Aipatzekoa da ingelesezko eredua eta Matx aipamena erabilita dagoen beherakada, baina noski, Matxek lortzen duen izen-entitatea gaztelaniaz dagoenez ulertzekoa da ingelesezko ereduan emaitzarik onegiak ez izatea. Originala erabilita, ordea, izen-entitatea motzagoa izanik, anbiguoagoa da eta honek ingelesezko ezagutza basean egotea ekar lezake. Hori da ondorioztatu duguna gaztelaniazko ereduan, Matx aipamena erabiltzean originala baino emaitza hobea duelako.

Denbora kontuan hartuta, onenak gaztelaniazko eredua eta guk zehaztutako aipamena erabiltzen duten esperimentuak dira. Honen arrazoia gaztelaniazko eredua ingelesezkoa baino txikiagoa izatea eta desanbiguatze prozesuan aipamena lortzearen urratsa aurrezten dugula da.

	Tresna	Denbora martxan jarri	Memoria
1	SP-0.6 ES eredua	~ 2 minutu	~ 7 GB
2	SP-0.6 EN eredua	~ 4 minutu	~ 13 GB
3	SP-0.7 ES eredua	~ 45 segundo	~ 5 GB
4	SP-0.7 EN eredua	~ 2 minutu	~ 9 GB

24 Taula: Erabilitako ereduen memoria eta zerbitzaria martxan jartzeko beharrezko denbora adierazten duen taula.

24 Taulan erabilitako ereduen zerbitzarien martxan jartzeko denbora eta memoria kostuak ikus daitezke. Bertan, *Spotlight*eko bertsio aldaketan izandako hobekuntzak ikus daitezke; batetik, ereduen tamaina murriztea eta bestetik, zerbitzaria martxan jartzeko denboraren murrizketa.

Azkarrena eta memoria kostu txikiena duena 3 kasua da eta motelena eta memoria kostu handiena duena 2 kasua. 22 eta 23 Tauletako emaitzak kontuan izanik eraginkorrenak direnak 0.7 bertsioko ereduak dira; hauekin 0.6 bertsioko ereduarekin lortutako emaitza guztiak hobetzen baitira.

5.2.5. UKBrekin desanbiguatzen

UKBrekin egin ditugun lehendabiziko esperimentuak izanik, modu automatikoan exekutatzeko scripta prestatu dugu. Script honek lana eta denbora aurrezten du, ingelesezko esperimentuetan bezala, beharrezkoak dituen argumentuak parametro bidez jasotzen baititu. Gainera, ebaluatzeko egin beharreko urratsak ere barnebiltzen ditu. Beraz, desanbiguatzeko eta ebaluatzeko scripta egin dugu, gaztelerazko edota ingelesezko baliabideak erabili ahal izango ditugularik.

Lehenengo esperimentu hau egiteko, 4.2.2 atalean azaldutako grafoak eta hiztegia eta 5.2.2. atalean lortutako sarrera fitxategiak erabili ditugu.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	UKB-AD grafoa	63.19	49.86	55.72	78.87	460	728	0.07
2	UKB-AU grafoa	64.84	51.35	57.32	79.20	474	731	0.03
3	UKB-AB grafoa	64.34	48.86	55.54	75.95	451	701	0.62

25 Taula: Gaztelaniazko grafo ezberdinetan egindako esperimentuetako emaitzak azaltzen dira.

25 Taulan sortutako grafoak erabilita lortutako emaitzak adierazten dira. Bertan ikus daiteke AU grafo ez zuzenduak dituela emaitzarik onenak ebaluazio-neurri guztietan, abiaduran izan ezik. Abiaduran azkarrena AB grafo ez-zuzendu murriztua nabarmen azkarragoa izan da.

Aipatzekoa da AD grafo zuzenduak daukan doitasuna hiruren artean txikiena izatea; nahiz eta informazio gehien duen grafoa izan. Estaldurari erreparatuz gero, hirurak 2.5 puntuko aldea dute gehienez. AU grafo ez zuzendua emaitzarik onenak jaso dituena izan da, ondoren, informazio gehien duen AD grafo zuzenduak eta azkenik, AB grafo murriztua.

Coverage ebaluazio-neurria onena duena ere AU grafo ez zuzendua da, AD grafo zuzendua 0.47 puntura izanik. AB grafo murriztua onenenetik 3.25 puntura gelditu da, informazio gutxiago edukitzearen eraginez.

Doitasuna, estaldura eta *Coverage* neurriak elkarrekin hartuta AD eta AB grafoen ezaugarrien eragina ikus daiteke. Grafo hauen ezberdintasuna nodoen arteko noranzko bakarreko loturak kentzea da, eta honen eragina argia da. Batetik, AB grafoaren Doitasuna ADrena baino handiagoa da, horren erantzule loturak dira; AB grafoan soilik zentzu bikoitza duten loturak dituenez, informazio zehatzagoa du. AD grafoak, ordea, lotura guztiak edukitzean zehaztasuna galtzen du, noranzko bakarreko loturen eraginez anbiguotasun handiagoa txertatzen delako. Baina, bestetik, AD grafoa erantzuteko ahalmen handiagoa dauka lotura gehiago edukitzeagatik. Horregatik nahiz eta AD grafo zuzenduak AB grafoa baino 27 erantzun gehiago eman, soilik 9 gehiago ondo emateko gai izan da.

5.2.6. Ezagutza-basea handitzen

*Spotlight*eko esperimentuetan bezala, UKBn ere ingelesezko ezagutza-basea erabili dugu. Ingelesezko ezagutza-basea handiagoa eta zabalagoa izanik emaitza hobeak lortzeko asmoz. Esan bezala ingelesezko ezagutza-basea oso handia denez, soilik AB grafo murriztua erabili dugu.

Grafo horretaz gain beste grafo bat ere sortu dugu esperimentu honetarako, ingelesezko AB eta gaztelerazko AU grafoak elkarren artean konbinatzean lortutakoa. Bi grafo hauek lotzeko aurretik erabili dugun *Wikipedia*k eskaintzen digun *interlingual* loturak erabili ditugu.

Baina arazotxo batekin aurkitzen gara; ingelesezko *Wikipedia*ko eta gaztelerazko *Wikipedia*ko izenburu berdinek ez diote erreferentzia egiten izen-entitate berberari. Adibidez, ingelesezko bertsioan Durango izenburua duen artikulua Mexikoko estatua da eta gaztelerazko bertsioan, ordea, Durango udalerria. Horrela zen behintzat esperimentua aurrera eraman genuen momentuan, gaur egungo *Wikipedia*n ez da hau gertatzen, hala ere guk ditugun *Wikipedia*ko erauzketetan bai. Arazoari aurre egiteko bi hizkuntzetako entitateak ezberdindu ditugu “ES@” eta “EN@” aurrizkien bidez.

Horrenbestez, hizkuntz ezberdinen entitateak aurrizkiekin jarri ditugu eta grafo ezberdin hauek konbinatzeko *interlingual* estekez baliatu gara. Gainera, desanbiguatzeko erabili dugun hiztegia ere eguneratu dugu hautagai posibleei “ES@” aurrizkia jarriz.

	Tresna	P	R	f1	C	SYS	OK	A/s
1	UKB-AU(ES) grafoa	64.84	51.35	57.32	79.20	731	474	0.03
2	UKB-AB(EN) grafoa	69.81	54.60	61.28	78.22	722	504	0.05
3	UKB-AB(EN) u AU(ES) grafoa	65.21	51.57	57.59	79.09	730	476	0.02

26 Taula: Ezagutza-basea ingelesezko ezagutzarekin aberastean lortutako emaitzak adierazten dira, aurreko esperimentuan edukitako emaitza onenekin alderatuta.

26 Taulan grafo berri hauek erabilita lortutako emaitzak eta aurreko onenak daude. Ezagutza-base handiagoa erabiltzean lortutako emaitzak aurreko esperimentuan, gaztelaniazko grafoa erabiltzen genuen esperimentuan, lortutakoak baino hobeak izan dira. Taulan doitasun altuena duena ingelesezko AB grafoa izan da, konbinaketa grafoa baino 4.6 puntu gehiago lortuta. Estalduran ere ingelesezko grafoa onena da, besteetatik 3 puntura.

Coverage ebaluazio-neurrian, ordea, gaztelaniazko grafoak lortu du emaitzarik onena; hala ere konbinaketa grafoa eskaera bakarrera dago eta ingelesezko grafoa 9 eskaeretara, *SYS* zutabean ikus daitekeen bezala. Abiadura aldetik ingelesezko grafoa azkarrena da, 0.05 aipamen segundoko desanbiguatuta; motelena konbinaketa grafoa, soilik 0.02 aipamen segundoko.

Beraz, lortutako emaitzetatik ondoriozta dezakegu ingelesezko ezagutza-basea dela onena. Aipatzekoa da konbinaketa grafoak lortutako emaitzak ez direla izan esperotakoak bezain onak.

	Tresna	Memoria
1	UKB-AU(ES) grafoa	2.1 GB
2	UKB-AB(EN) grafoa	3.3 GB
3	UKB-AB(EN) u AU(ES) grafoa	3.3 GB

27 Taula: Grafo ezberdinen memoria kostua adierazten da.

Esperimentuetan erabilitako grafoek duten memoria kostua 27 Taulan ikus daiteke, txikiena gaztelaniazko grafokoa da 2.1G eta beste biek kostu berdina dute, 3.3G. *Spotlight*eko ereduen memoria kostua adierazten duen 24 Taularekin alderatuz gero, UKBren memoria kostua txikiagoa da baina exekuzioan askoz motelagoa.

5.2.7. Konparaketa

Atal honetan *Spotlight*ekin eta UKBrekin lortutako emaitzarik onenak artearen egoerarekin konparatu ditugu.

Aipatzekoa da konparaketa honetarako erabilitako ebaluazio-neurria *Bcubed+f1* (*B+f1*) izan dela. Horrela egitearen arrazoia artearen egoerako sistemen emaitzak ebaluazio-neurri honetan soilik aurkitzea izan da, ebaluazio-neurri hau izan baita TAC 2012koa.

	Tresna	*B+f1*
1	SP-0.7 EN eredua	55.36
2	UKB-AB(EN) grafoa	50.00
3	basistech	54.4
4	lcc2012	62.6

28 Taula: Gaztelaniazko esperimentuen konparaketa artearen egoerarekin adierazten da

Artearen egoera adierazten duten sistemak *basistech* (J. Clarke et al.) eta *lcc2012* (S. Monahan & D. Carpenter, 2012) izan dira. Lehenengoa beste modelo batzuen irteeretan oinarritzen da eta, bigarrengoa zenbait ezagutza-baseetako dokumentuetan.

28 Taulan gure sistemak artearen egoerarekin alderatzen dira. Bertan ikus daiteke UKB sistema dela okerrena. *Spotlight* erabilita *basistech* sistema gainditzen dugu, ia-ia puntu bat aterata. Hala ere *lcc2012* sistematik 7 eta 12 puntura gaude.

5.3. Euskarazko esperimentuak

Euskarazko esperimentuetan soilik UKB tresna erabili dugu; euskarak dituen berezitasun lexiko morfologikoen eraginez *Spotlight* erabilgarri jartzea lan handia baita, euskarazko eredua sortzeko *Wikipedia* osoa lematizatu beharko genuke, besteak beste. Esperimentuak egiteko (I. Fernandez, 2012) tesian sortutako datu-multzoak erabili dira; garapen multzoa garapenerako eta test multzoa ebaluaziorako. Esperimentuen garapenean grafoek desanbiguatzeko duten gaitasuna ikuskatzea, lematizazioaren eragina neurtzea, eta desanbiguatu nahi dugun aipamena topatzeko ahalmena jorratu dira.

Euskarazko esperimentuak egitean sortutako ataza nagusien azalpena ondorengo ataletan adierazi dugu.

5.3.1. I. Fernandezen datu-multzoarekin lanean

Behin (I. Fernandez, 2012) tesi lanerako sortutako datu-multzoa lortuta, hau analizatu dugu. Analisiaren ondoren, datu-multzoa formatu egokian jarri behar izan dugu dagoeneko sortutako programak berrerabili ahal izateko eta erabilitako gainerako datu-multzoen itxura izan dezan.

Lan hau aurrera eramaten geundela, datu-multzoa osatuta zegoen dokumentuen kodeketa berdina ez zela konturatu ginen. Dokumentu batzuk *ISO* kodeketa erabiltzen zuten eta beste batzuk *UTF-8*. *UTF-8* kodeketa erabili dugu proiektu osoan zehar, horregatik *ISO* kodeketa zuten dokumentuak aldatu ditugu *script* lagungarria erabiliz.

Datu-multzoko identifikadoreek '#' karakterea (traola) erabiltzen dute. UKBrekin, software analisian esan bezala, hutsuneekin eta traolekin tentuz ibili behar gara. Arazoa konpontzeko identifikadoreetako traolak '_' bilakatu ditugu, modu honetan UKBk jaso behar dituen sarrera fitxategiak ondo interpretatu ahal izan ditu.

Aurreko lanak egin ostean, dokumentu bakar batean datu-multzoko eskaera guztiak ditugu. Identifikadorearekin, desanbiguatu beharreko aipamenarekin eta testuinguruarekin osatutako lerroetan, hots, TAC datu-multzoa dugun egitura berdinarekin.

Azkenik, urrezko patroiaren zuzentasuna egiaztatu dugu. Bertan, entitate batzuekin arazoak izan ditugu; nahiz eta urrezko patroian *Wikipedia*ko entitate bezala agertu horietariko batzuk ez ziren *Wikipedia*ko entitate egokiak. Hauetariko batzuk errore txikiek eraginda izan dira; hauek 29 Taulan ageri dira. Entitate hauek konpontzeko *Wikipedia*z eta haien testuinguruez baliatu gara. Horrela garapen datu-multzotik 4 entitate aldatu ditugu eta ebaluaziotik, bakarra.

DEV	
José_MaríAznar	José_María_Aznar
Haur_Hezkuntza	Haur_hezkuntza
Ibaetako_Kanpusa	Ibaetako_campus
Lezama_(Athletic)	Lezama_(kirol-instalakuntzak)
EVAL	
Cofidis	Cofidis_(txirrindularitza_taldea)

29 Taula: Ezkerreko entitateak datu-multzoko urre patroian gaizki etiketatutako entitateak dira; eskuinekoak, ordea, eskuz gainbegiratutako Wikipediako entitate zuzenak dira.

Aurretik egindako lanaren eraginez, I. Fernandezen datu-multzoa *Wikipedia* 2013ko bertsiora eguneratu dugu. Honek, era berean, ebaluazioa entitate bihurketa egin gabe egitea ahalbidetzen du.

5.3.2. UKBren baliabideak

Gaztelaniazko esperimentuekin bezala, euskarazko *Wikipedia* erabilita UKB beharrezkoak dituen grafoak eta hiztegia sortu ditugu. Horretarako euskarazko *Wikipedia*ko 2013ko abenduaren erauzketa erabili dugu.

Baliabideak egiteko gaztelaniazko esperimentuetan egindako urratsak eman dira. Hala ere, *Wikipedia* honen tamaina txikia dela eta, ez dugu aurreikusten aldaketa handiak egongo direnik grafo batetik bestera. Arrazoi horregatik AB grafo murriztua sortzea alde batera uztea erabaki dugu.

5.3.3. Datu-multzoa eta baliabideak ebaluatzeko algoritmo batzuk

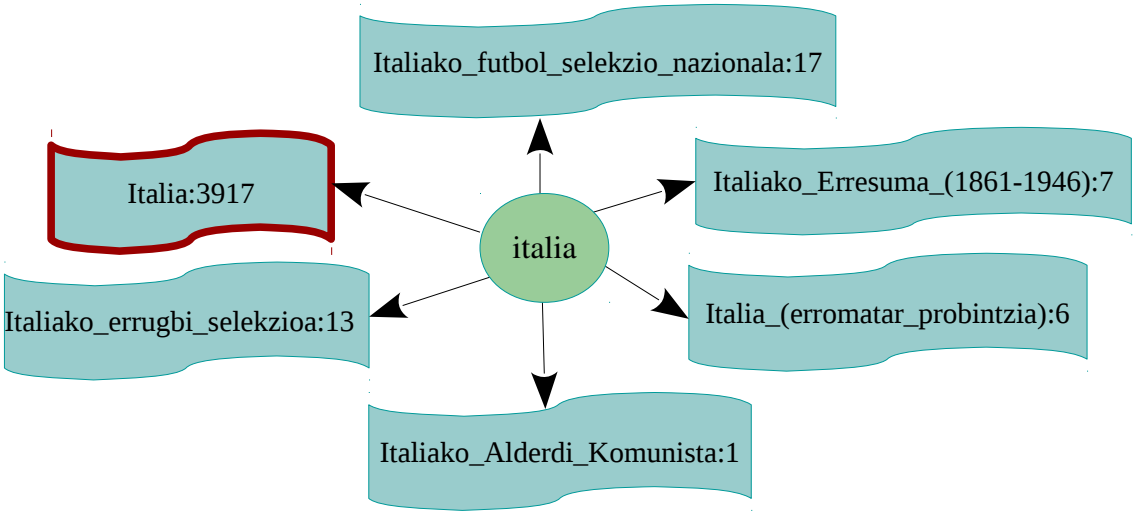
euskarazko datu-multzo honi buruzko konplexutasunaz eta erabiltzen ari garen baliabideen ahalmenaz informazio gutxi dugu. Horregatik *MFS* eta goi-borne algoritmoak erabili ditugu hauei buruzko informazio gehiago lortzeko asmoz.

MFS

MFS edo *Most Frequent Sense* desanbiguaziorako teknika sinple baina nahiko fidagarria da; hutsean desanbiguazio hautagaien artean agerpen gehien dituen entitate aukeratzen du.

Teknika honek esan bezala emaitza onak eskaintzen ditu. Eraginkortasun aldetik ezin hobea da; memoria kostu txikia eta azkarra baita. Gainera, euskarazko *Wikipedia* bezalako ezagutza-basea txikia edukiz gero, are eta gehiago.

Adibidez, 27 Irudian “italia” desanbiguatzerako orduan, inguruan dituen hautagaietatik ertz lodi gorria duena aukeratuko luke. Agerpen gehien dituen entitatea baita.



27 Irudia: Sortutako Euskarazko hiztegia modu grafikoan adierazita dago. Bertan, italia aingura Wikipediako zein entitateetara lotzen duen adierazten da. Entitatearen artikuluaren izenburuaren ondoren dagoen zenbakia aingura hori erabilita artikulu horretara lotutako kopurua da.

Goi-bornea

Goi-bornea edo *Upperbound*, erabiltzen ari garen baliabideen emaitzarik onena zein izan daitekeen jakiteko balio du. Modu honetan erabili ditugun baliabideen egokitasuna neur dezakegu.

Horretarako, script bat egikaritu dugu, zeinek desanbiguatu beharreko aipamena, hiztegia eta urrezko patroia erabiltzen duen. Hasteko, desanbiguatu beharreko aipamena hiztegiko ainguratzat hartzen dugu. Ondoren, aingura horrek dituen hautagaien artean urrezko patroiaren entitatea bilatzen du.

Adibidez, desanbiguatu beharreko aipamena “italia” baldin bada eta urrezko patroiko entitatea “Italiako_errugbi_selekzioa” izanik; 27 Irudian ikus dezakegu italia ainguraren 6 hautagaietako bat dela. Beraz, UKB emaitza egokia emateko gai izango litzateke. Aldiz, urrezko patroiko entitatea hautagaien artean agertuko ez balitz, UKB ezin izango luke emaitza egokirik eman.

5.3.4. Lematizazioa

Euskarazko izen entitateak deklinatuak ager daitezke testuinguruetan. Guk erabili ditugun baliabideetan terminoak deklinatuak edo lematizatuak agertzen dira. Euskarazko *Wikipedia*ren tamaina txikia izanik, aurreikusten dugu izen entitateak onartzen dituzten deklinabide asko eta asko ez daudela aingura bezala *Wikipedia*n.

Gainera idazleek ez dituzte berdin idazten artikuluak. Batzuek entitate lematizatuaren forma eta deklinazioa gidoiaz ezberdintzen dute. Honek aingurak jartzean deklinazioa kanpoan uztea dakar, 28 Irudian *Streaming* aipamenarekin ikus daitekeen bezala.

28 Irudia: Euskarazko Wikipediako artikulua ageri da, zeinetan Streaming ainguran deklinazioa gidoarekin egiten duen. Honen eraginez, Wikipediako erauzketa egitean Streaming aingura izango da eta ez Streaminga edota Streaming-a.

Horregatik 3.9 atalean azaldutako *Ixati* tresna erabili dugu datu-multzoko testuinguru garbiak lortzeko, hots deklinatu gabeko hitzez osatutako testuinguruak lortzeko.

5.3.5. UKBrekin desanbiguatzen

Behin datu-multzoa prest dugula, *Wikipedia*ko 2013ko erauzketatik sortutako hiztegiarekin eta grafoekin esperimentuak egin ditugu. Horretarako aurreko esperimentuetan bezala *script* lagungarri bat erabili dugu desanbiguazioa eta ebaluazioa automatikoki egin dezan.

Euskarazko esperimentuak egin ahal izateko UKBren sarrera prestatu behar dugu. Horretarako 5.2 ataleko esperimentuetan erabilitako *script*ak, hots, UKB erabiltzeko *script*ak berrerabili ditugu. Gainera 5.3.3 atalean azaldutako datu-multzoa eta ebaluatzeko baliatutako algoritmoak erabili ditugu, informazio osagarria edukitzeko.

Esperimentu honetan desanbiguatu beharreko aipamena aurrez badakigu eta 5.3.4 atalean lortutako testuinguru ezberdinak erabili ditugu.

	Tresna	Lem	P	R	f1	C	SYS	OK	A/s
1	UKB-AD grafoa	Ez	89.07	84.63	86.79	95.02	439	391	3
2	UKB-AU grafoa	Ez	88.86	84.63	86.70	95.24	440	391	2
3	UKB-AD grafoa	Bai	90.21	85.71	87.90	95.02	439	396	3
4	UKB-AU grafoa	Bai	90.00	85.71	87.80	95.24	440	396	2
5	MFS		89.32	85.06	87.14	95.24	440	393	
6	Goi-bornea		100	90.69	95.12	90.69	419	419	

30 Taula: Euskarazko lehenengo esperimentuaren emaitzak. "Lem" lematizatuta esan nahi du eta lematizatuta dagoen edo ez adierazteko erabili dugu. MFS eta Goi-borne kasuetan ez dugu adierazi, honek ez baitie eragiten.

Euskarazko lehenengo esperimentuaren emaitzak 30 Taulan adierazten dira. Hau aztertuta ikus daiteke ez daudela ezberdintasun nabariegirik esperimentuen emaitzen artean. *SYS* eta *OK* neurriek sistemak itzulitako erantzunak eta sistemak ondo erantzundako kopuruak dira. Hauei begiratuta ohar gaitezke AD eta AU grafoak erabiltzean lortzen diren ezberdintasuna itzulitako emaitzan dagoela eta eskaera bakarreko aldea dago.

Testuinguru lematizatuak erabiltzean, aldatzen dena sistemek ondo erantzundako eskaera kopurua da. Oraingo honetan 5 eskaeretako ezberdintasuna dago. Lematizatzean, eskaera gehiago ondo erantzuteko arrazoia testuinguru moduan dauden izen-aipamenen lema zuzena erabiltzea izan da.

AD grafo zuzendua AU grafo ez-zuzendua baino azkarragoa da, memoria kostu ezberdina baitute.

MFS algoritmoarekin lortutako emaitzak testuinguru normalen eta lematizatuen artean daude. Euskarazko ezagutza-base txikia delako lortzen ditu honek horren emaitza onak, aipamenak ez direnez oso anbiguoak, gehien agertzen den izen-entitatea itzulita aukera asko baitaude erantzun egokia emateko.

Goi-borneak lortutako emaitzekin ikus dezakegu hautatutako desanbiguatzeko aipamena ezagututa lor daitezkeen emaitzarik onenak. Hautatik ez gaude urruti, 419 eskaeretatik 396 lortu baititugu, 23 eskaerako ezberdintasuna dago.

	Tresna	Memoria
1	UKB-AD grafoa	192MB
2	UKB-AU grafoa	233MB

31 Taula: Euskarazko esperimentuetan UKBk erabilitako grafoen memoria kostua.

1 Grafo zuzenduak memoria kostu txikiagoa duela ikus daiteke 31 Taulan. Bien memoria ezberdintasuna grafo bakoitza sortzean nodoen erlazioen ezberdintasunak dakar. Hemen argi ikusten da 4.2.2 atalean azaldutako, UKBk grafo ez-zuzenduen loturak lotura bikoitz moduan jartzen dituela, lotura bat noranzko bakoitzean.

5.3.6. Aipamena

Aurreko esperimentuan desanbiguatu beharreko aipamena zein zen ezagutzen genuen. Oraingo honetan, aldiz, aipamen anbiguoa identifikatu eta, ondoren, desanbiguatu egin dugu. Horretarako, aipamen anbiguoa identifikatzeko, Matx erabili dugu. Modu honetan, UKBren aipamen identifikatzailea edo *spotte*a Matx izan da, *Spotlight*en kasuan *LingPipeSpotter*ra izan den moduan.

Matxen eginkizuna garapen multzoko eskaera ezberdinen testuinguruak prozesatzea izango da, aipamenak identifikatzeko. Beti bezala, Matxek identifikatutako aipamenen artean datu-multzoko eskaerak zehazten duen izen-entitatea barne duen aipamenik luzeena hautatu du. Modu honetan, identifikatzeko gai izan garen aipamenen eskaerekin datu-multzoko azpi-multzo bat lortzen dugu. 32 Taulan adierazten dira azpi-multzo hauen eskaera kopuruak.

Aurreko esperimentuan bezala 5.3.4 atalean lortutako testuinguru ezberdinak erabili ditugu, beraz, aurreko paragrafoan zehaztutako prozedura bi testuinguru ezberdinekin egin dugu. Gainera, Matx tresnako bi aldaera erabili ditugu; lehenengo eta bigarren heuristikoak hain zuzen.

	Konbinaketa	Identifikatutako eskaerak
1	Testuinguruak lematizatu gabe eta Matx 2. heur.	332
2	Testuinguruak lematizatu gabe eta Matx 1. heur.	335
3	Testuinguruak lematizatuta eta Matx 2. heur.	379
4	Testuinguruak lematizatuta eta Matx 1. heur.	472

32 Taula: Parametroen konbinaketa ezberdinak erabilita identifikatutako eskaera kopuruak adierazten dira. Gogoratu datu-multzo honek orotara 532 eskaera dituela.

32 Taulan ikus daitekeen moduan, konbinaketa ezberdinen eraginez lortzen ditugun eskaera kopuruak ezberdinak dira. Lematizatu gabeko testuinguruak eta 2. heuristikoa erabilita lortutako eskaera kopurua txikia da, 1. heuristikoa erabilita baino 3 eskaera gutxiago itzultzeko gai izan baita. Horregatik 1. konbinaketa ez dugu esperimentuetan erabiliko, ezberdintasun nahikorik ez baitaute.

5.3.5 esperimentuan grafo ezberdinek erakutsi duten antzekotasuna kontuan izanik, AD grafo zuzenduarekin esperimentu hau egitea erabaki dugu. Gainera grafo hau bien arteko azkarrena eta memoria kostu txikiena duena da.

Identifikatutako eskaera kopuruak eragin zuzena edukiko du hurrengo taulan adierazten diren emaitzetan.

	Tresna	Lem	Heur	P	R	f1	C	SYS	OK
1	UKB-AD grafoa	Ez	2	73.20	46.10	56.57	62.99	291	213
2	MFS	Ez	2	65.29	41.13	50.46	62.46	291	190
3	Goi-bornea	Ez	2	100	50.22	66.86	50.22	232	232
4	UKB-AD grafoa	Bai	2	91.62	71.00	80.00	77.49	358	328
5	MFS	Bai	2	91.62	71.00	80.00	77.49	358	328
6	Goi-bornea	Bai	2	100	75.32	85.93	75.32	348	348
7	UKB-AD grafoa	Bai	1	83.37	75.97	79.50	91.13	421	351
8	MFS	Bai	1	81.24	74.03	77.46	91.13	421	342
9	Goi-bornea	Bai	1	100	81.17	89.61	81.17	375	375

33 Taula: Euskarazko 2. esperimentuaren emaitzak adierazten dira. "Heur" heuristikoa, modu honetan 1. edo 2. heuristikoa erabili den adierazteko erabili dugu.

33 Taulan euskarazko 2. esperimentuaren emaitzak adierazten dira. Bertan, argi ikusten da doitasun altuena duen konbinaketa lematizatutako testuingurua eta Matxen 2. heuristikoa erabilitakoa dela. Bestalde, txarrena lematizatu gabeko testuinguruak lortzen ditu, izen-entitate asko deklinatuta egoteagatik gertatzen da. Horregatik lematizatzean lortzen dira emaitza hobeak. Honetaz gain, 2. heuristikoak erabiltzeak dakarren ziurtasun handiagoa ikus dezakegu; nahiz eta 1. heuristikoak baino erantzuteko gaitasun, *coverage*, txikiagoa duen, ondoren adieraziko dugun moduan.

Estaldurari erreparatuz gero, lematizatutako testuinguruak erabilitako esperimentuek lortu dute emaitzarik onenak; bereziki Matxen 1. heuristikoa erabiltzen duenak. Txarrena, doitasunean bezala, lematizatu gabeko esperimentuek izan dira.

Coverage ebaluazio-neurria aztertuta, erlazio zuzena ikusten da konbinaketa bakoitza identifikatzeko gai izan den eskaera kopuruarekin, hau da, aipamen anbiguoak identifikatzeko erabilitako konbinaketak goi-langa jartzen du. Argiago esanda, lematizatu gabeko testuinguruak eta 1. heuristikoa erabiltzearen konbinaketak 335 eskaera identifikatzeko gai da, 32 Taulan ikus daitekeenez, horregatik konbinaketa hau erabiltzean ezin izango ditu 335 aipamen baino gehiago desanbiguatu. Horregatik, onena aurretik gehien identifikatutako konbinaketak lortzen du, hots lematizatutako testuinguruak eta Matxen 1. heuristikoak erabiltzen duena.

MFS algoritmoak lortutako emaitzak UKBek lortutakoekin alderatzean, konturatzen gara lehenengo honen emaitzak ez direla inoiz hobeak. Soilik lematizatutako testuinguruak eta Matxen 2. heuristikoa erabiltzean lortzen ditu UKBren emaitza berdinak.

5.3.7. Ebaluazioa

Behin garapen fasea amaituta, euskarazko esperimentuetatik lortutako konbinaketa onena erabilita test datu-multzoan ebaluatu dugu. Erabilitako konbinaketa AD grafoa eta lematizatutako testuingurua erabiltzearena izan da.

Ondoren, (I. Fernandez, 2012) tesian UKB erabilita lortutako emaitzekin alderatuko genuen. Baina honek ebaluatutako datu-multzoa EzNIL eta NILen gainean egin du; guk, ordea, EzNILetan.

	Tresna	P	R	f1	C	SYS	OK
1	UKB-AD grafoa eta testuinguru lem.	92.60	88.79	90.65	95.88	419	388
2	I. Fernandez	76.25	75.80	76.02		497	379

34 Taula: Gure sistemak lortutako emaitzak I. Fernandezen tesian lortutako emaitzekin test datu-multzoan izandako emaitzak adierazten dira. 437 EzNIL entitate daude eta 63 NIL.

Berarekin kontaktuan jarri ondoren, emaitzak egon behar ziren katalogoa dagoeneko existitzen ez dela konturatu ginen. Honen eraginez, 34 Taulan adierazten diren emaitzak zuzenean konparagarriak ez dira.

6. ONDORIOAK ETA ETORKIZUNEKO LANAK

Azkenengo atal honetan ondorio orokorrez eta etorkizuneko lanez arituko gara. Ondorio atalean proiektuaren ondorioez, egindako denbora estimazioen gaineko ondorioez eta ondorio pertsonalez arituko gara. Etorkizuneko lanean proiektu honetan sakontzen jarraitzeko lana azalduko dugu.

6.1. Ondorioak

Ondorioak atala hiru azpi-atal ditu; proiektuaren ondorioak, egindako estimazioen alderaketa errealitatearekin eta ondorio pertsonalak. Lehenengo azpi-atalean ebaluazioan lortutako emaitzen ondorioak izan ezik, ondorio orokorrak. Estimazioak eta errealitatea atalean, PHDan egindako estimazioak errealitatean egindakoarekin alderatuko ditugu. Eta azkenik, proiektua egitean proiektuak egiteak eragindako ondorioak aztertuko ditugu.

6.1.1. Proiektuaren ondorioak

Ingelesezko esperimentuetatik *Spotlight*ek duen azkartasuna da ondorio positiboena. Honetaz gain, ordura arte estaldura txikiko sistematzat genuen, baina emaitzetan ikusi dugu ez dela horrela. Tresna honek doitasun balio handiak lortzen ditu, beraz, esan dezakegu emaitzen fidagarritasuna baldin bada helburua tresna hau erabiltzea gomendatzen da. Bestalde desanbiguatu beharreko aipamena zehatz dezakegunez *coverage* handia lor daiteke, hots, *Spotlight*ek duen erantzuteko gaitasuna handi dezakegu. Matx erabiltzeari esker, *Spotlight*en doitasuna eta *coverage*a handitu dugu, aipamen luzeena hautatzeak eraginda.

Gaztelaniazko esperimentuetan, ordea, *Spotlight*ek duen gaztelaniazko eredua ingeles eredua baino okerragoa dela ikusi dugu, baliteke ebaluatzeko erabilitako datu-multzoa *cross-lingual* (gaztelania eta ingelesa) izatea eragin pixka bat edukitzea[14]. Baina ziurrenik ingeles eredua hobea izateko arrazoi nagusia honek duen ezagutza-basea gaztelaniazko ereduarena baino handiagoa izatea da. UKBrekin izandako emaitzetatik ingelesezko grafoa erabiltzea da aukerarik onena. Honen arrazoia *Spotlight*entzat emandako berdina izan daiteke, ingelesezko grafoak ezagutza-base handiagoa izatea. Aipatzekoa da ingelesezko eta gaztelaniazko grafoak konbinatzean ez dugula esperotako hobekuntza izan, hala ere gaztelaniazko grafoa bakarrik erabilita baino emaitza hobeak izan ditu. Bi tresnak konparatzean *Spotlight* hobea dela ondoriozta dezakegu, gainera artearen egoera adierazten duen sistema batek baino emaitza hobeak lortu ditugula aipagarria da. Aipatzekoa da Matxek eskaintzen digun aipamenak erabilita ez direla *Spotlight*en *spotter*raren emaitzak hobetu, ingelesezko esperimentuetan gertatu ez bezala.

Bestetik, euskarazko esperimentuetan (I. Fernandez, 2012) tesi lanean sortutako datu-multzoa eguneratu dugu *Wikipedia*ko 2013 bertsiora. Gainera, lematizatzeak dakarren onurak ikusi ditugu, lema zuzena jakinda emaitzak asko handitzen baitira. Honekin batera, euskarazko *Wikipedia* artikuluetan ageri diren ainguren garrantzia ikusi dugu, izen-entitate asko eta asko ez baitaude deklinatuta. Azkenik, euskarazko ezagutza-base txikia izanik *MFS* algoritmoak lortzen dituen emaitza onak nabarmentzekoak dira. Era berean sortutako grafoek izandako emaitzen antzekotasuna aurreko arrazoi berdinak eragiten du. Beraz, komenigarria ikusten dugu euskarazko *Wikipedia* handitu beharra dagoela, adibidez, artikulu berriak sortuz eta dagoeneko sortuta dauden artikuluen arteko loturak ainguren bidez eginez.

14 Gogoratu datu-multzo honek 1991 gaztelaniazko eta 75 ingelesezko testuinguru dituela.

Softwarearen analisia egitean, *Spotlight*en funtzionamenduaren ezagutza sakonagoa behar dugula ikusi dugu. Aurreikusten baitugu *Spotlight*en kodean edota ereduetan aldaketak eginda honen emaitzak hobetzeko aukera dugula. Hala ere, *Spotlight* izen-entitateak estekatzeko sistema nahiko ona eta azkarra dela ondoriozta dezakegu. Horren froga proiektu honetan lortutako emaitzak dira. Gainera, IXAk etorkizunean tresna honekin lan egin ahal izateko beharrezko informazioa lortu eta dokumentatu dugu.

Denbora zehatza izanik proiektua egiteko, bileretan proposatutako ideia batzuk alde batera utzi behar izan dira. Hauek egiteko beharrezko denbora ez baikenuen. Etorkizuneko lanak atalean azalduko dira egiteke geratutako ideiak.

Proiektu honetan lortutako emaitzak eta ondorioak *QTLeap* proiektuan eta zenbait artikulutan lagungarri izan dira. Hauen artean aipatzekoa (A. Barrena et al., 2015) IkerGazte topaketetarako prestatutakoa.

6.1.2. Estimazioak eta errealitatea

Atal honetan PHDan egindako estimazioak errealitatean gertatutakoarekin konparatuko dugu. Horretarako 35 Taulaz lagunduko gara.

Atazak	Estimazioak	Errealitatea
Guztira	384 ordu	414 ordu
Prozesu taktikoak	84 ordu	80 ordu
K- Kudeaketa		
K1- Bilerak	50 ordu	50 ordu
K2- Artxiboen kudeaketa	5 ordu	7 ordu
K3- Lan gaztiguen kudeaketa	4 ordu	2 ordu
P- Planifikazioa		
P1- Proiektuaren Helburu Dokumentua (PHD) egin	25 ordu	21 ordu
Prozesu operatiboak	160 ordu	174 ordu
I- Iterazioak (7 iterazio)		
I1- 1.Iterazioa	50 ordu	55 ordu
I2- 2.Iterazioa	30 ordu	34 ordu
I3- 3.Iterazioa	10 ordu	12 ordu
I4- 4.Iterazioa	15 ordu	17 ordu
I5- 5.Iterazioa	20 ordu	20 ordu
I6- 6.Iterazioa	25 ordu	25 ordu
I7- 7.Iterazioa	10 ordu	10 ordu
Dokumentazioa	90 ordu	110 ordu
M- Memoria	80 ordu	100 ordu
A- Aurkezpen publikoa	10 ordu	10 ordu
Formakuntza	50 ordu	50 ordu

35 Taula: Egindako denbora estimazioak eta errealitatean gertatutakoak.

Nahiko zuzenak izan dira egindako denbora estimazioak. Horren erakusle proiektua egikaritzean izandako desbiderapen totala soilik estimatutakoa baion 30 ordu gehiagokoa izan dela.

Prozesu taktikoetan 4 ordu aurreztu ditugu, honen zergatia PHD dokumentua egitean aurreztutako denboragatik. Kudeaketan planifikatutakoarekin aldaketak egon dira baina euren artean berdindu dira. Adibidez, planifikatuak baino bilera gutxiago egin ditugu, baina egindako bilera batzuk luzeak izan direnez ordu kopuru berdina egon gara bileretan.

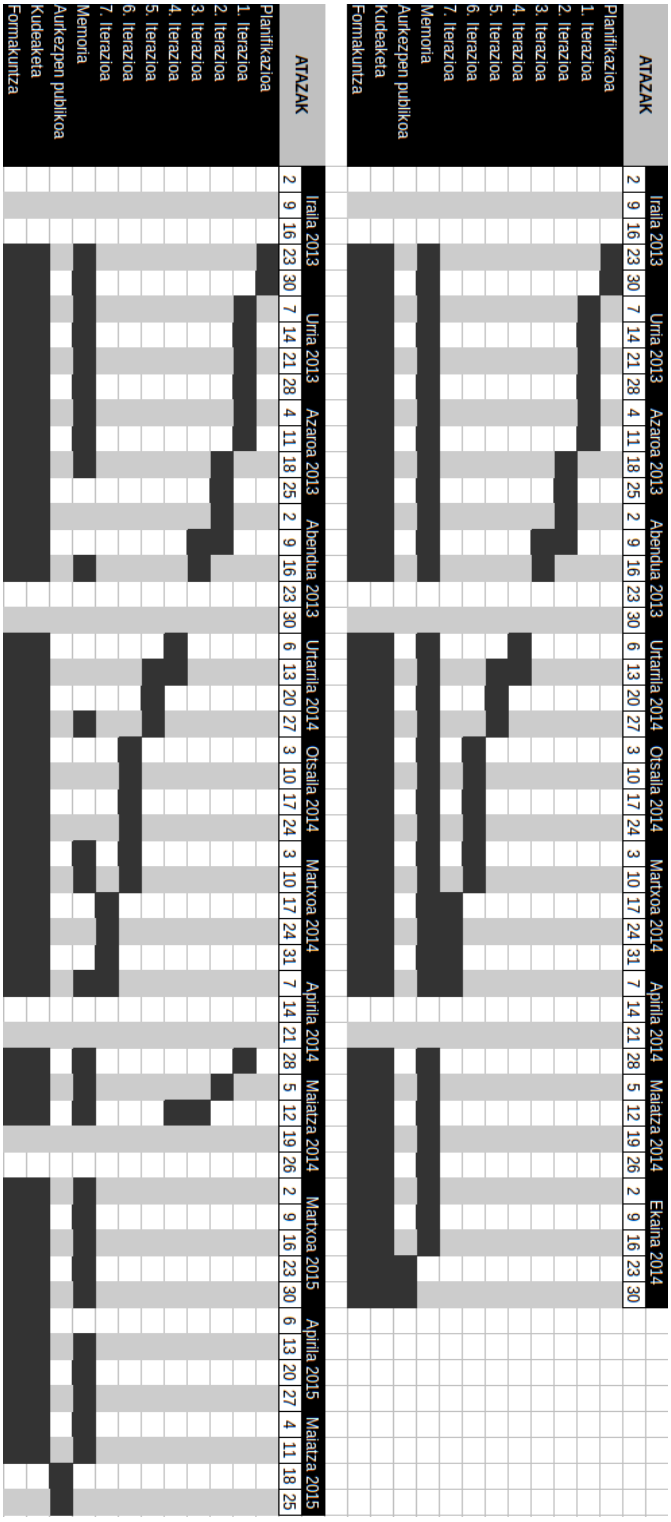
Prozesu operatiboetan, ordea, estimatutakoa baino 14 ordu gehiago behar izan ditugu. Honen zergatia proiektua egin bitartean *Spotlight*ek bertsio berria kaleratu izana da. Horregatik, bertsio berria analizatu eta *Spotlight* erabiltzen zuten esperimentuak errepikatu ditugu.

Dokumentazioan gertatu da desbiderapen handiena, planifikatutakoa baino 20 ordu gehiago erabili baititugu. Proiektuaren memoria ez nuenez guztiz egitearekin batera idatzi eta, gainera, dezente atzeratu egin nuenez, berriz ere memoria idazterako garaian egindakoa birgogoratu eta eraginkor idazteko ohitura falta eduki dut. Bestalde, esperotakoa baino memoria luzeagoa atera izana ere eragina izan du.

29 Irudian, bi *gantt* diagrama daude lehenengoak estimatutako kronologia adierazten du eta bigarrengoak, aldiz, errealitatean gertaturikoa. Ikus dezakegunez, plangintzan eta prozesu operatiboetan (hasieran) estimatutako kronologia bete dugu, baita kudeaketa eta formakuntza ere. Memoria pixkanaka-pixkanaka proiektua egin ahala betetzen joatea estimatu genuen, baina hori ez da posiblea izan. Hautazkoak genituen hiru ikasgaien lan karga eta kurtsoan zehar beste esparru batzuetan hartutako konpromisoak direla medio. Hala ere, iterazio bakoitza amaitzean zenbait laburpen eta emaitzen batura egin ditugu, memoria egiterako garaian errazago bete ahal izateko.

Gainera, apirilean *Spotlight*en bertsio berria ateratzean, *Spotlight*ekin zerikusia zuten iterazioetan berriz ere lan egin dugu, apirilaren azken hastean eta maiatzaren lehenengo bi aste bitartean. Horregatik, soilik memoriarekin aritu beharrean, aldi berean *Spotlighte*n bertsio berria analizatu, bertsio berri honen emaitzak lortu eta memoriaren idazketa egin ditugu.

Horren ondoren, ikusita proiektua ez genuela epe barruan amaituko alde batera uztea erabaki nuen. 2015eko martxoan proiektuaren memoria idazketarekin jarraitu nuen. Apirilaren 6ko astean ez genuen lanik aurreratu, aisialdi talde bateko kide izanik udalekuak baikenituen. Gainera, martxoaren 25etik maiatzaren 3ra bitartean Android ikastaro bat egiten ibili naiz. Honen eraginez memoriaren idazketa luzatu eta aurkezpenaren prestaketa atzeratu egin dut.



29 Irudia: Gantt diagramak. Goikoa estimatutakoa eta behekoa erreala.

6.1.3. Ondorio pertsonalak

2013/14ko ikasturtearen hasieran KBP proiektua zehaztu nuenean, argi nuen ikerketarekin zerikusia zuen proiektua hautatuko nuela. Karrera amaitu baino lehenago ikerketa bat egin, ikerketa talde baten funtzionamendua eta honetan sortzen den giroa ezagutu nahi bainuen.

Aitorrek eta Enekok *QTLeap* proiektu europarraren baitan lan egiteko aukera eman zidaten, izen-entitateen estekatze munduan ikertzeko. Gaia interesgarria egin zitzaidan eta aurrera egin genuen proiektuarekin. Egia esanda proiektuaren hasiera gogorra egin zitzaidan, bileretan hitz egindakoaren informazioa topatzeko zailtasunak aurkitu bainituen. Baina zailtasun hauek berehala gainditu nituen zuzendarien eta Anderren laguntzari esker. Edozein zalantza argitzeko prestutasuna adierazi baitzuten.

Nire proiektua *QTLeap* proiektuaren barnean kokatzen denez, talde lanean aritu ginen. Honek talde dinamikaren garrantzia ikustaraztea ekarri zidan. Taldea giro onean murgilduta baldin badago, denek egindako lanetik etekin handiagoak atera daitezke. Gainera esperientzia ederra lortu dut talde dinamika; lanak eta ardurak guztion artean banatu behar baikenituen eta erabakiak denon artean hartu.

Bestalde, 2013/14 ikasturtean 3 ikasgai izan ditut proiektuaz gain, honen eraginez Donostian ez bizitzea erabaki eta gurasoen etxera itzuli nintzen. Honek aurreko urteetan nuen lan egiteko dinamika apurtu egin dit; gero eta lan karga gutxiago eduki are eta lan gutxiago egitea eragin baitu. Gainera, Gasteizera berriz ere itzultzean beste zenbait gauzetan jarduteko gogoak piztu dizkit. Arrazoi hauengatik proiektuaren memoria alde batera uztea ekarri du. Hala ere, 2015 urteko martxoan berriz ere proiektuaren memoria idazten jarraitu nuen, egindako hausnarketaren eraginez.

Horren eraginez, proiektua egiteko ohitura berreskuratzea lortu nuen, egin nahi nituen gauzak alde batera utzi gabe. Honek nire burua ezagutzea eta etorkizunean nola jokatzen lagunduko dit.

Proiektua egiteak ezagutzen ez nituen programazio lengoaiak ikastea, estekatze sistemen funtzionamendu orokorra ezagutzea eta *Wikipedia*ren egitura ulertzea ekarri dit.

Amaitzean, ikerkuntzan gehiago egiteko gogoekin geratu naiz, baina lotuegi nago Gasteizekin eta bertan dudanarekin atzean uzteko.

6.2. Etorkizuneko lana

KBPa epe eta ordu kopuru zehatz batzuetan egin beharreko proiektua da, horregatik proiektua egin bitartean ordu gehiegi eskatzen dituzten lanak baztertu ditugu.

Egiteke geratu zaigun lana bi multzotan banatu dugu; batetik *Spotlight* tresnaren inguruko lana eta, bestetik, UKBrekin erlazionatutako lanak.

6.2.1. *Spotlight*en inguruko etorkizuneko lanak

Proiektuak hiru hizkuntzekin egindako esperimentuez osatzen da, QTLeap proiektuak beharrezkoak zituen beharrak asetzeko intentzioarekin. *Spotlight* tresna oso aprobetxagarria dela ikusi dugu, gainera mantentze eta berritze lanak egiten dituzten lan taldea du atzetik.

Horren eraginez, *Spotlight*entzat euskarazko eredua sortzea eta euskararen berezitasun lexiko-morfologikoak gainditzeko *Spotlight*en egin beharreko aldaketak egiteke gelditu dira. Aurreikusitako lan handia ikusita, honek KBP batean egin beharreko haina lan dauka.

Horretaz gain, *Spotlight*en kodean aldaketak egin ahal izateko beharrezkoak diren tresnaren ezagutza sakonagoa izateko lana egiteke dago oraindik. Honi esker, *Spotlight*ek gaztelaniako esperimentuetan erakutsitako emaitzak hobetzea aurreikusten dugu.

Gainera, *Spotlight*en bertsio berrian eginiko abiaduraren hobekuntzaz gehiago ikertzea faltatu zaigu. Bertan egindakoa IXAk dituen tresnatan aplikatu ahal izateko informazio interesgarria baita.

6.2.2. UKBren gaineko etorkizuneko lanak

UKB tresnak lorturiko emaitzak eta *Spotlight*enak alderatuta, argi dago desanbiguatzeko abiadura motela duela. Horregatik, *Spotlight*ek bertsio berrian egindako aldaketak edota beste tresna batzuen ikerketatik lortutako inplementazio teknikak erabilita UKB bera azkartzeko lana egiteke geratu zaigu.

Bestalde euskarazko ezagutza-basea txikia eta erabilitako hiztegiak lematizatu gabeko aingurez beteta daude. Hiztegi honen erantzuteko ahalmena hobetzeko asmoz, zenbait moldaketa egiteke gelditu zaizkigu. Adibidez, lema berdina duten aingura ezberdinen artikuluak komunak izatea. Modu honetan, “EEUU” edo “EEUUren” aingurek eskaintzen dituzten hautagaiak berdinak izango lirateke.

BIBLIOGRAFIA

Liburuak

Astigarraga A., Gojenola K., Sarasola K. eta Soroa A. 2009. *TAPE Testu-analisirako PERL tresnak*. UEU

Fernandez I. 2012. *Entitate-Izenak Euskaraz: Identifikazioa, Sailkapena, Itzulpena eta Desanbiguazioa*. Tesi txostena.

Barrena A. 2013. *Testuak informazio gehigarriarekin aberasten, entitate-izenen ezagutze eta desanbiguazioa.* Hizkuntzaren Azterketa eta Prozesamendua Master bukaerako proiektua.

Artikuluak

Agirre, E., Soroa, A.: *Personalizing PageRank for Word Sense Disambiguation*. In: Proceedings of the 12th Conference of the European chapter of the Association for Computational Linguistics, pp. 33–41 (2009)

Aduriz I., Aranzabe M., Arriola J., Diaz-De-Ilarraza A., Gojenola K., et al.: *A Cascaded Syntactic Analyser for Basque*. LNCS Series. Springer Verlag, pp.124-135, 2004.

Daiber J., Jakob M., Hokamp C., Mendes P. N.: *Improving Efficiency and Accuracy in Multilingual Entity Extraction.* Proceedings of the 9th International Conference on Semantic Systems (I-Semantics). Graz, Austria, 4–6 September 2013.

Hachey B., Radford W., Nothman J., Honnibal M., Curran J. R.: *Evaluating Entity Linking with Wikipedia.* In: Artificial Intelligence 194, pp. 130-150, (2013).

Agirre E., Lopez de Lacalle O., Soroa A.: *Random walks for knowledge-based word sense disambiguation*. In: Computational Linguistics 40, pp. 57–88. 2014.

Barrena A., Agirre E., Perez de Viñaspre J., Soroa, A.: *Izen-aipamenak desanbiguatu eta Wikipediara lotzen.* In: IkerGazte 2015.

Mendes P. N., Jakob M., García-Silva A., Bizer C.: *Dbpedia spotlight: Shedding light on the web of documents.* In: I-Semantics 2011 Proceedings of 7th International Conference on Semantic Systems. Graz, Austria. 2011.

Han X. & Sun L.: *An entity-topic model for entity linking*. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL ’12, pages 105–115, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

Houlsby N. & Ciaramita M.: *A Scalable Gibbs Sampler for Probabilistic Entity Linking.* In: Proceedings of ECIR, 335--346, 2014.

Clarke J., Merhav Y., Suleiman G., Zheng S. Murgatroyd D.: *Basis Technology at TAC 2012 Entity Linking.* In Procedings of the Text Analysis Conference, 2012.

Monahan S., Carpenter D.: *Lorify: A Knowledge Base from Scratch.* In Procedings of the Text Analysis Conference, 2012.

Barrena A., Agirre E. Perez de Viñaspre J. & Soroa A.: *Izen-aipamenak desanbiguatu eta Wikipediara lotzen.* In IkerGazte, 2015.

Internet loturak

- UKB: http://ixa2.si.ehu.es/ukb/
- *Spotlight*: http://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki
- *Shell* programazioa:

  http://persoal.citius.usc.es/tf.pena/ASR/Tema_2html/node20.html
- *Wikipedia*:
  - Euskal Wikipedia: eu.wikipedia.org
  - Wikipedia ingelesez: en.wikipedia.org
  - Wikipedia gaztelaniaz: es.wikipedia.org