Ondare Kulturala Wikipediara lotzen: Entitateen desanbiguazioaren ekarpenak

Karrera Bukaerako Proiektua

2012.eko uztailaren 10

Ander Barrena Madinabeitia

Gainbegiraleak:
Eneko Agirre - Aitor Soroa

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Gaien Aurkibidea

Irudien Zerrenda

Taulen Zerrenda

1 Kapitulua

Sarrera

Ondare kulturala, kultura baten adierazpenen multzoa da. Hau da, kultura hau existitu denaren aztarna edo herentzia. Ondarea, elementu ukigarri (eraikuntza, eskultura, liburuak eta abar) eta ukiezinez (tradizioa, folklorea) osatua dago. Proiektu honetan, ondare ukigarrian zentratuko gara, argazki, pintura eta idatzizko laginetan batez ere. Gaur egun, ondare kultural hau bildu eta digitalizatzen diharduten hainbat erakunde aurki daitezke, horietako bat Europeana da.

Europeana[1] Europear batasuneko kide diren 27 estatuen instituzio kulturalen kontribuzio digitalen liburutegia da. Euskarri digitalean dauden milioika elementuz osatua dago eta hau da hain zuzen Europeanaren ezaugarri garrantzitsuenetariko bat.

Hala ere, erabiltzaile arruntaren ikuspuntutik, elementu askok informazio falta handia dute. Hauen deskribapenak motzak dira eta ez dute informazio gehigarririk eskaintzen. Adibidez, Europeanan "Mona Lisa" artelan famatua bilatuz, honi dagokion Europeana elementuaren[2] (ikus 1.1 irudia) deskribapena oso urria dela ikus daiteke. Gainera "Aufbewahrung/Standort: Musée National du Louvre (Paris)"deskribapenak ez dio elementuari erreferentzia egiten, erakusgai dagoen museoari baizik.

Baina ondare kulturaleko elementuei buruzko Wikipediako artikuluek, informazio asko eta erlazionatutako elementuetara estekak eskaintzen dituzte. Aurreko adibidearekin jarraituz "Mona Lisa"-ri buruzko Wikipedia artikuluak[3] (ikus 1.2 irudia) deskribapen aberats eta zehatza eskaintzen du. Gainera egilearen informazioa, artelanaren inguruko historia eta erlazionaturiko

[1] http://www.europeana.eu

[2] http://www.europeana.eu/portal/record/08501/F29E726C1E32B3B49D5166E1B7F5434AFA6404B7.html?start=12&query=mona+lisa

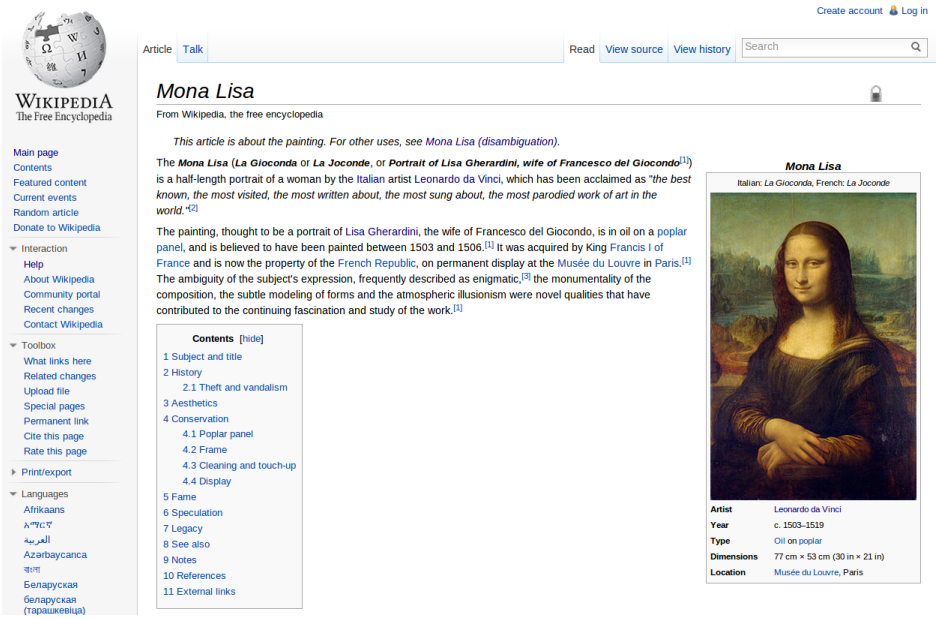
[3] http://en.wikipedia.org/wiki/Mona_Lisa

Irudia 1.1: Europeanako "Mona Lisa" artelanaren elementua. Deskribapenak ez du artelanari buruz ezer esaten eta ez du informazio gehigarririk eskaintzen.

artikulu zerrendak eskaintzen ditu.

Ondare kulturaleko elementuen informazioa aberastea ez da gauza berria, [1] artikuluan, LEMMO tresna eskaintzen dute Europeanako elementuak kanpoko informazioaz eskuz aberasteko. Karrera bukaerako proiektu honetan, aurrekoak artikuluak ez bezala, sistema automatiko bidezko aberasketaren ikerketa eta ebaluazioa egingo da. Lehenik ausaz aukeraturiko elementuen lagin bat eskuz etiketatu da dagokion Wikipedia artikuluarekin eta ondoren sistema automatikoen emaitzak ebaluatu dira lagin honen aurka. Etiketatzeko irizpide batzuk finkatu dira ikerketa honetan: elementuaren instantzia definitzen duen artikulua aurkitu nahi izan da eta ez erlazionatutako artikulua. Artikuluak, ondare kulturaleko elementuak deskribatzen duena deskribatu behar du.

Ondare kulturaleko elementu bati Wikipedia artikulu bat esleitzeko orduan, posibleak diren artikuluen artean aukera egokia egitean dago gakoa. Beraz bi pausu jarraitu behar dira: Lehenik hautagai zerrenda bat sortu eta bigarrenik hauek desanbiguatu behar dira artikulu bakarrarekin geratuz.



Irudia 1.2: Wikipediako "Mona Lisa"-ren artikulua. Deskribapen luze eta aberatsa eskaintzan du, egile eta modeloaren informazio gehigarria, erlazioa duten elementuetara estekak eta abar.

Ataza hau entitate-izenen desanbiguazioa izenaz ezagutzen da.

Adibide berdinarekin jarraituz, "Mona Lisa"-ren Europeana elementuari dagokion artikulua itzultzeko prozesua jarraian datorrena da. Lehenik Wikipediak 13 artikulu[4] eskaintzen ditu "Mona Lisa" izenarekin erreferentzia daitezkeenak. Aukeren artean, abesti, opera, aktoresa edo kantari bat aurki daitezke. Bigarrenik hautagai hauen artean desanbiguazio algoritmo batek artelana[5] aukeratu beharko luke. Alor honen inguruan lan asko egina dago (ikus [2, 3, 4, 5, 6, 7, 8, 9] artikuluak) baina oraindik ere arazo ireki bat da.

[4] http://en.wikipedia.org/wiki/Mona_Lisa_(disambiguation)
[5] http://en.wikipedia.org/wiki/Mona_Lisa

2 Kapitulua

Proiektuaren Helburuak

Proiektu honek bi helburu nagusi ditu:

1. Europeanako elementuak, sistema automatiko bidez, Wikipedia artikuluekin aberastea posible den ikertzea eta ebaluatzea. Ausaz aukeraturiko elementuen lagin etiketatua, aurrerantzean urre-patroia deituko dena, elementuen etiketatu egokia izango da. Ataza honetan, sistemek elementu bakoitzarentzat proposaturiko artikuluak urre-patroiaren aurka ebaluatuko dira.

2. Europeanako elementuak Wikipedia artikuluekin aberasteko sistema bat garatzea eta ebaluatzea. Sistema garatu ondoren urre-patroiaren aurka ebaluatuko da.

Bi helburu nagusi hauez gain, urre-patroia bera ebaluatu da eskuz eginiko beste etiketatu baten aurka. Batez ere, etiketatu garaian suertatu diren arazo eta zalantzak argitzeko asmoz.

3 Kapitulua

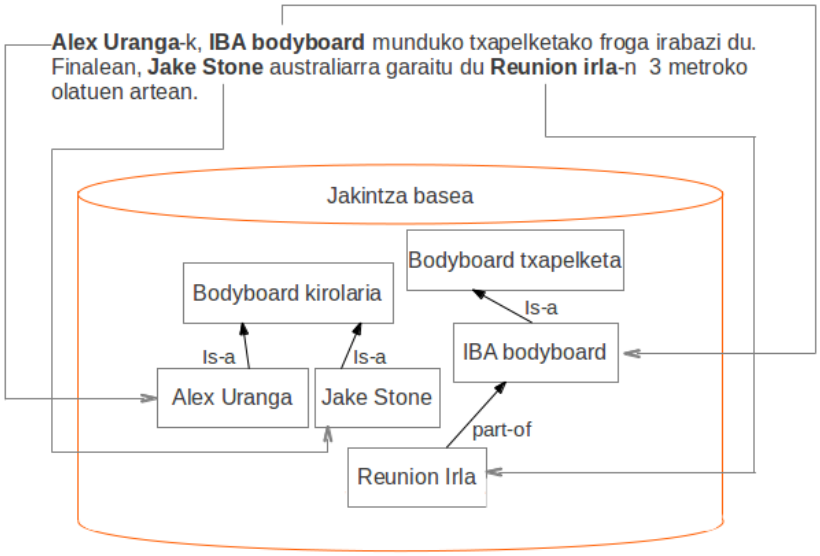
Aurrekarien azterketa

Lehenik eta behin, proiektu hau aurrera eramateko beharrezkoak diren aurrekariak aztertuko dira. Hasteko entitateak eta ezagutza baseak azalduko dira. Jarraian Wikipediak proiektu honetan duen garrantziaz hitz egingo da. Gainera Europeana eta hau inguratzen duten proiektu Europarren azalpen bat egingo da Ondare Kulturalaren arloan kokatzeko. Ondoren Europeanako elementuak artikuluekin lotzeko Wikipedia Miner sistema automatikoa azalduko da. Honen ostean, Europeanako elementuak Wikipedia artikuluekin aberasteko sistema garatzeko erabili diren aurrekariak aztertuko dira.

3.1 Entitateak ezagutza basearekin lotzen

Wikipedia artikuluei buruz hitz egiten denean, entitateei erreferentzia egiten zaie aldi berean. Entitateak, ontologia, semantika edo logika sistema batek existitzen direla froga dezaketen edozer gauza dira. Entitate hauek, ezagutza base bateko parte izango dira eta beste entitateekin loturak izango dituzte (ikus 3.1 irudia). Irudiko testuan entitateei erreferentzia egiten dieten izenak identifikatu eta ezagutza baseko entitateekin lotzean, esaldiaren esanahia ulergarriagoa bihurtzen da. Ezagutza baseko informazioaz badakigu "Alex Uranga" eta "Jake Stone" bodyboarder-ak direla eta "Reunion Irla"-ko froga "IBA" txapelketako parte dela. Modu honetan esaldi hau "Bodyboard/kirola" testuinguruan sailka daiteke. Baina entitateak ezagutza basearekin lotzean bi arazo nagusirekin topo egin daiteke.

Demagun testuan "Alex Uranga" agertu beharrean, "Alex","Alex U." edo "Uranga" agertzen dela eta kasu hauetan ez litzateke hain erraza izango dagokion entitatearekin lotzea. Bestalde, "Reunion irla"-k bodyboard txapelketako frogari erreferentzia egiten dio kasu honetan, baina Frantziako departamentua den Irla ere izan daiteke entitate helburua.



Irudia 3.1: Bertan ikus daiteke, testu arrunt bateko entitateen erreferentziak identifikatu eta ezagutza basearekin nola lotzen diren.

Proiektuaren testuinguruan arazo nagusia bigarrena izango da, izen berdinarekin entitate ezberdinak erreferentzia daitezkeela. Wikipediarentzat adibidez "La Gioconda" izenak erreferentzia ditzakeen entitate ezberdinak 3.2 irudian agertzen direnak dira. Izen berdinarekin artelana, modeloa, opera eta liburua erreferentzia daitezke. Aukeren artean, dagokion kasurako entitate egokia aukeratzeko atazari, entitate-izenen desanbiguazioa deritzo.

3.2 Wikipedia

Proiektu honetako entitateen ezagutza basea Wikipediako entitateek osatzen dute. Entitate eta Wikipedia artikuluen arteko lotura zuzena da: Wikipedian, "http://en.wikipedia.org/wiki/Mona_Lisa" artikulua bada, entitatea "Mona_Lisa" izango da.

Wikipedia[1], Wikimedia Fundation[2]-en entziklopedia eleanitza eta eduki askekoa da. Bertako artikuluak mundu osoko erabiltzaileek idazten dituzte eta bakoitzaren identifikadore unibokoa titulua da. Honen bitartez artikuluaren kontzeptua deskribatzen da eta kontzeptuen aldaerak edo formak, 3.2.2 ataleko birbideratze eta 3.2.1 ataleko desanbiguazio orrien bitartez lotzen dira artikulu nagusira.

[1] http://www.wikipedia.org
[2] http://en.wikipedia.org/wiki/Wikimedia_Foundation

Irudia 3.2: "La Gioconda" izenak Wikipedian erreferentzia ditzakeen entitate posibleak.

Artikuluen barnean beste artikuluetara doazen estekak, 3.2.3 ataleko aingura bitartez egiten dira eta artikuluan izen batez identifikatzen dira. Aingura izenak, sarrerak ez bezala errepikatuak egon daitezke artikulu ezberdinetan. Ezaugarri hauek, entitate-izenenkin lan egiteko informazio-iturri gisa oso interesgarriak dira. Gainera entitate-izenen desanbiguaziorako baliabide ezin hobea da Wikipedia.

3.2.1 Desanbiguazio orriak

Wikipediako 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.

3.3 irudian ikus daiteke "Michael Jordan" izenarentzat Wikipediako desanbiguazio orriak eskaintzen dituen artikuluen zerrenda. Saskibaloiko jokalariaz gain, unibertsitateko maisu, futbolari eta politikoaren artikuluetara

Irudia 3.3: Michael Jordan-en Wikipediako desanbiguazio orria.

loturak agertzen dira.

3.2.2 Birbideratzeak

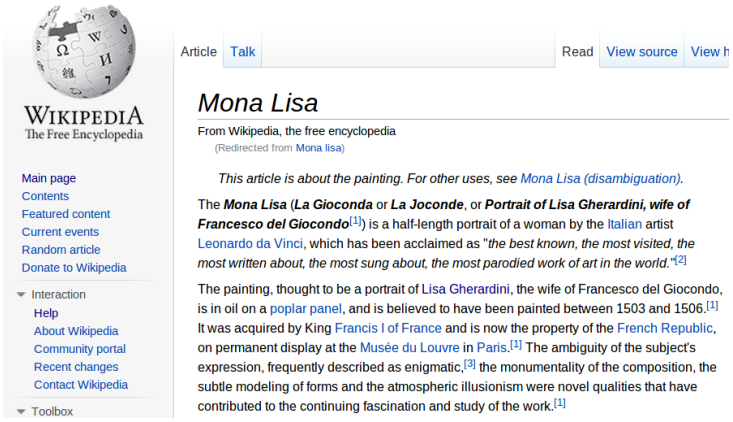
Wikipediaren orrialde asko birbideratze bidez atzitzen dira, adibidez, "UK"[3] artikulua birbideratze artikulu bat da, hain zuzen, "United Kingdom"[4] artikulura. Kasu hauetan, "UK" birbideratze orriak "United Kingdom" orrialdea ebazten duela esaten da. Orri hauen bitartez, entitate berari izen ezberdinen bitartez deitzeko arazoari aurre egiten zaio eta pluralizatze edo erlazionatutako hitzen erabilerari irtenbidea ematen zaie.

3.2.3 Aingurak

Wikipediaren antolaketan, betebehar garrantzitsua daukate artikuluen arteko aingura testuek. Aingurak, Wikipedia artikuluetan agertzen diren hyperestekak dira. Aingura testuan agertzen diren hitzek, gehienetan, erreferentziatu duten artikuluari buruzko informazio esanguratsua eskaintzen dute eta ez dute zertan artikuluaren izen berdina eduki behar. 3.4 irudian ikus daiteke "Mona Lisa"[5]-ren artikuluan, hyper-esteka moduan dauden aingura ezberdinak: Italian, Leonardo da Vinci, Lisa Gherardini ...

Aingura testuetan agertzen diren hitzak aztertuz, "Lisa Gherardini"-k adibidez "Lisa del Giocondo"[6] artikulura erreferentzia egiten du. Propietate hau, oso garrantzitsua izango da proiektuko helburuetako bat betetzeko. Izan ere, Europeanako elementuak Wikipedia artikuluekin aberasteko sistema garatzeko orduan, garrantzia handia emango zaio propietate honi.

[3] http://en.wikipedia.org/wiki/UK
[4] http://en.wikipedia.org/wiki/United_Kingdom
[5] http://en.wikipedia.org/wiki/Mona_lisa
[6] http://en.wikipedia.org/wiki/Lisa_del_Giocondo



Irudia 3.4: Mona Lisaren Wikipedia artikulua eta aingura testuak.

3.3 Wikipediaren 2011-ko Maiatzak 25-eko iraulketa

Wikipediaren ezagutza web-orrialdetik kontsulta daiteke, baina gure ikerketarako erabilera erosoagoa izan dadin informazio guzti hau internetetik jaisteko aukera dago. Wikipediak, web orrian duen informazio guztia XML erraldoi batean eskuratzeko aukera ematen du iraulketa bitartez. Iraulketak, egin diren datarekin identifikatzen dira eta XML formatuan kudeatzen du Wikipedia osoko informazioa. Proiektu honetan erabili den iraulketa 2011ko Maiatzak 25-eko Ingelesezko bertsioa da. 3.5 irudian ikus daiteke XML erraldoi honen lagin bat, informazio hau erabiltzeko beharrezkoa da erauzi eta kudeatuko duen tresna bat, 3.4 ataleko Wikipedia Miner sistema hain zuzen.

3.4 Wikipedia Miner sistema

Wikipedia Miner[7] sistema, Wikipediak eskaintzen dituen ezagutza aberatsak kudeatzeko tresna bat da. Horretarako, aurreko atalean aipatu den iraulketa bat jasoz, informazio hau erauzi eta antolatzen du. Horretaz gain erauzketaren ondoren, informazio guzti hau erabili daiteke kontsultak egiteko. Wikipedia Minerrek, edozein kontsultarekin erlazionatuak dauden Wikipedia artikuluak bueltatzen ditu. Tresna honek, kontsultak linean[8] egiteko aukera mugatua dauka. 3.6 irudian ikus daiteke "Mona Lisa" kontsultarentzat

[7] http://wikipedia-miner.cms.waikato.ac.nz/index.html
[8] http://wikipedia-miner.cms.waikato.ac.nz/demos/search/

```
 </page>
  <page>
    <title>List of Atlas Shrugged characters</title>
    <id>359</id>
    <revision>
      <id>425867426</id>
      <timestamp>2011-04-25T18:41:39Z</timestamp>
      <contributor>
        <username>EricCable</username>
        <id>7125848</id>
      </contributor>
      <minor />
      <comment>/* Dagny Taggart */</comment>
      <text xml:space="preserve">This is a list of characters in [[Ayn
Rand]]'s novel ''[[Atlas Shrugged]].''

==Major characters==
The following are major characters from the novel.&lt;ref&gt;Characters
are listed as &quot;major&quot; if they meet one of the following
criteria:
*they are listed as &quot;major&quot; characters in a widely available
study guide, such as [[CliffsNotes]], [[SparkNotes]], or [[Gale
(Cengage)|Gale's]] ''Novels for Students'';
*they are listed as &quot;primary heroic&quot; or &quot;arch-
villain&quot; characters in Gladstein's ''The New Ayn Rand Companion'';
*they are the focus of an essay in a scholarly book about the novel,
such as ''Essays on Ayn Rand's Atlas Shrugged'' or ''Ayn Rand's Atlas
Shrugged''.&lt;/ref&gt;

===Dagny Taggart===
[[Image:Taylor-Schilling-as-Dagny-Taggart.jpg|thumb|upright=0.56|right|
[[Taylor Schilling]] as Dagny Taggart in the 2011 film.]]
Dagny Taggart is the [[protagonist]] of the novel. She is Vice-
President in Charge of Operations for Taggart Transcontinental, under
her brother, James Taggart. However, due to James' incompetence, it is
Dagny who is responsible for all the workings of the railroad.
......
```

Irudia 3.5: 2011-ko Maiatzak 25-eko Ingelesezko Wikipedia iraulketaren lagina. Bertan "List of Atlas Shrugged characters" artikuluaren informazioa agertzen da. Informazio hau modu eroso batean erabili ahal izateko beharrezkoa da informazio hau erauzi eta kudeatuko duen sistema bat erabiltzea.

itzultzen dituen Wikipedia artikuluak. Bertan ikus daiteke emaitzak desanbiguatuak daudela goitik behera ordenatuak.

Desanbiguazioa [5] artikuluan jasotako metodoaz egiten du eta itzultzen duen artikulu bakoitzari pisu bat esleitzen dio artikuluen artean sailkapen bat eginez. Wikipedia iraulketa eta Wikipedia Miner erausleaz lokalean lan egitea ezinezkoa izan daiteke informazio eta lan karga altuak direla eta. Beraz, lan guztia zerbitzarietan egin da. Hurrengo azpi atalean, erausleak iraulketa nola kudeatu duen azalduko da Wikipedia Miner-en azpiegitura ulertzeko.

Irudia 3.6: Wikipedia Miner-ek linean egindako, "Mona Lisa" kontsultaren emaitzak. Ikus daiteke emaitzak desanbiguatuak itzultzen dituela.

3.4.1 Wikipedia Miner-en azpiegitura

Lehen esan bezala erausleak Wikipediako iraulketa (ikus 3.5 irudia) jasotzen du. XML formatuan dagoen fitxategi erraldoi honetatik informazio guztia jaso eta elkarrekin erlazionatuak dauden fitxategietan oinarritutako azpiegitura batean bihurtzen du. Funtsean, Wikipedia artikuluei zenbaki bat esleitzen die identifikatzaile bezala erabiliko duena. Ondoren, identifikatzaile hauen bitartez CSV (comma-separated values) fitxategiak sortzen ditu artikuluen arteko erlazioekin. Lehenik, artikuluak antolatzeko erabiltzen duen egitura aztertuko da.

Wikipediako artikulu bakoitza "identifikatzailea, titulua, mota" hirukoteaz erreferentzia egiten dio. Mota aldagaiak 4 balio ditu non 1 artikulua den, 2 kategoria, 3 birbideratzea eta 4 desanbiguazio orria. Jarraian dagoen adibidean ikus daiteke "identifikatzaile, titulua, mota" lagin bat desanbiguazio orri, birbideratze eta artikulu batera erreferentzia egiten duten hiru

sarrerekin.

- 19709347,"La Gioconda",4
- 1924249,"La Gioconda (opera)",1
- 10101187,"Gioconda Salvadori",3

Aingura guztiak kudeatzeko, "aingura, helburu_identifikatzaile, frekuentzia" patroiaz baliatzen da. Honela aingura eta artikulu arteko lotura identifikadore bitartez burutzen du. Azpian ikus daiteke aingura ezberdinetatik desanbiguazio orrira edo artikulura loturak eta hauen frekuentziak. Adibidez ""La Gioconda",1924249,189"-rekin "La Gioconda" aingurarekin 1924249 identifikatzailea duen "La Gioconda (opera)" artikulura 189 aldiz erreferentzia egin dela esan nahi du. Hemen ikus daiteke, Wikipedia Miner sistemak, 3.2.3 atalean aipatu den propietate garrantzitsua nola kudeatzen duen. Gainera propietatea neurtzen duen frekuentzia gehitzen dio erlazioari.

- "La Gioconda",19709347,5
- "La Gioconda (disambiguation)",19709347,0
- "La Gioconda",1924249,189
- "La Gioconda (film)",1924249,0
- "La Gioconda (opera)",1924249,5

Birbideratze eta Desanbiguazio orrien kasuetan, "identifikatzaile, helburu_identifikatzaile" bezala kudeatzen ditu. Adibidez, 19709347 "La Gioconda" desanbiguazio artikulua, 1924249 "La Gioconda (opera)" artikulura bideratua dagoela adierazteko "19709347,1924249" errepresentazioa erabiltzen du. Birbideratzeak modu berdinean kudeatzen ditu.

Azpiegitura honekin, Wikipedia artikuluek elkarren artean dituzten loturak ebatziz informazio asko atera daiteke. Bide batez, helburuetan aipatu den sistema garatzeko, azpi egitura hau erabili da. Honetarako, aingura eta entitateen arteko erlazioak dituen hiztegi bat erabili da. Hiztegi hau izango da desanbiguazio metodoa garatzeko oinarri nagusia. Jarraian hiztegi hau eraikitzeko aurrera eraman diren pausuak azalduko dira.

3.5 Standford-UBC sistema, hiztegia eraikitzen

Europeanako elementuak Wikipedia artikuluekin aberasteko sistema bat garatzeko, lehen pausua hiztegi bat sortzea izan da. Hiztegia eraikitzeko [10] artikuluan aipatzen diren oinarriak erabili dira. Ainguretan oinarrituz, artikulu batekiko duen erlazioaren probabilitateaz sorturiko hiztegia sortzen da. Aingura eta artikuluaren arteko erlazioa, aingura hori, artikulu batera erreferentzia moduan agertu den kopuruaren kontaketaz puntuatzen da. Informazio guzti hau, Wikipedia Miner-ek erauzi duen azpiegituran aurkitzen da lehen aipatu den bezala.

Hiztegiaren egitura, lerroko, aingura eta entitate zerrenda dira hutsuneaz banandua. Entitate zerrenda, aingura-testu berdinetik erreferentzia duten entitate ezberdinez osatzen da. Entitate zerrendako entitate bakoitza entitate kontaketa parea da ":" bereizia. Entitate eta aingura-izenak azpimarraz bereiziak doaz hitz bat baino gehiago duten kasuetan. 3.1 taulan ikus daiteke hiztegiaren egitura lagin batean.

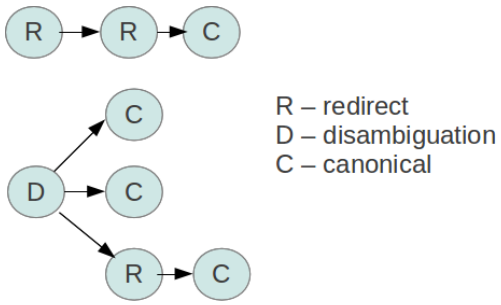
aingura	entitatea:kontaketa zerrenda
la_gioconda	**La_Gioconda_(opera):203** Mona_Lisa:7 Lisa_del_Giocondo:7 ...
mona_lisa	**Mona_Lisa:364** Mona_Lisa_(film):81 Mona_Lisa_(Nat_King_Cole_song):52 ...
the_mona_lisa	Mona_Lisa:6
mona_lisa_albums	Mona_Lisa_albums:1

Taula 3.1: Hiztegiaren lagina eta zuzeneko desanbiguazio adibideak. Bertan ikus daiteke aingura-testu berdinarentzat erreferentzia duten artikuluak. "la_gioconda" aingura 203 aldiz erabili da "La_Gioconda_(opera)" entitateari erreferentzia egiteko.

3.1 taulan ikus daitekeen hiztegiaren laginean, zuzenean desanbiguatu daiteke izen bat. Adibidez testu batean agertu den "La Gioconda" izena hiztegian aingura moduan bilatuz "la_gioconda", kontaketa altuena duen "La_Gioconda_(opera)" entitatea bueltatuz egin daiteke ataza hau. Edo "mona_lisa" izenarentzat adibidez, 364 alditan erreferentzia duen artelana bueltatuz. Hiztegia eraikitzean, gerta daiteke, aingura batek birbideratze edo desanbiguazio orrialde batera erreferentzia egitea. Hau gertatzen den

kasuan, bideratze hauek ebatzi behar dira hiztegia entitate kanonikoez soilik osatua egon dadin.

3.5.1 Entitate kanonikoa



Irudia 3.7: Birbideratze eta desanbiguazio orrien ebazpenean jarraitu beharreko pausoen adibideak ikus daitezke irudian. Goiko atalean birbideratze (R) batetik, birbideratze batera lotura eta azken honetatik artikulu kanonikora (C) loturak agertzen dira. Azpiko atalean berriz, desanbiguazio (D) orri bateko aukeretan, bi artikulu kanonikoetara eta birbideratze batera erreferentzia ikus daiteke.

Hiztegiko entitate zerrendetan birbideratze eta desanbiguazio orriak ager ez daitezen, hauek ebatzi behar dira entitate kanonikoa bilatuz. Entitate kanonikoa, 3.7 irudian ikusten diren zuhaitzetako hostoetako entitateak dira. Ebazpen hau burutzeko Wikipedia Miner-ek erauzitako fitxategiak erabili dira birbideratze edo desanbiguazio orrietatik, artikulu kanonikoetara loturak dituztenak hain zuzen. Hasiera batean birbideratze orriaren helburu artikulua jartzea nahikoa zela pentsatu zen, baina baliteke birbideratze orri batek beste batera erreferentzia egitea eta azken honek artikulu kanonikora. Edo are eta okerrago, birbideratze bat desanbiguazio orri batera. 3.7 irudian ikus daiteke entitate edo artikulu kanonikoaren bilaketarako posibleak izan daitezkeen bi egoera.

Artikuluen kanonikoen ebazpenerako, artikuluen arteko dependentzia zuhaitzak sortu dira birbideratze eta desanbiguazio orrietan oinarrituak.

3.6 Europeana eta Paths

Orain arte, Europeanako elementuei buruz eta hauei esleitu beharreko Wikipedia artikuluei buruz solastu da. Jarraian, Europeana-ri buruzko datu batzuk azalduko dira eta Wikipedia artikuluez aberasteko arrazoiaren bultzatzailea izan den PATHS proiektua azalduko da.

Europeana[9] Europear batasuneko kide diren 27 estatuen instituzio kulturalen kontribuzio digitalen liburutegia da. Atzipen askea eskaintzen du bere eduki guztietara eta 2008ko azaroaren 20 jaio zen. Bere edukien artean, liburuak, pelikulak, arte-lanak, egunkariak, multimedia fitxategiak, mapak eta beste hainbat elementu aurki daitezke bilduma ezberdinetan banatuak. Honen helburua, web orriaren bitartez, erabiltzaileei proiektu honetako parte diren herrialdeen kultura aberastasuna eskura jartzea da.

Proiektuaren helburuetan azaldu den urre-patroia (ausaz aukeraturiko Europeana elementuen lagin etiketatua) osatzen duten elementuak, Scran eta Culture Grid izeneko bildumetatik ausaz hartu dira. Scran bilduma 310.800 elementuz osatua dago, fitxategi hauek, museo, galeria eta beste hainbat iturri ezberdinetatik gehituak dira. Culture Grid bilduma berriz, Erresuma batuko 40 bilduma ezberdinetako 547.000 elementuz osatua dago. Europeanako elementuak Europeana Semantic Elements Specifications (ESE) bidez kudeatzen dira, hauek hurrengo azpi-atalean azalduko dira.

3.6.1 Europeana Semantic Elements Specifications (ESE)

```
<record>
  <dc:identifier>http://www.picturethepast.org.uk/frontend.php?keywords=
                 Ref_No_increment;EQUALS;NCCW001197</dc:identifier>
  <europeana:uri>http://www.europeana.eu/resolve/record/09405/
                 C052AA1727D9C258801CF676473953A0861A47C0</europeana:uri>
  <dc:title>The Major Oak</dc:title>
  <dc:source>Picture the Past OAI feed</dc:source>
  <dc:contributor>North East Midland Photographic Record</dc:contributor>
  <dc:description>The largest Oak tree in England, perhaps in the world,
                  this famous tree has withstood lightning, the drying-out
                  of its roots and even a recent fire. The hollow tree has
                  a circumference of 32 feet and the spread of its branches
                  makes a ring 260 feet round.</dc:description>
  <dcterms:isPartOf>Picture the Past</dcterms:isPartOf>
  <dc:language>EN-GB</dc:language>
  <dc:publisher>North East Midland Photographic Record</dc:publisher>
  <dc:subject>Robin_Hood</dc:subject>
  <dc:type>Image</dc:type>
  <dc:format>JPEG/IMAGE</dc:format>
  <europeana:provider>CultureGrid</europeana:provider>
  <europeana:hasObject>true</europeana:hasObject>
  <europeana:country>uk</europeana:country>
  <europeana:type>IMAGE</europeana:type>
  <europeana:language>en</europeana:language>
</record>
```

Irudia 3.8: Europeanako ESE fitxategi baten adibidea. "The Mayor Oak" elementuaren titulu, deskribapen eta irudiari buruzko informazioarekin.

[9] http://www.europeana.eu/portal/

ESE fitxategiak XML formatuan daude eta 3.8 irudian ikus daitekeen itxura dute. Fitxategietan <record> etiketen bidez bildumetako elementuak banatzen dira. Etiketa garrantzitsuenak aztertuz, <dc:identifier> etiketak elementu bakoitzaren identifikatzailea dauka, Europeana osoan bakarra izango da eta elementua gainontzeko elementuetatik bereiziko du. <dc:title> eta <dc:description> etiketetan berriz, titulua eta deskribapenak daude.

3.6.2 PATHS (Personalized Access To cultural Heritage Spaces)

PATHS[10] proiektua, Europeana bezalako ondare kulturaleko bilduma erraldoietan, bilduma digitalen arteko bisitaldi gidatu eta pertsonalizatuak eskaintzeko aplikazio bat garatzean datza. Informazio kantitate handietan erabiltzaile arruntak galtzeko duen erraztasunaren aurrean soluzio bezala agertu zen Europeana-Paths lotura hau. Paths edo bisitaldi gidatu hauen bitartez, erabiltzaileei ondare kultural guztia aztertzeko bide berriak eskaintzea da helburua.

Bisitaldi baten ibilbidean erlazionatutako elementuak eta beste erabiltzaileek egindako bisitaldiak ere gomendatuko dira. Ibilbideak, modu ezberdinetan antola daitezke: gertakari historiko baten inguruan (gerra hotza), artelan eta egile baten inguruan (Picasso eta bere artelanak), toki baten inguruan (Venezia) edo pertsona baten inguruan (Muhammad Ali) adibidez.

Ibilbide honetako elementu bakoitza Wikipedia artikuluekin aberastea proposatu zen eta horretarako Wikipedia Miner-ek elementu bakoitzarentzat proposatutako artikuluak, elementuari txertatu zitzaizkion ESE fitxategian. Horretarako, Wikipedia Miner-i elementu bakoitzak *dc:title*, *dc:description*, *dc:subject* eremuan duten testua eman eta irteeran bueltatutako artikuluak gehitu ziren. ESE eta Wikipedia Miner-ek esleitutako artikuluen txertaketatik ESEPaths formatu berria (ikus 3.9) sortu zen.

ESEPaths fitxategi berrietako Wikipedia artikuluak Wikipedia Miner-ek esleitutako pisuaz datoz (confidence). Gainera zein eremutatik (field) eta eremu honetako zein hitz-multzotik (hasiera eta bukaera markekin definituak) esleitu diren gehitzen du. 3.9 irudian ikus daiteke hirugarren dagoen "http://en.wikipedia.org/wiki/Major Oak" artikulua, tituluko 4-13 karaktereen artetik esleitu dela, "Major Oak" azpimultzotik hain zuzen.

[10] `http://group.europeana.eu/web/guest/details-paths/`

```
<record>
 <dc:identifier>http://www.picturethepast.org.uk/frontend.php?keywords=
               Ref_No_increment;EQUALS;NCCW001197</dc:identifier>
 <europeana:uri>http://www.europeana.eu/resolve/record/09405/
           C052AA1727D9C258801CF676473953A0861A47C0</europeana:uri>
 <dc:title>The Major Oak</dc:title>
 <dc:source>Picture the Past OAI feed</dc:source>
 <dc:contributor>North East Midland Photographic Record</dc:contributor>
 <dc:description>The largest Oak tree in England, perhaps in the world,
                 this famous tree has withstood lightning, the drying-out
                 of its roots and even a recent fire. The hollow tree has
                 a circumference of 32 feet and the spread of its branches
                 makes a ring 260 feet round.</dc:description>
 <dcterms:isPartOf>Picture the Past</dcterms:isPartOf>
 <dc:language>EN-GB</dc:language>
 <dc:publisher>North East Midland Photographic Record</dc:publisher>
 <dc:subject>Robin_Hood</dc:subject>
 <dc:type>Image</dc:type>
 <dc:format>JPEG/IMAGE</dc:format>
 <europeana:object>http://www.peoplesnetwork.gov.uk/dpp/resource/2210977/
                   stream/thumbnail_image_jpeg</europeana:object>
 <europeana:provider>CultureGrid</europeana:provider>
 <europeana:hasObject>true</europeana:hasObject>
 <europeana:country>uk</europeana:country>
 <europeana:type>IMAGE</europeana:type>
 <europeana:language>en</europeana:language>
 <paths:background_link source="wikipedia" start_offset="0" end_offset="9"
  field="dc:title" field_no="0" confidence="0.017"
  method="wikipedia-miner-1.2.0">http://en.wikipedia.org/wiki/TheMajor(Hellsing)
 </paths:background_link>
 <paths:background_link source="wikipedia" start_offset="4" end_offset="9"
  field="dc:title" field_no="0" confidence="0.017"
  method="wikipedia-miner-1.2.0">http://en.wikipedia.org/wiki/Major
 </paths:background_link>
 <paths:background_link source="wikipedia" start_offset="4" end_offset="13"
  field="dc:title" field_no="0" confidence="0.130"
  method="wikipedia-miner-1.2.0">http://en.wikipedia.org/wiki/Major Oak
 </paths:background_link>
 <paths:background_link source="wikipedia" start_offset="10" end_offset="13"
  field="dc:title" field_no="0" confidence="0.231"
  method="wikipedia-miner-1.2.0">http://en.wikipedia.org/wiki/Oak
 </paths:background_link>
 <paths:background_link source="wikipedia" start_offset="10" end_offset="13"
  field="dc:title" field_no="0" confidence="0.017"
  method="wikipedia-miner-1.2.0">http://en.wikipedia.org/wiki/Quercus robur
 </paths:background_link>
</record>
```

Irudia 3.9: ESEPaths fitxategi baten adibidea. "The Mayor Oak" elementuaren titulu, deskribapen eta irudiari buruzko informazioaz gain berrikuntza moduan background linkak agertzen dira. Bertan Wikipedia Miner-ek titulutik proposaturiko artikuluak agertzen dira.

3.7 Aurrekarien azterketaren laburpena

Aurrekarien atala amaitzeko, helburuak berrikusiko ditugu eta hauek burutzeko aurrekariek eskaintzen dutena laburbilduko da.

Lehen helburua, Europeanako elementuak, sistema automatiko bidez, Wikipedia artikuluekin aberastea posible den ikertzea eta ebaluatzea da. Hau egiteko Europeanako elementuak Wikipedia Miner-ekin aberastu ditugu.

Bigarren helburua, Europeanako elementuak Wikipedia artikuluekin aberasteko sistema bat garatzea eta ebaluatzea da. Honetarako, sistemaren oinarria izango den hiztegia prest dago.

Oraindik sistema hauek ebaluatzeko urre-patroia sortu eta ebaluatu behar da eta garatuko den sistemaren oinarriak besterik ez dira azaldu. Hurrengo atalean, urre-patroia nola sortu eta ebaluatu den azalduko da. Ondoren sistemaren garapena azalduko da eta azkenik bi sistemak (Wikipedia Miner eta garatu dena) urre-patroiaren aurka ebaluatuko dira.

4 Kapitulua

Esperimentuaren diseinua

Europeanako elementuak, sistema automatiko bidez, Wikipedia artikuluekin aberastea posible den ikertzeko egingo den esperimentuaren diseinua azalduko da atal honetan. Baina lehenik esperimentuak burutzeko behar diren datuak prestatu behar dira.

Esperimentu guztiak Europeanako 400 elementuren inguruan egingo dira. Lehenik ausaz aukeratutako dira elementuak eta eskuz etiketatuko dira dagokion Wikipedia artikuluarekin. Eskuz eginiko etiketatua elkarrengandik independenteak diren 3 pertsonako bi taldek burutu dituzte. Alde batetik Donostiako EHU-ko taldeak etiketatutakoa dago eta bestetik Sheffield-eko taldeak eginikoa. Beraz, etiketatu bikoitza dago, elementu bakoitzarentzat bi artikulu. Donostian egindako etiketatua urre-patroi moduan erabili da eta Sheffield-ekoa berriz, urre-patroiaren ebaluaziorako. Etiketatuan, elementuaren instantzia definitzen duen artikulua aurkitu nahi izan da eta ez erlazionatutako artikulua. Artikuluak, ondare kulturaleko elementuak deskribatzen duena deskribatu behar du.

Bestalde, Europeanako elementuak Wikipedia artikuluekin aberasteko sistemaren garatzeko oinarriak ezarri dira aurreko atalean. Atal honetan sistema hau nola garatu den azalduko da.

Esperimentua egiteko behar den guztia azaldu ondoren, kapitulu honen amaieran esperimentuaren pausoak azalduko dira.

4.1 Eskuzko etiketatua: Urre-patroiaren sorkuntza eta ebaluazioa

Ausaz Europeanako 400 elementu hartuz, esperimentu guztietarako elementuak finkatu ziren. Elementu hauek bi bilduma ezberdinetatik hartu ziren Scran bildumatik erdia, 200 elementu eta Culture Grid bildumatik beste er-

dia. Guztiak, aipaturiko bi bilduma hauen gainetik dagoen bilduma handiago bateko kide dira. Horrela eskuzko etiketatua erraztuko dute, elkarren artean erlazionatuak baitaude. 4.1 taulan ikus daitezke ausazko elementu ezberdinen mota eta kopurua. Bertan ikusten den moduan gehienak argazkiak diren arren, txanponak eta margolanak ere ageri dira.

Mota	kopurua
Argazkiak	276
Txanponak eta tresnak	57
Liburuak edo liburuxkak	24
Besteak	21
Margolanak	14
Ikus entzunezkoak	8
Guztira	400

Taula 4.1: Ausaz harturiko Europeanako 400 elementuen motak

4.1.1 Eskuz eginiko etiketatuaren metodologia

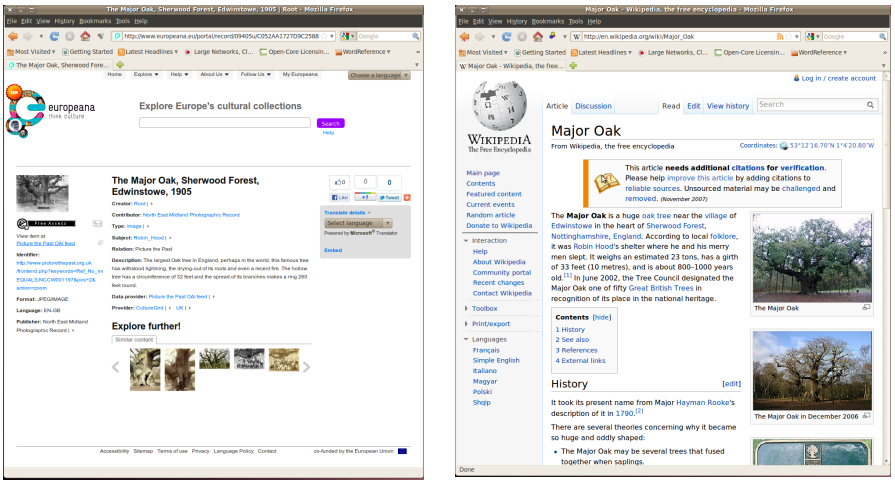
Etiketatua burutzeko elementu bakoitzari eskuz artikulua esleitu beharra zegoen. Horretarako, Europeanako ESE fitxategietako informazio osoa erabili dute bi taldeetako etiketatzaileek, titulutik hasita, argazki eta deskribapen testuetaraino. Etiketatzaileen lana elementu hauei Ingelesezko Wikipediako artikulua lotzea da eta artikulurik esleitzeko ez dagoenean NIL bueltatzea. Ataza honen emaitza, 4.2 taulan ikusten den itxura duten bi fitxategi dira, talde bakoitzeko bat. Hauek elementuaren identifikatzailea <dc:identifier> eta NIL edo artikulua dute.

<dc:identifier>	NIL edo Artikulua
000-000-701-088-C	http://en.wikipedia.org/wiki/Reo_Speed-Wagon
000-299-995-321-C	NIL
000-000-577-216-C	http://en.wikipedia.org/wiki/Heart_scarab
000-000-559-364-C	NIL

Taula 4.2: Eskuz eginiko etiketatuaren lagina, irudian ikus daiteke lau elementuetako birentzat artikulua esleitu dela eta beste bien kasuan NIL. NIL-ak esleitzeko artikulurik ez dagoela islatzen du.

Etiketatua burutzeko irizpide nagusia hau da: "Wikipedia artikuluak eta elementuak objektu bera deskribatu beharra dute. Argazkien kasuan, artikulua subjektuari buruzkoa izan behar du , hau da, argazkiko pertsonaia edo

lekua". Arau hau jarraituz, Europeana elementu bati esleitu zaion Wikipedia artikulua ikus daiteke 4.1 irudian.



Irudia 4.1: Europeanako "The Major Oak"elementua ezkerrean eta honi dagokion Wikipedia artikulua eskuinean

Atalaren hasieran aipatu bezala, Donostian egindako etiketatua urre-patroi moduan erabili da eta Sheffield-ekoa berriz, eskuz eginiko etiketatzaileen arteko adostasuna ebaluatzeko. Etiketatzaileek, nahi gabe birbideratze edo desanbiguazio orrialde bat esleitu duten kasurako 3.5.1 ataleko metodo berdinarekin artikulu kanonikoak ebaztu dira.

4.1.2 Urre-patroiaren etiketatuaren emaitzak

Eskuz eginiko etiketatuaren ondoren, urre-patroiko 400 elementuetatik 89-k dute artikulua esleitua %22-a, beraz NIL kasuak asko dira %78 hain zuzen (ikus 4.3 taula). Kontuan hartuta etiketatua burutzeko irizpidea zorrotza dela etiketatzaileak gai izan dira elementu sorta batentzat artikulua esleitzeko. Honek esan nahi du ondare kulturaleko elementuak Wikipediako artikuluekin aberastea posible dela %22-ak informazio gehigarria lortu dezakeelako.

Bildumak	Scran	Culture Grid	Guztira
Elementu kopurua	200	200	400
NIL	140	171	311
Artikulua dute	60	29	89

Taula 4.3: Urre-patroiaren elementuen kontaketak. Aipagarria da 400 elementuetatik 311 NIL direla.

Jarraian urre-patroia ebaluatuko da Sheffield-eko taldearen etiketatuaren aurka, modu honetan, etiketatzaileen arteko adostasuna neurtuko da. Bide batez etiketatzeko irizpidea egokia den jakiteko.

4.1.3 Etiketatzaileen arteko adostasuna

Atal honetan, etiketatzaile taldeen artean dagoen adostasuna aztertuko da, hau da, EHU-ko taldearen urre-patroia eta Sheffield-eko etiketatuaren arteko adostasuna. 4.4 taulan ikus daitekeen moduan bi taldeen artean %92.5 eta %80.0-ko adostasun portzentaje altuak lortu dira, Culture Grid eta Scran bildumetan hurrenez hurren. Balio hauek etiketatzaileek artikulu berdina edo biok NIL bezala etiketatu diren kasuen kontaketak dira. Beraz esan daiteke artikulua esleitu edo NIL esleitu erabakitzean adostasuna altua dela.

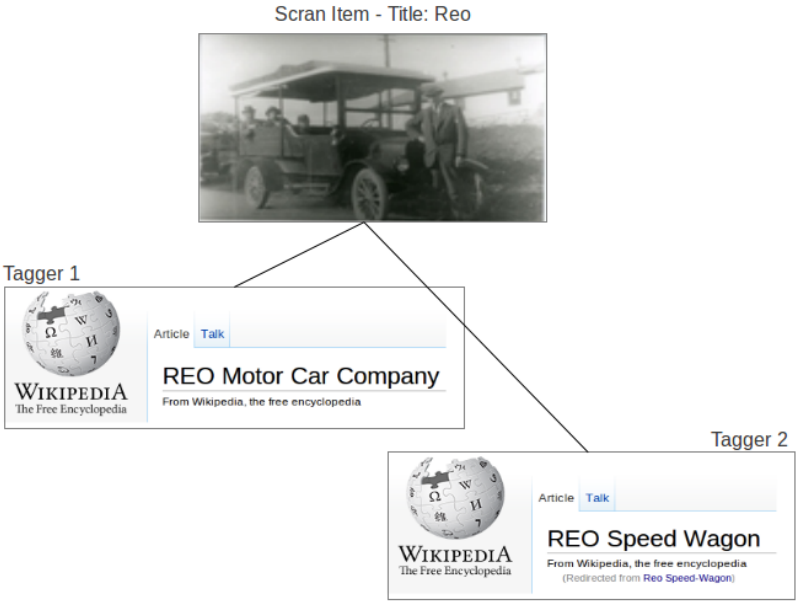
NIL kasuak asko direla kontuan hartuta, etiketatzaileek artikulua esleitu duten 58 kasuen artean (22 Culture Grid-en eta 36 Scran-en) %90.9 eta %94.4-ko adostasun aipagarria lortu da.

	Scran	**Culture Grid**
Adostasuna guztira	80.0%	92.5%
Adostasuna: Biok NIL	126	165
Desadostasuna: Batek NIL	38	13
Adostasuna: Artikulu berdina	34	20
Desadostasuna: Artikulu desberdina	2	2

Taula 4.4: Etiketatzaileen arteko adostasunaren emaitzak: Lehen lerroak 400 itemen arteko adostasun portzentajea erakusten du. Bigarren eta hirugarren lerroek NIL kontaketak. Azken bi lerroek, artikuluak aukeratu diren item-en kontaketak erakusten dituzte.

Adostasun maila altua izan den arren, bi taldeek artikulu desberdinak aukeratu dituzten 4 (2 Scran-en eta 2 Culture Grid-en) kasu daude. Jarrian horietako bat aztertuko da. Elementu hau 'REO' izeneko kamioi baten argazkiari buruzkoa da eta 4.2 irudian ikus daiteke etiketatzaileen desadostasuna. EHU-ko etiketatzaileek Reo Speed-Wagon[1] izeneko kamioi modeloa bezala etiketatu dute. Sheffield-eko taldeak berriz, REO Motor Car Company[2] bezala, kotxe hau komertzializatu zuten enpresako artikuluarekin. Esan beharra dago bi artikuluak onargarriak direla etiketatzeko irizpidearekiko eta elkarren artean oso antzekoak. Beraz, etiketatzeko irizpidea zorrotza izan arren, ikusi da interpretazio ezberdinak eduki ditzakeela.

[1] `http://en.wikipedia.org/wiki/ReoSpeed-Wagon`
[2] `http://en.wikipedia.org/wiki/REO_Motor_Car_Company`



Irudia 4.2: Etiketatzaile taldeen arteko desadostasun adibidea: "REO" elementuarentzat etiketatzaile talde bakoitzak artikulu ezberdina aukeratu duen arren elkarren artean oso antzerakoak dira.

4.1.4 Eskuz eginiko etiketatuaren analisia

Etiketatzeko irizpidea hau dela gogoratuz: "Wikipedia artikuluak eta elementuak objektu bera deskribatu beharra dute. Argazkien kasuan, artikulua subjektuari buruzkoa izan behar du , hau da, argazkiko pertsonaia edo lekua". Eskuz eginiko etiketatuaren analisia eginez, elementuaren arabera etiketatzeko irizpideak interpretazio ezberdinak izan ditzakeela ikusi da. Beraz zorroztasun maila bat gehitzea beharrezkoa dela pentsatu da ondorengo irizpideak gehituz:

- Elementua txanpon bat denean, Wikipedia artikuluak txanpona bera deskribatu behar du eta bestela NIL etiketatu. Adibidez txanpon honentzat "Denarius, of Lucius Marcius Censorinus"[3] deskribatzen duen artikulua[4] existitzen da eta etiketatu egokia litzateke. Irizpide hau elementu mota guztietara zabaltzen da, hauek kontzeptu baten instantzia moduan kontsideratuz.

- Toki edo pertsona baten argazkia denean, hauek anonimoak diren ka-

[3] http://www.europeana.eu/portal/record/00401/2FCF4C116A23D5F179CEE72DC9CAEE2A02721F79.html

[4] http://en.wikipedia.org/wiki/Denarius_of_L._Censorinus

suan ez dira etiketatuak izango. Adibidez "futbolaria", "eliza", "hiria" edo "oinezkoak" bezalako kasuei dagozkien artikuluak ez dira onartuko.

- Elementuaren tituluan pertsona eta toki baten agerpena baldin badago, lehentasuna pertsonaren artikuluak izan beharko du. Hala ere etiketatu bikoiztua egitea planteatu daiteke.

Irizpide berri hauek etorkizunean etiketatzeko pentsatuak dauden arren, urre-patroian eta Sheffield-eko taldearen etiketatuan aplikatu dira. Izan ere lehen etiketatua burutu ondoren etiketatzaileek elementu batzuentzat etiketa zalantzan jarri baitzuten. Beraz zalantza hauek argitzeko definitu ziren irizpide berriak. Atal honekin amaituz, urre-patroia prest dago sistema automatikoek itzuliko dituzten emaitzak ebaluatzeko. Hurrengo atalean, garatuko den sistemaren inplementazioa azalduko da eta aurrerago Wikipedia Miner-ekin batera ebaluatuko da.

5 Kapitulua

Europeanako elementuak Wikipedia artikuluekin aberasteko sistemaren garapena

Aurrekarietako 3.5 atalean, aingura-entitate hiztegia nola eraiki den azaldu da. Honekin jarraituz sistema automatikoek 400 elementuen ESE fitxategietako <dc:title> eremuetako testua sarrera bezala jasoko dutela aipatu da. Garatuko den sistemak, titulu honetatik, hiztegian agertzen diren aingurak aurkitu behar ditu. Ondoren sistemaren emaitza hau izango da: aingura hauei hiztegian lotuak dauden hautagaien zerrendako entitateak. Lehenengo pausua beraz tituluan hiztegiko aingurak bilatzea izango da.

5.1 Elementuen tituluetan hiztegiko aingurak identifikatzen

Ataza honetan, esperimentuetarako erabili diren 400 item-en tituluetan hiztegiko aingurak aurkitzea da helburua. Horretarako perl-en programatutako script bat erabili da. Honek, sarrera bezala jasotzen duen testu batean aingurak bilatzen ditu tituluko azpi-kate guztietan. Demagun, "Leonardo da Vinci margolari famatuaren Mona Lisa artelana" titulua duen elementua aztertu nahi dugula. Titulu honetan aurkituko lituzkeen aingurak, "Leonardo da Vinci","Leonardo", "Mona Lisa", "Lisa" eta "Mona" dira, guzti hauek aingura bat baitute hiztegian. Ondoren aingura bakoitzari hiztegian lotuak dauden entitateen zerrendak elementu honen entitate posible bezala itzuliko lituzke sistemak.

Beraz, azken pausu honekin, Europeanako elementuak Wikipedia artikuluekin aberasteko sistema bat osatu da. Baina oraindik, Wikipedia Miner-

ekin lehiatzeko esleitzen dituen entitateak pisatu behar dira.

5.2 Hiztegiak bueltatzen dituen entitateak pisatzeko metodologia

Elementu bakoitzarentzat, sistemak itzultzen dituen artikuluak, pisatu nahi dira. Balio hauek esleitzeko [6] artikuluan azaltzen den algoritmoan oinarrituak egongo dira pisuak. Lehenik algoritmo hau azalduko da eta ondoren sistemarentzat erabili dena.

5.2.1 Xianpei Han eta Lee Sun-en algoritmoa

Entitate-izenenen desanbiguaziorako algoritmo honek, c testu batean agertzen den s izen aipamenetik, hautagaiak izan daitezkeen $e1, e2, e3...$ entitateak sortzen ditu. Hautagaiak sortzeko hiztegi bat erabiltzen du. Hautagaiak izan daitezkeen entitate bakoitzeko, hiru probabilitateren biderkaduraz sailkapen bat egiten du. Sailkapen hau egiteko erabiltzen diren probabilitateen balioak gure sistemak entitateei pisu bat esleitzeko erabiliko dira.

- Lehen biderkagaiak, e entitatea zein ospetsua den neurtzen du. Entitate batzuk beste batzuk baino ospetsuagoak dira ezagutza basean. Adibidez Wikipedian "Karlos Argiñano", "Juanito Oiarzabal" baino ospetsuagoa da. Beraz adibide honetan "Karlos Argiñanok" probabilitate altuagoa lortuko du. Hau $P(e)$ probabilitate banaketa bezala izendatuko da.

- Bigarren biderkagaiak, s izenak e entitatea izendatzeko duen probabilitatea neurtzen du. Hau da, entitate hori aipatzeko erabili den izena askotan erabilia izan baldin bada probabilitate handiagoa izango du atal honetan. Hau $P(s|e)$ probabilitate banaketa bezala izendatuko da.

- Hirugarren biderkagaiak c testu batean e entitateak azaltzeko duen probabilitatea neurtzen du. Sukaldaritzari buruz hizketan dabilen testu batean "Karlos Argiñanok"probabilitate handiagoa lortuko du beste edozeinetan baino. Hau $P(c|e)$ probabilitate banaketa bezala izendatuko da.

Lehen esan bezala hiru probabilitateen biderketa bidez $P(s, c, e) = P(e)P(s|e)P(c|e)$ hautagaiak diren entitateak sailkatuko dira. Probabilitate altuena lortzen duena izango da s izenarekin c testuan agertu den aipamenari dagokion e

entitatea. Beraz entitate irabazlea aurreko formularen maximoa lortzen duena izango da.

$$e = \underset{e}{\operatorname{argmax}}\, P(s, c, e) = \underset{e}{\operatorname{argmax}}\, P(e)P(s|e)P(c|e)$$

5.2.2 Pisuak esleitzeko sistema propioa

Entitate bakoitzari pisu bat esleitzeko oinarria aurreko ataleko algoritmoaren probabilitateetan dago. Baina aldaketa txiki batzuekin. Elementu bakoitzarentzat dugun informazio guztia honen titulua da, ez dugu testuingururik. Beraz, pisatzeko ditugun baliabideak, hiztegiko aingura bezala identifikatu den tituluko s azpi-katea eta hiztegitik s-rentzat sortu diren e entitate zerrenda dira.

Beraz pisatzeko algoritmoaren formula $P(s, e) = P(e)P(s|e)$ laburtua izango da.

$P(e)$ kalkulatzeko formula hau erabiliko da:

$$P(e) = \frac{Count(e)}{|M|}$$

$Count(e)$ bidez, e entitateak ezagutza basera erreferentzia duen aldien kontaketa da, $|M|$ kontaketa guztien batura. Balio guzti hauek hiztegitik lortu daitezke. $Count(e)$ kalkulatzeko, entitate hau edozein ainguratik agertu den kontaketen batura da. $|M|$ berriz entitate guztiek erreferentzia duten aldien batura.

$P(s|e)$ kalkulatzeko formula hau erabiliko da:

$$P(s|e) = \frac{Count(e, s)}{\sum_{s} Count(e, s)}$$

$Count(e, s)$ bidez, s izenarekin e entitateari erreferentzia egin den aldien kontaketa da. $\sum_s Count(e, s)$, e entitatea s izen ezberdinekin guztira agertu den aldien kontaketa da, beraz $Count(e)$-ren berdina. Beraz $P(s|e)$-ren formula hau litzateke aldaketak eginez:

$$P(s|e) = \frac{Count(e, s)}{Count(e)}$$

Azkenik $P(s, e) = P(e)P(s|e)$ kalkulatzeko aurreko bi formulak biderkatzea besterik ez da gelditzen:

$$P(s,e) = P(e)P(s|e) = \frac{Count(e)}{|M|}\frac{Count(e,s)}{Count(e)}$$

$Count(e)$ sinplifikatuz formularen azken itxura:

$$P(s,e) = P(e)P(s|e) = \frac{Count(e,s)}{|M|}$$

$P(s,e)$ izango da, hiztegiko aingura bezala identifikatu den elementuaren tituluko s azpi-katetik eratorri den e entitatearen pisua. Formula hau ezagunagoa da, "egiantza handieneko estimazio" izenaz.

5.2.3 Pisaketaren adibidea

Adibidez, demagun elementu baten tituluan "la Gioconda" s azpi-katea aurkitu dela eta hau hiztegiko aingura da. Hauek dira hiztegiak proposatzen dituen e entitateak: "Mona Lisa" artelana, "La Gioconda, opera" eta "Lisa del Giocondo" artelaneko modeloa. Jarraian entitate bakoitzaren $P(e)$ pisaketak:

1. Mona Lisa $P(e) = 4.11 * 10^{-06}$
2. La Gioconda, opera $P(e) = 2.05 * 10^{-06}$
3. Lisa del Giocondo $P(e) = 2.97 * 10^{-07}$

Pisaketa irizpide hau bakarrik kontuan izanda, Mona Lisa edo La Gioconda arte lanari dagokion entitateak du pisu handiena. Esan beharra dago, sailkapen honetan lehenengo entitateak bigarrenaren probabilitatea bikoizten duela. Hau normala da, arte lanak operak baino askoz gehiagotan du erreferentzia Wikipedian. Ondoren $P(s|e)$ soilik erabiliz pisaketen balioak:

1. La Gioconda, opera $P(s|e) = 2.15 * 10^{-07}$
2. Lisa del Giocondo $P(s|e) = 5.11 * 10^{-08}$
3. Mona Lisa $P(s|e) = 3.69 * 10^{-09}$

Bertan ikus daiteke "la Gioconda" izenarekin ezagunagoak direla opera eta modeloa artelana baino. Esan beharra dago, arte lana "Mona Lisa" izenarekin ezagunagoa dela eta hau da emaitza hauen zergatia. Azkenik bi banaketak konbinatuz, hauek dira pisuen azken balioak:

1. La Gioconda, opera $P(s,e) = P(e)P(s|e) = 4.41 * 10^{-13}$

2. Mona Lisa $P(s, e) = P(e)P(s|e) = 1.52 * 10^{-14}$
3. Lisa del Giocondo $P(s, e) = P(e)P(s|e) = 1.52 * 10^{-14}$

Orain ikus daiteke, pisuak esleitzerako orduan, Mona Lisa arte lana besteak baino entitate ospetsuagoa den arren, e entitatea eratorri den s azpikatearen garrantzia.

5.3 Esperimentuaren egitura

Puntu honetan, esperimentua burutzeko elementu guztiak prest daude. Alde batetik urre-patroia dago eta 400 elementuetako bakoitzarentzat esleitu beharko litzatekeen artikulua du, artikulurik esleitu ez den kasuan NIL. Bi sistema automatikoek, 400 elementuei, entitate zerrendak esleitu dizkiete eta artikulurik esleitu ezin izan duten kasuetan NIL bueltatu dute.

Ebaluazioa bi zatitan banatuko da. Lehenik sistemak artikulu egokia zerrendako hautagaien artean sortzeko gai den ebaluatuko da. Hau da, urre-patroian dagoen artikulua zerrendetan dagoen aztertuko da. Bigarrenik heuristiko bidez sistemek itzuli dituzten zerrendako artikulu bakarra aukeratuko da. Ondoren, aukeratu den artikulu hau urre-patroiko berdina den aztertuko da. Honetarako garrantzi handia izango dute sistemek artikuluei esleitu dizkieten pisuek.

6 Kapitulua

Sistema automatikoen ebaluazioa eta analisia

Atal honetan sistema automatikoek Wikipedia artikuluak esleitzeko duten gaitasuna ebaluatuko da. Bi sistemak, Wikipedia Miner eta xede berdinerako garatu den sistema dira. Hemendik aurrera Europeanako elementuak Wikipedia artikuluekin aberasteko garatu den sistema "EWA" bezala izenpetuko da eta Wikipedia Miner "WM" bezala.

Sistemek esleitu dituzten artikulu eta NIL kopuruak aztertuz (ikus 6.1 taula), WM-ek 400 elementuetatik 381 elementuri artikuluak esleitu dizkio, EWA-k aldiz 399 elementuri. Ikus daitekeenez sistemek elementuen tituluetatik hautagaiak sortzeko erraztasuna dute. Baina hau ez da abantaila bat, urre-patroian 89 artikuluk bakarrik dute artikulua esleitua.

Sistema	**WM**	**EWA**	**Urre-patroia**
Elementu kopurua	400	400	400
NIL	19	1	311
Artikulua dute	381	399	89

Taula 6.1: WM, EWA eta Urre-patroiaren NIL edo artikulu kontaketak. Aipagarria da, sistemek artikulua esleitzeko duten erraztasuna.

6.1 taulako balioak, elementuei artikulu bat gutxienez esleitu dioten zenbaketak dira. Sistemek elementu bakoitzari artikulu bat baino gehiago esleitu diote eta kontaketak guztira eginda, WM-ek 1732 artikulu esleitu ditu eta EWA-k 88502. Honen arrazoia sinplea da, EWA-ren kasuan, elementuen tituluetatik atera daitezkeen hiztegiko aingurak asko dira eta aingura hauek, hiztegian entitate asko dituzte lotuak.

Ebaluazioa bi zatitan banatu da. Lehenik sistemek sortzen dituzten hautagaien artean hautagai egokia sortzeko duten gaitasuna aztertuko da. On-

doren heuristiko bidez zerrendako artikulu bakarra aukeratuz, hau egokia den ebaluatuko da. Emaitzak bitan banatu ditugu, alde batetik Scran-eko 200 elementuak eta bestetik Culture Grid-ekoak.

Emaitzetan, zehaztasuna (Urre-patroian dagoen emaitza lortzen duten kasuak, zati elementu guztiak), doitasuna (urre-patroiko artikulu bera lortu duten elementuen kontaketa, zati sistemak artikulua esleitu dion kasuen kontaketa), estaldura (urre-patroiko artikulu bera lortu duten elementuen kontaketa, zati urre-patroian artikulua esleituta duten elementuen kontaketa) eta f1 (doitasun eta estalduraren arteko batez besteko harmonikoa) erabiliko dira.

6.1 Sistema automatikoek hautagai egokia sortzeko duten gaitasunaren ebaluazioa

Azterketa honetan, algoritmo perfektu baten bitartez zerrendako artikulu egokia aukeratzeko gai izango bagina bezala ebaluatu nahi izan dira sistema automatikoak. Horretarako, Wikipedia artikulu egokia hautagaien artean dagoen ebaluatzeko bi orakulu diseinatu dira:

- Lehenengo Orakuluak, elementu bakoitzarentzat sistemak itzuli duen artikulu zerrendatik egokia aukeratzen du (dagoen kasuan), bestela ez du artikulurik bueltatzen.
- Bigarren Orakuluak, urre-patroian elementuak artikulurik esleitua ez badauka (NIL da) NIL itzultzen du, sistemak artikulua esleitu dionean ere. Bestela lehenengo orakulua aplikatzen du: Artikulu sortatik egokia aukeratu eta hau itzultzen du, posiblea den kasuan.

6.2 eta 6.3 tauletan ikus daitekeen moduan, lehenengo orakuluaren zehaztasuna oso baxua da (0.115 eta 0.24 bilduma eta sistemaren arabera). Honekin, sistemek elementu gehienentzat artikuluak bueltatu dituztela egiztatzen da, eskuz eginiko etiketatzean ez bezala, hemen %22-ak bakarrik dute artikulua. Doitasunarekin berdina gertatzen da, sistemak artikulua bueltatu duen elementu askotan eskuz NIL etiketatu baita. Hala ere, estaldura balioak altuak dira, %65 eta %75.9 balioen artean. Balio hauek, estalduraren goi bornea erakusten dute, hau da, lehenengo orakuluak irteerako zerrendatik artikulu egokia aukeratuko balu bezala. Beraz urre-patroiko artikulua duten 89 elementuentzat, WM-ek 60 (39 Scran-en eta 21 Culture Grid-en) asmatu ditzake eta EWA-k 62 (40 Scran-en eta 22 Culture Grid-en) gehienez.

EWA-k, WM-ekin alderatuz, artikulu kantitate handiagoa bueltatzen du elementu bakoitzarentzat. Beraz NIL kasurik ez egotean doitasuna txikiagoa

da. Hala ere, artikulu zerrenda zabalagoa bueltatuz, aukera gehiago dago artikulu egokia bertan aurkitzeko, hau da estaldura balio altuagoen zergatia.

Wikipedia Miner	zehaztasuna	doitasuna	estaldura	f1
1. Orak. Scran	0.240	0.206	0.650	0.313
1. Orak. CGrid	0.240	0.122	0.724	0.209
2. Orak. Scran	0.895	0.672	0.650	0.661
2. Orak. CGrid	0.960	0.750	0.724	0.737

Taula 6.2: WM-entzat orakuluen emaitzak

EWA	zehaztasuna	doitasuna	estaldura	f1
1. Orak. Scran	0.200	0.200	0.667	0.308
1. Orak. CGrid	0.115	0.111	0.759	0.193
2. Orak. Scran	0.900	0.667	0.667	0.667
2. Orak. CGrid	0.965	0.759	0.759	0.759

Taula 6.3: EWA sistemarentzat orakuluen emaitzak

Bigarren orakulua aztertuz, argi geratzen da NIL kasuen antzematean dagoela gakoa. Sistemak eta bigarren orakuluaren bitartez NIL-ak bueltatuz, zehaztasun maila oso altuak dira %89.5 eta %96.5 artean. Gainontzeko neurketek gora egin dute NIL-ak antzemateko garrantzia azpimarratuz eta estaldurak dagoenetan jarraitzen du bigarren orakulu honetan. Esan beharra dago, atal honetan, EWA-k emaitza hobeak lortu dituela WM-ekin alderatuz.

Orakuluen emaitzetatik atera daitekeen ondorio nagusia, Europeanako elementuei artikuluak automatikoki esleitzea posible dela da. Esan beharra dago sistemek tituluko informazioa soilik izan dutela ataza hau burutzeko eta eskuz etiketatu dituztenak elementuaren informazio guztia (irudia barne). Irudia, etiketatzaileen esanetan, erabakigarria izan da NIL kasuak etiketatzeko garaian. Etorkizunean, sistemen emaitzak hobetzeko, elementuaren informazio gehiago jasotzea egokia litzateke. Azkenik, NIL kasuen antzemateko sistema on bat ere erabakigarria izango litzateke emaitzak hobetzeko.

6.2 Sistema automatikoak eta heuristikoen bidezko artikuluen hautaketa

Oraingo honetan, entitate-izenen desanbiguazioan urrats bat emateko, artikulu zerrenda beharrean artikulu bakarra itzultzeko saiakera egingo da. Sistemek esleitzen dituzten artikulu zerrendetatik, artikulu bakarra edo "onena"

aukeratzeko heuristiko batzuen azterketa egin da. Heuristiko posibleen artean hauek aukeratu ziren azterketa bat egiteko:

- PisuMax heuristikoa: Sistemak altuen pisatutako artikuluak du lehentasuna.
- Luzeena heuristikoa: Tituluko azpimultzo luzeenetik esleitu den artikuluak du lehentasuna.
- Ezkerrekoa heuristikoa: Tituluan ezker aldeko azpimultzotik esleitutako artikuluak du lehentasuna.

	PisuMax	**Luzeena**	**Ezkerrekoa**	**Maximoa**
WM	42	36	3	60
EWA	7	41	2	62

Taula 6.4: WM eta EWA-k heuristiko ezberdinekin artikulu egokia

Orakuluek diotenez, urre-patroian artikulua duten 89 elementuetatik, WM-ek 60-rentzat lortu dezake artikulu egokia eta EWA-k 62-rentzat. 6.4 taulan ikus daiteke sistemek altuen pisatutako artikuluak aukeratuz, 42 artikulu asmatzen dituela batak, baina 7 besteak. Luzeena heuristikoaz berriz, EWA-k 41 asma ditzake eta WM-ek 36. Ezkerrekoa heuristikoaren emaitzak ez dira bat ere aipagarriak.

Azterketa honen ondoren, sistema bakoitzarentzat 3 heuristikoak modu jakin batean aplikatzen dituen algoritmo bat erabili da. WM-ek itzulitako emaitzetan, pisuak garrantzia duela ikusita, algoritmo hau aplikatu da artikulu bakarra aukeratzeko:

1. PisuMax heuristikoa: Sistemak altuen pisatutako artikuluak du lehentasuna.
2. Luzeena heuristikoa: Tituluko azpimultzo luzeenetik esleitu den artikuluak du lehentasuna.
3. Ezkerrekoa heuristikoa: Tituluan ezker aldeko azpimultzotik esleitutakoak du lehentasuna.

Elementu baten artikulu zerrendan pisu handienarekin berdinketa dagoen kasuan, azpimultzo luzeenetik esleitu dena aukeratzen da. Oraindik berdinketarik balego ezker aldeko azpimultzotik esleitutakoa hartuko da.

EWA-rentzat, tituluko azpimultzo luzeenetik esleitu den artikuluak garrantzia duenez, hau da aplikatu den algoritmoa:

1. Luzeena heuristikoa: Tituluko azpimultzo luzeenetik esleitu den artikuluak du lehentasuna.

2. Ezkerrekoa heuristikoa: Tituluan ezker aldeko azpimultzotik esleitutakoak du lehentasuna.

3. PisuMax heuristikoa: Sistemak altuen pisatutako artikuluak du lehentasuna.

Heuristikoren baten berdinketak ematen direnean, WM-i aplikatu zaion algoritmoan bezala hurrengo heuristikoa aplikatuko da. 6.5 taulan ikus daiteke algoritmo hauek aplikatzean WM-ek artikulu bat gehiago asmatzea lortu duela PisuMax heuristikoaz alderatuz. EWA-ren emaitzetan algoritmoak ez du aldaketarik suposatu Luzeena heuristikoarekiko.

	Heuristikoen algoritmoa	**Maximoa**
WM	43	60
EWA	41	62

Taula 6.5: WM eta EWA-k hautaketa algoritmoekin lortutako emaitzak

Bi sistementzat emaitza onenak ematen dituzten algoritmoen arteko ezberdintasunak aztertuz, algoritmoetan PisuMax heuristikoaren kokapena aipatzekoa da. EWA-k Wikipedian entitateak dituen agerpen kopuruetan oinarritutako pisaketa erabiltzen du beraz beharrezkoa da tituluaren azpimultzoetan aingura luzeena aukeratzea lehenik, hau izaten baita emaitza egokia. WK-ek bueltatzen dituen pisaketetan aldiz, guzti hau kontuan hartuta dago bere algoritmoan. Izan ere lehenik tituluan desanbiguatzeko hautagai posibleak bilatzen baititu.

Hauek izan dira heuristikoekin eta algoritmoarekin lortu diren emaitzak. Hurrengo azpi-atalean, algoritmoaren emaitzak sakonduko dira Scran eta Culture Grid bildumentzat urre-patroi osoaren gainean.

6.2.1 Sistema automatikoak eta heuristikoen bidezko artikuluen hautaketaren emaitzak

WM sistemarentzat heuristikoen algoritmoak izan dituen emaitzetan (ikus 6.6 taula) NIL-en antzemateak eragina dauka zehaztasun eta doitasun baxuen balioetan berriz ere. Estaldura begiratuz, ikus daiteke bi bildumentzat %48.3 lortu dela, honek esan nahi du heuristikoen algoritmo sinple hauekin, urre-patroian artikulua duten elementuen erdientzat artikulu egokia aukeratu daitekeela.

WM	zehaztasuna	doitasuna	estaldura	f1
Scran	0.190	0.153	0.483	0.233
Culture Grid	0.205	0.081	0.483	0.139

Taula 6.6: Heuristikoen algoritmoa WM-en emaitzetan aplikatuak

6.7 taula begiratuz, EWA-rentzat zehaztasun eta doitasun balioak behera egiten dute WM-ekin alderatuz. Estaldurak %41.7 eta %55.2-ko balioak ematen ditu bildumaren arabera eta aurreko sistemaren antzerako balioak ditu.

EWA	zehaztasuna	doitasuna	estaldura	f1
Scran	0.125	0.125	0.417	0.192
Culture Grid	0.085	0.080	0.552	0.140

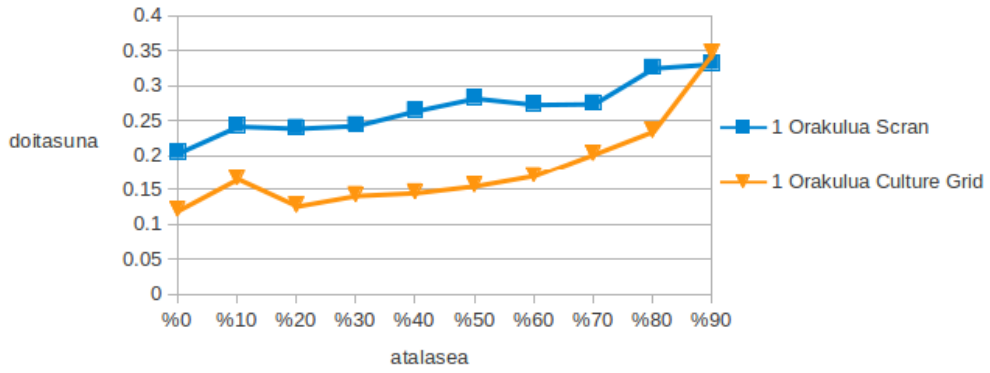
Taula 6.7: Heuristikoen algoritmoa EWA-ren emaitzetan aplikatuak

Emaitza hauetatik berriz azpimarratzekoa da urre-patroian artikulua duten elementuen erdientzat artikulu egokia aukeratu daitekeela heuristikoen algoritmo sinple honen bitartez. Hala ere NIL kasuen antzematean sakondu beharra berriz argi geratu da. Horretarako, pisu baxua duten artikuluak baztertuz NIL hauen antzematean aurrera pausu bat emateko asmoz, WM-ek pisatzen dituen artikuluen azterketa bat egin da.

6.3 Wikipedia Miner-ek esleitzen dituen pisuen azterketa, NIL-en antzematearen lehen pausuak

Sistemek itzultzen dituzten artikulu kopuru altua ikusita (WM-ek 1732 artikulu esleitu ditu eta EWA-k 88502) pisuaren arabera baztertuz, NIL kasuak antzemateko saiakera bat egin da. WM sistemak artikuluei esleitzen dizkion pisuetan oinarritu gara honetarako. Sistema honek itzulitako artikuluen pisuak 0.973 eta 0.002 artean daudela ikusi da. Batez-bestekoa Scran eta Culture Grid bildumetan 0.48 eta 0.477 dira hurrenez hurren. Maximo eta minimoaren artean hamar tarte definitu dira eta tarte bakoitzean pisu horretatik behera dauden artikuluak baztertu dira. Lehenengo orakuluaren doitasunaren laguntzaz, sistemak itzuli duen sortatik egokia aukeratu eta bestela NIL itzuliz egin da azterketa.

6.1 irudiko grafikoan, pisaketa bidez artikuluak baztertzea posible dela ikus daiteke, batez ere, doitasunak gora egiten baitu grafikoaren eskuineko



Irudia 6.1: Lehenengo orakuluaren iragarpenaz baliatuz, pisuen arabera artikuluak baztertuz NIL-en antzematean aurrera pausu bat egiteko azterketa ikusten da. Doitasunak grafikoaren eskuinaldean gora egiten du, beraz pisu handieneko artikuluekin geratuz eta gainontzekoak baztertuz emaitzak onak direla erakusten du.

balioetan. Azken finean pisaketa altuko artikuluekin geldituz eta gainontzekoak baztertuz doitasuna igotzen da baina hala ere goi bornetik behera aurkitzen da balio hau. Bigarren orakuluak %67 eta %75-eko balioak ditu eta oraingoa %35 iritsi da gehienez. Honen arrazoia, Wikipedia Miner-ek itzulitako artikuluen pisuen %50-ak balio maximoaren %10aren azpitik daudela da. Artikulu hauek 6.6 taulan erabili den algoritmoa aplikatuz baztertuak gelditzen dira, lehenengo orakuluak ontzat eman dituen arren.

Beraz azterketa honen ondoren, NIL kasuak antzemateko sistema bat garatzeko aurrekariak aztertu dira. Hau egiteko oinarria, artikuluen pisuen bidezko bazterketa izan daiteke.

7 Kapitulua

Ondorioak eta etorkizuneko ildoak

Dokumentu honekin amaitzeko eta proiektuan ateratako ondorioak azaltzeko, proiektuaren helburuak gogoratuko dira. Helburu bat: Europeanako elementuak, sistema automatiko bidez, Wikipedia artikuluekin aberastea posible den ikertzea eta ebaluatzea zen. Emaitzetan, urre-patroiko %22-ari artikulua esleitzea posible dela ikusi da eta kontuan hartuz Europeanaren elementu kopuru erraldoia, ezin da balio hau gutxietsi. Gainera %22 honetatik sistema automatikoen bidez %75.9 automatikoki lotzea posible dela ikusi da, sistemak itzulitako aukeren artean algoritmo perfektu batekin aukera egokia eginez. Hemen proposatutako heuristikoen algoritmoekin lortu diren emaitza onenen estaldura balioak %48.3 eta %55.2 dira, baina oraindik hobetzea badago %75.9-ko balioa lortu arte. Esan beharra dago, sistemek titulua soilik jasotzen dutela sarrera gisa, beraz deskribapen eta subjektuko informazioa gehituz, balio hauek hobetzea lor daitekeela pentsatzen da.

Bigarren helburua, Europeanako elementuak Wikipedia artikuluekin aberasteko sistema bat garatzea eta ebaluatzea zen. Hau ere lortu egin da EWA sistemarekin. Esan beharra dago WM-ekin alderatuz ez dituela emaitza txarrak lortu eta alor batzuetan emaitza hobeak lortu dituela. Adibidez, sistema automatikoen bidez %75.9 automatikoki lotzea posible dela esaten dugunean EWA-ren balioa da, WM %72.4-ra iritsi da neurketa horretan.

Ikerketarekin jarraitzeko ebaluazio datu multzoa handiagotu nahi den kasurako ere, eskuz etiketatzeko garaian suertatu diren arazo eta zalantzak argitu dira. Irizpide berri batzuk proposatuz, artikulua aukeratzeko garaian zorroztasun maila bat gehitu da.

Laburbilduz, Europeana Wikipedia artikuluen informazioaz aberastu daitekeela egiaztatu da eta hau egiteko sistema bat garatu daitekeela ere. Bestalde sistemek itzultzen dituzten artikulu zerrendak murrizteko lehen pausuek,

etorkizunerako lana bultzatzen dute. Etorkizunean NIL kasuak antzemateko sistema on batekin eta hautaketarako algoritmoen hobekuntzarekin, sistema on bat garatzeko aukera ikusten da.

Amaitzeko, dokumentuaren eranskin bezala, ikerketa honen inguruan idatzi den "Matching Cultural Heritage items to Wikipedia" artikulua dator. Bertan proiektu honetarako egin den lan guztia modu labur batean azaltzen da. "The eigth international conference on Language Resources and Evaluation LREC-2012" konferentziarako onartu zuten eta Istanbulen egin den konferentzian, oso arrera ona izan duen poster baten bitartez aurkeztua izan da.

Bibliografia

[1] Haslhofer, B.; Roochi, E. M.; Gay, M.; Simon, R. Augmenting europeana content with linked data resources. Proceedings of the 6th International Conference on Semantic Systems, pp 40:1–40:3, New York, NY, USA, 2010.

[2] Bunescu, R. C.; Pasca, M. Using encyclopedic knowledge for named entity disambiguation. in *EACL*. The Association for Computer Linguistics, 2006.

[3] Cucerzan, S. Large-scale named entity disambiguation based on wikipedia data. in *EMNLP-CoNLL*, pp 708–716. ACL, 2007.

[4] Mihalcea, R.; Csomai, A. Wikify!: linking documents to encyclopedic knowledge. in *CIKM*, Silva, M. J.; Laender, A. H. F.; Baeza-Yates, R. A.; McGuinness, D. L.; Olstad, B.; Olsen, Ø. H.; Falcão, A. O., Eds., pp 233–242. ACM, 2007.

[5] Milne, D.; Witten, I. H. Learning to link with wikipedia. in *Proceeding of CIKM '08*, pp 509–518, New York, NY, USA, 2008.

[6] Han, X.; Sun, L. A generative entity-mention model for linking entities with knowledge base. in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp 945–954, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[7] Hoffart, J.; Yosef, M. A.; Bordino, I.; Fürstenau, H.; Pinkal, M.; Spaniol, M.; Taneva, B.; Thater, S.; Weikum, G. Robust disambiguation of named entities in text. in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp 782–792, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[8] Gottipati, S.; Jiang, J. Linking entities to a knowledge base with query expansion. in *Proceedings of the 2011 Conference on Empirical Methods*

in Natural Language Processing, pp 804–813, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[9] Ji, H.; Grishman, R. Knowledge base population: Successful approaches and challenges. in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp 1148–1158, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[10] Chang, A. X.; Spitkovsky, V. I.; Yeh, E.; Agirre, E.; Manning, C. D. Stanford-ubc entity linking at tac-kbp. in *Proceedings of TAC 2010*, Gaithersburg, Maryland, USA, November 2010.

Publikazioak

- **Matching Cultural Heritage items to Wikipedia**
  Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando, Mark Stevenson
  The eigth international conference on Language Resources and Evaluation LREC-2012 **2012**,
  https://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1341334636/publikoak/lrec.pdf

Eranskina

Matching Cultural Heritage items to Wikipedia

Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando, Mark Stevenson

IXA NLP Group, University of the Basque Country, Donostia, Basque Country,
{e.agirre,abarrena014,oier.lopezdelacalle,a.soroa}@ehu.es
Natural Language Processing Group, Sheffield University, Regent Court, 211 Portobello, Sheffield, UK
{s.fernando,r.m.stevenson}@sheffield.ac.uk

Abstract

Digitised Cultural Heritage (CH) items usually have short descriptions and lack rich contextual information. Wikipedia articles, on the contrary, include in-depth descriptions and links to related articles, which motivate the enrichment of CH items with information from Wikipedia. In this paper we explore the feasibility of finding matching articles in Wikipedia for a given Cultural Heritage item. We manually annotated a random sample of items from Europeana, and performed a qualitative and quantitative study of the issues and problems that arise, showing that each kind of CH item is different and needs a nuanced definition of what "matching article" means. In addition, we test a well-known wikification (aka entity linking) algorithm on the task. Our results indicate that a substantial number of items can be effectively linked to their corresponding Wikipedia article.

1. Introduction

Current efforts for the digitisation of Cultural Heritage are providing common users with access to vast amount of materials. Europeana[1], for instance, is incorporating millions of digitised Cultural Heritage (CH) items from Europe's archives, museums, libraries and audio visual collections and providing access through a single portal. The main strength of Europeana lays in the vast number of items it contains. Sometimes, though, this quantity comes at the cost of a restricted amount of metadata, with many items having very short descriptions and a lack of rich contextual information. Wikipedia, in contrast, offers in-depth descriptions and links to related articles for many CH items, and is thus a natural target for automatic enrichment of CH items.

Enriching CH items with information from Wikipedia or other external resources is not novel. In (Haslhofer et al., 2010), for instance, the authors also acknowledge the interest of enriching CH items. They present the LEMMO framework, a tool to help users annotate Europeana items with external resources (i.e. Web pages, Dbpedia entries, etc.), thus extending Europeana items with user-contributed annotations.

In contrast to their work, our research aims to provide an evaluation of automatic annotation, and not only a description of an interface for manual annotation. We thus annotated a random sample of items, and performed a qualitative and quantitative study of the issues and problems that arise, showing that each kind of CH item is different and needs a nuanced definition of what "matching article" means. We also show that Wikipedia articles cover a substantial number of items.

Our research aims at finding Wikipedia articles that match the content of each target CH item. Note that this is more restrictive than finding Wikipedia articles that are related, as the matching article needs to describe the same CH object described in the target item. This problem is closely linked to Wikification, the process where a flat piece of text is enriched with links to the articles which are explicitly mentioned in the text. The process involves two inter-related steps: to choose which are the potential articles mentioned in the text, and to disambiguate them. For instance, assume that the famous Mona Lisa painting has been digitised and published as a CH item. In Wikipedia there are 11 articles which can be referred to as Mona Lisa[2], ranging from songs to a movie, and including actresses, singers and even a crater in Venus. In the first step of Wikification the algorithm would retrieve the 11 articles, and in the disambiguation step, the algorithm would select the painting[3]. Although a relatively recent concept, there is now a flurry of activity around this problem (Bunescu and Pasca, 2006; Cucerzan, 2007; Mihalcea and Csomai, 2007; Milne and Witten, 2008; Han and Sun, 2011; Hoffart et al., 2011; Gottipati and Jiang, 2011; Ji and Grishman, 2011). We tested a well-known method (Milne and Witten, 2008) and our own in-house system on the task.

The paper is structured as follows. We begin by describing Europeana and the target collections. Section 3 presents the methodology for the manual annotation, followed by a Section analysing the annotated dataset. In Section 5 we describe the wikification systems used and the results when it is evaluated on our dataset. Finally, Section 7 draws the conclusions and outlines future work.

2. Europeana and the target collections

Europeana[4] is the prototype website of the European digital library. Europeana incorporates over 20 million digitised items from Europe's archives, museums, libraries and audio visual collections and provides access to them through a single portal. The need for personalised user services has been recognised from the early stages of Europeana's development. The items are supplied by over 1,500 institutions, including the British Library, the Louvre and other local museums, who have provided digitised items from their collections. We have focused on two of these collections:

[1] http://www.europeana.eu

[2] http://en.wikipedia.org/wiki/Mona_Lisa_(disambiguation)

[3] http://en.wikipedia.org/wiki/Mona_Lisa

[4] http://www.europeana.eu

```
<record>
  <dc:identifier>http://www.picturethepast.org.uk/frontend.php?keywords=Ref_No_increment;EQUALS;NCCW001197</dc:identifier>
  <europeana:uri>http://www.europeana.eu/resolve/record/09405/C052AA1727D9C258801CF676473953A0861A47C0</europeana:uri>
  <dc:title>The Major Oak</dc:title>
  <dc:source>Picture the Past OAI feed</dc:source>
  <dc:contributor>North East Midland Photographic Record</dc:contributor>
  <dc:description>The largest Oak tree in England, perhaps in the world, this famous tree has withstood lightning,
                  the drying-out of its roots and even a recent fire. The hollow tree has a circumference of 32 feet
                  and the spread of its branches makes a ring 260 feet round.</dc:description>
  <dcterms:isPartOf>Picture the Past</dcterms:isPartOf>
  <dc:language>EN-GB</dc:language>
  <dc:publisher>North East Midland Photographic Record</dc:publisher>
  <dc:subject>Robin_Hood</dc:subject>
  <dc:type>Image</dc:type>
  <dc:format>JPEG/IMAGE</dc:format>
  <europeana:provider>CultureGrid</europeana:provider>
  <europeana:hasObject>true</europeana:hasObject>
  <europeana:country>uk</europeana:country>
  <europeana:type>IMAGE</europeana:type>
  <europeana:language>en</europeana:language>
</record>
```

Figure 1: Example of an ESE record from Europeana.

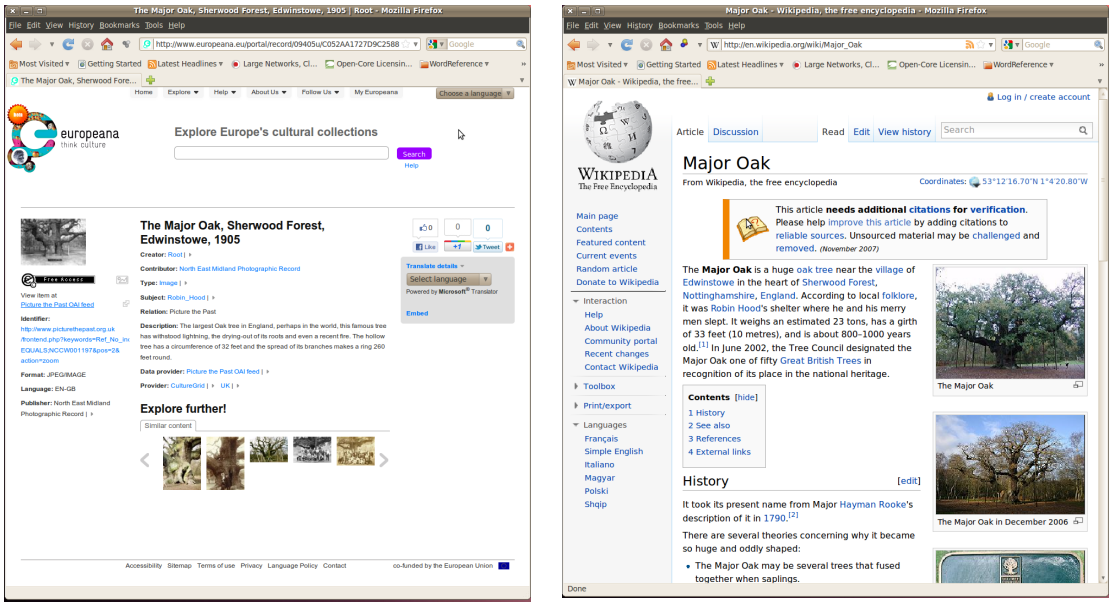


Figure 2: The Europeana item referring to a picture of the "The Major Oak" taken in 1905 (left), and the Wikipedia article on the same tree.

Culture Grid[5] and Scran[6].

Culture Grid (**Cgrid** for short) contains over one million items from 40 different UK collections including national and regional museums and libraries. The Scran collection is an online resource containing images and media from different museums, galleries and archives in Scotland.

The Europeana item records are associated with metadata which is extracted from the original collection though a process known as "ingestion". This paper uses a version of this metadata stored in a format known as Europeana Semantic Elements (ESE)[7]. Figure 1 shows an example of an ESE record describing a photograph of a well known tree, "The Major Oak". We focus on the `dc:title` and `dc:description` fields of the ESE records since the information they contain is relatively consistent (compared to other fields) and they generally contain enough text to work with. Figure 2 shows the item for the picture of "The Major Oak" as shown in the Europeana interface and the corresponding Wikipedia article referring to the same tree.

The combined collections contain approximately $858,000$ items, with $547,000$ items in Cgrid, and $310,800$ in Scran. Most of the items (99%) have a title ("dc:title"), which has 6 tokens on average, but only 68% have any description ("dc:description" field), with 27 words on average.

3. Methodology for a manually annotated dataset

We selected a random subset comprising 400 items from the Scran and Cgrid collections in Europeana. The items were then ordered according to the subcollections they came from, so the annotators had a relatively coherent set of items, coming from a relatively small number of collections

[5] `http://www.culturegrid.org.uk`

[6] `http://www.scran.ac.uk`

[7] `http://version1.europeana.eu/web/guest/technical-requirements`

Type	Count
Photographs	276
Coins and Artifacts	57
Books, booklets etc	24
Other	21
Paintings	14
Audio and Video	8
Total	400

Table 1: Types of Europeana items in the sample.

such as "The National Museum Record"[8], "The portables Antiquities Scheme"[9] or Scran. [10]. Table 1 shows the type of the items in the sample. The majority are photographs, but there are also other types such as paintings or antique coins.

The annotators were given the records with all the metadata (see Figure 1). They could also access the item as shown in the Europeana interface (see Figure 2) and they had to return the URL of a single English Wikipedia article (see Figure 2) matching the item, or NIL if they could not find any matching entry. The definition of a matching entry provided to the annotators was: "the Wikipedia article and the item must describe the same particular object. In the case of photographs, the article must be about the subject of the photograph, e.g a particular person or location." Note that this definition of matching tries to find equivalent items and articles, and thus does not consider other kinds of relations between item and Wikipedia article, such as for example linking an item to the article about "photography" because it's a photograph, or linking an item to the article of the author.

4. Analysis of the annotated dataset

The random subset of 400 items was independently tagged by two groups of annotators, one in Donostia and another in Sheffield, each one comprising three persons. As a result, the subset was annotated twice and two tags were obtained for each item. We chose one group's answers as gold standard, and used the other for calculating Inter Annotator Agreement (IAA) figures, as explained in Section 4.2.

According to the gold standard, 89 items were successfully linked to Wikipedia articles (22% of the sample). Given that the method for matching entries was very strict it came as a surprise that the annotators were able to identify a matching article for so many items. This result suggests that the task of matching Cultural Heritage elements to external resources such as Wikipedia can have a real impact in the richness of the descriptions for that 22% of the sample.

The remainder of this section describes the normalisation of URLs from Wikipedia to a canonical Wikipedia URL, followed by an analysis of agreement between annotators and qualitative analysis about what the annotators consider a "matching article" to be.

[8] http://viewfinder.english-heritage.org.uk/

[9] http://finds.org.uk/database/

[10] http://www.scran.ac.uk/

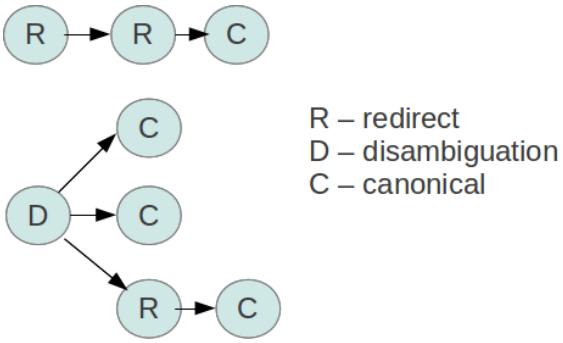


Figure 3: Normalisation flow.

4.1. Normalisation

Wikipedia articles are often accessed following so called redirect pages. For instances, the page "UK"[11] is a redirect page pointing to the Wikipedia article "United Kingdom"[12]. In such cases, we say that the redirect page "UK" resolves to the article "United Kingdom". Redirect pages fulfil many purposes like dealing with alternate names, plurals and closely related words.

When analysing the annotated items we found some discrepancies between annotators due to redirect pages: one annotator tagged the item with the "normal" article whereas other annotator used a redirect page resolving to the same article. We thus normalised the annotator results and resolved all redirect pages. In principle, it is enough to build a mapping among redirect pages and the articles they refer to. However, the process is further complicated due the fact that some redirects resolve to pages which are also redirects. Even worse, sometime redirect pages resolve to disambiguation pages (pages pointing to all possible meanings of a string) which can themself refer to other redirect pages. Figure 3 shows two examples of the normalisation flow. The upper part of the figure shows a redirect resolving to another redirect which finally links to the desired article (the *canonical* article). The lower part shows an example of a disambiguation page referring to many pages; two of them are canonical pages but one page is a redirect which links to a canonical article.

The normalisation script thus builds a dependency tree between redirects, disambiguation pages and final articles. Then, it associates each article with a canonical link. Normal articles map to themselves; redirect pages map to the canonical page and disambiguation pages map to a set of possible canonical pages. Note that for our particular dataset no annotator chose a disambiguation page.

4.2. Inter Annotator Agreement

The overall Inter-Annotator Agreement (IAA) between the two tags available for each item are very high: 92.5% in the Cgrid collection and 80.0% in Scran (see Table 2). The agreement takes into account the items which were not associated with an article (i.e. tagged as NIL).

Given the high number of items with NIL, we also computed the IAA for items that were linked to an article for

[11] http://en.wikipedia.org/wiki/UK

[12] http://en.wikipedia.org/wiki/United_Kingdom

Match	Scran	Cgrid
Overall IAA	80.0%	92.5%
Agreement: Both NIL	126	165
Disagreement: One NIL	38	13
Agreement: Same article	34	20
Disagreement: Different articles	2	2

Table 2: Inter Annotator Agreement figures. The first row shows the percentage over all 400 items. The second and third rows show the numbers of items for which one of the annotations was NIL. The final two rows show the numbers of items where both annotators chose an article.

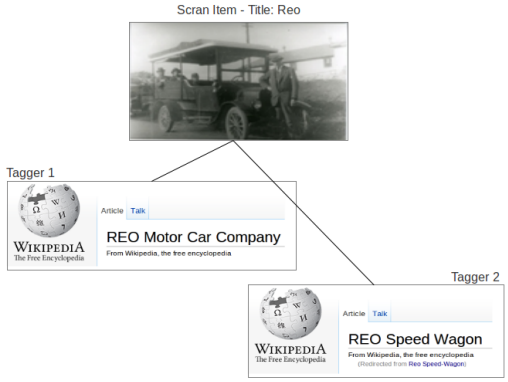


Figure 4: An example where the annotators did not agree. In this case both articles were acceptable.

both tags (22 items in Cgrid and 36 for Scran). The agreement for these items is even higher: 90.9% for Cgrid and 94.4% for Scran. We analysed the few cases where the taggers had both returned an article but would not agree, and found that in all cases both articles were acceptable, and very close in meaning. For instance, in one case an item about a light motor truck named 'Reo', manufactured by REO Motor Car Company was linked to the article about the truck model[13] by one tagger, and to the article about the company[14] by the other tagger (see Figure 4).

Most of the disagreements were due to one tagger not returning any article (NIL) and the other tagger choosing one article. We analysed these disagreements and in general the articles were relevant, and thus well linked. In a few cases, the article is not appropriate, although close. For instance, the item titled "Glyndebourne Opera Company present Le Comte Ory" containing one picture of a performance was linked to an article about an opera festival hold in Glyndebourne[15].

Overall the high IAA numbers (both overall and for items with no NILs) show that the annotation is reliable and that the task itself is well-defined.

[13] http://en.wikipedia.org/wiki/Reo_Speed-Wagon

[14] http://en.wikipedia.org/wiki/REO_Motor_Car_Company

[15] http://en.wikipedia.org/wiki/Glyndebourne_Festival_Opera

4.3. Qualitative analysis

Analysis of the annotations and the feedback received from the annotators suggested that the interpretation of "matching article" varied depending on the typology of the cultural item, as follows:

- If the item is a coin then the Wikipedia article judged as matching described the same kind of coin. For instance, the coin at http://www.europeana.eu/portal/record/09405v/0A9BB0DE9630F20665E36F10366069FDA3DAEA0D.html has no entry in Wikipedia, and therefore a NIL match would be returned. However, it is useful to consider cultural heritage items as instances of particular concepts. For instance, we can find an item about a particular antique coin like "Denarius, of Lucius Marcius Censorinus"[16]. As there is an article about this particular kind of coin[17], the annotators chose to link both.

- If the item is a picture of a particular location or person, that location or person was the subject of the matching Wikipedia article. For instance, for the item entitled "Hampton Court"[18] the matching article is http://en.wikipedia.org/wiki/Hampton_Court_Palace, but for the item "St. Leonards Church" [19] there was no matching article (even if the street is mentioned in the article on the town where it's located "Sunningwell"[20]). The same applies to people. For instance, the matching article for item "Albert Ball, Trent College"[21] s http://en.wikipedia.org/wiki/Albert_Ball. The same applies to organisations like soccer clubs. Note that pictures mentioning anonymous people (e.g. peasants) or locations can never be linked.

- If the title of the item mentions a person and a location annotators chose to focus on the person, as it's usually the focus of the picture. In the future, we would like to consider allowing double annotation.

- Many pictures are nearly 100 years old so there was sometimes a mismatch between the item of the picture and the more recent Wikipedia article.

5. Evaluating automatic systems

This section describes the evaluation of two automatic systems for linking Europeana items to Wikipedia when run on our dataset. Both systems take raw text as input, identify the possible anchors and link each to a Wikipedia article.

[16] http://www.europeana.eu/portal/record/00401/2FCF4C116A23D5F179CEE72DC9CAEE2A02721F79.html. Note that Europeana links change over time.

[17] http://en.wikipedia.org/wiki/Denarius_of_L._Censorinus

[18] http://www.europeana.eu/portal/record/09405r/7303A4578E3AE78F72EC75CB1F02DE47ECAFFE91.html

[19] http://www.europeana.eu/portal/record/09405o/F9C5A09A56B9C54DE0FCC9B53716716AAC751312.html

[20] http://en.wikipedia.org/wiki/Sunningwell#Parish_church

[21] http://www.europeana.eu/portal/record/09405u/03AD6F4A73D75F4BC5748E8AD2BA7096D45C7534.html

Wminer	acc	prec	recall	F1
Oracle1 Scran	0.240	0.206	0.650	0.313
Oracle1 Cgrid	0.240	0.122	0.724	0.209
Oracle2 Scran	0.895	0.672	0.650	0.661
Oracle2 Cgrid	0.960	0.750	0.724	0.737

Table 3: Oracle results for Wminer on Scran & Cgrid

Dict	acc	prec	recall	F1
Oracle1 Scran	0.200	0.200	0.667	0.308
Oracle1 Cgrid	0.115	0.111	0.759	0.193
Oracle2 Scran	0.900	0.667	0.667	0.667
Oracle2 Cgrid	0.965	0.759	0.759	0.759

Table 4: Oracle results for Dict on Scran & Cgrid

The experiments were carried out by providing each system with the text in the `dc:title` elements of the items. The Wikipedia Miner toolkit (**Wminer** for short)[22] links entities found in a text to Wikipedia articles. The toolkit uses the method first presented in (Milne and Witten, 2008) which disambiguates terms by combining three features: the conditional probability of the article given the term (for example, the term "apple" is more likely to link to the article about the fruit than the one about the computer company), the probability of two terms appearing in Wikipedia as a collocation, and a vector-based similarity metric inspired by Normalized Google Distance (but using the links made to each Wikipedia article rather than Google's search results).

The second system uses a implementation similar to the dictionary method described in (Chang et al., 2010), which we refer to as **dict**. This method creates a dictionary containing information about the probability of a string matching a Wikipedia article. Each association between a string and article is scored by counting the number of times that the string appeared as the anchor text of an article's incoming hyperlinks. Note that such dictionaries can disambiguate any of the dictionary's keys directly by simply returning the highest-scoring article. We used the 2011 Wikipedia dump to construct the dictionary and are currently improving the linking algorithm to improve results using this approach.

5.1. Oracle results

In order to evaluate the automatic linking systems, we take the annotations of the first team as our gold standard (GS). We report separate results for the 200 items from Scran and the 200 items from Cgrid. We report accuracy (the ratio of items which get the same label as in the GS divided by the total number of items), precision (the ratio of items correctly linked to an article divided by the total number of items linked by the system), recall (the ratio of items correctly linked to an article divided by the total number of items of items linked to articles in the GS) and F1, the harmonic mean of precision and recall. Note that accuracy takes into account whether the system correctly assigns NIL, while the rest of measures only take into account items linked to articles (and thus discard items tagged as NIL).

Given the text in the title, a linking algorithm will return a set of articles, weighted according to the relevance assigned by the algorithm.

We first analysed whether the automatic linking algorithms are able to find matching articles, that is, whether the target matching article is contained among the articles they return. We are also interested in determining the upperbound in performance for a linking system which chose the correct matching article among the articles returned by the automatic systems. We set up two oracles:

- Oracle1: given a set of articles suggested by the wikifier for the item, choose the correct one (if available), otherwise return any article.
- Oracle2: if an item has no linked article in the GS (i.e. it was annotated as NIL) return NIL, regardless the output of the automatic system. Otherwise apply Oracle1, that is, given a set of articles suggested by the wikifier for the item, choose the correct one, if available

Tables 3 and 4 show the results for each oracle generated by the automatic systems. The accuracy of Oracle1 is very low (between 0.115 and 0.240, depending on the collection and system). The reason for this is that automatic systems suggest a matching article for most items while human annotators are much more selective and only link 22% of the items. The precision is also low for the same reason, as most of the articles returned by the systems were assigned NIL by the annotators. However, recall is high, ranging from 0.650 to 0.759. These figures are the upperbound for the recall of any automatic system built on the output of those wikifiers since the oracle selects all of the correct mappings which they return.

The Dict wikifier tends to return more articles than Wminer, in fact Dict always returns an article, and thus has a lower precision. The articles returned by Dict contain the correct article more often than Wminer as demonstrated by the higher recall figure on each of the collections.

Finally, the results for Oracle2 demonstrate the importance of choosing when to return NIL since a system which returns NIL with perfect accuracy (such as Oracle2) achieves high accuracy (between 0.895 and 0.965). The precision, recall and F-measures would also be high, with Dict generally outperforming Wminer by a small margin.

These results demonstrate that it is feasible to construct a system for automatically linking items to their matching Wikipedia entities based on the output of the Wminer and Dict methods. It is worth noting that we only use only used the text in the title for each item. The annotators mentioned that they used the information in the whole item, including the accompanying picture. This information was often an important factor in the annotators' decision to return NIL. In the future, we would like to explore whether performance could be improved by making use of information from other fields in the items.

[22] `http://wdm.cs.waikato.ac.nz/`

Wminer	acc	prec	recall	F1
Scran	0.190	0.153	0.483	0.233
Cgrid	0.205	0.081	0.483	0.139

Table 5: Results of applying the heuristics over the articles proposed by Wminer on Scran and Cgrid.

Dict	acc	prec	recall	F1
Scran	0.125	0.125	0.417	0.192
Cgrid	0.085	0.080	0.552	0.140

Table 6: Results of applying the heuristics over the articles proposed by Dict on Scran and Cgrid.

5.2. Article Selection Heuristics

We now explore a simple method for selecting the correct article from the set returned by the two methods. Information about the weights returned by each system is used alongside the start and end offsets of the words that were matched to the wikipedia article. In this preliminary study a simple algorithm based on the following set of heuristics is tested:

- Articles with high weights are preferred
- Articles matching longer strings are preferred
- Articles that match the start of the title are preferred

For Wminer the article with the highest weight is chosen first. In the case of ties the article with the longest matching string is chosen. If there is still a tie the article which matches closer to the start of the title is chosen.
The results for this heuristic are shown in Table 5. The main reason for the low accuracy and precision figures is that Wminer returns articles for items tagged as NIL. Recall is higher, 0.483 in both collections, showing that such a simple heuristic is able to select the correct article for nearly 50% of the items that have a corresponding Wikipedia article.
The Dict approach is somewhat different from Wminer since it returns a context-independent weight which is not comparable between articles and consequently the articles are set up in a different order. The articles with the longest match is chosen first and if there is a tie the one which matches closest to the start of the item title is selected. The weights returned by Dict are only used if there is still a tie. Results are shown in Table 6 which shows that the accuracy and precision figures are lower than those obtained using Wminer. The recall varies between the two collections and is lower for Scran than Cgrid.

5.3. Thresholding on weights

Given the over-generation of links for items, we analysed the effect of using the weight returned by Wminer to discard low scoring articles. The weights returned by Wminer ranged from 0.973 to 0.002, with an average of 0.480 on Scran and 0.477 on Cgrid. Ten thresholds lying between these values were selected. At each point we discard all articles with weights below the threshold. In this study we were interested in the ability to correctly identify cases

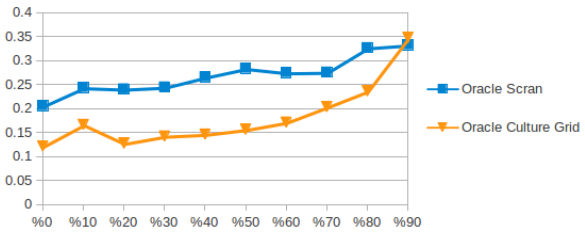


Figure 5: Precision of Oracle1 applied to Wminer weights filtered using various thresholds.

when there is not suitable article (i.e where the annotators selected NIL) as well as identifying the correct article and consequently Oracle1 is applied, i.e. for items linked to Wikipedia articles we choose the correct article if available among the choices returned by the system.
Figure 5 shows that Wikiminer weights are in principle useful to decide when to return NIL as precision raises for higher thresholds. However, the best precision that is achieved when the thresholds are applied is still well below the upperbound (precision for oracle2 is 67% and 75% for Scran and Cgrid respectively).
After the experiment we have seen that around 50% of the Wminer articles get weights in the lowest threshold band (under 10% of the maximum value). This explains why applying the heuristic used in Table 5 did not improve the results. It turns out that many correct articles have very low Wminer weights, and thus are discarded by the heuristic (but chosen by the Oracle1).

6. Conclusions and future work

In this paper we have performed an analysis of the issues that arise when Cultural Heritage items from Europeana are matched with Wikipedia articles. We have shown that up to 22% of items in Europeana can be matched with a counterpart in Wikipedia, a remarkable proportion when the vast number of items in Europeana is considered.
A well-know Wikification algorithm (Wikipedia miner) and an in-house method (Dict) were applied. It was found that up to 75.9% of the items matching a Wikipedia article could be linked automatically, given a perfect algorithm for choosing the correct one among the articles returned by the systems. A simple heuristic based on the weights returned by the systems, length and position in the title attains recall of 48.3% with Wminer and up to 55.2% with Dict (depending on the collection). The results are high for such a simple system, although the 75.9% upperbound shows that there is room for improvement. Note that we only used the text in the title, and an analysis of the text in the description could allow to find more and better matching articles.
We believe that the results reported in this paper are promising, and show potential for deploying a system which suggests Wikipedia articles for Europeana items. The main practical hurdle seems to devise a method which is able to decide when to abstain from returning an article, as there is a high ratio of items which do not have a corresponding

Wikipedia article and the automatic systems tend to always suggest articles. An initial study based on using the weights returned by Wminer showed promising results.

In future we plan to build a system which detects when to return NIL as well as improving techniques for selecting the correct article from those selected. We plan to achieve this by making use of more of the metadata associated with the item, and not only the title.

In addition, we also found that it could be useful to allow linking to subsections of Wikipedia articles, e.g in the case of streets or churches that are described inside the article of a town. For instance one of the Europeana items refers to Sunningwell parish church [23] and the article about Sunningwell includes a section on it[24].

Finally, in addition to identifying the best matching Wikipedia article it would also be interesting to identify related articles based on a fixed typology. For instance, in the case of an item showing the picture of a location, such as a monument or church, the system could return the article referring to the town in which the picture was taken.

Acknowledgments

The research leading to these results was carried out as part of the PATHS project (`http://paths-project.eu`) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082 and KNOW2 project (TIN2009-14715-C04-01). We want to thank the anonymous reviewers for their comments.

7. References

Bunescu, R. C. and Pasca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *EACL*. The Association for Computer Linguistics.

Chang, A. X., Spitkovsky, V. I., Yeh, E., Agirre, E., and Manning, C. D. (2010). Stanford-ubc entity linking at tac-kbp. In *Proceedings of TAC 2010*, Gaithersburg, Maryland, USA.

Cucerzan, S. (2007). Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL*, pages 708–716. ACL.

Gottipati, S. and Jiang, J. (2011). Linking entities to a knowledge base with query expansion. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 804–813, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Han, X. and Sun, L. (2011). A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 945–954, Portland, Oregon, USA. Association for Computational Linguistics.

Haslhofer, B., Roochi, E. M., Gay, M., and Simon, R. (2010). Augmenting europeana content with linked data resources. Proceedings of the 6th International Conference on Semantic Systems, pages 40:1–40:3, New York, NY, USA.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Ji, H. and Grishman, R. (2011). Knowledge base population: Successful approaches and challenges. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1148–1158, Portland, Oregon, USA. Association for Computational Linguistics.

Mihalcea, R. and Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. In Silva, M. J., Laender, A. H. F., Baeza-Yates, R. A., McGuinness, D. L., Olstad, B., Olsen, Ø. H., and Falcão, A. O., editors, *CIKM*, pages 233–242. ACM.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceeding of CIKM '08*, pages 509–518, New York, NY, USA.

[23] `http://www.europeana.eu/portal/record/09405o/9215A3E5F9C4586ABB01D3EACFBA0B239AACDED4.html?query=Sunningwell`

[24] `http://en.wikipedia.org/wiki/Sunningwell#parish_church`