EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Doktorego-tesia

Entitate izendunen desanbiguazioa ezagutza-base erraldoien arabera

Ander Barrena Madinabeitia

2016

EUSKAL HERRIKO UNIBERTSITATEA
Lengoaia eta Sistema Informatikoak Saila

Entitate izendunen desanbiguazioa ezagutza-base erraldoien arabera

Ander Barrena Madinabeitiak Eneko Agirre Bengoa eta Aitor Soroa Etxaberen zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2016ko abendua.

Eskerrak

Lehenik eta behin, nire tesiko zuzendariak eskertu nahi nituzke. Batetik, asteroko bileretan artikuluak komentatzen eta autoreak kritikatzen pasa ditugun orduengatik. Bestetik, arloaren egoera hobetzen duen artikulu bakoitza erronka berri bat bilakatzeagatik. Gainera, *Emacs*ek bizitza errazten dizula erakusteagatik, eta nola ez, terminaleko ingurune zuri-beltzean nahi den guztia egin daitekeela erakusteagatik. Baina batez ere, eskainitako laguntza guztiarengatik, horregatik ez balitz tesi hau ez litzateke posible izango eta. Horregatik guztiagatik, eta ahaztu zaizkidan gauzengatik: Eskerrik asko Aitor eta Eneko!

Bigarrenik, familia eskertu nahiko nuke. Aita, ama eta anaiari nire tesian interesa azaltzeagatik, *disanbigueixon* behin eta berriro esateagatik eta aldioro noiz bukatuko dudan galdetzeagatik (batez ere, aurkezpenera eta lunchera etortzeko duten gogoagatik). Nire amonari aurkezpenera etortzeko arropa bost hilabete lehenago prestatzeagatik eta aurkezpena ikusteko duen gogoagatik. Osaba-izebei eta lehengusu-lehengusinei, familiako doktorea izango denari animoak bidaltzeagatik. Bizkaiko familiari eskerrak eman nahi dizkiot, aipamen berezia eginez *Ga* eta *Owa* deitzen didazuen bikotetxoari.

Arrasateko lagun eta koadrilakoei eskerrak eman nahi dizkiet, tesiak iraun duen denbora honetan ez ditut askotan bisitatu, baina beti hor egongo direla badakit. Tesi honen bukaera, behar den bezala, zuekin ospatu nahiko nuke. Bestalde, nire pixukideak izan zareten guztioi ere eskerrak eman nahi dizkizuet. Bereziki, hasieratik eta orain arte elkarrekin jarraitu dugun hirukoteari.

Ixakideei ere eskerrak eman nahi dizkiet. Uneoro laguntzeko prest egon

zaretelako eta lanerako giro ezin hobea sortu duzuelako. Batez ere, 318 bulegoko faunari, txisteen listoia gero eta beherago jartzeagatik, eta platano bat bulegoan jatea "taboo" bat bihurtzeagatik.

Tesi hau iraun duen denboran Donostian ezagutu dudan jende guztiari eskerrak eman nahi dizkiot. Batez ere, olatuak nirekin partekatu dituzuen horiei, zuekin ñoñostiar bat gehiago sentitu naizelako. Negu hotzetan, dardarka, Ondarretako olatuan momentu ahaztezinak igaro ditugun kortxero guztiek, aipamen berezia merezi dute. Gezurra badirudi ere, tesirako inspirazio momentu gehienak uretan gertatu dira. Besterik gabe, tuboooo!

Azkenik, eta nola ez, eskerrik asko Nerea. Egiten dudana ulertzeko egin duzun esfortzuagatik, jarri duzun interesagatik eta momentu guztietan nire alboan egoteagatik.

Guzti-guztioi, berriz ere, eskerrik asko!

Esker instituzionalak

Euskal Herriko Unibertsitateari, ikerketa-lan hau egiteko emandako ikertzaileak prestatzeko bekarengatik.

Laburpena

Gaur egun, interneten nabigatzeko orduan, ia-ia ezinbestekoak dira bilatzaileak, eta guztietatik ezagunena Google da. Bilatzaileek egungo arrakastaren zati handi bat ezagutza-baseen ustiaketatik eskuratu dute. Izan ere, bilaketa semantikoekin kontsulta soilak ezagutza-baseetako informazioaz aberasteko gai dira. Esate baterako, musika talde bati buruzko informazioa bilatzean, bere diskografia edo partaideetara esteka gehigarriak eskaintzen dituzte. Herrialde bateko lehendakariari buruzko informazioa bilatzean, lehendakari izandakoen estekak edo lurralde horretako informazio gehigarria eskaintzen dute. Hala ere, gaur egun pil-pilean dauden bilaketa semantikoen arrakasta kolokan jarriko duen arazoa existitzen da. Termino anbiguoek ezagutza-baseetatik eskuratuko den informazioaren egokitasuna baldintzatuko dute. Batez ere, arazo handienak izen berezien edo entitate izendunen aipamenek sortuko dituzte.

Tesi-lan honen helburu nagusia entitate izendunen desanbiguazioa (EID) aztertu, eta hau burutzeko teknika berriak proposatzea da. EID sistemek testuetako izen-aipamenak desanbiguatu, eta ezagutza-baseetako entitateekin lotuko dituzte. Izen-aipamenen izaera anbiguoa dela eta, hainbat entitate izendatu ditzakete. Gainera, entitate berdina hainbat izen ezberdinekin izendatu daiteke, beraz, aipamen hauek egoki desanbiguatzea tesiaren gakoa izango da.

Horretarako, lehenik, arloaren egoeraren oinarri diren bi desanbiguazio eredu aztertuko dira. Batetik, ezagutza-baseen egituraz baliatzen den eredu

globala, eta bestetik, aipamenaren testuinguruko hitzen informazioa ustiatzen duen eredu lokala. Ondoren, bi informazio iturriak modu osagarrian konbinatuko dira. Konbinazioak arloaren egoerako emaitzak hainbat datumultzo ezberdinetan gainditukо ditu, eta gainontzekoetan pareko emaitzak lortuko ditu.

Bigarrenik, edozein desanbiguazio-sistema hobetzeko helburuarekin ideia berritzaileak proposatu, aztertu eta ebaluatu dira. Batetik, diskurtso, bilduma eta agerkidetza mailan entitateen portaera aztertu da, entitateek patroi jakin bat betetzen dutela baieztatuz. Ondoren, patroi horretan oinarrituz eredu globalaren, lokalaren eta beste EID sistema baten emaitzak modu adierazgarrian hobetu dira. Bestetik, eredu lokala kanpotiko corpusetatik eskuratutako ezagutzarekin elikatu da. Ekarpen honekin kanpo-ezagutza honen kalitatea ebaluatu da sistemari egiten dion ekarpena justifikatuz. Gainera, eredu lokalaren emaitzak hobetzea lortu da, berriz ere arloaren egoerako balioak eskuratuz.

Tesia artikuluen bilduma gisa aurkeztuko da. Sarrera eta arloaren egoera azaldu ondoren, tesiaren oinarri diren ingelesezko lau artikulu erantsiko dira. Azkenik, lau artikuluetan jorratu diren gaiak biltzeko ondorio orokorrak planteatuko dira.

Gaien aurkibidea

1

Sarrera

Tesi-lan hau hizkuntzaren prozesamenduaren alorrean kokatzen da. Alor honetako gakoa hala beharrez, testu hutsa makinak ulertzea da. Testuaren ulermenerako karaktere-kateak identifikatu, eta hauek dagokien adierekin lotzea ezinbestekoa da. Adibidez, ingelesez idatzitako testu honetan:

Three of the greatest guitarrist
started their career in a single band:
Clapton, Beck and Page.[1]

Esate baterako, *band* hitza musika taldea edo gorputza babesteko oihal zatia izan daiteke (Agirre and Edmonds 2007,Navigli 2009, Agirre *et al.* 2014). Hitz batek bere testuinguruan duen adiera zuzena esleitzeari hitzen adiera desanbiguazioa esaten zaio (HAD). Adieren artean egokia aukeratu beharrean, karaktere-kateak zuzenean Wikipediako artikulu egokietara lotzea Wikifikazio gisa ezagutzen da (Mihalcea and Csomai 2007; Milne and Witten 2008b). Wikifikazioak testuaren ulermenetik haratago, Wikipediako ezagutza entziklopedikoa gehituz, testuaren aberasketa dakar.

Hala ere, arreta berezia eskatzen duten karaktere-kateak existitzen dira, izen-aipamenak hain zuzen ere. Izen propioa duten pertsona, leku edo erakundeak testuetan anbiguetate oso handia duten izen-aipamenekin azaltzen dira. Adibidez, adibidean azaldu den *Beck* aipamena `Jeff_Beck`[2] edo

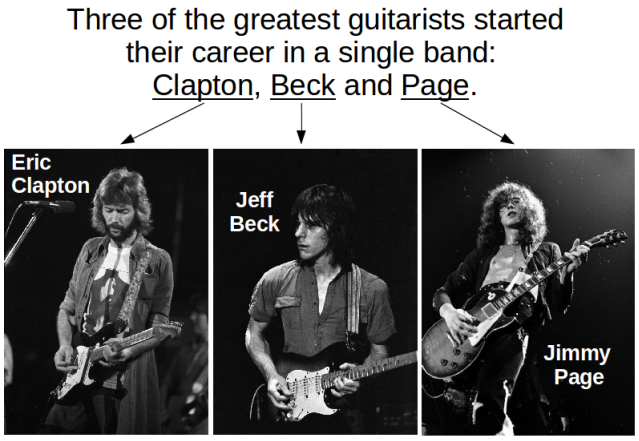
[1] Tesian ingelesezko datuak eta baliabideak erabiliz ikertu da, horregatik, adibide guztiak ingelesez daude.

[2] `http://en.wikipedia.org/wiki/Jeff_Beck`

`Beck_Hansen`[3] musikarien artean desanbiguatzea, oraindik ere erronka handia izaten jarraitzen du (Milne and Witten 2008b). Ataza hau entitate izendunen desanbiguazio gisa ezagutzen da.

1.1 Entitate izendunen desanbiguazioa (EID)

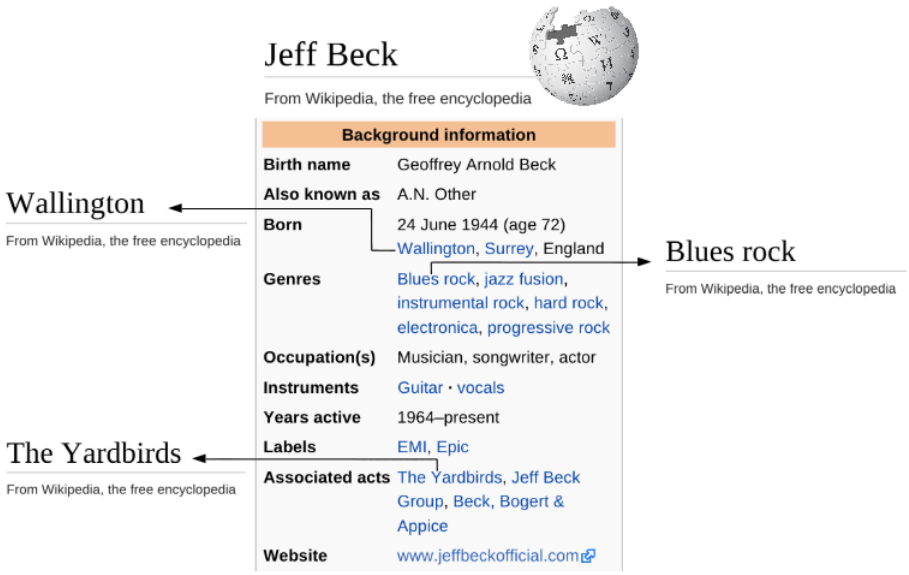
EID atazak edozein testutan azaltzen diren izen-aipamenak ezagutza-baseko entitateekin lotzean datza. Helburua 1.1 irudian azaltzen den adibidean *Clapton*, *Beck* eta *Page* izen-aipamenak `Eric_Clapton`, `Jeff_Beck` eta `Jimmy _Page` entitateetara lotzea da:



1.1 irudia – Testuko izen-aipamenak ezagutza-baseko entitateetara lotzen.

Entitateak izaera errepikaezina duten pertsona, leku edo erakundeak dira, eta ezagutza-baseetako instantziak dira. Ezagutza-baseek entitateak eta entitateekin erlazioa duten informazio egituratua gordetzen dute, horien adibide DBpedia (Bizer *et al.* 2009)[4], Wikidata[5], BabelNet (Navigli and Ponzetto 2012a) edo Freebase (Bollacker *et al.* 2008)[6] dira. Askotan, ezagutza hau modu erdi automatikoan eskuratzen da Wikipediako artikulu eta infotauletan[7] oinarrituta. Adibidez, `Jeff_Beck` artikuluaren infotaulan aurkitu

[3] `http://en.wikipedia.org/wiki/Beck`
[4] `http://wiki.dbpedia.org/`
[5] `http://www.wikidata.org/wiki/Wikidata:Main_Page`
[6] `http://wiki.freebase.com/wiki/Main_Page`
[7] Infotaulak wikipediako artikuluetan goi-eskuinaldean azaltzen informazio kutxak dira.

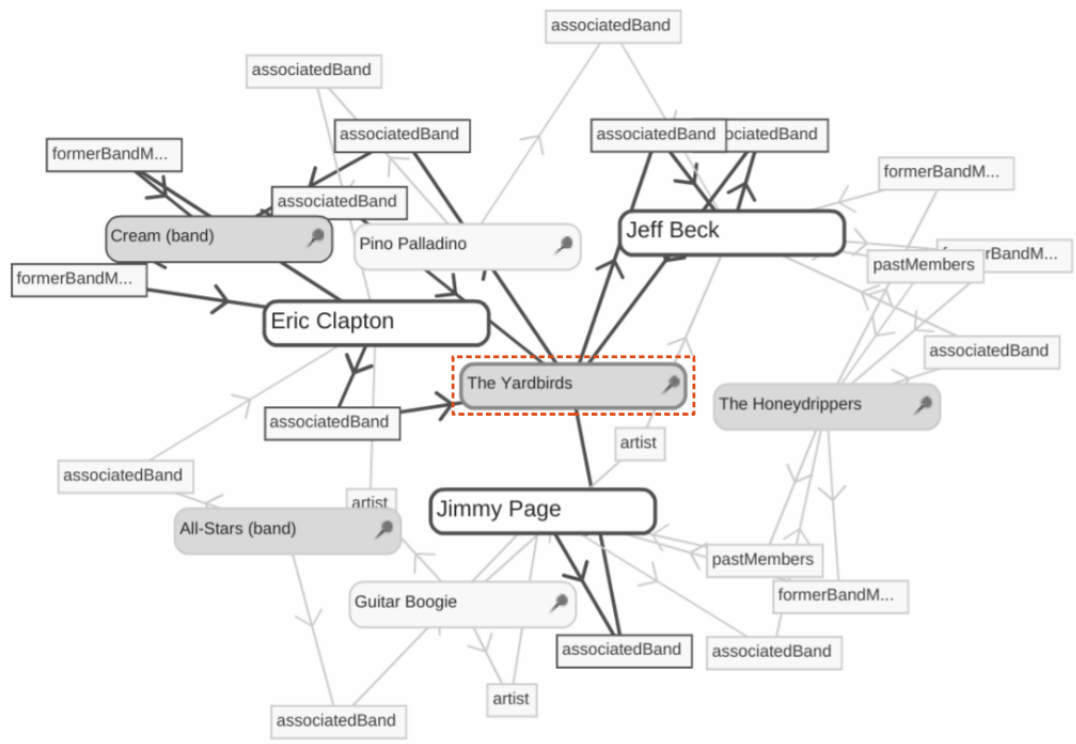


1.2 irudia – Wikipedian `Jeff_Beck` entitatearen infotaula, eta beste entitateetara estekak. (Iturria: `http://www.wikipedia.org` Data: 2016-10-24)

daitekeen informazioa hau da: jaiotze data eta lekua, musika estiloak, instrumentuak, sariak, erlazionatutako taldeak etab (ikus 1.2 irudia). Informazio hau Wikipediako orrietara estekatuta dago, beraz entitateen arteko erlazioak eskuratzeko, eta bide batez, ezagutza-baseen egitura sortzeko baliabide ezin hobea dira. Esate baterako, infotaulatik erlazio hauek erauzi daitezke:

- `Jeff_Beck` < *jaioterria* > `Wallington`
- `Jeff_Beck` < *musika estiloa* > `Blues_Rock`
- `Jeff_Beck` < *erlazionatutako taldea* > `The_YardBirds`

Artikulu eta infotauletako erlazio erauzketak, ezagutza-basea grafo gisa errepresentatzera eramaten du. Grafo honetan `Eric_Clapton`, `Jeff_Beck` eta `Jimmy_Page` entitateak, eta musikarekin zer ikusia duten erlazioak aukeratuz 1.3 irudian ikus daitekeen azpigrafoa sortu daiteke. Musikariek talde askotan parte hartu duten arren, hiruak talde bakarrean jo zuten elkarrekin.

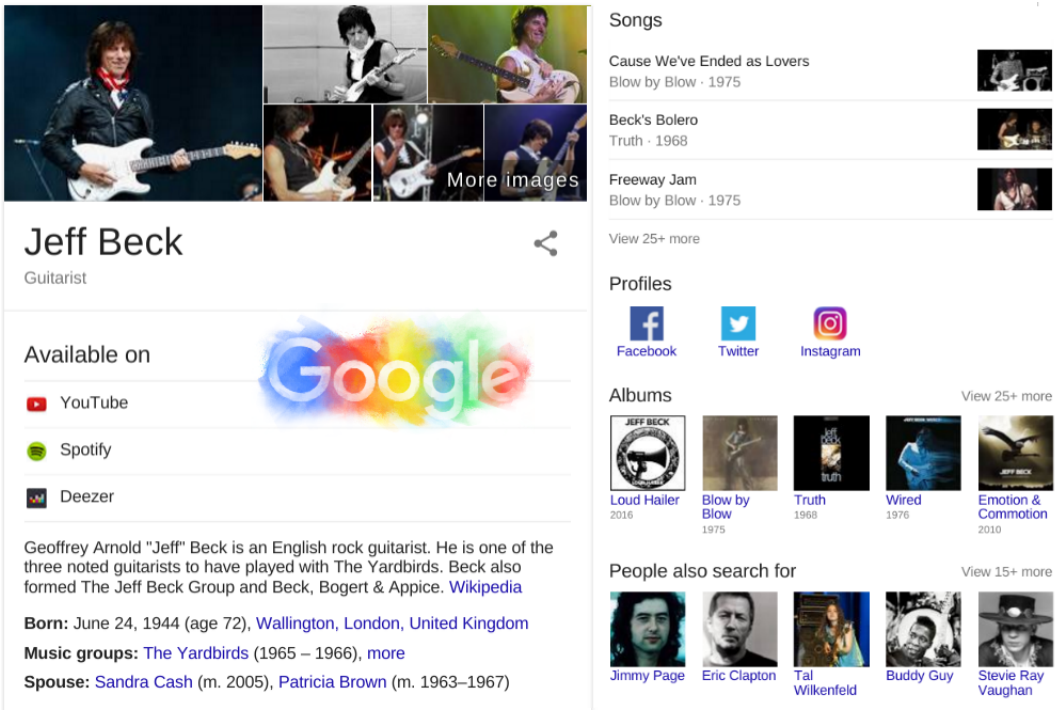


1.3 irudia – DBpediako ezagutza-basean musikari eta musika taldeen arteko erlazioak erakusten dituen diagrama. (Iturria: `http://www.visualdataweb.org/relfinder/relfinder.php` Data: 2016-07-11)

Talde hau zein den jakiteko hiru entitateen artean dauden erlazioei esker `The_Yardbirds`[8] dela egiaztatu daiteke.

Ezagutza-baseen erabileraren adibide argia Google Knowledge Graph[9] da. Google 2012. urtetik aurrera bilatzailearen kontsulten emaitzak bilaketa semantikoekin aberasten hasi zen. Bilaketa semantikoek ezagutza-baseetako informazioa ustiatzen dute. Horren adibide, Googlen *Jeff Beck* bilatuz gero ohiko web orriak bueltatzeaz gain, entitate horri lotutako informazio gehigarria azaltzen da (ikus 1.4 irudia). Informazio hau Wikipedia eta beste ezagutza-baseetatik eskuratzen dute, bide batez erabiltzaileari informazio guztia klik bakarrera hurreratuz. Bertan ikus daiteke nola kontsulta soil horrek ezagutza-baseari esker eskuratu duen informazio guztia: informazio

[8] `http://en.wikipedia.org/wiki/The_Yardbirds`
[9] `http://en.wikipedia.org/wiki/Knowledge_Graph`



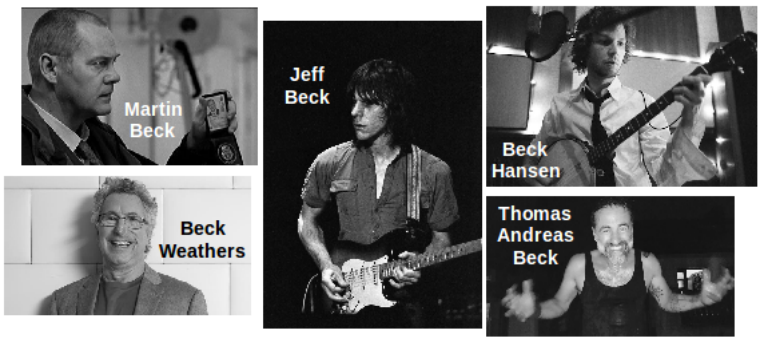
1.4 irudia – Google Knowledge Graph. (Iturria: `http://www.google.com` Data: 2016-09-11)

pertsonala, bere abesti ospetsuenen zerrenda, sare sozialetako orrietara estekak, diskografia...

Hala ere, bilaketa semantikoen arrakasta kolokan jarriko duen arazoa izen-aipamenen anbiguetatea da. Esate baterako, *Beck* izenak hainbat entitate erreferentziatu ditzake, adibidez `Jeff_Beck` edo `Beck_Hansen` musikariak, baina *Beck* izeneko mendi, laku edo irlak ere badaude.[10] Berdina gertatzen da *Clapton*[11] eta *Page*[12] izenekin. Beraz, 1.1 irudiko izen-aipamenak `Eric_Clapton`, `Jeff_Beck` eta `Jimmy_Page` entitateetara lotzen ez badira, testu horrentzat ezagutza-baseetatik eskuratuko den informazioa ez da egokia izango.

Izen-aipamenak desanbiguatzeko EID sistemek jarraituko dituzten urratsak hiru dira:

[10] `http://en.wikipedia.org/wiki/Beck_(disambiguation)`
[11] `http://en.wikipedia.org/wiki/Clapton`
[12] `http://en.wikipedia.org/wiki/Page`



1.5 irudia – Beck izen-aipamenarentzat Wikipediatik lortuko liratekeen entitate-hautagaien zerrenda.

1. urratsa: testuan azaltzen diren izen-aipamenak identifikatzea da. Tesilan honetan eskuz identifikatutakoak erabiliko dira (1.1 irudian azpimarratuak dauden *Clapton*, *Beck* eta *Page* adibidez).

2. urratsa: izen-aipamenak izanda, entitate-hautagaiak sortzea. Adibidez, *Beck* aipamenarentzat, beste batzuen artean, 1.5 irudian azaltzen diren entitate-hautagaiak sortuko dira.

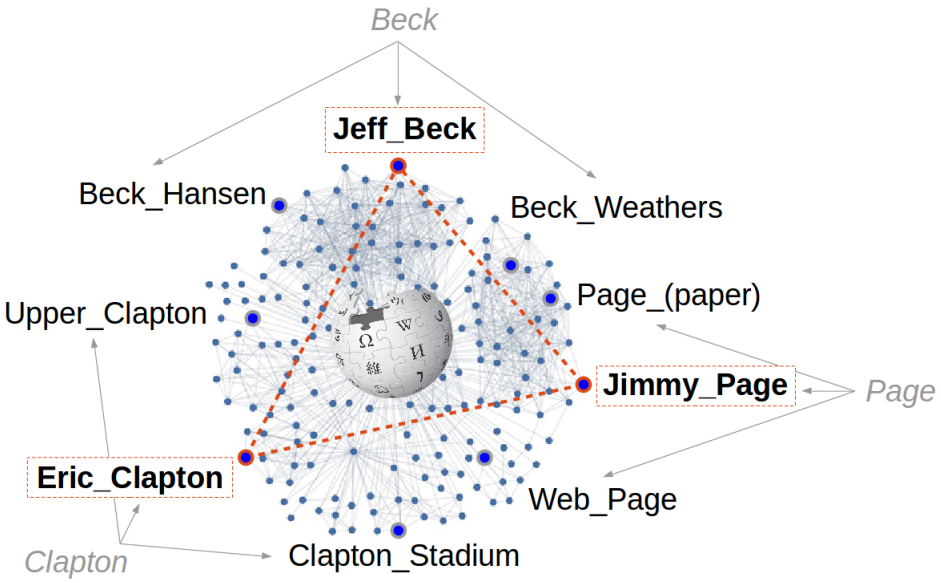
3. urratsa: izen-aipamenaren desanbiguazioa da. Kasuan kasu, hautagaien artean dagokion entitate egokia aukeratzea izango da. Tesi honen aportazioak urrats honetan egingo dira.

1.2 Oinarrizko EID sistemak

EID sistemen desanbiguazio-urratsean jarriko dugu arreta tesi honetan, alegia, testuko izen-aipamenak identifikatu eta bakoitzarentzat hautagai posibleak eskuratu ondoren. Hautagai egokia aukeratzeko desanbiguazio sistemak bi azpimultzo nagusitan banatzen dira (Ratinov *et al.* 2011), algoritmo globalak eta algoritmo lokalak, hain zuzen ere. Bi metodoak aipamena inguratzen duen testuinguruan oinarritzen dira.

1.2.1 Algoritmo Globalak

Testuinguruko aipamenetatik abiatuz, hautagai guztien artean koherenteak direnak aukeratzen dituzten algoritmoei, algoritmo global esaten zaie (Agirre



1.6 irudia – Entitateak Wikipediako hiperesteken grafoan. Marra etenak entitateen artean hiperestekak daudela errepresentatzen dute. Geziak aipamen eta hautagaien arteko erlazioa.

et al. 2014). Demagun, 1.1 irudiko adibidean *Page* `Jimmy_Page` gitarjolea dela dakigula. Hori kontuan izanda, erraza litzateke beste biak ere gitarjoleak direla jakitea. Ideia hau ezagutza-basea grafo gisa errepresentatuz gauzatu daiteke: adibidez Wikipedia oinarri hartuta adabegiak artikuluak izango dira eta ertzak artikuluen arteko hiperestekak (ikus 1.6 irudia). Hiperesteken grafoak argi erakusten du hautagaien konbinazio koherenteena zein den, hiru musikariena hain zuzen. Beraien artean ageri baitira hiperesteka zuzenak. Grafoaren egitura ustiatzeko algoritmoaren berezitasunak 3. eta 4. kapituluetan sakonduko dira.

1.2.2 Algoritmo Lokalak

Algoritmok lokalak izen-aipamena inguratzen duten testuinguruko hitzetan oinarritzen dira. Algoritmo batzuek testuingurua errepresentatzeko ordena gabeko hitz multzoak erabiltzen dituzte (Han and Sun 2011). Adibidez, 1.1 adibidean *Beck* desanbiguatzeko bere testuinguruko hitzekin multzo bat sortuko da.

- *Beck* $\{three, of, the, greatest, guitarists...clapton, and, page\}$

Ondoren, *Beck* aipamenaren entitate-hautagaietako bakoitzarentzat (`Beck_Hansen`, `Jeff_Beck` eta `Beck_Weathers` adibidez) ezagutza-baseko testuinguruekin hitzen multzoak sortuko dira:

- `Beck_Hansen` $\{musician, album, cover,$ **guitarists**$...\}$.
- `Jeff_Beck` $\{$**guitarists**, **page**, **clapton**, $jimmy...\}$.
- `Beck_Weathers` $\{everest, disaster, jenkins, richard...\}$.

Algoritmo lokalek izen-aipamenaren hitz multzoa eta entitate-hautagaien hitz multzoen artean dagoen antzekotasuna erabiltzen dute. Kasu honetan, beltzez nabarmendu dira desanbiguatu nahi den multzoarekiko amankomunean dituzten hitzak. Adibidean erraz ikus daiteke multzo antzekoena `Jeff_Beck` entitatearena dela. Ezaugarri lokalak ustiatzeko algoritmoaren berezitasunak 3. eta 5. kapituluetan sakonduko dira.

1.3 Motibazioa eta ekarpenak

Tesi-lan honen motibazio nagusia EID sistemen azterketa eta metodo berrien proposamenak dira. Honekin, ezagutza-baseen erabilera egokia eta bilaketa semantikoen arrakasta bultzatu nahi da. Motibazio honi lotutako ekarpenak bi izango dira:

- Lehenik, ezaugarri globalak ustiatzeko eredu berritzailea planteatu eta ebaluatu da, ezaugarri globalak EID atazan aplikatuz. Ekarpen honek ezagutza-basea grafo gisa errepresentatuz izen-aipamenak desanbiguatzea ahalbidetzen du. Gainera, grafoa eraikitzeko metodologiak desanbiguazioan duen ekarpena neurtu da (4. kapitulua).
- Bigarrenik, ezaugarri lokalak ustiatzen dituen sistema bat oinarritzat hartu, eta algoritmo globalarekin konbinatu da. Ekarpen honekin, batetik, eredu lokalen ekarpena neurtu da, eta, bestetik, bi informazio iturriak osagarriak direla erakutsi da. Konbinaketa honi esker arloaren egoerako emaitzak lortu dira (5. kapitulua).

Tesiaren bigarren motibazioa EID sistemak dituzten gabeziak gainditzeko edo emaitzak hobetzeko edozein sistemak barneratu ditzakeen teknika eta ezaugarri gehigarriak ikertzea da. Horretarako, orain arte arloaren egoeran jorratu ez diren bi ideia berritzaile proposatu, aztertu eta ebaluatu dira. Motibazio honi lotutako ekarpenak hauek dira:

- Lehenik, diskurtso, bilduma eta agerkidetza mailan entitateen portaera aztertu da, eta entitateek kasu guztietan patroi jakin bat betetzen dutela ikusi da. Ondoren, propietate hau hiru EID sistemetan barneratu da, eta azkenik, emaitzetan hobekuntza esanguratsuak dituztela ikusi da (6. kapitulua).

- Bigarrenik, etiketatu gabeko kanpotiko corpusetatik ezagutza eskuratu, eta desanbiguazioan egin dezakeen ekarpena aztertu da. Horretarako, EID sistema lokala informazio honekin elikatu da. Sistemak kanpo-ezagutza barneratuz emaitzak hobetzen dituela ikusi da. Baliabide honen kalitatea ere ebaluatu da, bide batez, kanpotiko ezagutzak dakarren ekarpena justifikatzeko (7. kapitulua).

Bestalde, tesi honetan hainbat baliabide ekoiztu dira:

- EID burutzeko softwarea eta baliabideak sortu dira. Alde batetik, algoritmo globalak erabiliz, beste aldetik, algoritmo lokalak uztartuz, eta azkenik, biak konbinatuz. Ekarpen hauek 4. eta 5. kapituluetan aztertzen dira.

- Entitateen portaera eta propietateak aztertzeko datu-multzoak prestatu dira. Ekarpen hau 6. kapituluko emaitzak frogatzeko erabili da.

- EID sistemak kanpo-ezagutzaz elikatzeko baliabideak sortu dira. Ekarpen hau 7. kapituluan ekoiztu da.

1.4 Tesiaren egitura eta osatzen duten argitalpenak

Tesia artikuluen bilduma gisa aurkeztuko da. Sarrera eta arloaren egoeraren atalak azaldu ondoren, hurrengo lau kapituluak (4,5,6 eta 7) ingelesez argitaratu diren artikuluak dira. Tesiaren egitura orokorra mantentzeko artikuluen formatuak aldatu dira. Gainera, formulak eta terminologia bateratu egin dira dokumentuaren osotasuna mantentzeko.

- **1. kapitulua** Sarrera.
- **2. kapitulua** Wikipedia, entitateen ezagutza-basea.
  - Atal honek tesi honen ardatz nagusia den Wikipedia aztertuko du. Izan ere, garatu diren sistemek erabiliko duten ezagutza-basea eta baliabide nagusia da.
- **3. kapitulua** Arloaren egoera.
  - Kapitulu honetan arloaren egoera errepasatu, eta EID sistemak sakonago aztertuko dira. Gainera, tesi honetan garatu diren sistema eta baliabideak ulertu ahal izateko azalpenak emango dira.
- **4. kapitulua** Algoritmo globalak, ausazko ibilbideak Wikipedia grafoan.

  Eneko Agirre, Ander Barrena and Aitor Soroa. ***Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation.*** *arXiv.org CoRR 2015.*

  - Artikuluen bildumako lehen artikuluan Wikipediatik erauzitako grafo ezberdinak aztertuko dira antzekotasun eta desanbiguazio atazetan. Horretarako, Wikipediatik hiperesteka ezberdinen ekarpena ebaluatuko da. Eredu globalak EID atazan aplikatuko dira hainbat datu-multzo ezberdinetan emaitzak emanez.
- **5. kapitulua** Algoritmo globalak eta lokalak konbinatzen.

  Ander Barrena, Aitor Soroa and Eneko Agirre. ***Combining Mention Context and Hyperlinks from Wikipedia for Named Entity Disambiguation.*** *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics *SEM 2015.* Denver, Colorado, USA. 2015.

  - Bigarren artikuluak tesi honetan garatu den algoritmo lokala aztertuko du, ordenarik gabeko hitz multzoetan oinarritzen dena. Bestalde, aurreko artikuluan erabili den eredu globalarekin konbinatuko da. Konbinazio honek arloaren egoerako emaitzak gaindituko ditu hainbat datu-multzotan, eta besteetan sistema hoberenen besteko emaitzak lortuko ditu.

- **6. kapitulua** Entitate bakarra diskurtsoan eta agerkidetzan.

  Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Peñas and Aitor Soroa. ***"One Entity per Discourse" and "One Entity per Collocation" Improve Named-Entity Disambiguation.*** *Proceedings of the 25th International Conference on Computational Linguistics COLING 2014.* Dublin, Ireland. 2014.

  – Hirugarren artikuluak entitateek corpusetan betetzen duten propietate bat aztertuko du. Propietate honen arabera, testu batean izen-aipamen berdina behin baino gehiagotan azaltzen bada, %96-98an entitate berdina erreferentziatuko du. Azterketa hau testuetatik agerkidetza sintaktikoetara zabalduko da, propietateak %91-98an betetzen jarraitzen duela ikusiz. Azkenik propietate hau sistema ezberdinen desanbiguazioaren emaitzan aplikatuko da, testuan aipamen berdinari usuen azaldu den entitatea esleituz. Teknika erraz honek aztertu diren sistema guztietan emaitzak modu adierazgarrian hobetuko ditu.

- **7. kapitulua** EID sistemak kanpo-ezagutzaz elikatzen.

  Ander Barrena, Aitor Soroa and Eneko Agirre. ***Alleviating Poor Context with Background Knowledge for Named Entity Disambiguation.*** *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ACL 2016.* Berlin, Germany. 2016

  – Laugarren artikuluak eredu lokala oinarritzat hartu, eta kanpotiko corpusetatik lorturiko informazioaz elikatuko du. Kanpo-ezagutzak sistemaren emaitzak hobetuko ditu arloaren egoerako emaitzak lortuz. Bestalde, informazio gehigarri honen azterketa sakona egingo da bere erabilgarritasuna frogatzeko.

- **8. kapitulua** Ondorioak eta etorkizuneko ildoak.

  – Kapitulu honetan ideia orokorrak laburbildu eta ondorio nagusiak azalduko dira. Gainera, etorkizunerako lanak planteatuko dira egindako ikerketak aurrera jarrai dezan.

1.5 Bestelako argitalpenak eta kolaborazioak

Jarraian EID atazaren inguruan egin diren gainontzeko argitalpenak eta kolaborazioak zerrendatuko dira. Bide batez, tesi honetan garatu diren EID sistemen aplikazio errealak azalduko dira.

- **UKP-UBC Entity Linking at TAC-KBP.**

  Nicolai Erbs, Eneko Agirre, Aitor Soroa, Ander Barrena, Ugaitz Etxebarria, Iryna Gurevych, Torsten Zesch. ***UKP-UBC Entity Linking at TAC-KBP.*** *Text Analysis Conference, Knowledge Base Population 2012.* Mariland, USA. 2012.

  – Artikulu honek Text Analisys Conference - Knowledge Base Population *TAC-KBP*[13] konferentzian *Entity Linking* atazan argitaratutako sistemaren oinarriak laburbiltzen ditu. Konferentzia honetako helburua testu hutsetik entitate-izendunen ezagutza-baseak eraiki eta aberasteko sistemak garatzea da. Horretarako, txapelketa ezberdinak antolatzen dituzte. Ikerlari talde ezberdinek beraien sistemen emaitzak eman, eta parte hartzaile guztien artean sailkapena egiten dute. Argitalpen honetan urtero antolatzen den *Entity Linking* atazan 2012. urtean aurkeztu zen sistemaren oinarriak azaltzen dira.

[13] `http://tac.nist.gov/`

- **UBC Entity Linking at TAC-KBP 2013: random forests for high accuracy.**

  Ander Barrena, Eneko Agirre and Aitor Soroa. ***UBC Entity Linking at TAC-KBP 2013: random forests for high accuracy.*** *Text Analysis Conference, Knowledge Base Population 2013.* Mariland, USA. 2013.

  – Ildo beretik, 2013. urtean antolatu zen *TAC-KBP* Entity Linking atazako sistemaren nondik norakoak azaltzen dira. Tesi honetan azaltzen diren sistema global eta lokalaren konbinaketa bitartez, atazako bigarren emaitzarik onena lortu zen. Sistemak ausazko erabaki-zuhaitzetan oinarritutako algoritmoa erabiltzen du bi algoritmoen ezaugarriak konbinatzeko.

- **UBC Entity Discovery and Linking and Diagnostic Entity Linking at TAC-KBP 2014.**

  Ander Barrena, Eneko Agirre and Aitor Soroa. ***UBC Entity Discovery and Linking and Diagnostic Entity Linking at TAC-KBP 2014.*** *Text Analysis Conference, Knowledge Base Population 2014.* Mariland, USA. 2014.

  – Oraingoan, 2014. urtean *TAC-KBP* Entity Linking atazarako garatu zen sistemaren artikulua aurkezten da. Urte honetako atazan aipamenen identifikazioa, desanbiguazioa eta klaseetan sailkatzea eskatzen zuten. Aurreko urtean aurkeztutako sistema hedatu zen atazako beharretara. Emaitzetan ikusi zen sistemak atazaren beharrei taxuz erantzun ziola. Azpimarratzekoa da desanbiguazioari dagokionez sistemak emaitza onak lortu zituela.

- **UBC Entity Recognition and Disambiguation at ERD 2014.**

  Ander Barrena, Eneko Agirre and Aitor Soroa. ***UBC Entity Recognition and Disambiguation at ERD 2014.*** *Entity Recognition and Disambiguation Challenge - ERD 2014.* Gold Coast, Australia. 2014.

  - ERD 2014[14] txapelketaren helburua, entitateen aipamen identifikazioa eta desanbiguazioa bultzatzea da. Horretarako, txapelketa bat antolatu zen bi azpimultzo nagusitan banatua. Batetik testu motzen desanbiguazioa, hau da, bilatzaileetan egiten diren kontsultena. Bestetik, testu luzeen edo arrunten desanbiguazioa. Ataza honen gakoa desanbiguazioa denbora tarte baten barruan egitea da, 20 segundo testu motzetan eta 60 segundo luzeetan. Aurkeztutako sistemak 6. eta 10. postuak lortu zituen aipatutako atazetan. Sistema hau algoritmo globaletan oinarritzen da desanbiguazioa burutzeko.

- **Izen-aipamenak desanbiguatu eta Wikipediara lotzen.**

  Ander Barrena, Eneko Agirre, Jokin Perez de Viñaspre eta Aitor Soroa. ***Izen-aipamenak desanbiguatu eta Wikipediara lotzen.*** *Ikergazte - Firts Conference For Basque Researchers 2015.* Durango, Basque Country. 2015.

  - Artikulu hau 5 kapituluan aurkeztuko den artikuluaren euskarazko bertsioa da. Horretaz aparte, euskarazko EID sistema sortzeko lehen urratsak eman dira. Artikulua Ikergazte[15] konferentzian aurkeztu zen.

- **Matching Cultural Heritage items to Wikipedia.**

  Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando and Mark Stevenson. ***Matching Cultural Heritage items to Wikipedia.*** *Proceedings of the eighth international conference on Language Resources and Evaluation LREC-2012* Istanbul, Turkey. 2012.

  - Argitalpen honek ondare kulturaleko elementuak sistema automatiko bidez Wikipedia artikuluekin aberastea posible den ikertu eta ebaluatzen du. Horretako aipamenen identifikazio eta desanbiguaziorako oinarrizko sistema azaltzen du. Bide batez EID atazaren aplikazio erreal baten prototipoa erakusten du.

[14] `http://www.msr-waypoint.net/en-us/um/people/minchang/publication/ERD2014.pdf`

[15] `http://www.ueu.eus/ikergazte/`

- **PATHSenrich: a Web Service Prototype for Automatic Cultural Heritage Item Enrichment.**

  Agirre E., Barrena A., Fernandez K., Miranda E., Otegi A. and Soroa A. ***PATHSenrich: a Web Service Prototype for Automatic Cultural Heritage Item Enrichment.*** *Research and Advanced Technology for Digital Libraries, International Conference on Theory and Practice of Digital Libraries, TPDL 2013.* Valletta, Malta. 2013.

  – Artikulu honek aurreko artikuluaren prototipoa oinarritzat hartuta, aplikazio erreala aurkezten du.

- **Lexical semantics, Basque and Spanish in QTLeap: Quality Translation by Deep Language Engineering Approaches.**

  Eneko Agirre, Iñaki Alegria, Nora Aranberri, Mikel Artetxe, Ander Barrena, Antonio Branco, Arantza Diaz de Ilarraza, Koldo Gojenola, Gorka Labaka, Arantxa Otegi and Kepa Sarasola. ***Lexical semantics, Basque and Spanish in QTLeap: Quality Translation by Deep Language Engineering Approaches.*** *Procesamiento del Lenguaje natural SEPLN 2015.* Alicante, Spain. 2015.

  – Artikulu honetan itzulpen automatikoa hobetzeko helburuaz semantika eta analisi sintaktiko sakona erabiltzen dira. Semantikaren arloan EID sistemak erabiltzen dira itzulpenaren kalitatea hobetzeko.

Tesi honek iraun duen denboran hainbat proiektutan parte hartu da, EID sistemak hainbat domeinutan aplikatuz.

- **PATHS: Personalized Access To cultural Heritage Spaces,**[16] Proiektu hau, Europeana[17] gisako ondare kulturaleko bilduma erraldoietan, bilduma digitalen arteko bisitaldi gidatu eta pertsonalizatuak eskaintzeko aplikazioa garatzean datza. Ibilbide honetako elementuak Wikipedia artikuluekin aberasten dituen EID sistema garatu zen.

[16] `http://group.europeana.eu/web/guest/details-paths/`
[17] `http://www.europeana.eu/portal/en`

- **READERS: Evaluation And DEvelopment of Reading Systems,**[18] Proiektu honen helburua kanpo-ezagutza automatikoki eskuratzeko asmoz, testu kantitate erraldoiak irakurtzeko sistema garatzea da. Horretarako, testuetako izen-aipamenak ezagutza-baseetara lotzen dira.

- **QTLeap: Quality Translation by deep Language Engineering Approaches,**[19] Proiektu honek itzulpen automatikoaren kalitatea hobetzeko asmotan, hizkuntzaren sakoneko ingeniaritza metodologiak aztertzen ditu . Batez ere, izenen itzulpenaren kalitatea hobetzeko, itzuli nahi den testuan izenak desanbiguatuak izatea eskatzen da.

- **TUNER: Automatic domain adaptation for semantic processing.** Proiektu honetan domeinu egokitzapen automatikoa semantikaren prozesamenduan burutzen da. Horretarako, ezinbesteko ataza izango da EID.

1.6 Aurrekariak Ixa taldean

Tesi honen ikerketa IXA taldean[20] egin da. Talde honek Euskal Herriko Unibertsitatean hizkuntzaren prozesamenduan dihardu lanean. Bere ikerketaren ardatza euskarara bideratzen duen arren, beste hizkuntzetan ere adituak dira. Tesi-lan hau IXA taldean kokatzeko, honekin zerikusia duten beste hiru tesi aipatuko dira jarraian.

EID atazak erlazio estua du hitzen adiera desanbiguazioarekin. Hitzek eta izen-aipamenek testuinguruaren arabera adiera ezberdina dute. HAD atazan hainbat lan egin da IXA taldean. Adibidez (Martinez 2004) tesian, tresna bat garatzeko lehen urratsak eman ziren. Horretarako, ezaugarri sintaktiko, semantiko eta domeinukoak erabiltzen ziren. Bestalde, (Lopez de Lacalle 2009) tesian kernel metodo ezberdinak uztartu ziren domeinu aldaketei aurre egiteko.

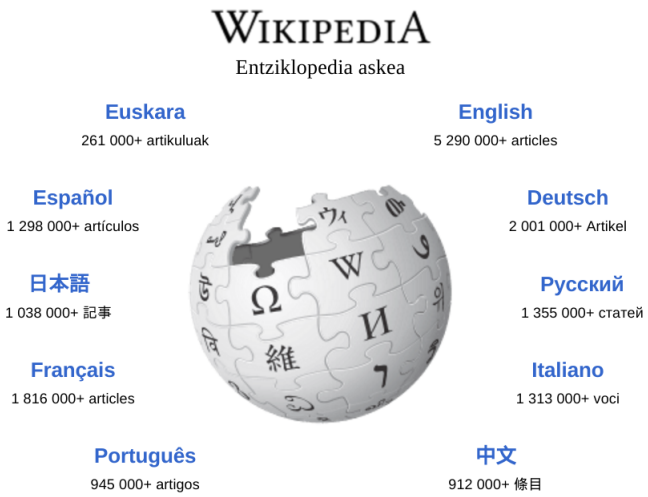
Tesi-lan honen aurrekari gisa, (Fernandez 2012) tesiak euskarazko entitateak automatikoki lantzen ditu. Horretarako, entitate izenak identifikatu, sailkatu, itzuli eta desanbiguatzen ditu. Baliabide urriko hizkuntza izanik, baliabideen berrerabilpenean eta metodo ez-gainbegiratuetan arreta jartzen

[18] `http://www.chistera.eu/projects/readers`
[19] `http://qtleap.eu/`
[20] `ixa.si.ehu.es`

du. Metodo sinpleen konbinaketak hobesten ditu metodo sofistikatuen aurrean. Azkenik, euskararen ezaugarri morfosintaktikoen eragina aztertzen du ataza honetan.

2

Wikipedia, entitateen ezagutza-basea



Atal honetan tesi honetako ardatz nagusia aztertuko da, Wikipedia hain zuzen. Batetik, EID sistemek erabiliko duten ezagutza-basea da, eta desanbiguatuko diren izen-aipamenak Wikipediako artikuluetara lotuko dira. Bestetik, EID sistemek erabiliko duten ezagutza bertatik eskuratuko da.

Wikipedia[1] Wikimedia Foundation[2]-en entziklopedia eleanitza eta eduki askekoa da. Bertako artikuluak mundu osoko erabiltzaileek idazten dituzte,

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

eta orduro eguneratzen diharduen baliabidea da. Gaur egun, 294 hizkuntza ezberdinetan idatzitako 41 milioi artikuluk osatzen dute.[3]

Jarraian, Wikipediako egitura osatzen duten ezaugarriak banan banan azalduko dira: artikuluak, aingurak, birbideratze-orriak eta desanbiguazio-orriak hurrenez hurren.

2.1 Artikuluak

Artikuluak edo sarrerak eduki entziklopedikoa duten orriak dira, eta kontzeptuak edo entitateak deskribatzen dituzte. Kontzeptuak objektu baten irudikapen abstraktuak dira, adibidez musikariak (`Musician`) edo musika taldeak (`Musical_ensemble`). Entitateak existitzen diren (edo ziren) pertsona, leku edo erakundeak dira, adibidez, `Jeff_Beck` eta `The_Yardbirds`.

Artikuluen identifikadore unibokoa titulua da, eta tituluaren aldaerak edo formak, birbideratze-orrien eta desanbiguazio-orrien bitartez artikulura lotzen dira. Birbideratze eta desanbiguazio-orriak ez dira artikuluak kontsideratzen, ez baitute eduki entziklopedikoa eskaintzen.

2.2 Aingurak

Artikuluen edukian beste artikuluetara doazen estekak aingura bitartez egiten dira. Wikipedian artikulu-tituluak ez bezala, aingura-testuak errepikatuak egon daitezke. Aingura-testuak erreferentziatu duten artikuluaren tituluaz agertu daitezke. Adibidez, 2.1 irudian Wikipediako artikulu baten edukia ikus daiteke. *The Yardbirds* aingurak `The_Yardbirds` artikulura estekatzen du. Hala ere, aingura-testuetan artikulua izendatzeko modu eta aldaera ezberdinak azaldu daitezke. Esate baterako, adibideko testuan *rock* aingurak `Rock_music` artikulura estekatzen du.

2.3 Birbideratze-orriak

Birbideratze-orriak artikulu-tituluen aldaerak edo formak artikulura lotzeko beste bide bat dira. Erabiltzailea ahalik eta azkarren artikulura bideratzeko

[3]Data 2016-11-24

Jeff Beck

From Wikipedia, the free encyclopedia

Geoffrey Arnold "Jeff" Beck (born 24 June 1944) is an English rock guitarist. He is one of the three noted guitarists to have played with The Yardbirds (the other two being Eric Clapton and Jimmy Page). Beck also formed The Jeff Beck Group and Beck, Bogert & Appice.

Much of Beck's recorded output has been instrumental, with a focus on innovative sound, and his releases have spanned genres ranging from blues rock, hard rock, jazz fusion, and an additional blend of guitar-rock and electronica. Although he recorded two hit albums (in 1975 and 1976) as a solo act, Beck has not established or maintained the sustained commercial success of many of his contemporaries and bandmates.[1][2] Beck appears on albums by Mick Jagger, Tina Turner, Morrissey, Jon Bon Jovi, Malcolm McLaren, Kate Bush, Roger Waters, Donovan, Stevie Wonder, Les Paul, Zucchero, Cyndi Lauper, Brian May, Stanley Clarke, Screaming Lord Sutch and ZZ Top.

2.1 irudia – Wikipedian `Jeff_Beck` artikuluaren lehen bi parrafoak. Urdinez dauden hitz-kateak beste artikuluetara aingurak dira.

sortuak dira. Birbideratzeen bidez artikuluak izendatzeko sinonimoak, laburdurak, izen-aldaerak edo errore ortografikoak kudeatzen dira. Adibidez, `Geoffrey_A._Beck` eta `Geoffrey_Beck` orriak birbideratze-orriak dira, hain zuzen, `Jeff_Beck` artikulura birbideratzen dutenak.

2.4 Desanbiguazio-orriak

Wikipediako desanbiguazio-orriek adiera ezberdinen artean bereizteko loturak eskaintzen dituzte, eta artikulu-titulua oso anbiguoa den kasuetan erabiltzen dira. Adibidez, `Beck` artikuluak Beck Hansen musikaria deskribatzen du, baina titulua oso anbiguoa da. Horregatik, `Beck_(disambiguation)` orriak beste adiera posibleak eskaintzen ditu (ikus 2.2 irudia). Desanbiguazio-orrian adiera lehenetsia `Beck` (Beck Hansen) da, baina bestelako artikuluetara estekak azaltzen dira: `Beck_Weathers`, `Beck_Lakes` edo `River_Beck`. Hala ere, desanbiguazio-orrietan ez dira adiera guztiak azalduko, ez dira exhaustiboak. Adibidez, `Jeff_Beck` artikulua ez da estekatuta azaltzen.

Beck (disambiguation)

From Wikipedia, the free encyclopedia

Beck (born 1970) is an American singer and songwriter.

Beck may also refer to:

People [edit]

- Beck (surname), including a list of people with the family name
- Beck Weathers (born 1946), American pathologist and survivor of the 1996 Mount Everest disaster

Places [edit]

- Beck (stream), an English dialect word for a small stream
- Beck Lakes, twin lakes in California
- River Beck, in South London, England
- Mount Beck, Mac. Robertson Land, Antarctica
- Beck Peak, Ross Dependency, Antarctica
- Cape Beck, Black Island, Ross Dependency, Antarctica

2.2 irudia – Wikipedian `Beck_(disambiguation)` desanbiguazio-orriaren adibidea.

2.5 Wikipediatik informazioa erauzten

Tesi honetan landuko diren EID sistemek erabiliko duten ezagutza Wikipediatik eskuratuko da. Bertako ezaugarriak ustiatze aldera, ezagutza lau baliabidetan errepresentatuko da: ezagutza-basea, hiztegia, grafoa eta artikuluen testuinguru-bildumak izenekoak.

2.5.1 Ezagutza-basea

Tesian landuko diren EID sistemen desanbiguazioaren emaitza Wikipediako entitate edo kontzeptuak dira. Honenbestez, Wikipediako artikulu bildumak ezagutza-basea errepresentatuko du.

Ezagutza-baseen artean DBpedia (Bizer *et al.* 2009), Wikidata, BabelNet (Navigli and Ponzetto 2012a) edo Freebase (Bollacker *et al.* 2008) aurkitu daitezke, eta hauen oinarria ere Wikipedia da. Beraz, tesi honetako EID sistemen irteerak zuzenean lotu daitezke ezagutza-base hauetara.

2.5.2 Hiztegia

EID sistemek testuetan azaltzen diren izen-aipamenak Wikipediako artikulu egokietara lotuko dituzte. Horretarako, testua eta artikuluen arteko zubia eraikiko da, hiztegia hain zuzen ere (Chang *et al.* 2010).

Hiztegiaren egitura sarrera eta erlazionatutako artikuluen zerrendaz osatzen da. Sarrerak letra xehe eta azpimarraz elkarturiko hitz-kateek osatuko dute. Hiztegian sarrera eta artikuluen erlazioa maiztasunarekin puntuatuko da, eta hau litzateke egituraren adibidea:

sarrera `artikulua1`:maiztasuna `artikulua2`:maiztasuna

Hiztegia eraikitzeko lehenik eta behin, Wikipediako artikulu bakoitza bere tituluaz erlazionatuko da. Sarrera sortzean parentesi arteko hitzak egongo balira, hauek kenduko lirateke. Adibidez:

- *jeff_ beck* `Jeff_Beck:1`
- *eric_ clapton* `Eric_Clapton:1`
- *jimmy_ page* `Jimmy_Page:1`
- ...

Ondoren, Wikipediako birbideratze-orrien tituluak birbideratzen dituzten artikuluekin erlazionatuko dira (ikus 2.3 ataleko birbideratze-orrien adibideak):

- *geoffrey_ a._ beck* `Jeff_Beck:1` ←
- *geoffrey_ beck* `Jeff_Beck:1` ←
- *jeff_ beck* `Jeff_Beck:1`
- *eric_ clapton* `Eric_Clapton:1`
- *jimmy_ page* `Jimmy_Page:1`
- ...

Wikipediako desanbiguazio-orrien tituluak orrian estekatzen diren adiera guztiekin erlazionatuko dira (ikus 2.4 ataleko desanbiguazio-orriaren adibidea)[4]:

[4] `Beck_(disambiguation)` titulua sarrera bihurtzean, letra xehez eta parentesi artekoa kenduz egingo da → *beck*.

- *beck* `Beck:1` `Beck_(surname):1` `Beck_Weathers:1` ... ←
- *geoffrey_ a._ beck* `Jeff_Beck:1`
- *geoffrey_ beck* `Jeff_Beck:1`
- *jeff_ beck* `Jeff_Beck:1`
- *eric_ clapton* `Eric_Clapton:1`
- *jimmy_ page* `Jimmy_Page:1`
- *...*

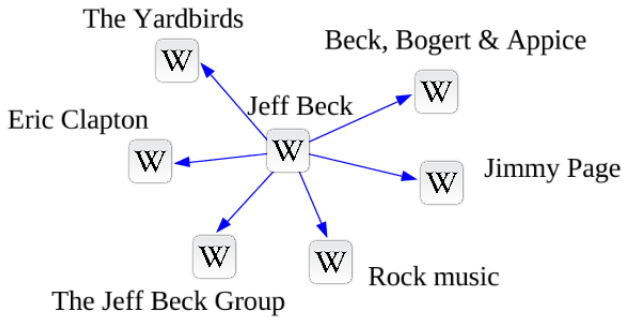
Azkenik, Wikipediako artikuluetako aingurak sarrera-artikulu erlazio gisa erabiliz, hiztegia osatuko da:

- *beck* `Beck:1555` `Beck_(manga):70` ... `Jeff_Beck:3` ... ←
- *geoffrey_ a._ beck* `Jeff_Beck:1`
- *geoffrey_ beck* `Jeff_Beck:1`
- *jeff_ beck* `Jeff_Beck:1210` `The_Jeff_Beck_Group:1` ... ←
- *eric_ clapton* `Eric_Clapton:3339` `Eric_Clapton_(album):28` ... ←
- *jimmy_ page* `Jimmy_Page:1418` `Jimmy_Page_(footballer):3` ... ←
- *...*

2.5.3 Hiperesteken grafoa

Wikipediako hiperesteken egitura grafo gisa errepresentatu daiteke, aurrerago grafo honek duen informazioa ustiatzeko helburuarekin. Algoritmo globaletan oinarritzen diren sistema gehienek hiperesteken grafoa edo honen azpimultzo bat erabiltzen dute (Alhelbawy and Gaizauskas 2014, Moro *et al.* 2014, Chisholm and Hachey 2015, Pershina *et al.* 2015).

Grafoa eraikitzeko urratsak bi dira: adabegiak artikuluak dira, eta ertzak beraien artean existitzen diren hiperestekak. Demagun 2.1. irudian ikusten den `Jeff_Beck` artikuluko edukia aztertzen dela. Adibidean, lehen parrafoa hartu, eta hiperesteka edo aingura bakoitza, dagokion artikuluarekin lotuz, 2.3 azaltzen den azpigrafo zuzendua sortzen da. Wikipedia goitik behera prozesatuz, eta artikulu bakoitzean azaltzen diren estekak dagokion artikulura lotuz, Wikipediako grafoa sortzen da.

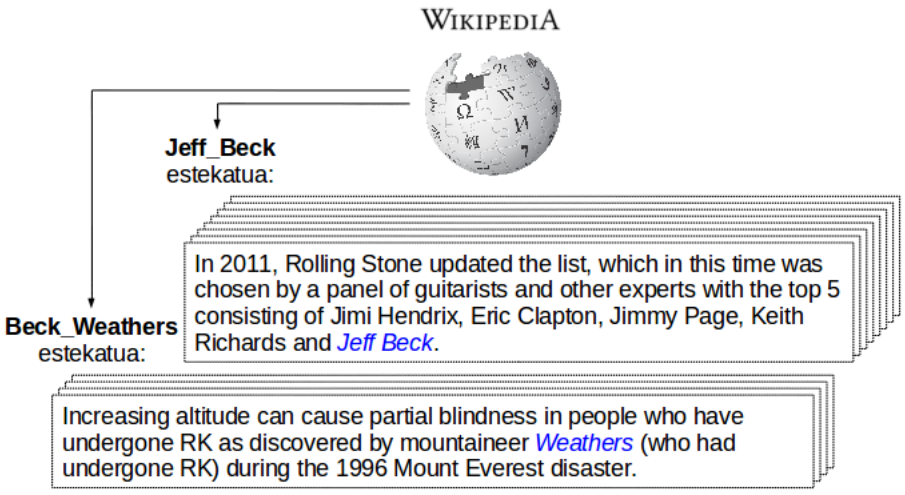


2.3 irudia – Wikipediako `Jeff_Beck` entitatearen artikuluko lehen parrafoko hiperestekekin sortutako grafo zuzendua.

2.5.4 Artikuluen testuinguru-bildumak

Wikipedia ikasketa-corpus gisa errepresentatu daiteke, ikasketa-automatikoa aplikatzeko helburuarekin (Milne and Witten 2008b). Baliabide hau artikuluen testuinguru-bildumak gisa izendatuko da. Orokorrean, algoritmo lokaletan oinarritutako sistema gehienak erabiltzen duten baliabidea da (Han and Sun 2011, Houlsby and Ciaramita 2014, Lazic *et al.* 2015).

Artikulu bakoitza estekatuta azaldu deneko testuinguru-bildumak eginez, Wikipedia ikasketa-corpus gisa errepresentatzen da. Horrela, artikulu bakoitza zein aingura-testurekin eta zein testuingurutan azaldu den erakutsiko da. Baliabidearen adibidea `Jeff_Beck` eta `Beck_Weathers` artikuluentzat 2.4 irudian ikus daiteke.



2.4 irudia – Wikipediako `Jeff_Beck` eta `Beck_Weathers` artikuluen testuinguru-bildumen adibideak. Beltzez testuingurua eta urdinez aiguratestua ikus daitezke.

3

Arloaren egoera

Atal honetan arloaren egoeraren azterketa egingo da, arreta berezia jarriz algoritmo global eta lokaletan. Horretarako, teknika hauek aplikatzen dituzten bi artikulu sakonduko dira. Izan ere, hauek izango dira tesi honetan garatuko diren EID sistemen oinarri nagusiak. Ondoren, arloaren egoerako EID sistemen azterketa orokorra egingo da. Sistemetako askok bi algoritmoen ezaugarriak konbinatzen dituztela ikusiko da.

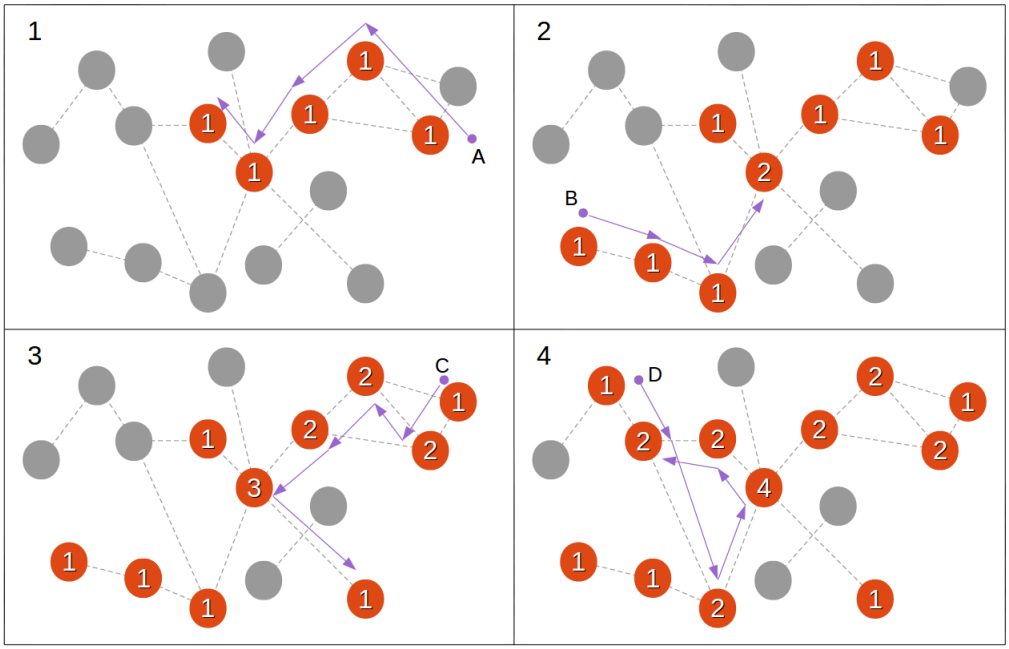
Arloaren egoerako EID sistemek eskuz identifikatutako izen-aipamenak erabiltzen dituzte, eta hautagaien-sorkuntza hiztegi bidez egiten dute.[1] Orokorrean, ez da arreta berezia jartzen hautagaien-sorkuntzaren atalean, naiz eta, sistemaren eraginkortasunarentzat oso garrantzitsua izan.

3.1 Ausazko ibilbideak: algoritmo globalak

EID sistema asko algoritmo globaletan oinarritzen dira izen-aipamenak desanbiguatzeko (Agirre and Soroa 2009, Hoffart *et al.* 2011, Alhelbawy and Gaizauskas 2014, Moro *et al.* 2014, Pershina *et al.* 2015). Algoritmo globalak, orokorrean, sistema ez-gainbegiratuak dira, eta ezagutza-baseen egitura erabiltzen dute desanbiguazioa burutzeko.

Horretarako, ezagutza-basetik erauzitako grafoan zehar ausazko ibilbideak algoritmoa aplikatu daiteke. Hau azaltzeko (Agirre *et al.* 2014) artikuluko *PageRank* algoritmoa (Brin and Page 1998) erabiliko da.

[1] Izen-aipamena hiztegiko sarreran bilatu, eta erlazionatutako artikuluak (entitateak edo kontzeptuak) hautagaiak izatera pasako dira.



3.1 irudia – Ausazko ibilbidea 4 urratsetan errepresentatu da: A,B,C eta D puntuetan hasten direnak. Urrats bakoitzak ausazko ibilbidea erakusten du, ibiltariak ausazko salto bat egitea erabaki duen arte. Zenbakiek adabegi bakoitza zenbat aldiz bisitatu den adierazten dute.

PageRank algoritmoaren ideia nagusia grafoko adabegien garrantzia kuantifikatzea da, grafoaren egituran duten garrantzia erlatiboa kontuan izanda. Demagun grafoan zehar ausazko ibilbide bat egiten dela. Demagun ere ausazko ibiltariak, adabegi bat bisitatzen duen bakoitzean bi aukera dituela: ausazko ibilbidean jarraitu edo ausaz grafoko beste adabegi batera salto egin (ikus 3.1). Ausazko ibilbidean zehar ibiltaria behin baino gehiagotan pasatuko da ondo konektatuta dauden adabegietatik, eta ez hainbeste konexio gutxi dituztenetatik. Ibilbide honetatik probabilitate bat zenbatetsi daiteke grafoko adabegi bakoitzeko.

Izan bitez N adabegidun G grafoa, eta d_i, i adabegitik beste adabegietara doazen esteka kopurua. Demagun M matrizea $N \times N$ tamainakoa dela, eta $M_{ij} = \frac{1}{d_i}$ baldin eta esteka bat existitzen bada $i$tik $j$tara. N elementu dituen P bektorearen kalkulua ekuazio honek definituko luke:

$$\mathbf{P} = cM\mathbf{P} + (1 - c)v \tag{3.1}$$

Formularen lehen batukariak ibiltariaren ausazko ibilbidea errepresentatzen du, eta bigarrenak, ibiltariak ausaz grafoko edozein adabegitara salto egiteko probabilitatea. $v = N \times 1$ bektorearen balioak $1/N$ra hasieratuz, grafoko edozein adabegietara salto egiteko probabilitatea uniformea da. Batukarien garrantzia c koefizienteak zehazten du, eta orokorrean, ibiltariak ausazko ibilbidean jarraitzea hobesten duten balioak ezartzen dira. *PageRank* algoritmoa 3.1 ekuazioa iteratiboki eta konbergitu arte exekutatuz kalkulatzen da. Ausazko jauziak egin ezean, ezin da bermatu P bektorearen kalkuluak konbergituko duenik.

Orain arteko azalpenean v bektorea $1/N$ balio uniformeekin definitu da. Beraz, ibiltariak grafoko edozein adabegitara probabilitate berdinarekin salto egingo du. Baina bektore hau moldatuz, ibiltariaren ausazko saltoak adabegi jakin batzuetara bideratu daitezke, *PageRank* algoritmoaren bertsio pertsonalizatua exekutatuz (Personalized *PageRank* edo PPR).

3.1.1 Ausazko ibilbideak EID atazan

EID atazari berriz eutsiz, ausazko ibilbide pertsonalizatuak erabiliz izen-aipamenak desanbiguatzeko adibide bat planteatuko da. Demagun, testu berean *Clapton*, *Beck* eta *Page* aipamenak agertzen direla, eta hiztegitik artikulu-hautagaien zerrendak sortzen direla:

- *Clapton* → `Eric_Clapton`, `Upper_Clapton` eta `Clapton_Stadium`
- *Beck* → `Beck_Weathers`, `Beck_Hansen` eta `Jeff_Beck`
- *Page* → `Page_(paper)`, `Jimmy_Page` eta `Web_Page`

Izan bedi Wikipediako grafoa non adabegiak artikuluak diren, eta ertzak artikuluen arteko hiperestekak. *Beck* desanbiguatzeko, nahikoa litzateke ausazko saltoak *Clapton* eta *Page* izen-aipamenen artikulu-hautagaietara bideratzea. Honek, ausazko ibilbideen pertsonalizazioa testuinguruaren arabera gidatzen du. Azkenik, ausazko ibilbide pertsonalizatuaren emaitza `Beck_Weathers`, `Beck_Hansen` eta `Jeff_Beck` artikuluentzat jaso, eta lortu duten probabilitate banaketaren arabera sailkatuko lirateke.

3.2 Hitz multzoak eta eredu sortzailea: algoritmo lokalak

EID alorrean algoritmo lokalen erabilera, eta batez ere, hitz multzoena, nabarmena izan da (Hoffart *et al.* 2011, Han and Sun 2012, He *et al.* 2013, Lazic *et al.* 2015). Izan ere, teknika erraz eta eraginkorrak dira. Orokorrean, eredu gainbegiratuak izaten dira eta gehienetan, Wikipediako aingura-artikulu erlazioa erabiltzen dute ikasketa burutzeko (Milne and Witten 2008b).

Jarraian, (Han and Sun 2011) artikulua oinarritzat hartuta, eredu lokal bat azalduko da. Hasteko, eredu sortzaile bat planteatuko da hiru urratsetan definituko dena.

(1) Izan bedi e entitate bat, eredu sortzaileak ezagutza-baseko entitateen banaketatik sortzeko gaitasuna duena ($P(e)$).
(2) Demagun e entitateak, bere burua izendatzeko erabili diren s izen-aipamenen banaketa sortzeko gai dela ($P(s|e)$).
(3) Azkenik, izan bedi e entitateak sortu ditzakeen c testuinguruen banaketa bat ($P(c|e)$).

Honenbestez, eredu sortzailearen azalpenean oinarrituta, s izen-aipamena eta c testuingurua e entitateak sortzeko duen probabilitatea zenbatetsiko da. Horretarako, jarraian datorren formula erabiliko da:

$$P(s, c, e) = P(e)P(s|e)P(c|e) \tag{3.2}$$

EID atazan eredua aplikatzeko, 3.2 formulan probabilitate altuena lortzen duen e entitatea, c testuinguruko s izen-aipamenari lotuko diogun entitatea izango da:

$$\arg\max_e P(c, s, e) = \arg\max_e P(e)P(s|e)P(c|e) \tag{3.3}$$

Ereduak hiru ezagutza barneratzen ditu, $P(e)$ entitatearen-ezagutza, $P(s|e)$ izenen-ezagutza eta $P(c|e)$ testuinguruaren-ezagutza. Lehenengo bi ezagutzak aldez aurretiko probabilitateetan oinarritzen dira, eta hiztegiko maiztasunekin zenbatezten dira. Ezagutza lokala hirugarren ezagutzak errepresentatuko du, eta horretarako, hitz multzoak erabiliko dira. Hitz multzoak sortzeko artikuluen testuinguru-bildumak erabiltzen dira. Jarraian, ezagutza bakoitza nola zenbatetsi azalduko da.

Entitatearen ezagutzak e entitatea sortzeko probabilitatea erakutsiko du. Gero eta gehiagotan aipatua izan, orduan eta balio altuagoa jasoko du entitateak. $P(e)$ probabilitatea zenbatesteko jarraian datorren formula erabiliko da:

$$P(e) = \frac{C(e) + 1}{|M| + N} \tag{3.4}$$

Egiantza handieneko zenbatezketa erabiliz, entitatea hiztegian zenbat aldiz aipatu den zenbatu ($C(e)$), eta normalizatuko da. Entitatea inoiz erreferentziatu ez bada, probabilitateak 0 eman ez dezan, +1 izeneko leunketa aplikatuko da. $|M|$ eta N balioak aipamen kopuru totala eta entitate kopurua dira, hurrenez hurren.

Izenaren ezagutzak e entitatea izanda s izen-aipamena sortzeko gaitasuna erakutsiko du. $P(s|e)$ probabilitatea jarraian datorren formulak zenbatetsiko du:

$$P(s|e) = \frac{C(e, s) + 1}{\sum_{s} C(e, s) + S} \tag{3.5}$$

Berriz ere egiantza handieneko zenbatezketaz e entitatea s aipamenarekin zenbat aldiz aipatu den zenbatu ($C(e, s)$), eta normalizatuko da. Leunketa metodo berdina aplikatuz, 0 probabilitateak ekidituko dira. S, e entitatea izendatzeko erabili diren izen ezberdinen kopurua da.

Testuinguruaren ezagutzak e entitateak c testuingurua sortzeko duen gaitasuna modelatuko du, eta $P(c|e)$ gisa izendatuko da. Eredu honek $P(c|e)$ balio altua emango dio baldin eta e entitatea askotan azaldu bada c testuinguruan.

c testuingurua n terminodun $w_1, w_2, w_3...$ hitz multzo gisa errepresentatzen da. $P(c|e)$ banaketa w termino bakoitza e entitateak sortzeko duen probabilitatearen biderkaduraz kalkulatzen da:

$$P(c|e) = P(w_1|e)P(w_2|e)...P(w_n|e) \tag{3.6}$$

$P(w|e)$ak zenbatesteko termino horrek e entitatearen testuinguruan azaltzeko duen probabilitatea kalkulatuko da. Zenbatezketa hau 2 kapituluko 2.5.4 atalean azaldu diren artikuluen testuinguru-bildumetatik eskuratzen

da. Entitate bakoitzarentzat testuinguru guztiak batu, eta hitz multzo bat sortuko da. Jarraian, terminoak zenbatzen dira, hitzen ordena kontuan hartu gabe ($C_e(w)$). Azkenik, egiantza handieneko zenbatezketaz terminoen probabilitateak kalkulatzen dira, eta hauek $P_e(w)$ gisa izendatuko dira:

$$P_e(w) = \frac{C(e,w)}{\sum_w C(e,w)} \tag{3.7}$$

Esate baterako, `Jeff_Beck` eta `Beck_Weathers` entitateen $P_e(w)$ zenbatezketak, *Clapton* eta *Everest* terminoentzat hauek lirateke:

$$P_{\texttt{Jeff\_Beck}}(Clapton) = 0.00255$$
$$P_{\texttt{Jeff\_Beck}}(Everest) = 0$$

$$P_{\texttt{Beck\_Weathers}}(Clapton) = 0$$
$$P_{\texttt{Beck\_Weathers}}(Everest) = 0.00896$$

Adibidean ikusten denez, `Beck_Weathers` entitateak ezin du *Clapton* terminoa sortu, Wikipedian ez baita inoiz estekatua azaldu bere testuinguruan *Clapton* hitzarekin. Berdina gertatzen da `Jeff_Beck` eta *Everest* terminoarekin. $P_e(w)$ banaketa termino horrek Wikipedia osoan azaltzeko duen probabilitaterekin leunduko da, berriz ere 0 probabilitateak ekiditeko. Leunketa $P_w(w)$ gisa gehituko da eta bere ekarpena λ parametroaz mugatua egongo da (Jelinek and Mercer 1980):

$$P(w|e) = \lambda P_e(w) + (1-\lambda)P_w(w) \tag{3.8}$$

3.2.1 Hitz multzoak eta eredu sortzailea EID atazan

Eredu sortzailea osatzen duten hiru ezagutzak nola zenbatetsi azaldu da, jarraian, EID atazari eutsiz, eredu sortzailea erabiliz *Beck* desanbiguatuko da testu honetan:

Three of the greatest guitarrist started
their career in a single band:
Clapton, Beck and Page.

Formulazioan erabili den terminologia aplikatuz:

- s: Desanbiguatu nahi den izen-aipamena izango da. Kasu honetan *Beck*.

- c: Izen-aipamenaren testuingurua da, eta bertan aurkitzen diren terminoek osatzen duten hitz multzoa dira.
  $[Three, of, the, greatest...single, band, Clapton, and, Page]$

- e: Desanbiguatu nahi den s izen-aipamenaren artikulu-hautagaiak izango dira. Adibidez, `Jeff_Beck` eta `Beck_Weathers`.

Beraz, c testuinguruan desanbiguatu nahi den s izen-aipamena, 3.3 ekuazioan probabilitate handiena lortzen duen e entitatera lotuko da.

3.3 Bestelako EID algoritmoak

Atal honetan EID atazako arloaren egoerako artikulu esanguratsuenak aztertuko dira. Bide batez, tesian garatu diren sistemekin batera, arloaren egoerako emaitza onenak lortzeko bidean lehiatuko diren sistemak dira. EID atazako ekarpenak eredu global eta lokaletan banatu ohi dira. Hala ere, arloaren egoerako sistema askok hauen konbinaketak planteatzen dituztela ikusiko da.

EID atazan lehen urratsak (Bunescu and Pasca 2006) artikuluan ematen direla esan daiteke. Lan honek Wikipediaren egitura aztertzen du, eta lehenengo aldiz, EID atazarako ikasketa corpus gisa duen ahalmena erakusten du. Arreta berezia eskaintzen diote Wikipedia osatzen duten hiperesteka, birbideratze eta desanbiguazio-orriei. Sistemaren oinarria Wikipedian entrenatutako sostengu bektoreen makina da.

Aurreko artikuluan oinarrituta (Mihalcea and Csomai 2007; Milne and Witten 2008b) lanek testua Wikipediako ezagutzaz nola aberastu azaltzen dute, hain zuzen ere, Wikifikazio ataza definitzen dute. Lehenak desanbiguaziorako eredu global eta lokalak aurkezten ditu, baina beste terminologia bat erabiliz, HAD atazan *knowledge-based methods* eta *data-driven* gisa definitzen direnak. Ondoren, biak konbinatzen ditu bozka bidezko metodo bat erabiliz. Bigarrenak ikasketa automatikoan oinarritutako sistema aurkezten du, antzekotasun eta probabilitate ezberdinak konbinatzen dituena. Ezaugarri nagusi gisa (Milne and Witten 2008a) artikuluan definitzen den antzekotasun balioa erabiltzen dute. Hitz gutxitan, balio honek entitateek Wikipedian dituzten

hiperesteka zuzenak erabiliz, bi entitateen arteko antzekotasuna kalkulatzen du.

Algoritmo "global" eta "lokal" terminologia (Ratinov *et al.* 2011) artikuluak erabiltzen du lehenengo aldiz. Ordutik aurrera arloaren egoeran erabili den terminologia bihurtu da. Artikulu honek algoritmoen aldaerak eta ezaugarriak aztertzen ditu bakoitzaren ekarpen nagusiak azpimarratuz. Informazio lokala bi modutan errepresentatzen dute. Batetik, izen-aipamena inguratzen duten terminoek osatzen duten hitzekin. Bestetik, izen-aipamena agertu den dokumentu osoa testuinguru gisa hartuz. Informazio globala errepresentatzeko (Milne and Witten 2008a) artikuluan egin den gisa, Wikipediako esteka zuzenetatik eskuratutako antzekotasun balioetan oinarritzen dira. Azkenik, informazio guztia sostengu bektoreen makinak entrenatuz konbinatzen dute.

(Hoffart *et al.* 2011) artikuluan aldez-aurretiko probabilitateekin eta antzekotasun balioekin testuko izen-aipamenentzat ezaugarri lokalak kalkulatzen dituzte. Artikulu honetan ere (Milne and Witten 2008a) laneko antzekotasun balioak barneratzen dituzte koherentzia neurri gisa. Ondoren, entitatehautagaiekin azpigrafoaren adabegiak definitzen dituzte, eta Wikipediako esteka zuzenak erabiliz informazio globala errepresentatzen dute. (Suchanek *et al.* 2008) laneko YAGO ontologiako ezagutza aipamenen sorkuntza egiteko eta grafoaren koherentzia ustiatzen duten moduluak elikatzeko erabiltzen dute.

(Hoffart *et al.* 2012a) lanean entitateen arteko antzekotasun balio berritzailea planteatu eta EID atazan aplikatzen dute. Artikuluaren helburua esteka zuzenetan oinarritutako antzekotasun balioak dituzten gabeziak gainditu, eta eredu berria planteatzea da. Horretarako, entitateak hitz-kate jakin batzuekin erlazionatzen dituzte, eta ondoren, hitz-kate hauen teilakatzean oinarritzen dira. Antzekotasun balioa ebaluatzeko (Hoffart *et al.* 2011) artikuluan definitzen den sistema oinarritzat hartu, eta (Milne and Witten 2008a) artikuluan definitutako antzekotasun balioa ordezkatzen dute. Antzekotasun balio berriarekin sistemaren eraginkortasuna hobetzen dute.

(Houlsby and Ciaramita 2014) artikuluak Wikipedian oinarritutako eredu probabilistiko lokala azaltzen du. Horretarako, gai-ereduak (*Topic Models*) erabiltzen ditu, kasu honetan, Wikipediako artikulu bakoitzak gai bat errepresentatzen duelarik. Algoritmoa *Tagme* (Ferragina and Scaiella 2012) sistemarekin hasieratzen dute. Beraz, hitz gutxitan esanda, *Tagme* sistemaren ekarpen globala eta gai-ereduen informazio lokala konbinatzen dute.

(Chisholm and Hachey 2015) artikuluan web-eko hiperesteken informa-

zioa erabiltzen dute desanbiguazioa burutzeko. Bi urratsetan oinarritzen den eredu gainbegiratua entrenatzen dute, eredu lokal eta globalak konbinatuz. Artikulu honen ekarpen nagusia Wikipediako eta web-eko hiperesteken konbinaketa da.

Ausazko ibilbideetan oinarritzen diren artikuluei dagokienean (Moro *et al.* 2014) artikuluan HAD eta EID ataza batera burutzen duen sistema aurkezten dute, eta (Navigli and Ponzetto 2012a) laneko BabelNet[2] ezagutza-basearen gainean desanbiguatzen dute. Hasteko, testuan azaltzen diren izen-aipamen eta hitzekin, ezagutza-basean erreferentziatzen dituzten adabegiak biltzen dituzte. Ondoren, adabegi hauek ezagutza-basean elkarren artean dituzten hiperesteka zuzenen bidez azpigrafoa sortzen dute. Azkenik, ausazko ibilbide pertsonalizatu konplexu bat aurkezten dute desanbiguazioa burutzeko. Artikulu honetan HAD eta EID atazak batera egitearen ekarpena azpimarratzen dute.

(Alhelbawy and Gaizauskas 2014) artikuluan eredu globalen eta lokalen konbinazio bat aurkezten da. Informazio lokal gisa, entitate-hautagaien Wikipediako deskribapenak eta izen-aipamenaren testuinguruaren arteko antzekotasuna erabiltzen dituzte. Horretaz gain, izen-aipamen eta entitate-hautagaien tituluen hitzen arteko antzekotasun balioak barneratzen dituzte. Ondoren, Wikipediako esteka zuzenak erabiliz testua osatzen duten izen-aipamenetatik grafoa eraikitzen dute. Azkenik, *PageRank* algoritmoa exekutatzen dute grafoko adabegiek duten informazio lokala kontuan hartuta.

(Pershina *et al.* 2015) artikuluak, ildo berean jarraituz, esteka zuzenak erabiliz ausazko ibilbideen algoritmo global eraginkorra aurkezten dute. Bide batez, ausazko ibilbideetan aldez-aurretiko hiru probabilitateen ekarpena konparatzen du. Gainera, murriztapen batzuen bidez algoritmoaren emaitzak hobetzeko gai da. Horretarako, entitate-hautagaiek ausazko ibilbidean egiten duten ekarpena mugatzen dute.

(Hachey *et al.* 2011) lanak esteka zuzenetatik haratago pausu bat ematen dute, grafoa eraikitzean entitate-hautagaiak lotzeko bi esteketako loturak eginez. Literaturan esteka zuzenetatik haratago grafoa eraiki duten artikulu bakarrenetakoa da. Bestalde, testuinguruen antzekotasunean oinarritutako informazio lokala ausazko ibilbideekin konbinatzen dute.

Eredu lokaletan bakarrik oinarritzen diren sistemei dagokienean, 3.2 atalean sakondu den (Han and Sun 2011) artikulua aurkitzen da. Lan hori onarritzat hartuta (Daiber *et al.* 2013) artikuluan *DBpedia Spotlight* sis-

[2]Babelnet Wikipedia eta WordNet konbinatzen dituen ezagutza-basea da.

tema eleanitza aurkezten da. Sistema honen abantaila nagusiak efizientzia eta eleaniztasuna dira, izan ere, 9 hizkuntza ezberdinetan izen-aipamenak identifikatu eta desanbiguatzeko gai dira. Hori gutxi balitz, sistemak garatzaile eta erabiltzaileen partetik duen jarraipena eta arreta aipatzekoak dira.[3]

(Lazic *et al.* 2015) artikuluan eredu lokal probabilistiko bat azaltzen dute, eta gainera, kanpo-ezagutza barneratzen dute etiketatu gabeko corpusetatik. (Han and Sun 2011) artikuluan azaldu den eredu sortzailetik haratago, ereduaren zenbatezketa hobea erakusten dute. Parametroen zenbatezketa hobeak emaitzetan islada dute, izan ere, eredu lokalak soilik erabiliz emaitza onak erakusten dituzte.

EID atazaren arloaren egoeran sistema ezberdin asko argitaratzen dira. Orokorrean eredu gainbegiratuen bidez ezaugarriak konbinatzeko metodo ezberdinak aztertzen dituzte. Atal honetan, tesiarekin zer ikusi zuzena dutenak soilik aztertu dira. Amaitzeko, TAC-KBP eta ERD gisako txapelketetan aurkeztu diren sistema guztien berezitasunak (McNamee and Dang 2009,Ji *et al.* 2010,Ji *et al.* 2014,Carmel *et al.* 2014) artikuluetan aurkitu daitezke.

3.4 Datu-multzoak eta ebaluazio-metrikak

EID sistemen eraginkortasuna ebaluatzeko hainbat datu-multzo eskuragarri daude. Adibidez, 2009. urtetik aurrera urtero ospatzen den TAC-KBP txapelketako *Entity Linking* atazarako sortutakoak. Helburua ingelesezko testu ezberdinetako izen-aipamenak Wikipediako azpimultzo batetik sortutako ezagutza-basera lotzea da. Datu-multzo hauetan berrietako, foroetako edo interneteko web orrietako testuak biltzen dira, eta bertan azaltzen diren izen-aipamenak dagokion entitatearekin eskuz etiketatuta daude. Tesi honetako emaitzak ebaluatzeko konferentzia honetako 6 datu-multzoak erabiliko dira, 2009 eta 2014 urte bitartekoak (aurrerantzean *TAC09-TAC14*).

Bestalde, tesi honetan garatuko diren sistemak *AIDA* eta *KORE*[4] deituriko beste bi datu-multzoetan ere ebaluatuko dira. Lehenak, berrietatik eskuratu diren dokumentuak biltzen ditu. Bigarrenak, testu oso motzak eta izen oso anbiguoz osaturiko dokumentuak. *AIDA*[5] datu-multzoa ikasketa,

[3] `https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki`

[4] `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/`

[5] CoNLL2003 entitate-izendunen identifikaziorako datu-multzoan, izen-aipamenak da-

Datu-multzoa	**Izen-aipamenak**	**Anbiguotasuna**
AIDA-testa	4792	75
AIDA-testb	4485	79
AIDA-train	18541	75
KORE	144	81
TAC09	1675	26
TAC10	1020	46
TAC11	1121	38
TAC12	1177	58
TAC13	1183	51
TAC14	2817	51

3.1 taula – Datu-multzoen ezaugarriak: izen-aipamen kopuruak eta hiztegiaren araberako batezbesteko anbiguotasuna.

garapen eta test azpimultzoetan banatzen da. Hauen izenak hurrenez hurren *AIDA-train*, *AIDA-testa* eta *AIDA-testb* dira.

3.1. taulan datu-multzo bakoitzak urre-patroian dituen izen-aipamen kopurua ikus daiteke. Gainera, hiztegiaren arabera aipamenek batezbeste Wikipediako zenbat artikulu esleituak dituzten erakusten da.

Ebaluazio-metrikei dagokionez, doitasun, estaldura eta F1 neurri estandarrak erabili dira. Doitasuna sistemak egoki desanbiguatu dituen eta sistemak desanbiguatu dituen izen-aipamemen kopuruen arteko zatiketa da. Estaldura egoki desanbiguatuak eta urre-patroiko izen-aipamenen kopuruen arteko zatiketa da. F-neurria doitasuna eta estalduraren arteko batezbesteko harmoniko gisa definitzen da.

Hala ere, tesi honetan gehien erabili den metrika micro-zehaztasuna izenekoa da. Micro-zehaztasuna ondo desanbiguatutako izen-aipamen kopurua eta urre-patroiko aipamen kopuru totalaren arteko zatiketaz kalkulatzen da.[6] Dena dela, *TAC11-tik TAC14*-ra bitarteko datu-multzoen ebaluazioan micro-zehaztasunaz batera bCubed+ (Amigó *et al.* 2009) metrika ere erabiltzen da. Metrika honek egoki lotu diren eta talde egokian multzokatuak dauden izen-aipamenen F-neurria kalkulatzen du.

Metrika hau erabiltzearen arrazoia ezagutza-basean entitaterik erreferentziatzen ez dituzten izen-aipamenak dira. Aipamen hauek NIL entitatera

gokion entitatearekin etiketatu ziren.

[6]Estalduraren berdina da.

lotzen dira. Testu batean azaldu diren izen-aipamenek NIL berdinari erreferentzia egin ahal diote, beraz, zenbaki batekin identifikatzen dira, NIL001 adibidez. Ezagutza-baseetan existitzen ez diren NIL hauen multzokatzea ebaluatzeko bCubed+ metrika erabiltzen da. Hala ere, tesi honetan ez da NIL-en atala landuko.

Argibideak eta terminologia bateratua

Jarraian datozen 4 kapituluek ingelesez argitaratutako artikuluen bilduma osatuko dute. Terminologia bateratzeko eta ulergarritasuna errazteko jarraian datozen argibideak garrantzitsuak dira:

- Aurrerantzean algoritmo globalen erabilerari ausazko ibilbideen bitartez erreferentzia egiteko **Personalized PageRank** edo **PPR** laburdura erabiliko da.
- Algoritmo lokalen erabilerari hitz multzoen bitartez erreferentzia egiteko $\mathbf{p}(\mathbf{c}|\mathbf{e})$ laburdura erabiliko da. Orokorrean hiztegiko aldez-aurretiko probabilitateekin konbinatua azalduko da $\mathbf{p}(\mathbf{e})\mathbf{p}(\mathbf{s}|\mathbf{e})\mathbf{p}(\mathbf{c}|\mathbf{e})$ formatuan (konbinaketaren azalpenak 5. eta 7. kapituluetan sakonduko dira).

4

Algoritmo globalak, ausazko ibilbideak Wikipedia grafoan

Kapitulu honek artikulu bildumaren lehen artikulua azalduko du. Laburbilduz, Wikipediako hiperestekak aztertzen dira hainbat aspektu ezberdinetan. Horretarako, ausazko ibilbideak (Personalized PageRank edo PPR artikuluan) algoritmoa aplikatuko da. Helburua EID eta antzekotasun atazetarako Wikipediako hiperesteka erlazio optimoena aurkitzea da. Tesi honetako lana EID atazan zentratu da, horregatik, antzekotasunaren atalak tesitik kanpo geldituko lirateke. Jarraian, artikuluaren jatorrizko fitxa eta ingelesezko bertsioa:

Eneko Agirre, Ander Barrena and Aitor Soroa. ***Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation.*** *arXiv.org CoRR 2015.*

Hyperlinks and other relations in Wikipedia are a extraordinary resource which is still not fully understood. In this paper we study the different types of links in Wikipedia, and contrast the use of the full graph with respect to just direct links. We apply a well-known random walk algorithm on two tasks, word relatedness and named-entity disambiguation. We show that using the full graph is more effective than just direct links by a large margin, that non-reciprocal links harm performance, and that there is no benefit from categories and infoboxes, with coherent results on both tasks. We set new state-of-the-art figures for systems based on Wikipedia links, comparable to systems exploiting several information sources and/or supervised machine

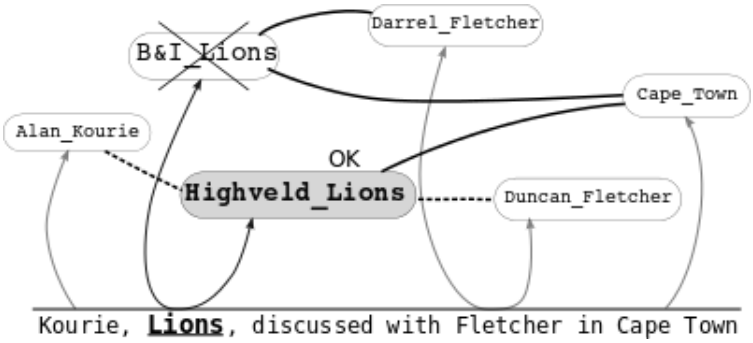
learning. Our approach is open source, with instruction to reproduce results, and amenable to be integrated with complementary text-based methods.

4.1 Introduction

Hyperlinks and other relations between concepts and instances in Wikipedia have been successfully used in semantic tasks (Milne and Witten 2013). Still, many questions about the best way to leverage those links remain unanswered. For instance, methods using direct hyperlinks alone would wrongly disambiguate *Lions* in Figure 4.1 to `B&I_Lions`, a rugby team from Britain and Ireland, as it shares two direct links to potential referents in the context (`Darrel_Fletcher`, a British football player, and `Cape_Town`, the city where the team suffered some memorable defeats), while `Highveld_Lions`, a cricket team from South Africa, has only one. When considering the whole graph of hyperlinks we find that the cricket team is related to two cricketers named `Alan_Kourie` and `Duncan_Fletcher` and could thus pick the right entity for *Lions* in this context. In this paper we will study this and other questions about the use of hyperlinks in word relatedness (Gabrilovich and Markovitch 2007) and named-entity disambiguation, NED (Hachey *et al.* 2012).

Previous work on this area has typically focused on novel algorithms which work on a specific mix of resource, information source, task and test dataset (cf. Sect. 4.7). In the case of NED, the evaluation of the disambiguation component is confounded by interactions with mention spotting and candidate generation. With very few exceptions, there is little analysis of components and alternatives, and it is very difficult to learn any insight beyond the fact that the mix under study attained certain performance on the target dataset[1]. The number of algorithms and datasets is growing by the day, with no well-established single benchmark, and the fact that some systems are developed on test data, coupled with reproducibility problems (Fokkens *et al.*, 2013, on word relatedness), makes it very difficult to know where the area stands. There is a need for clear points of reference which allow to understand where each information source and algorithm stands with respect to other alternatives.

[1]See Hachey *et al.* 2012 and García *et al.* 2014 for two exceptions on NED. The first is limited to a single dataset, the second explores methods based on direct links, which we extend to using the full graph.



4.1 Figure – Simplified example motivating the use of the full graph. It shows the disambiguation of *Lions* in "*Alan Kourie, CEO of the Lions franchise, had discussions with Fletcher in Cape Town*". Each mention is linked to the candidate entities by arrows, e.g. `B&I_Lions` and `Highveld_Lions` for *Lions*. Solid lines correspond to direct hyperlinks and dashed lines to a path of several links. An algorithm using direct links alone would incorrectly output `B&I_Lions`, while one using the full graph would correctly choose `Highveld_Lions`.

We thus depart from previous work, seeking to set such a point of reference, and focus on a single knowledge source (hyperlinks in Wikipedia) with a clear research objective: given a well-established random walk algorithm (Personalized PageRank (Haveliwala 2002)) we explore sources of links and filtering methods, and contrast the use of the full graph with respect to using just direct links. We follow a clear development/test/analysis methodology, evaluating on a extensive range of both relatedness and NED datasets. The results are confirmed in both tasks, yielding more support to the findings in this research. All software and data are publicly available, with instructions to obtain out-of-the-box replicability[2].

The contributions of our research are the following: (1) We show for the first time that performing random walks over the full graph is preferable than considering only direct links. (2) We study several sources of links, showing that non-reciprocal links hurt and that the contribution of the category structure and links in infoboxes is residual. (3) We set the new state-of-the-art for systems based on Wikipedia links for *both* word relatedness and named-entity disambiguation. The results are close to the best systems to date, which use several information sources and/or supervised machine learn-

[2] `http://ixa2.si.ehu.es/ukb/README.wiki.txt`

ing techniques, and specialize on either relatedness or disambiguation. Our work shows that a careful analysis of varieties of graphs using a well-known random walk algorithm pays off more than most ad-hoc algorithms.

The article is structured as follows. We first present previous work, followed by the different options to build hyperlink graphs. Sect. 4.4 reviews random walks for relatedness and NED. Sect. 4.5 sets the experimental methodology, followed by the analysis and results on development data (Sect. 4.6) and the comparison to the state of the art (Sect. 4.7). Finally, Sect. 5.7 draws the conclusions.

4.2 Previous work

The irruption of Wikipedia has opened up enormous opportunities for natural language processing (Hovy *et al.* 2013), with many derived knowledge-bases, including DBpedia (Bizer *et al.* 2009), Freebase (Bollacker *et al.* 2008), and BabelNet (Navigli and Ponzetto 2012a), to name a few. These resources have been successfully used on semantic processing tasks like word relatedness, named-entity disambiguation (NED), also known as entity linking, and the closely related Wikification. Broadly speaking, Wikipedia-based approaches to those tasks can be split between those using the text in the articles (e.g., Gabrilovich and Markovitch, 2007) and those using the links between articles (e.g., Guo et al., 2011).

Relatedness systems take two words and return a high number if the two words are similar or closely related[3] (e.g. *professor* - *student*), and a low number otherwise (e.g. *professor* - *cucumber*). Evaluation is performed comparing the returned values to those by humans (Rubenstein and Goodenough 1965).

In NED (Hachey *et al.* 2012) the input is a mention of a named-entity in context and the output is the appropriate instance from Wikipedia, DBpedia or Freebase (cf. Figure 4.1). Wikification is similar (Mihalcea and Csomai 2007), but target terms include common nouns and only relevant terms are disambiguated. Note that the disambiguation component in Wikification and NED can be the same.

Our work focuses on relatedness and NED. We favored NED over Wikification because of the larger number of systems and evaluation datasets, but

[3]Relatedness is more general than similarity. For the sake of simplicity, we will talk about relatedness on this paper.

our conclusions are applicable to Wikification, as well as other Wikipedia-derived resources.

In this section we will focus on previous work using Wikipedia links for relatedness, NED and Wikification. Although relatedness and disambiguation are closely related (relatedness to context terms is an important disambiguation clue for NED), most of the systems are evaluated in either relatedness or NED, with few exceptions, like WikiMiner (Milne and Witten 2013), KORE (Hoffart *et al.* 2012a) and the one presented in this paper.

Milne and Witten (Milne and Witten 2008a) are the first to use hyperlinks between articles for relatedness. They compare two articles according to the number of incoming links that they have in common (i.e. overlap of direct-links) based on Normalized Google Distance (NGD), combined with several heuristics and collocation strength. In later work (Milne and Witten 2013), they incorporated machine learning. The authors also apply their technique to NED (Milne and Witten 2008b), using their relatedness measures to train a supervised classifier. Unfortunately they do not present results of their link-based method alone, so we decided to reimplement it (cf. Sect. 4.6). We show that, under the same conditions, using the full-graph is more effective in both tasks. We also run their out-of-the-box system[4] on the same datasets as ours (cf. Sect. 4.7), with results below ours.

Apart from hyperlinks between articles, other works on relatedness use the category structure (Strube and Ponzetto 2006; Ponzetto and Strube 2007, 2011) to run path-based relatedness algorithms which had been successful on WordNet (Pedersen *et al.* 2004), or use relations in infoboxes (Nastase and Strube 2013). In all cases, they obtain performance figures well below hyperlink-based systems (cf. Sect. 4.7). We will explore the contribution of such relations (cf. Sect. 4.3), incorporating them to the hyperlink graph.

Attempts to use the whole graph of hyperlinks for relatedness have been reported before. Yeh et al. (Yeh *et al.* 2009) obtained very low results on relatedness using an algorithm based on random walks similar to ours. Similar in spirit, Yazdani and Popescu-Belis (Yazdani and Popescu-Belis 2013) built a graph derived from the Freebase Wikipedia Extraction dataset, which is derived but richer than Wikipedia. Even if they mix hyperlinks with textual similarity, their results are lower than ours. One of the key differences with these systems is that we remove non-reciprocal links (cf. Sect. 4.3).

Regarding link-based methods for NED, there is only one system which

[4] `https://sourceforge.net/projects/wikipedia-miner/`

relies exclusively on hyperlinks. Guo et al. (Guo *et al.* 2011) use direct hyperlinks between the target entity and the mentions in the context, counting the number of such links. We show that the use of the full graph produces better results.

The rest of NED systems present complex combinations. Lemahnn et al. (Lehmann *et al.* 2010) present a supervised system combining features based on hyperlinks, categories, text similarity and relations from infoboxes. Despite their complex and rich system, we will show that they perform worse than our system. (Hachey *et al.* 2011) explored hyperlinks beyond direct links for NED, building subgraphs for each context using paths of length two departing from the context terms, combined with text-based relatedness. We will show that the full graph is more effective than limiting the distance to two, and report better results than their system. Several authors have included direct links using the aforementioned NGD in their combined systems (Ratinov *et al.* 2011; Hoffart *et al.* 2011). Unfortunately, they do no report separate results for the NGD component. In very recent work (García *et al.* 2014) compare NGD with several other algorithms using direct links, but do not explore the full graph, or try to characterize links. We will see that their results are well below ours (cf. Sect. 4.7).

Graph-based algorithms for relatedness and disambiguation have been successfully used on other resources, particularly WordNet. Hughes and Ramage (Hughes and Ramage 2007) were the first presenting a random walk algorithm over the WordNet graph. Agirre et al. (Agirre *et al.* 2010) improved over their results using a similar random walk algorithm on several variations of WordNet relations, reporting the best results to date among WordNet-based algorithms. The same algorithm was used for word sense disambiguation (Agirre *et al.* 2014), also reporting state-of-the-art results. We use the same open source software in our experiments. As an alternative to random walks, Tsatsaronis et al. (Tsatsaronis *et al.* 2010) use a path-based system over the WordNet relation graph.

In more recent work (Navigli and Ponzetto 2012b, Pilehvar *et al.* 2013), the authors present two relatedness algorithms for BabelNet, an enriched version of WordNet including articles from Wikipedia, hyperlinks and cross-lingual relations from non-English Wikipedias. In related work, Moro et al. (Moro *et al.* 2014) present a multi-step NED algorithm on BabelNet, building semantic graphs for each context. We will show that Wikipedia hyperlinks alone are able to provide similar performance on both tasks.

4.3 Building Wikipedia Graphs

Wikipedia pages can be classified into main articles, category pages, redirects and disambiguation pages. Given a Wikipedia dump (a snapshot from April 4, 2013), we mine links between articles, between articles and category pages, as well as the links between category pages (the category structure). Our graphs include a directed edge from one article to another iff the text of the first article contains a hyperlink to the second article. In addition, we also include hyperlinks in infoboxes.

The graph contains two types of nodes (articles and categories) and three types of directed edges: hyperlinks from article to article (**H**), infobox links from article to article (**I**), links from article to category and links from category to category (**C**).

We constructed several graphs using different combinations of nodes and edges. In addition to the directed versions (**D**) we also constructed an undirected version (**U**), and a reduced graph which only contains links which are reciprocal (**R**), that is, we add a pair of edges between $a1$ and $a2$ if and only if there exists a hyperlink from $a1$ to $a2$ and from $a2$ to $a1$. **Reciprocal** links capture the intuition that both articles are relevant to each other, and tackle issues with links to low relevance articles, e.g. links to articles on specific years like 1984. Some authors weight links according to their relevance (Milne and Witten 2013). Our heuristic to keep only reciprocal links can be seen as a simpler, yet effective, method to avoid low relevance links.

Table 4.1 gives the number of nodes and edges in some selected graphs. The graph with less edges is the one with reciprocal hyperlinks **HR**, and the graphs with most edges are those with undirected edges, as each edge is modeled as two directed edges[5]. The number of nodes is similar in all, except for the infobox graphs (infoboxes are only available for a few articles), and the reciprocal graph **HR**, as relatively few nodes have reciprocal edges.

4.3.1 Building the dictionary

In order to link running text to the articles in the graph, we use a dictionary, i.e., a static association between string mentions with all possible articles the mention can refer to.

[5]This was done in order to combine undirected and reciprocal edges, and could be avoided in other cases.

Graph	Edges	Nodes	RG	TAC09_{200}
Cd	18,803K	4,873K	51.1 † ‡	49.5 † ‡
Cu	37,598K	4,873K	72.9 † ‡	65.5 † ‡
Id	6,572K	1,860K	43.1 † ‡	57.0 † ‡
Iu	12,692K	1,860K	52.8 † ‡	65.5 † ‡
Hd	90,674K	4,103K	75.1 † ‡	65.0 † ‡
Hu	165,258K	4,103K	76.6 ‡	66.0 † ‡
Hr	16,338K	2,955K	**88.4**	**68.5**
HrCu	53,005K	4,898K	78.2 ‡	67.5 ‡
HrIu	26,394K	3,273K	82.9 ‡	68.0 ‡
HrCuIu	63,184K	4,900K	75.6 † ‡	67.5 ‡

4.1 Table – Statistics for selected graphs and results on development data for relatedness (RG, Spearman) and NED (TAC09_{200}, accuracy) with default parameters (see text). See Sect. 4.4.1 for abbreviations. † for stat. significant differences with Hr in either RG or TAC09_{200}. ‡ for stat. signif. when comparing on all relatedness or NED datasets.

Article	Freq.	Prob.
Gotham_City	32	0.38
Gotham_(magazine)	15	0.18
...		
New_York_City	1	0.01
Gotham_Records	1	0.01

4.2 Table – Partial view of dictionary entry for "gotham". The probability is calculated as the ratio between the frequency and the total count.

We built our dictionary from the same Wikipedia dump, using article titles, redirections, disambiguation pages, and anchor text. Mention strings are lowercased and all text between parentheses is removed. If an anchor links to a disambiguation page, the text is associated with all possible articles the disambiguation page points to. Each association between a mention and article is scored with the prior probability, estimated as the number of times that the mention occurs in an anchor divided by the total number of occurrences of the mention as anchor. Note that our dictionary can disambiguate any mention, just returning the highest-scoring article. Table 4.2 partially shows a sample entry in our dictionary.

Drink		Alcohol	
Drink	.124	Alcohol	.145
Alcoholic_beverage	.036	Alcoholic_beverage	.026
Drinking	.028	Ethanol	.018
Coffee	.020	Alkene	.006
Tea	.017	Alcoholism	.006

4.3 Table – Sample of the probability distribution returned by PPR for two words. Top five articles shown.

4.4 Random Walks

The PageRank random walk algorithm (Brin and Page 1998) is a method for ranking the vertices in a graph according to their relative structural importance. PageRank can be viewed as the result of a random walk process, where the final rank of node i represents the probability of a random walk over the graph ending on node i, at a sufficiently large time.

Personalized PageRank (PPR) is a variation of PageRank (Haveliwala 2002), where the query of the user defines the importance of each node, biasing the resulting PageRank score to prefer nodes in the vicinity of the query nodes. The query bias is also called the teleport vector. PPR has been successfully used on the WordNet graph for relatedness (Hughes and Ramage 2007; Agirre *et al.* 2010) and WSD (Agirre and Soroa 2009; Agirre *et al.* 2014). In our experiments we use UKB version 2.1[6], an open source software for relatedness and disambiguation based on PPR. For the sake of space, we will skip the details, and refer the reader to those papers. PPR has two parameters: the number of **iterations**, and the **damping factor**, which controls the relative weight of the teleport vector.

4.4.1 Random walks on Wikipedia

Given a dictionary and graph derived from Wikipedia (cf. Sect. 4.3), PPR expects a set of mentions, i.e., a set of strings which can be linked to Wikipedia articles via the dictionary. The method first initializes the teleport vector: for each mention in the input, the articles in the respective dictionary entry are set with an initial probability, and the rest of articles are set to zero.

[6] `http://ixa2.si.ehu.es/ukb`

We explored two options to set the initial probability of each article: the uniform probability or the **prior** probability in the dictionary. When an article appears in the dictionary entry for two mentions, the initial probability is summed up. In a second step, we apply PPR for a number of iterations, producing a probability distribution over Wikipedia articles in the form of a PPR vector (PPV).

The probability vector can be used for both relatedness and NED. For **relatedness** we produce a PPV vector for each of the words to be compared, using the single word as input mention. The relatedness between the target words is computed as the cosine between the respective PPV vectors. In order to speed up the computation, we can reduce the size of the PPV vectors, setting to zero all values below rank k after ordering the values in decreasing order.

Table 4.3 shows the top 5 articles in the PPV vectors of two sample words. The relatedness between pairs Drink and Alcohol would be non-zero, as their respective vectors contain common articles.

For **NED** the input comprises the target entity mention and its context, defined as the set of mentions occurring within a 101 token window centered in the target. In order to extract mentions to articles in Wikipedia from the context, we match the longest strings in our dictionary as we scan tokens from left to right. We then initialize the teleport probability with all articles referred by the mentions. After computing Personalized PageRank, we output the article with highest rank in PPV among the possible articles for the target entity mention. Figure 4.1 shows an example of NED.

If the prior is being used to initialize weights, we multiply the prior probability with the Pagerank probabilities before computing the final ranks. In the rare cases[7] where no known mention is found in the context, we return the node with the highest prior.

Note that our NED and relatedness algorithms are related. NED is using using relatedness, as Pagerank probabilities are capturing how related is each candidate article to the context of the mention. Following the first-order and second-order co-occurrence abstraction (Islam and Inkpen, 2006; Agirre and Edmonds, 2007, Ch. 6), we can interpret that we do NED using first-order relatedness, while our relatedness uses second-order relatedness.

Figure 4.2 summarizes all parameters mentioned so far, as well as their default values, which were set following previous work (Agirre *et al.* 2010, 2014).

[7]Less than 3% of instances.

1. **Graphs** in Table 4.1 (default: **Hr**)
2. Number of **iterations** in PageRank
   $\boldsymbol{i} \in \{1, 2, 3, 4, 5, 10, 15 \ldots 50\}$ (default: **30**)
3. **Damping factor** in PageRank:
   $\boldsymbol{\alpha} \in \{0.75, 0.8, 0.85, 0.90, 0.95, 0.99\}$ (default: **0.85**)
4. Initializing with **prior** or not (**P** or **¬P**) (default: **P**)
5. Relatedness: number of values in PPV:
   $\boldsymbol{k} \in \{100, 200, 500, 1000, 2000, 5000, 10000\}$ (default: **5000**)

4.2 Figure – Summary of variants and parameters as well as the default values for each of them.

Name	**Reference**	#
RG	Rubenstein and Goodenough 1965	65
MC	Miller and Charles 1991	30
353	Gabrilovich and Markovitch 2007	353
TSA	Radinsky *et al.* 2011	287
KORE	Hoffart *et al.* 2012a	420
TAC09	McNamee *et al.* 2010	1675
TAC10	`http://www.nist.gov/tac/`	1020
TAC13	`http://www.nist.gov/tac/`	1183
AIDA	Hoffart *et al.* 2011	4401
KORE	Hoffart *et al.* 2012a	143

4.4 Table – Summary of relatedness (top) and NED (bottom) datasets. Rightmost column for number of instances.

4.5 Experimental methodology

We summarize the datasets used in Table 4.4. RG, MC and 353 are the most used relatedness datasets to date, with TSA and KORE being more recent datasets where some top-ranking systems have been evaluated. Word relatedness datasets were lemmatized and lowercased, except for KORE, which is an entity relatedness dataset where the input comprises article titles[8]. Following common practice rank-correlation (Spearman) was used for evaluation.

Regarding NED, the TAC Entity Linking competition is held annually.

[8]We had to manually adjust the articles in KORE, as the exact title depends on the Wikipedia version. We missed 3 for our 2013 version, which could slightly degrade our results.

Due to its popularity it is useful to set the state of the art. We selected the datasets in 2009 and 2010, as they have been used to evaluate several top ranking systems, as well as the 2013 dataset, which is the most recent. In addition, we also provide results for AIDA, the largest and only dataset providing annotations for all entities in the documents, and KORE, a recent, very small dataset focusing on difficult mentions and short contexts. Evaluation was performed using accuracy, the ratio between correctly disambiguated instances and the total number of instances that have a link to an entity in the knowledge base[9]. Each dataset uses a different Wikipedia version, but fortunately Wikipedia keeps redirects from older article titles to the new version. As customary in the task, we automatically map the articles returned by our system to the version used in the gold standard.

Following standard practice in NED, we do not evaluate mention detection[10], that is, the datasets already specify which are the target mentions. Note that TAC provides so called "queries" which can be substrings of the full mention, e.g. "Smith" for a mention like "John Smith"). Given a mention, we devised the following heuristics to improve candidate generation: (1) remove substring contained in parenthesis from the mention, then check dictionary, (2) if not found, remove "the" if first token in the mention, then check dictionary, (3) if not found, remove middle token if mention contains three tokens, then check dictionary, (4) if not found, search for a matching entity using the Wikipedia API[11]. The heuristics provide an improvement of around 4 points on development. Later analysis showed that these heuristics seem to be only relevant on the TAC datasets, because of the way the query strings are designed, but not on AIDA or KORE.

4.5.1 Development and test

We wanted to follow a standard experimental design, with a clear development/test split for each task. Unfortunately there is no standard split in the literature, and the choice is difficult: The development dataset should be representative enough to draw conclusions on different alternatives and parameters, but at the same time the most relevant datasets in the literature

[9]Corresponds to non-NIL accuracy at TAC-KBP (also called KB accuracy) and Micro P@1.0 in (Hoffart *et al.* 2011)

[10]See (Cornolti *et al.* 2013) for a framework to evaluate both mention detection and disambiguation.

[11]`http://en.wikipedia.org/w/api.php`

should be left for testing, in order to have enough points for comparison. In addition, some recent algorithms suposedly setting the state of the art are only tested on newly produced datasets. Note also that relatedness datasets are small, making it difficult to find statistically significant differences.

In order to strike a balance between the need for in-depth analysis and fair comparison to previous results, we decided to focus on the two oldest datasets from each task for development and analysis: RG for relatedness and a subset of 200 polysemic instances from TAC09 for NED (TAC09_{200})[12]. The rest will be used for test, where the parameters have been set on development. Given the need for significant conclusions, we re-checked the main conclusions drawn from development data using the aggregation of all test datasets, but **only after** the comparison to the state of the art had been performed. This way we ensure both a fair comparison with the state of the art and a well-grounded analysis.

We performed significance tests using Fisher's z-transformation for relatedness (Press *et al.*, 2002, equation 14.5.10), and paired bootstrap resampling for NED (Noreen 1989), accepting differences with p-value < 0.05. Given the small size of the datasets, when necessary, we also report statistical significance when joining all datasets as just mentioned.

4.6 Studying the graph and parameters

In this section we study the performance of the different graphs and parameters on the two development datasets, RG and TAC09_{200}. The next section reports the results on the test sets for the best parameters, alongside state-of-the-art system results.

As mentioned in Sect. 4.4.1, PPR has several parameters and variants (cf. Figure 4.2). We first checked exhaustively all possible combinations for different graphs, with the rest of parameters set to **default values**. We then optimized each of the parameters in turn, seeking to answer the following questions:

Which links help most? Table 4.1 shows the results for selected graphs. The first seven rows present the results for each edge source in isolation, both using directed and undirected edges. Categories and infoboxes suffer from producing smaller graphs, with the hyperlinks yielding the best results. The undirected versions improve over directed links in all cases, with the use of

[12]The dataset in `http://ixa2.si.ehu.es/ukb/README.wiki.txt` includes the subset.

Graph	Param.	RG	Param.	TAC09_{200}
Hr	**default**	**88.4**	**default**	**68.5**
Hr	¬P	87.0	¬P	49.0 †
Hr	α0.85	88.4	α0.85	68.5
Hr	i30	88.4	i15	68.5
Hr	k5000	88.4	–	–

4.5 Table – Parameters: Summary of results on development data for relatedness (RG, Spearman correlation) and NED (TAC09_{200}, accuracy) for several parameters using Hr graph. Parameters are set to default values (see text) except for the one noted explicitly. † for statistical significant differences with respect to default.

reciprocal edges for hyperlinks obtaining the best results overall (the graphs with reciprocal edges for categories and infoboxes were too small and we omit them). The trend is the same in both relatedness and NED, highlighting the robustness of these results.

Regarding combined graphs, we report the most significant combinations. The reciprocal graph of hyperlinks outperforms all combinations (including the combinations which were omitted), showing that categories and infoboxes do not help or even degrade slightly the results. The differences are statistically significant (either on the individual datasets or in the aggregation on all datasets) in all cases, confirming that Hr is significantly better.

The degradation or lack of improvement when using infoboxes is surprising. We hypothesized that it could be caused by non-reciprocal links in HrIu. In fact, removing non-reciprocal links from HrIu improved results slightly on NED, matching those of Hr. This lack of improvement with infoboxes, even when removing non-reciprocal links, can be explained by the fact that only 5% of reciprocal links in Iu are not in Hr. It seems that this additional 5% is not helping in this particular dataset. Regarding categories, the category structure is mostly a tree, which is a structure where random walks do not seem to be effective, as already observed in (Agirre *et al.* 2014) for WordNet.

Is initialization of random walks important? The second row in Table 4.5 reports the result when using uniform distributions when initializing the random walks (instead of prior probabilities). The results degrade in both datasets, the difference being significant only for NED. This was later confirmed in the rest of relatedness and NED datasets: using prior probabil-

Graph	Method	RG	TAC09_{200}
HR	NGD	81.8 ‡	57.5†
HR	PPR (1 iter.)	43.4 † ‡	60.5† ‡
HR	PPR (2 iter.)	78.3 ‡	66.0† ‡
HR	PPR *default*	**88.4**	**68.5**

4.6 Table – Result when using single links, compared to the use of the full graph on development data. We reimplemented NGD. † for stat. signif. difference with PPR. ‡ for stat. signif. using all datasets.

Graph	Method	Year	RG	TAC09_{200}
HR	PPR *default*	2010	86.3	68.5
HR	PPR *default*	2011	85.6	**70.5**
HR	PPR *default*	2013	**88.4**	68.5

4.7 Table – PPR using different Wikipedia versions

ities for initialization improves results in all cases, but it is only significant in NED datasets. These results show that relatedness is less sensitive to changes in the distribution of meanings, that is, using the more informative prior distributions of meaning only improves results slightly. NED, on the contrary, is more sensitive, as the distribution of senses affects dramatically the performance.

Is the value of α and i important? The best α on both datasets was obtained with default values (cf. Table 4.5), in agreement with related work using WordNet (Agirre *et al.* 2010). The lowest number of iterations where convergence was obtained were 30 and 15, respectively, although as few as 5 iterations yielded very similar performance (87.1 on relatedness, 68.0 on NED).

Is the size of the vector, k, important for relatedness? The best performance was attained for the default k, with minor variations for $k > 1000$.

Is the full graph helping? When the PPR algorithm does a single iteration, we can interpret that it is ranking all entities using direct links. When doing two iterations, we can loosely say that it is using links at distance two, and so on. Table 4.6 shows that PPR is able to take profit from the full graph well beyond 2 iterations, specially in relatedness. These results were confirmed in the full set of datasets, with statistically significant differences in all cases.

In addition, we reimplemented the relatedness and NED algorithms based on NGD over direct links (Milne and Witten 2008a, b), allowing to compare them to PPR on the same experimental conditions. We first developed the relatedness algorithm[13]. Table 4.6 reports the best variant, which outperforms the 0.64 on RG reported in their paper. We followed a similar methodology for NED[14]. Table 4.6 shows the results for NGD, which performs worse than PPR. This trend was confirmed on the full set of datasets for relatedness and NED with statistical significance in all cases except KORE, which is the smallest NED dataset. Figure 4.1 illustrates why the use of longer paths is beneficial. In fact, NGD returns 0.14 for `B&I_Lions` and 0.13 for `Highveld_Lions`, but PPR correctly returns 0.05 and 0.75, respectively.

How important is the Wikipedia version? Table 4.7 shows that the versions we tested are not affecting the results dramatically, and that using the last version does not yield better results in NED. Perhaps the larger size and number of hyperlinks of newer versions would only affect new articles and rare articles, but not the ones present in TAC09_{200}. We kept using 2013 for test.

What is the efficiency of the algorithm? The initialization takes around 5 minutes[15], where most of the time is spent loading the dictionary into memory, 4m50s. Using a database instead, initialization takes 10s. Memory requirements for HR were 4.7 Gb, down to 1.1 Gb when using the database. The main bottleneck of our system is the computation of Personalized PageRank, each iteration taking around 0.60 seconds. We are currently checking fast approximations for Pagerank, and plan to improve efficiency.

[13]In order to replicate the NGD relatedness algorithm, we checked the open source code available, exploring the use of inlinks and outlinks and the use of maximum pairwise article relatedness. We also realized that the use of priors ("commonness" according to the terminology in the paper) was hurting, so we dropped it. We checked both reciprocal and unidirectional versions of the hyperlink graph, with better results for the reciprocal graph.

[14]We checked both reciprocal and undirected graphs with similar results, combined with prior (similar results), weighted terms in the context (with improvement) and checked the use of ambiguous mentions in the context (marginal improvement). Reported results correspond to reciprocal, combination with prior, weighting terms and using only monosemous mentions.

[15]Time measured in a single server with Xeon E7-4830 8 core processors, 2130 MHz, 64 GB RAM.

4.7 Comparison to related work

In the previous section we presented several results on the same experimental conditions. We now use the graph and parametrization which yield the best results on development (default parameters with HR). Comparison to the state of the art is complicated by many systems reporting results on different datasets, which causes the tables in this section to be rather sparse. The comparison for relatedness is straightforward, but, in NED, it is not possible to factor out the impact of the candidate generation step. Given the fact that our candidate generation procedure is not particularly sophisticated, we don't think this is a decisive factor in favour of our results.

Table 4.8 and 4.9 report the results of the best systems on both tasks. Given that several systems were developed on test data, we also report our results on RG and TAC2009, marking all such results (see caption of tables for details). We split the results in both tables in three sets: top rows for systems using link and graph information alone, middle rows for link- and graph-based systems using WordNet and/or Wikipedia, and bottom rows for more complex systems. We report the results of our system repeatedly in each set of rows, for easier comparison. Our main focus is on the top rows, which show the superiority of our results with respect to other systems using Wikipedia links and graphs. The middle and bottom rows show the relation to the state of the art.

For easier exposition, we will examine the results by row section simultaneously on relatedness and NED. The **top rows** in Table 4.8 report four relatedness systems which have already been presented in Sect. 4.2, showing that our system is best in all five datasets. Note that the (Milne and Witten 2013) row was obtained running their publicly available system with the supervised Machine Learning component turned off (see below for the results using SUP). The top rows of table 4.9 report the most frequent baseline (as produced by our dictionary) and three link-based systems (cf. Sect. 4.2), showing that our method is best in all five datasets. These results show that the use of the full graph as devised in this paper is a winning strategy.

The relatedness results in the **middle rows** of Table 4.8 include several systems using WordNet and/or Wikipedia (cf. Sect. 4.2), including the system in (Agirre *et al.* 2010), which we run out-of-the-box with default values. To date, link-based systems using WordNet had reported stronger results than their counterparts on Wikipedia, but the table shows that our Wikipedia-based results are the strongest on all relatedness datasets but one

(MC, the smallest dataset, with only 30 pairs). In addition, the table shows our results when combining random walks on Wikipedia and WordNet[16], which yields improvements in most datasets. In the counterpart for NED in Table 4.9, Moro et al. (Moro *et al.* 2014) outperform our system, specially in the smaller KORE (143 instances), but note that they use a richer graph which combines WordNet, the English Wikipedia and hyperlinks from other language Wikipedias.

Finally, the **bottom rows** in both tables report the best systems to date. For lack of space, we cannot review systems not using Wikipedia links. Regarding relatedness, we can see that our combination of WordNet and Wikipedia would rank second in all datasets, with only one single system (based on corpora) beating our system in more than one dataset (Radinsky *et al.* 2011). Regarding NED, our system ranks first in the TAC datasets, including the best systems that participated in the TAC competitions (Varma *et al.* 2009; Lehmann *et al.* 2010; Cucerzan and Sil 2013), and second to (Moro *et al.* 2014) on AIDA and KORE.

4.8 Conclusions and Future Work

This work departs from previous work based on Wikipedia and derived resources, as it focuses on a single knowledge source (links in Wikipedia) with a clear research objective: given a well-established random walk algorithm we explored which sources of links and filtering methods are useful, contrasting the use of the full graph with respect to using just direct links. We follow a clear development/test/analysis methodology, evaluating on a extensive range of both relatedness and NED datasets. All software and data are publicly available, with instructions to obtain out-of-the-box replicability[17].

We show for the first time that random walks over the full graph of links improve over direct links. We studied several variations of sources of links, showing that non-reciprocal links hurt and that the contribution of the category structure and relations in infoboxes is residual. This paper sets a new state-of-the-art for systems based on Wikipedia links on both word relatedness and named-entity disambiguation datasets. The results are close to those of the best combined systems, which specialize on either relatedness or disambiguation, use several information sources and/or supervised machine

[16]We multiply the scores of PPR on Wikipedia and WordNet.

[17]`http://ixa2.si.ehu.es/ukb/README.wiki.txt`

learning techniques. This work shows that a careful analysis of varieties of graphs using a well-known random walk algorithm pays off more than most ad-hoc algorithms proposed up to date.

For the future, we would like to explore ways to filter out informative hyperlinks, perhaps weighting edges according to their relevance, and would also like to speed up the random-walk computations.

This article showed the potential of the graph of hyperlinks. We would like to explore combinations with other sources of information and algorithms, perhaps using supervised machine learning. For relatedness, we already showed improvement when combining with random walks over WordNet, but would like to explore tighter integration (Pilehvar *et al.* 2013). For NED, local methods (Ratinov *et al.* 2011; Han and Sun 2011), global optimization strategies based on keyphrases in context like KORE (Hoffart *et al.* 2012a) and doing NED jointly with word sense disambiguation (Moro *et al.* 2014), all are complementary to our method and thus promising directions.

	Source		RG	353	TSA	MC	KORE
Ponzetto and Strube 2011	Wiki11	c	75.0*				
Nastase and Strube 2013	Wiki13	ci	67.0				
Milne and Witten 2013	Wiki13	la	69.5r	59.7r	35.8r	77.2r	65.9r
Yeh *et al.* 2009	Wiki09	g		48.5			
PPR *default* HR	**Wiki13**	g	0 **88.4***	1 **72.8**	1 **64.1**	1 **81.0**	1 **66.2**
Agirre *et al.* 2010	WNet	g	1 **86.2r**	68.5	45.4r	3 85.2r	
Tsatsaronis *et al.* 2010	WNet	g	86.1	61.0			
Navigli and Ponzetto 2012b	WNet+Wiki12 (cl)	g+CL		65.0		1 **90.0**	
Pilehvar *et al.* 2013	WNet+Wiki13	g	86.8*				
PPR *default* HR	**Wiki13**	g	0 **88.4***	2 **72.8**	1 **64.1**	4 **81.0**	1 **66.2**
PPR *default* HR	**WNet+Wiki13**	g	0 **91.8***	1 **78.5**	2 **62.9**	2 **87.6**	1 **66.2**
Gabrilovich and Markovitch 2007	Wiki07	t	82.0	75.0	59.0	73.0	
Hoffart *et al.* 2012a	Wiki12	t					0 69.8*
Yazdani and Popescu-Belis 2013	Freebase	gt		70.0*			
Radinsky *et al.* 2011	Time	C		1 **80.0**	1 **63.0**		
Baroni *et al.* 2014	Corpus	C	84.0*	71.0			
Agirre *et al.* 2009	WNet+Corpus	Cg+SUP	0 **96.0**x	78.0x			
Milne and Witten 2013	Wiki13	la+SUP	83.5r	74.0x	52.8r	81.3r	1 **66.5**r
PPR *default* HR	**WNet+Wiki13**	g	0 **91.8***	2 **78.5**	2 **62.9**	2 **87.6**	2 **66.2**

4.8 Table – Spearman results for relatedness systems. The source column includes **codes** for information used (**t** for article text, **l** for direct hyperlinks, **g** for hyperlink graph, **c** for categories, **i** for infoboxes, **a** for anchor text) and other information sources (**CL** for crosslingual links, **C** for corpora, **SUP** for supervised Machine Learning). The results include the following codes: ***** for best reported result among several variants, **x** for cross-validation result, **r** for third-party system ran by us. We also include the rank of our PPR system in each group or rows, including the systems above it (excluding * and x systems, which get rank 0 if they are top rank).

System	Source		TAC09	TAC10	TAC13	AIDA	KORE
MFS baseline	Wiki13	l	68.3	73.7	72.7	69.0	36.4
Guo *et al.* 2011	Wiki10	l	1 **74.0**	74.1			
Milne and Witten 2013	Wiki13	la	57.4r	58.5r	37.1r	56.0r	35.7r
García *et al.* 2014	Wiki12	l		76.6			
PPR *default* HR	Wiki13	g	0 **78.8***	1 **83.6**	1 **81.7**	1 **80.0**	1 **60.8**
Moro *et al.* 2014	WNet+Wiki13	g+CL				1 **82.1**	1 **71.5**
PPR *default* HR	Wiki13	g	0 **78.8***	1 **83.6**	1 **81.7**	2 **80.0**	2 **60.8**
Bunescu and Pasca 2006	Wiki11	tc	0 **83.8**ra*	68.4ra			
Cucerzan 2007	Wiki11	tc	0 **83.5**ra*	78.4ra		51.0ro	
Hachey *et al.* 2011	Wiki11	tcg		79.8*			
Hoffart *et al.* 2012a	Wiki12	t				0 81.8*	0 64.6*
Hoffart *et al.* 2011	Wiki11	tli+SUP				0 81.8*	
Milne and Witten 2013	Wiki13	la+SUP	57.5r	63.4r	40.0r	55.6r	37.1r
Best TAC KBP system	—	—	1 **76.5**	80.6	77.7		
PPR *default* HR	Wiki13	g	0 **78.8***	1 **83.6**	1 **81.7**	2 **80.0**	2 **60.8**

4.9 Table – Accuracy of NED systems, using the same codes as in in Table 4.8. Some early systems have been re-implemented and tested by others: **ra** for (Hachey *et al.* 2012), **ro** (Hoffart *et al.* 2011). We report rank of our PPR system in each group or rows, including systems above (excluding * systems, which get rank 0 if they are top rank).

5

Algoritmo globalak eta lokalak konbinatzen

Kapitulu honek artikulu bildumaren bigarren artikulua azalduko du. Laburbilduz, eredu sortzaileen ikuspuntutik testuingurua modelatuko duen eredu lokala azalduko da. Ondoren, aurreko artikuluan garatu den eredu globala konbinatuko da. Artikuluko eredu global eta lokalen konbinaketak eredu osagarri eta egonkorra erakutsiko du, jarraian, jatorrizko fitxa eta ingelesezko bertsioa:

Ander Barrena, Aitor Soroa and Eneko Agirre. ***Combining Mention Context and Hyperlinks from Wikipedia for Named Entity Disambiguation.*** *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics *SEM 2015.* Denver, Colorado, USA. 2015.

Named entity disambiguation is the task of linking entity mentions to their intended referent, as represented in a Knowledge Base, usually derived from Wikipedia. In this paper, we combine local mention context and global hyperlink structure from Wikipedia in a probabilistic framework. Our results show that the two models of context, namely, words in the context and hyperlink pathways to other entities in the context, are complementary. We test our method in eight datasets, improving the state-of-the-art results in five, without any tuning, showing that it is robust to out-of-domain scenarios. When tuning combination weights, we match the best reported results on the widely-used AIDA-CoNLL test-b.

5.1 Introduction

Linking mentions occurring in documents to a knowledge base is the main goal of Entity Linking or Named Entity Disambiguation (NED). This problem has attracted a great number of papers in the NLP and IR communities, and a large number of techniques, including local context and global inference (Ratinov *et al.* 2011). We propose to use a probabilistic framework that combines entity popularity, name popularity, local mention context and global hyperlink structure, relying on information in Wikipedia alone. Entity and name popularity are useful disambiguation clues in the absence of any context. The local mention context provides direct clues (in the form of words in context) to disambiguate each mention separately. The hyperlink structure of Wikipedia provides a global coherence measure for all entities mentioned in the same context.

The advantages of our method with respect to other alternatives are as follows: (1) It does not involve a large number of methods and classifier combination. (2) The method learns the parameters directly from Wikipedia so no additional hand-labeled data and training is needed. (3) We combine the global hyperlink structure of Wikipedia with a local bag-of-words probabilistic model in an intuitive and complementary way. (4) The absence of training allows for robust results in out-of-domain scenarios.

The evaluation of NED is fragmented, with several popular shared tasks, such as TAC-KBP[1], ERD[2] or NEEL[3]. Other evaluation datasets include AIDA-CoNLL and KORE50[4], which are very common in NED evaluation. Note that each dataset poses different problems. For instance AIDA-CoNLL is composed of news, and systems need to disambiguate all occurring mentions. TAC includes news and discussion forums, and focuses on a large number of mentions for a handful of challenging strings. KORE50 includes short sentences with very ambiguous mentions. Unfortunately, there is no standard dataset, and many contributions in this area report results in just one or two datasets. We report our results on eight datasets, improving the state-of-the-art results on five.

[1] `http://www.nist.gov/tac/2014/KBP/`

[2] `http://web-ngram.research.microsoft.com/erd2014/`

[3] `http://www.scc.lancs.ac.uk/microposts2015/challenge/index.html`

[4] `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/`

5.2 Resources

The knowledge used by our Bayesian network comes from Wikipedia. We extract three information resources to perform the disambiguation: a dictionary, textual contexts and a graph.

The dictionary is an association between strings and Wikipedia articles. We construct the dictionary using article titles, redirections, disambiguation pages, and anchor text. If the mention links to a disambiguation page, it is associated with all possible articles the disambiguation page points to. Each association between a string and article is scored with the prior probability, estimated as the number of times that the mention occurs in the anchor text of an article divided by the total number of occurrences of the mention. We choose candidate entities for disambiguation by just assigning all entities linked to the mention in the dictionary.

In addition we build a graph using the Wikipedia link structure, where entities are nodes and edges are anchor links among entities from Wikipedia. We used the third-party dictionary and graph described in (Agirre *et al.* 2015), which is publicly available[5].

Finally, we extract textual contexts for all the possible candidate entities from a Wikipedia dump. We collect all the anchors including a link to each entity in Wikipedia, and extract a context of 50 words around the anchor link.

5.3 A Generative Bayesian Network

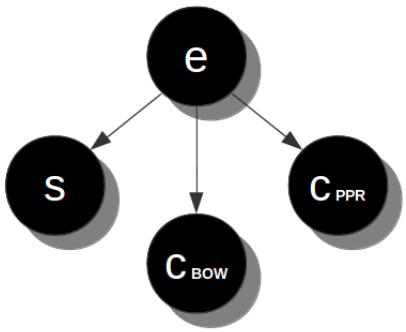
Given a mention s occurring in context c, our system ranks each of the candidate entities e. Figure 7.2 shows the dependencies among the different variables. Note that context probability is given by two different resources.

Candidate entities are ranked combining evidences from 4 different probability distributions, which we call entity knowledge $P(e)$, name knowledge $P(s|e)$, context knowledge $P(c_{\text{bow}}|e)$ and graph knowledge $P(c_{\text{ppr}}|e)$ respectively.

Entity knowledge $P(e)$ represents the probability of generating entity e, and is estimated as follows:

$$P(e) = \frac{C(e) + 1}{|M| + N}$$

[5] `http://ixa2.si.ehu.eus/ukb/ukb-wiki.tar.bz2`



5.1 Figure – Dependencies among variables in a Bayesian network. The network gives as a result this formula: $P(s, c_{\text{bow}}, c_{\text{ppr}}, e) = P(e)P(s|e)P(c_{\text{bow}}|e)P(c_{\text{ppr}}|e)$.

where $C(e)$ describes the entity popularity, e.g., the number of times the entity e is referenced within Wikipedia, $|M|$ is the number of entity mentions and N is the total number of entities in Wikipedia. As can be seen, the estimation is smoothed using the *add-one* method.

Name knowledge $P(s|e)$ represents the probability of generating a particular string s given the entity e, and is estimated as follows:

$$P(s|e) = \theta\frac{C(e,s)}{C(e)} + (1-\theta)\frac{C(s)}{|M|}$$

where $C(e, s)$ is the number of times mention s is used to refer entity e and $C(s)$ is the number of times mention s is used as an anchor. θ parameter is set to 0.9 according to development experiments done in the AIDA-CoNLL development set (also known as AIDA-CoNLL test-a, cf. Section 5.4).

The context knowledge is modeled in two different ways. In the bag-of-words model, $P(c_{\text{bow}}|e)$ represents the probability of generating context $c = \{w_1, w_2, \ldots, w_n\}$ given the entity e, and is estimated as follows:

$$P(c_{\text{bow}}|e) = P(w_1|e)P(w_2|e)...P(w_n|e)$$

where $P(w|e)$ is estimated as:

$$P(w|e) = \lambda P_e(w) + (1-\lambda)P_w(w)$$

$P_e(w)$ is the maximum likelihood estimation of each word w in the context of e entity. Context words are smoothed by $P_w(w)$ that is the likelihood of words in the whole Wikipedia. λ parameter is set to 0.9 according to development experiments done in AIDA-CoNLL test-a.

The graph knowledge is estimated using personalized Pagerank. We used the probabilities returned by UKB[6] (Agirre *et al.* 2015). This software returns $P(e|c_{\text{ppr}})$[7] the probability of visiting a candidate entity when performing a random walk on the Wikipedia graph starting in the entity mentions in the context. In order to introduce it in the generative model, we must first convert it to $P(c_{\text{ppr}}|e)$. We use Bayes formula to estimate the probability:

$$P(c_{\text{ppr}}|e) = P(e|c_{\text{ppr}})P(c_{\text{ppr}})/P(e)$$

Finally, the *Full Model* combines all evidences to find the entity that maximizes the following formula:

$$e = \arg\max_e P(s, c_{\text{bow}}, c_{\text{ppr}}, e) = \arg\max_e P(e)P(s|e)P(c_{\text{bow}}|e)P(c_{\text{ppr}}|e)$$

5.4 Experiments

We tested our algorithms on a wide range of datasets: AIDA-CoNLL test-b (Hoffart *et al.* 2011), KORE50 (Hoffart *et al.* 2012a) and six TAC-KBP[8] datasets corresponding to six years of the competition (AIDA, KORE and TAC hereafter). No corpus was used for training the parameters of the system, apart from Wikipedia, as explained in the previous sections.

We used gold-standard mentions and we evaluated only those mentions linked to a Wikipedia entity (ignoring so-called NIL cases). Depending on the dataset, we used the customary evaluation measure: micro-accuracy (AIDA, KORE, TAC09 and TAC10) or Bcubed+ (TAC11, TAC12, TAC13 and TAC14)[9].

Each gold standard uses a different Wikipedia version: 2010 for AIDA and KORE, 2008 for TAC. We use the Wikipedia dump from 25-5-2011 to build our resources, as this is close to the versions used at the time. We mapped

[6] `http://ixa2.si.ehu.es/ukb/`

[7] Note that, contrary to us, the results in (Agirre *et al.* 2015) multiply the Pagerank probability with the prior.

[8] `http://www.nist.gov/tac/publications/index.html`

[9] Note that TAC14 results correspond to the so-called Diagnostic Entity Linking task.

gold-standard entities to 2011 Wikipedia automatically, using redirects in the 2011 Wikipedia. This mapping could cause a small degradation of our results.

5.4.1 Results

The top 4 rows in table 5.2 show the performance of the different combinations among probabilities. The remaining row shows the best results reported to date on those datasets (see caption for details).

The results suggest that each probability contributes to the final score of the *Full Model*, shown on row 4, showing that both context models are complementary between each other[10]. The only exception is TAC13, where the bow model is best.

Our system obtains very good results in all datasets, excelling in TAC09-10-11-12-13, where it beats the state-of-the-art. The figures obtained by the *Full Model* on AIDA, KORE and TAC14 are close to the best results. Note that the table shows the results of the system reporting the best values for each dataset, that is, our system is compared not to one single system but to all those systems. For example, (Hoffart *et al.* 2012b) reported lower figures for KORE, 64.58. Regarding the results for TAC-KBP, the full task includes linking to the Knowledge Base and detecting and clustering NIL mentions. In order to make results comparable to those for in AIDA and KORE, the table reports the results for mentions which are linked to the Knowledge Base, that is, results where NIL mentions are discarded.

5.5 Adjusting the model to the data

We experimented with weighting the probabilities to adapt the *Full Model* mentioned above to a specific scenario. For the *Weighted Full Model*, we introduce the α, β, γ and δ parameters[11] as follows:

$$e = \arg\max_e P(s, c_{\text{bow}}, c_{\text{ppr}}, e) = \arg\max_e P(e)^\alpha P(s|e)^\beta P(c_{\text{bow}}|e)^\gamma P(c_{\text{ppr}}|e)^\delta$$

[10]The results of our combination involving the UKB software are not comparable to those reported by (Agirre *et al.* 2015), due to the different formulation of the probability distribution which involves the prior.

[11]$\alpha + \beta + \gamma + \delta = 1$

Test	AIDA
$P(e)P(s\|e)P(c_{\text{bow}}\|e)P(c_{\text{ppr}}\|e)$	83.28
$P(e)^{\alpha}P(s\|e)^{\beta}P(c_{\text{bow}}\|e)^{\gamma}P(c_{\text{ppr}}\|e)^{\delta}$	84.88
Moro *et al.* 2014	82.10
Hoffart *et al.* 2011	82.54
Houlsby and Ciaramita 2014	84.89

5.1 Table – Micro accuracy results for AIDA introducing the *Weighted Full Model* in row 2.

Weighting may change the optimal configuration for θ and λ, we thus optimized all parameters on the development set of AIDA, yielding $\theta = 0.9$, $\lambda = 0.7$, $\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.6$ and $\delta = 0.1$ performing a exhaustive grid search. The step size used in this experiment is 0.1. The parameters yielded high results for development, up to 83.48.

Table 5.1 summarizes the results of the *Weighted Full Model* for AIDA, showing that model reaches 84.88 points, a la par to the best micro accuracy reported by (Houlsby and Ciaramita 2014) and above those reported by (Hoffart *et al.* 2011; Moro *et al.* 2014) (respectively, 82.54[12] and 82.10). Unfortunately the parameter distribution seems to depend on the test dataset, as the same parameters failed to improve the results on the other datasets.

5.6 Related Work

The use of Wikipedia for named entity disambiguation is a common approach in this area. In the related field of Wikification, (Ratinov *et al.* 2011) introduced the supervised combination of a large number of global and local similarity measures. They learn weights for each of those measures training a supervised classifier on Wikipedia. Our approach is different in that we just combine four intuitive methods, without having to learn weights for them. Unfortunately they don't report results for NED.

(Moro *et al.* 2014) present a complex graph-based approach for NED and Word Sense Disambiguation which works on BabelNet, a complex combina-

[12]Note that values by (Hoffart *et al.* 2011) were reported on a subset of AIDA. The micro accuracy results reported in our table correspond to the latest best model from the AIDA web site: `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/`.

tion of several resources including, among others, Wikipedia, WordNet and Wiktionary. Our results are stronger over AIDA, but not on the smaller KORE.

(Hoffart *et al.* 2011) presents a robust method based on entity popularity and similarity measures, which are used to build a mention/entity graph. They include external knowledge from Yago, and train a classifier on the train part of AIDA, obtaining results comparable to ours. Given that we do not train on in-domain training corpora, we think our system is more robust.

The use of probabilistic models using Wikipedia for NED was introduced in (Han and Sun 2011). In this paper, we extend the model with a global model which takes the hyperlink structure of Wikipedia into account.

(Houlsby and Ciaramita 2014) presents a probabilistic method using topic models, where topics are associated to Wikipedia articles. They present strong results, but they need to initialize the sampler on another NED system, Tagme (Ferragina and Scaiella 2012). In some sense they also combine the knowledge in the graph with that of a local algorithm (Tagme), so their work is complementary to ours, but in their case the improvement obtained when using the graph is negligible. They only provide results on AIDA, and it is thus not possible to compare their robustness with that of our algorithm.

5.7 Conclusions and future work

Bayesian networks provide a principled method to combine knowledge sources. In this paper we combine popularity, name knowledge and two methods to model context: bag-of-words context, and hyperlink graph. The combination outperforms the state-of-the-art in five out of eight datasets, showing the robustness of the system in different domain and dataset types. Our results also show that in all but one dataset the combination outperforms individual models, indicating that bag-or-word context and graph context are complementary. We show that results can be further improved when tuning the weights on in-domain development corpora, matching the best results on the widely-used AIDA-CoNLL test-b.

Given that Bayesian networks can be further extended, we are exploring to introduce additional models of context into a Markov Random Field algorithm. Our current model assumes that the two models of context (bag or words and graph) are independent given e, and we would like to explore alternatives to relax this assumption.

Test	**AIDA**	**KORE**	**TAC09**	**TAC10**	**TAC11**	**TAC12**	**TAC13**	**TAC14**
$P(e)P(s\|e)$	67.54	35.42	67.04	76.96	67.83	46.20	66.54	62.01
$P(e)P(s\|e)P(c_{\text{bow}}\|e)$	75.05	60.42	77.19	85.20*	75.55	57.06	**74.56***	71.21
$P(e)P(s\|e)P(c_{\text{ppr}}\|e)$	76.83	54.86	79.40*	83.92*	79.75	70.13*	70.21	71.28
$P(e)P(s\|e)P(c_{\text{bow}}\|e)P(c_{\text{ppr}}\|e)$	**83.28**	**70.83**	**82.21***	**85.98***	**81.85***	**71.65***	73.99*	**76.48**
Best (state-of-the-art)	84.89	71.50	79.00	80.60	80.10	68.50	71.80	79.60

5.2 Table – Bold marks the best value among probability combinations, and * those results that overcome the best value reported in the state-of-the-art: (Houlsby and Ciaramita 2014) for AIDA, (Moro *et al.* 2014) for KORE, (Han and Sun 2011) for TAC09 and see TAC-KBP proceedings for the rest.

6

Entitate bakarra diskurtsoan eta agerkidetzan

Kapitulu honek artikulu bildumaren hirugarren artikulua azalduko du. Laburbilduz, dokumentu eta agerkidetza mailan izen-aipamen berdina behin baino gehiagotan azaltzen baldin bada, gehienetan entitate berdinari erreferentzia egingo diola ikusiko da. Lehenik, hipotesi hau hainbat corpusetan aztertuko da. Ondoren, hiru EID sistemetan aplikatuko da emaitzak hobetzen direla ikusiz. Sistema hauen artean, aurreko bi ataletako eredu globala eta lokala azaltzen dira. Bestalde, arloaren egoerako *Spotlight* izeneko sisteman ere ekarpenak ebaluatuko dira. Jarraian, artikuluaren jatorrizko fitxa eta ingelesezko bertsioa:

Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Peñas and Aitor Soroa. ***"One Entity per Discourse" and "One Entity per Collocation" Improve Named-Entity Disambiguation.*** *Proceedings of the 25th International Conference on Computational Linguistics COLING 2014.* Dublin, Ireland. 2014.

The "one sense per discourse" (*OSPD*) and "one sense per collocation" (*OSPC*) hypotheses have been very influential in Word Sense Disambiguation. The goal of this paper is twofold: (i) to explore whether these hypotheses hold for entities, that is, whether several mentions in the same discourse (or the same collocation) tend to refer to the same entity or not, and (ii) test their impact in Named-Entity Disambiguation (NED). Our experiments show consistent results on different collections and three state-of-the-art NED

system. *OSPD* hypothesis holds in around 96%-98% of documents whereas *OSPC* hypothesis holds in 91%-98% of collocations. Furthermore, a simple NED post-processing in which the majority entity is promoted, produces a gain in performance in all cases, reaching up to 8 absolute points of improvement in F-measure. These results show that NED systems would benefit of considering these hypotheses into their implementation.

6.1 Introduction

The "one sense per discourse" (*OSPD*) hypothesis was introduced by (Gale *et al.* 1992), and stated that a word tends to preserve its meaning when occurring multiple times in a discourse. They estimated that the probability of two occurrences of the same polysemous noun drawn from one document having the same sense to be around 94% for documents from Grolier encyclopedia, and 96% for documents from Brown, based on word senses from the Oxford Advanced Learner's Dictionary and a handful of examples. A few years later, (Krovetz 1998) reported 66% on larger corpora (SemCor and DSO) annotated with WordNet senses by third parties, but, unfortunately, he only reported how many polysemous nouns occurred with a single sense in **all** documents, not in each document. In the context of statistical machine translation, (Carpuat 2009) reported that, 80% of the time, words occurring multiple times in a source document are translated into a single word in the target language.

In the case of entities, *OSPD* is closely related to coreference, where the task is to find whether two different mentions (perhaps using different surface strings like *John* and *he*) in a document refer to the same entity or not. For instance, the coreference system presented by (Lee *et al.* 2013), uses a heuristic which links mentions in a document that share the same surface string: "This sieve [heuristic] accounts for approximately 16 CoNLL F1 points improvement, which proves that a significant percentage of mentions in text are indeed repetitions of previously seen concepts". Our paper actually quantifies the amount of those repetitions for entities, providing additional evidence for the heuristic.

The "one sense per collocation" (*OSPC*) hypothesis was introduced by (Yarowsky 1993), stating that a word tends to preserve its meaning when occurring with the same collocate. Yarowsky tested his hypothesis for several definitions of collocate, including positional collocates (word to left or

right) and syntactic collocations (governing verb of object, governing verb of subject, modifying adjective). He reported entropy on train data, as well as disambiguation performance on unseen data, with the precision ranging between 90% and 99% for a handful of words with two distinct homograph senses, like, e.g. *bass* or *colon*. In larger-scale research, (Martinez and Agirre 2000) measured the precision of similar collocations on corpora (Semcor and DSO) annotated by third parties with finer-grained senses from WordNet, reporting lower figures around 70%.

In this paper, we take a collocation to be a word (or multiword term) that co-occurs with the target named-entity more often than would be expected by chance. In our case we use syntactic dependencies to extract co-occurring terms.

These two hypotheses have been very influential, and have inspired multiple heuristics and methods in Word Sense Disambiguation research (Agirre and Edmonds, 2007, Chapters 5,7,10,11). In this work we are going to show that both hypotheses hold for named-entities as well, and that the hypotheses can be used to post-process the output of any Named-Entity Disambiguation system (NED) to improve its performance. NED, also known as Entity Linking, takes as input a named-entity mention in context and assigns it a specific entity from a given entity repository (Hachey *et al.* 2012; Daiber *et al.* 2013).

In the first part of this work we are going to test whether the two hypotheses hold for entity mentions with respect to a repository of entities extracted from Wikipedia. For instance, do all occurrences of mention *Abbott* in a document refer to the same entity? Do all occurrences of mention *CPI* as subject of verb *rise* refer to the same entity? Do all occurrences of *CDU* in relation to *Merkel* refer to the same entity? The examples in Figures 6.1 and 6.2 show evidence that this is indeed the case. The experiments aim at quantifying in which degree *OSPD* and *OSPC* hypotheses hold for entities[1].

In the second part of the paper, we will explore a simple method to incorporate *OSPD* and *OSPC* hypotheses to any existing NED system, showing their potential. After running the NED system, we take its output and observe, for each mention string, which is the entity returned most often for a given document (or collocation), assigning to all occurrences the majority entity. We tested the improvements with a freely available NED system

[1]For the sake of clarity we will also refer to *OSPD* and *OSPC* for entities as *OSPD* and *OSPC*.

Abbott Beefs Up Litigation Reserves NORTH CHICAGO, Ill. (AP) *Abbott* Laboratories Inc., bracing for a costly settlement in a federal investigation involving the prostate-cancer drug Lupron, said Friday it was increasing litigation reserves by $344 million. As part of the announcement, *Abbott* said it had restated its quarterly results and is now reporting a loss of $319.9 million for the first three months of this year rather than a profit. The move comes amid long-running negotiations between the U.S. Department of Justice and TAP Pharmaceutical Products, the 50-50 joint venture between *Abbott* and Takeda Chemical Industries of Japan that made Lupron. *Abbott* said in January ...

6.1 Figure – Example of *OSPD* for entities. All occurrences of *Abbott* refer to `Abbott_Laboratories`.

(Daiber *et al.* 2013), a reimplementation of a strong Bayesian NED system (Han and Sun 2011) and an in-house graph-based system. We got statistically significant improvements for all systems and "one sense" hypotheses that we tested, with a couple exceptions.

In order to check the *OSPD* and *OSPC* hypotheses for entities, we first looked into existing datasets. AIDA (Hoffart *et al.* 2011)[2] is a publicly available hand-tagged corpus based on the CoNLL named-entity recognition and disambiguation task dataset. AIDA contains links of all entity mentions in full documents, so it is a natural fit for *OSPD*. We estimated *OSPD* based on more than 4,000 mentions that occur multiple times in a document. For completeness, we also estimated *OSPD* at the collection level.

OSPD and *OSPC* are independent of each other, as one is applied at the document level and the other at the corpus level, focusing on the entities that occur with a specific collocation. Multiple occurrences of a target string in a document usually occur with different collocations, and conversely, multiple occurrences of a target string with a specific collocation typically occur in different documents. Note also that singletons (entities that are only mentioned once in a document) are not affected by *OSPD*, but could be affected by *OSPC*.

[2] `http://www.mpi-inf.mpg.de/yago-naga/aida/downloads.html`

***CPI* subject-of rise:**
... China's consumer price index, or *CPI*, rose 2.8 percent last December ...
... October, the *CPI* rose 1.35 percent, the core price index grew 1.13 percent ...
... month-on-month basis, March *CPI* rose 2.3 percent from February, ...
... in China, Hong Kong and Singapore, whose *CPI*s have rised 8.0 percent, ...
... The core *CPI* rose 0.2 percent, in line with Wall Street expectations ...
Angela Merkel has *CDU*:
... power with Merkel's *CDU* nationally in an uneasy "grand coalition" ...
... Michael Glos, also from the CSU, the sister party to Merkel's *CDU* ...
... Merkel's *CDU* had been able to rely on the CSU's strength in Bavaria ...
... but while her conservative *CDU* wanted new legal tools to do so, ...
... The new development has put a further strain on Merkel's *CDU* ...

6.2 Figure – Examples of *OSPC* for entities, showing five examples for a syntactic collocation (top row) and fie examples for a more specific proposition (bottom row). *CPI* might refer to `Comunist_Party_of_India` or `Consumer_Price_Index`, among others, but refers to the second in all cases. *CDU* can refer to the German `Christian_Democratic_Union` or `Catholic_Distance_University`, among others, but refers to the first in all cases.

In order to estimate *OSPC*, no available corpus existed, so we decided to base our dataset on the TAC-KBP 2009 Entity Linking dataset[3] (TAC09 for short) (Ji *et al.* 2010). The TAC09 dataset involves 138 mention strings, which have been annotated in several documents drawn primarily from Gigaword[4]. We extracted several syntactic collocations for those 138 mention strings from Gigaword, and hand-annotated them, yielding an estimate for the *OSPC*. Note that TAC09 only provides the annotation for a specific mention in a document, so we had to annotate by hand the rest of occurrences in the documents. For instance, we analyzed examples of *CPI* as subject of the verb *rise* (cf. Figure 6.2). Some of the syntactic collocations like the subjects of verb *has* seemed very uninformative, so we decided to also check the *OSPC* hypothesis on more specific collocations, involving more complete argument structures. For instance, we checked *ABC* occurring as subject of *has* with object *radio*. We call this more specific collocations *propositions* (Peñas and Hovy 2010).

[3] `http://www.nist.gov/tac/2013/KBP/EntityLinking/index.html`
[4] `http://catalog.ldc.upenn.edu/LDC2003T05`

The paper is structured as follows. We will first present the resources used in this study. Section 6.3 presents the results of *OSPD*. Section 6.3.1 extends *OSPD* when, instead of documents, we take the complete collection. Section 6.4 presents the study of *OSPC* both for syntactic dependencies and propositions. Section 6.5 presents the experiments where *OSPD* and *OSPC* are used to improve the performance of existing systems. Finally, we draw the conclusions and future work.

6.2 Resources used

AIDA is based on the corpus used in the CONLL named-entity recognition and classification task, where all entities in full documents had been linked to the referred Wikipedia articles (using the 2010 Wikipedia dump). We use the full AIDA dataset, with 1,393 documents, 34,140 disambiguated entity mentions, where 27,240 are linked to a Wikipedia article. All in all there are 6,877 distinct mention strings (types) which are linked at least once to a Wikipedia article. The rest refer to articles not in Wikipedia (NIL instances), and were discarded. This corpus covers news from a sample of a few days spanning from 1996-05-28 to 1996-12-07.

In order to prepare our dataset for *OSPC*, we chose the dataset of the TAC KBP 2009 Entity Linking competition, as this dataset have been extensively used in Entity Linking evaluation. In addition, the corpus used in the task was very large, allowing us to mine relevant collocations (see below). We manually annotated the occurrences in the extracted collocations, producing two datasets, one for each kind of collocation (cf. Section 6.4). Note that the TAC KBP organizers only annotated one specific mention in each target document. For completeness, we also tagged the rest of the occurrences of the target mentions in the documents, thus allowing us to provide *OSPD* estimated based on TAC09 data as well. This is the third dataset that we annotated by hand. The hand-annotation was performed by a single person, and later reviewed by the rest of the authors. The three annotation datasets are publicly available[5]. Hand-tagging is costly, so we tagged around 250 examples of syntactic collocations and around 250 examples of propositions.

Note that both AIDA and TAC09 contain mentions that were not linked to a Wikipedia article because the mention referred to an entity which was not listed in the entity inventory. We ignored all those cases (called NIL

[5] `http://ixa2.si.ehu.es/OEPDC`

NHasN	→	U.S. dollar
NPN	→	condition of anonymity
NVN	→	official tells AFP
NVNPN	→	article maintains interest within layout
NVPN	→	others steal from input
VNPN	→	includes link to website

6.1 Table – List of the six patterns used to extract propositions, with some examples.

cases), as we would need to investigate, for each NIL, which actual entity they refer to.

The collocations were extracted from the TAC KBP collection (Ji *et al.* 2010), comprising 1.7 million documents, 1.3 millions from newswire and 0.5 millions from the web. We have parsed them with the Stanford CoreNLP software (Klein and Manning 2003), obtaining around 650 million dependencies (De Marneffe and Manning 2008). We selected subject, object, prepositional complements and adjectival modifiers as the source for syntactic collocations. In order to provide more specific collocations, we implemented the syntactic patterns proposed in (Peñas and Hovy 2010), which produce so-called propositions. The result is a database with 16 million distinct propositions. Table 6.1 shows the six patterns used in this work, together with some examples.

In order to know whether a mention is ambiguous, we built a dictionary based on Wikipedia which lists, for each string mention, which entities it can refer to. We followed the construction method of (Spitkovsky and Chang 2012), which checked article titles, redirects, disambiguation pages and hyperlinks to find mention strings that can be used to refer to entities. Contrary to them, we could not access hyperlinks in the web, so we could use only those in Wikipedia. According to our dictionary, the ambiguity of the mentions that we are studying is very high, 26.4 entities on average for the mentions in AIDA, and 62.6 entities on average for the mentions in TAC09.

6.3 One entity per discourse

In order to estimate *OSPD* we divided the number of times a mention string referred to different entities in the document with the number of times a

	AIDA	**TAC09**
Mention-document pairs	4,265	334
Ambiguous pairs	170	6
OSPD	**96.0%**	**98.2%**

6.2 Table – One entity per discourse: per document statistics in AIDA and TAC09 datasets. Pairs stand for the number of unique mention-document pairs. The 4,265 pairs in AIDA correspond to 12,084 occurrences of mentions, and the 334 pairs in TAC09 correspond to 1,173 occurrences.

mention string occurred multiple times in the document. In the denominator and numerator we count each mention-document pair once.

Regarding AIDA, we found 12,084 occurrences of mentions which occurred more than once in a document, making 4,265 unique mention-document pairs[6] (cf. Table 6.2). In the vast majority of the cases those mentions refer to a single entity in the document, and only in 170 cases the mentions in the document refer to several entities. The last row in Table 6.2 shows the ratio between those values, 96.01%, showing that *OSPD* is strong in this dataset.

We also checked *OSPD* in the TAC09 dataset. Out of the 138 distinct mention strings used in the task, we discarded those only linked to NIL (that is, no corresponding Wikipedia article existed) and those which were not ambiguous (that is, they had only one entity in the dictionary, cf. Section 6.2). That leaves 105 mention strings, occurring 1,776 times in 918 different documents, which we annotated by hand. The 105 strings occurred 1,776 times in 918 documents. Removing the cases where the mention occurred only once, we were left with 1,173 occurrences, which make 334 unique mention-document pairs, of which only 6 occurred with more than one sense (rightmost row in Table 6.2). This yields an estimate for *OSPD* of 98.2%.

Finally, we also thought about measuring *OSPD* on the Wikipedia articles, where many mentions have been manually linked to their respective article. Unfortunately, we noted that Wikipedia guidelines explicitly prevent authors linking a mention multiple times: *Generally, a link should appear only once in an article, but if helpful for readers, links may be repeated in infoboxes, tables, image captions, footnotes, and at the first occurrence after*

[6]By unique mention-document pairs we mean that we only count once for a mention occurring multiple times in a document. For instance if mention *Smith* occurs 10 times in the whole corpus, 8 times in document A and 2 times in document B, we count two unique mention-document pairs.

	All mentions		First mention	
	AIDA	**TAC09**	**AIDA**	**TAC09**
Mention types	3,363	105	2,731	105
Ambiguous types	475	26	454	25
OSPD (collections)	**85.9%**	**75.2%**	**83.4%**	**76.2%**

6.3 Table – One entity per collection: statistics in AIDA and TAC09. In the first two columns ("All mentions") we consider all mention types (3,363 types in AIDA correspond to 23,726 occurrences of mentions, and 105 types in TAC09 correspond to 1,776 occurrences). In the second two columns ("First mention") we leave only the first mention of each document (in this case, there are 2,731 mention types in AIDA which correspond to 15,275 occurrences, and 105 types in TAC09 corresponding to 941 occurrences).

the lead[7]. The fact that Wikipedia editors did not explicitly state exceptions to the above rule (e.g. for cases where the word or phrase is used to refer to two different articles, thus breaking the *OSPD* hypothesis) is remarkable, and might indicate that Wikipedia editors had not felt the need to challenge the *OSPD* hypothesis.

6.3.1 One entity per collection

We took the opportunity to also explore "one entity per collection", which gives an idea of what is the spread of entities for whole document collections. In this case, there is no need to count mention-document pairs, as there is one single document, the collection, so we estimate the hypothesis according to mention types. The first two columns in table 6.3 shows that, overall, mentions which occurred more than once in the collection tend to refer to the same entity 85.9% of the time in AIDA, and 75.2% of the time in TAC09.

As we know that multiple mentions in a document tend to refer to one entity, the second two columns in table 6.3 offers the statistics when factoring out multiple occurrences of mention in a document, that is, leaving the first mention in each document. The statistics are very similar, with minor variations.

[7] `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking#What_generally_should_be_linked`

	Syn. coll.	Propositions
Mention-collocation pairs	58	61
Ambiguous pairs	5	1
OSPC	**91.4%**	**98.4%**

6.4 Table – One entity per collocation: statistics for syntactic collocations and propositions. The 58 mention-collocation pairs correspond to 262 occurrences, and the 61 mention-proposition pairs to 279.

We think that the lower estimate for TAC09 is an artifact of how the TAC KBP organizers set up the dataset, as they were explicitly looking for cases where the target string would refer to different entities, making the task more challenging for NED systems. This fact does not affect *OSPD* for documents, as those strings still tend to refer to a single entity per document, but given the need to find occurrences for different entities, the organizers (Ji *et al.* 2010) did focus on strings occurring with different entities across the document collection. This is in contrast with AIDA, where they tagged all named-entities occurring in the target documents. Had the organizers of TAC09 focused on a random choice of strings and documents, the one entity per collection would also hold to the high degree exhibited in AIDA, as the genre of most of the documents is also news (as in AIDA).

6.4 One entity per collocation

In order to estimate *OSPC* for **syntactic collocations**, we manually annotated several occurrences of the 138 mention strings of the TAC09 dataset. Hand-tagging mention entities is a costly process, so we chose (at random) one syntactic dependency relation for each of the 138 mention strings that occurred more than five times in the corpus. We then hand-tagged at random five occurrences of each collocation (cf. Figure 6.2). This method would provide a maximum of 5 examples for each of the 138 mentions, but after checking the minimum frequency of the collocations, the quality of the context, repeated sentences, mentions that are not ambiguous in the dictionary, and whether the mention could be attached to an entity in the database, the actual number was lower. All in all we found 58 mention-collocation pairs (262 occurrences) for syntactic collocations (cf. middle column in Table 6.4). Only 5 mentions referred to more than one entity per collocation, yielding

that *OSPC* for syntactic collocation is around 91.4%.

To gather the dataset for **propositions**, we followed the same method as for the syntactic collocations, that is, we chose (at random) one propositions involving one of the 138 mention strings that occurred more than five times in the corpus, and hand-tagged at random five occurrences of each proposition (cf. Figure 6.2. As with syntactic collocations, we also found a limited number of mentions filling the desired properties. That left 61 mention-collocation pairs (279 occurrences) for propositions (cf. right column in Table 6.4). Only 1 mention referred to more than one entity per proposition, yielding *OSPC* for propositions around 98.4%. This shows that the more specific the context is, the stronger is the link between mention and entity.

6.5 Improving performance

In order to check whether any of the "one sense" hypothesis above could improve the performance of a NED system, we followed a simple procedure: After running the NED system, we take its output and observe, for each mention string, which is the entity returned most often for a given document (or collocation), assigning to all occurrences the majority entity. In case of ties, we return the entity with the highest support from the NED system. We tested the improvements on three NED systems: the freely available DBpedia Spotlight, a reimplementation of a strong Bayesian NED system and a graph-based system.

DBpedia Spotlight is a freely available NED system (Daiber *et al.* 2013), based on a generative probabilistic model (Han and Sun 2011). Nowadays it is one of the most widely used NED systems and attains performances close to state-of-the-art (Daiber *et al.* 2013). We used the default values of the parameters for all the experiments in this paper.

We also tested an in-house reimplementation of the generative probabilistic model presented in (Han and Sun 2011) (represented as $p(e)p(s|e)p(c|e)$ formula hereafter). This is a state-of-the-art system which got the same accuracy as the best participant (72.0) when evaluated in the non-NIL subset of TAC13.

UKB is a freely-available system for performing Word Sense Disambiguation and Similarity based on random walks on graphs (Agirre *et al.* 2015) (PPR hereafter). Instead of using it on WordNet, we represented Wikipedia

Mention in context		**Entity**
Abbott Beefs Up Litigation ...	→	`Abbot_Kinney`
Abbott Laboratories Inc., bracing ...	→	`Abbott_Laboratories`
Abbott said it had restated ...	→	`Abbott_Laboratories`
between *Abbott* and Takeda ...	→	`Abbott_Laboratories`
Abbott said in January ...	→	`Abbott_Laboratories`

6.3 Figure – Applying *OSPD*: Each of the five occurrences of *Abbott* in the document in Figure 6.1 has been tagged independently by a NED systems, which return the correct entity in all but one case (precision 80%). Applying *OSPD* would return the correct entity `Abbott_Laboratories` in all cases, improving precision to 100%.

as a graph, where vertices are the wikipedia articles and edges represents bidirectional hyperlinks among Wikipedia pages, effectively implementing a NED system. We used a Wikipedia dump from 2013 in our experiments. PPR is a competitive, state-of-the-art system which attained a score of 69.0 when evaluated in the non-NIL subset of the TAC13 dataset.

The input of the systems is the context of each mention to be disambiguated, in the form of a 100 token window centered in the target mention. In NED, the identification of the correct mention to be disambiguated is part of the problem. AIDA does provide gold mentions, but TAC09 only provides a query string which might be just a substring of the real mention in the document. We treated both corpus in the same way. In the case of DBpedia Spotlight we use the built-in mention spotter. In the case of our in-house implementations, we use the longest string that matches a valid entity mention in the system, as given by the dictionary (cf. Section 6.3).

Some of the NED systems do not return an entity for all mentions, so we evaluate precision, recall and the harmonic mean (F1 measure). Statistical significance has been estimated using Wilcoxon. We reused the same corpora as in the previous sections for the evaluation, and also removed all NIL mentions (i.e. mentions which refer to an entity not in Wikipedia).

6.5.1 One entity per discourse

We report the improvements using *OSPD* for both **document** and **collection** levels. At the document level, we relabel mentions that occur multiple times in a document using the entity returned most times by the NED system

	AIDA			TAC09		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Spotlight	83.24	63.90	72.30	64.48	46.44	53.99
+ *OSPD* Discourse	**84.17**	70.01	76.44	**64.65**	**48.50**	**55.42**
+ *OSPD* Collection	84.02	**74.64**	**79.05**	56.24	47.98	51.78
PPR	70.09	69.03	69.55	67.70	67.64	67.67
+ *OSPD* Discourse	71.30	70.23	70.76	**70.21**	**70.21**	**70.21**
+ *OSPD* Collection	**75.79**	**74.64**	**75.21**	68.84	68.84	68.84
$\mathbf{p(e)p(s|e)p(c|e)}$	65.71	65.11	65.41	65.49	65.49	65.49
+ *OSPD* Discourse	67.77	67.37	67.57	66.27	66.27	66.27
+ *OSPD* Collection	**74.29**	**73.89**	**74.09**	**68.24**	**68.24**	**68.24**

6.5 Table – Applying *OSPD*: NED performance on AIDA and TAC09 *OSPD* datasets, including each of the three NED systems, and the results after applying *OSPD* at the document and collections levels. Bold marks best result for each system.

in that document. Figure 6.3 illustrates the idea for a NED system on the same sample document as in Figure 6.1. At the collection level, we relabel mentions using the entity returned most times by the NED systems in the whole collection.

Table 6.5 reports the results of the performance as evaluated on mentions occurring multiple times in the AIDA and TAC09 datasets. The numbers in the left part of the table correspond to the performance as evaluated on mentions occurring multiple times in AIDA documents. Note that the number of occurrences where *OSPD* at the collection level can be applied is larger (a superset of those for *OSPD* at the document level), as, for instance, a mention string occurring once in three different documents won't be affected by *OSPD* at the document level, but it could be relabeled at the collection level. We were especially interested in making the numbers between *OSPD* at the document and collection levels directly comparable, and therefore report the results on the same occurrences, that is, the occurrences where *OSPD* at the document level can be applied.

The results show a small but consistent improvement for *OSPD* at the document level in precision, recall and F1 for the three NED systems, around 1 or 2 absolute points. The improvements when applying *OSPD* at the collection level are also consistent, but remarkably larger, between 5 and 9 absolute points. All improvements are statistically significant (p-value below

***CPI* subject-of rise**	**Angela Merkel has *CDU*:**
Consumer_price_index	Christian_Democratic_Union_(Germany)
Consumer_price_index	Catholic_Distance_University
Communist_Party_of_India	Christian_Democratic_Union_(Germany)
Communist_Party_of_India	Christian_Democratic_Union_(Germany)
Consumer_price_index	Christian_Democratic_Union_(Germany)

6.4 Figure – Applying *OSPC*: A NED system system tagged each example in Figure 6.2 independently. For *CPI*, the precision is 60%, but after relabeling with *OSPC* it would be 100%. For *CDU*, the improvement is from 80% to 100%.

0.01).

Table 6.5 also reports the results after applying *OSPD* to TAC09 instances which occurred more than once in a document. Results for *OSPD* at document level and collection level follow the same methodology as for AIDA. The improvement at the collection level is not so consistent, with a loss in performance for Spotlight, a small improvement for PPR, and a larger improvement for $(p(e)p(s|e)p(c|e))$. All differences across the table are statistically significant (p-value below 0.01).

While the *OSPD* at the document level is strong in both corpora, Section 6.3.1 showed that the *OSPD* at the collection level is only strong in AIDA, with a much lower estimate in TAC09. This fact would explain why the improvement with *OSPD* at the collection level is not consistent. Following the rationale in Section 6.3.1, we think that had the organizers of the task chosen strings and documents at random, the improvement in TAC 2009 at the collection level would be also as high as in AIDA. The high improvement in AIDA at the collection level compared to the more modest improvement at the document level, despite having a lower *OSPD* estimate (cf. Section 6.3.1), could be caused by the fact that there are more occurrences and evidence in favor of the majority entity.

6.5.2 One entity per collocation

Figure 6.4 shows the application of *OSPC* to the output of a NED system to two sample collocations in our dataset. In this case, the application of *OSPC* would increase precision to 100%. The actual result on the datasets produced in Section 6.4 for syntactic collocations and propositions is reported on table

	Syntactic collocations			Propositions		
	prec.	**recall**	**F1**	**prec.**	**recall**	**F1**
Spotlight	82.46	66.41	73.57	74.67	60.22	66.67
+ *OSPC*	**82.63**	**67.18**	**74.11**	**74.79**	**62.72**	**68.23**
PPR	75.86	75.57	75.72	67.87	67.38	67.63
+ *OSPC*	**78.54**	**78.24**	**78.39**	**68.59**	**68.10**	**68.35**
p(e)p(s|e)p(c|e)	75.57	75.57	75.57	71.33	71.33	71.33
+ *OSPC*	**78.24**	**78.24**	**78.24**	**73.12**	**73.12**	**73.12**

6.6 Table – Applying *OSPC*: NED performance on TAC09, including each of the three NED systems, and the results after applying *OSPC* for syntactic collocations and propositions. Bold is used for best results for each system.

6.6.

Regarding syntactic collocations, table 6.6 shows that the improvement is small but consistent for the three systems on precision, recall and F1, ranging from 0.5 to 2.5 absolute points in F1 score. The results for propositions also show the same trend, with consistent improvements across the table. All differences in the two tables are statistically significant (p-value < 0.01), except for PPR.

6.6 Conclusions and future work

Our study shows that *OSPD* holds for 96%-98% (in the AIDA and TAC09 datasets, respectively) of the mentions that occur multiple times in documents. We also measured *OSPD* at the collection level (86% and 75%, respectively). *OSPC* holds for 91% of the mentions that occur multiple times in the syntactic collocations that we studied, and 98% of the mentions that occur multiple times in more specific collocations. We reused the publicly available AIDA dataset for estimating *OSPD*. In addition, we created a dataset to study *OSPC* based on the TAC KBP Entity Linking 2009 task dataset, which is publicly available[8].

We carefully chose to estimate both OPSD and *OSPC* on TAC09, in order to make the numbers between *OSPD* and *OSPC* comparable. The *OSPD* numbers for AIDA are very similar to those obtained on TAC09,

[8] `http://ixa2.si.ehu.es/OEPDC`

providing complementary evidence. Although the high estimate of *OSPD* for entities was somehow expected, the high estimate of *OSPC* for the syntactic collocations, especially the propositions, was somehow unexpected, given the high ambiguity rate of the discussed strings, and the fact that the ambiguity included similar entities, like for instance *ABC* which can refer, among other 190 entities, to the `American_Broadcasting_Company` or the `Australian_Broadcasting_Corporation`.

Our results also show that a simple application of the *OSPD* and *OSPC* hypotheses to the output of three different NED systems improves the results in all cases. Remarkably, the highest performance gain, 8 absolute points, was for *OSPD* at the collection level in the AIDA corpus.

The results presented here could be largely dependent on the domain and genre of the documents, as well as the definition of collocation. Our work is a strong basis for claiming that *OSPD* and *OSPC* hold for entities, but the evidence could be further extended exploring alternative operationalization of collocations and a larger breadth of genres and domains.

For the future we would like to check whether these hypotheses can be further used to improve current NED systems. The *OSPD* hypothesis can be used to jointly disambiguate all occurrences of a mention in a document. The *OSPC* hypothesis could be used to acquire important disambiguation features, or to perform large-scale joint entity linking. The *OSPD* for whole collections could be useful for documents on specific domains, and for domain adaptation scenarios.

7

EID sistemak kanpo-ezagutzaz elikatzen

Kapitulu honek artikulu bildumaren laugarren artikulua azalduko du. Helburua EID sistemek testuinguru urria duten izen-aipamenak desanbiguatzeko dituzten zailtasunak leuntzea da. Horretarako, eredu lokala kanpotiko corpusetatik eskuraturiko ezagutzaz elikatuko da. Sistemaren emaitzak hobetzea lortuko da, eta kanpo-ezagutzaren garrantzia ebaluatuko da. Jarraian, artikuluaren jatorrizko fitxa eta ingelesezko bertsioa:

Ander Barrena, Aitor Soroa and Eneko Agirre. ***Alleviating Poor Context with Background Knowledge for Named Entity Disambiguation.*** *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ACL 2016.* Berlin, Germany. 2016

Named Entity Disambiguation (NED) algorithms disambiguate mentions of named entities with respect to a knowledge-base, but sometimes the context might be poor or misleading. In this paper we introduce the acquisition of two kinds of background information to alleviate that problem: entity similarity and selectional preferences for syntactic positions. We show, using a generative Näive Bayes model for NED, that the additional sources of context are complementary, and improve results in the AIDA and TAC-KBP DEL 2014 datasets, yielding the third best and the best results, respectively. We provide examples and analysis which show the value of the acquired background information.

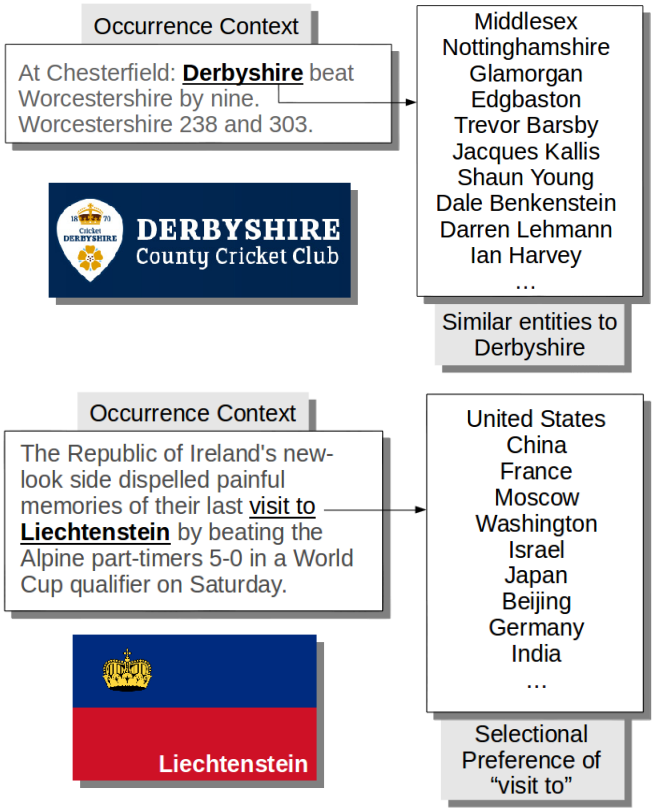
7.1 Introduction

The goal of Named Entity Disambiguation (NED) is to link each mention of named entities in a document to a knowledge-base of instances. The task is also known as Entity Linking or Entity Resolution (Bunescu and Pasca 2006; McNamee and Dang 2009; Hachey *et al.* 2012). NED is confounded by the ambiguity of named entity mentions. For instance, according to Wikipedia, *Liechtenstein* can refer to the micro-state, several towns, two castles or a national football team, among other instances. Another ambiguous entity is *Derbyshire* which can refer to a county in England or a cricket team. Most NED research use knowledge-bases derived or closely related to Wikipedia.

For a given mention in context, NED systems (Hachey *et al.* 2012; Lazic *et al.* 2015) typically rely on two models: (1) a mention module returns possible entities which can be referred to by the mention, ordered by prior probabilities; (2) a context model orders the entities according to the context of the mention, using features extracted from annotated training data. In addition, some systems check whether the entity is coherent with the rest of entities mentioned in the document, although (Lazic *et al.* 2015) shows that the coherence module is not required for top performance.

Figure 7.1 shows two real examples from the development dataset which contains text from News, where the clues in the context are too weak or misleading. In fact, two mentions in those examples (*Derbyshire* in the first and *Liechtenstein* in the second) are wrongly disambiguated by a bag-of-words context model.

In the first example, the context is very poor, and the system returns the *county* instead of the *cricket team*. In order to disambiguate it correctly one needs to be aware that *Derbyshire*, when occurring on News, is most notably associated with cricket. This background information can be acquired from large News corpora such as Reuters (Lewis *et al.* 2004), using distributional methods to construct a list of closely associated entities (Mikolov *et al.* 2013). Figure 7.1 shows entities which are distributionally similar to *Derbyshire*, ordered by similarity strength. Although the list might say nothing to someone not acquainted with cricket, all entities in the list are strongly related to cricket: Middlesex used to be a county in the UK that gives name to a cricket club, Nottinghamshire is a county hosting two powerful cricket and football teams, Edgbaston is a suburban area and a cricket ground, the most notable team to carry the name Glamorgan is Glamorgan County Cricket Club, Trevor Barsby is a cricketer, as are all other people in the distribu-



7.1 Figure – Two examples where NED systems fail, motivating our two background models: similar entities (top) and selectional preferences (bottom). The logos correspond to the gold label.

tional context. When using these similar entities as context, our system does return the correct entity for this mention.

In the second example, the words in the context lead the model to return the *football team* for *Liechtenstein*, instead of the *country*, without being aware that the nominal event "visit to" prefers locations arguments. This kind of background information, known as selections preferences, can be easily acquired from corpora (Erk 2007). Figure 7.1 shows the most frequent entities found as arguments of "visit to" in the Reuters corpus. When using these filler entities as context, the context model does return the correct entity for this mention.

In this article we explore the addition of two kinds of background infor-

mation induced from corpora to the usual context of occurrence: (1) given a mention we use distributionally similar entities as additional context; (2) given a mention and the syntactic dependencies in the context sentence, we use the selectional preferences of those syntactic dependencies as additional context. We test their contribution separately and combined, showing that they introduce complementary information.

Our contributions are the following: (1) we introduce novel background information to provide additional disambiguation context for NED; (2) we integrate this information in a Bayesian generative NED model; (3) we show that similar entities are useful when no textual context is present; (4) we show that selectional preferences are useful when limited context is present; (5) both kinds of background information help improve results of a NED system, yielding the state-of-the-art in the TAC-KBP DEL 2014 dataset and getting the third best results in the AIDA dataset; (6) we release both resources for free to facilitate reproducibility. [1]

The paper is structured as follows. We first introduce the method to acquire background information, followed by the NED system. Section 7.4 presents the evaluation datasets, Section 7.5 the development experiments and Section 7.6 the overall results. They are followed by related work, error analysis and the conclusions section.

7.2 Acquiring background information

We built our two background information resources from the Reuters corpus (Lewis *et al.* 2004), which comprises 250K documents. We chose this corpus because it is the one used to select the documents annotated in one of our gold standards (cf. Section 7.4). The documents in this corpus are tagged with categories, which we used to explore the influence of domains.

The documents were processed using a publicly available NLP pipeline, Ixa-pipes,[2] including tokenization, lematization, dependency tagging and NERC.

[1] `http://ixa2.si.ehu.es/anderbarrena/2016ACL_files.zip`
[2] `http://ixa2.si.ehu.es/ixa-pipes/`

7.2.1 Similar entity mentions

Distributional similarity is known to provide useful information regarding words that have similar co-occurrences. We used the popular word2vec[3] tool to produce vector representations for named entities in the Reuters corpus. In order to build a resource that yields similar entity mentions, we took all entity-mentions detected by the NERC tool and, if they were multi word entities, joined them into a single token replacing spaces with underscores, and appended a tag to each of them. We run word2vec with default parameters on the pre-processed corpus. We only keep the vectors for named entities, but note that the corpus contains both named entities and other words, as they are needed to properly model co-occurrences.

Given a named entity mention, we are thus able to retrieve the named entity mentions which are most similar in the distributional vector space. All in all, we built vectors for 95K named entity mentions. Figure 7.1 shows the ten most similar named entities for *Derbyshire* according to the vectors learned from the Reuters corpus. These similar mentions can be seen as a way to encode some notion of a topic-related most frequent sense prior.

7.2.2 Selectional Preferences

Selectional preferences model the intuition that arguments of predicates impose semantic constraints (or preferences) on the possible fillers for that argument position (Resnik 1996). In this work, we use the simplest model, where the selectional preference for an argument position is given by the frequency-weighted list of fillers (Erk 2007).

We extract dependency patterns as follows. After we parse Reuters with the Mate dependency parser (Bohnet 2010) integrated in IxaPipes, we extract $(H \xrightarrow{D} C)$ dependency triples, where D is one of the Subject, Object or Modifier dependencies[4] (SBJ, OBJ, MOD, respectively), H is the head word and C the dependent word. We extract fillers in both directions, that is, the set of fillers in the dependent position $\{C : (H \xrightarrow{D} C)\}$, but also the fillers in the head position $\{H : (H \xrightarrow{D} C)\}$. Each such configuration forms a template, $(H \xrightarrow{D} *)$ and $(* \xrightarrow{D} C)$.

[3] `https://code.google.com/archive/p/word2vec/`

[4] Labels are taken from the Penn Treebank `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`

In addition to triples (single dependency relations) we also extracted tuples involving two dependency relations in two flavors: $(H \xrightarrow{D_1} C_1 \xrightarrow{D_2} C_2)$ and $(C_1 \xleftarrow{D_1} H \xrightarrow{D_2} C_2)$. Templates and fillers are defined as done for single dependencies, but, in this case, we extract fillers in any of the three positions and we thus have three different templates for each flavor.

As dependency parsers work at the word level, we had to post-process the output to identify whether the word involved in the dependency was part of a named entity identified by the NERC algorithm. We only keep tuples which involve at least one name entity. Some examples for the three kinds of tuples follow, including the frequency of occurrence, with entities shown in bold:

(beat $\xrightarrow{SBJ}$ **Australia**) 141

(refugee $\xrightarrow{MOD}$ **Hutu**) 1681

(visit $\xrightarrow{MOD}$ to $\xrightarrow{MOD}$ **United States**) 257

(match $\xrightarrow{MOD}$ against $\xrightarrow{MOD}$ **Manchester United**) 12

(Spokesman $\xleftarrow{SBJ}$ tell $\xrightarrow{OBJ}$ **Reuters**) 1378

(**The Middle East** $\xleftarrow{MOD}$ process $\xrightarrow{MOD}$ peace) 1126

When disambiguating a mention of a named entity, we check whether the mention occurs on a known dependency template, and we extract the most frequent fillers of that dependency template. For instance, the bottom example in Figure 7.1 shows how *Liechtenstein* occurs as a filler of the template (visit $\xrightarrow{MOD}$ to $\xrightarrow{MOD}$ *), and we thus extract the selectional preference for this template, which includes, in the figure 7.1, the ten most frequent filler entities.

We extracted more than 4.3M unique tuples from Reuters, producing 2M templates and their respective fillers. The most frequent dependency was MOD, followed by SUBJ and OBJ [5] The selectional preferences include 400K different named entities as fillers.

Note that selectional preferences are different from dependency path features. Dependency path features refer to features in the immediate context of the entity mention, and are sometimes added as additional features of supervised classifiers. Selectional preferences are learnt collecting fillers in the same dependency path, but the fillers occur elsewhere in the corpus.

[5] 1.5M, 0.8M and 0.7M respectively

7.3 NED system

Our disambiguation system is a Näive Bayes model as initially introduced by (Han and Sun 2011), but adapted to integrate the background information extracted from the Reuters corpus. The model is trained using Wikipedia,[6] which is also used to generate the entity candidates for each mention.

Following usual practice, candidate generation is performed off-line by constructing an association between strings and Wikipedia articles, which we call dictionary. The association is performed using article titles, redirections, disambiguation pages, and textual anchors. Each association is scored with the number of times the string was used to refer to the article (Agirre *et al.* 2015). We also use Wikipedia to extract training mention contexts for all possible candidate entities. Mention contexts for an entity are built by collecting a window of 50 words surrounding any hyper link pointing to that entity.

Both training and test instances are pre-processed the same way: occurrence context is tokenized, multi-words occurring in the dictionary are collapsed as a single token (longest matches are preferred). All occurrences of the same target mention in a document are disambiguated collectively, as we merge all contexts of the multiple mentions into one, following the one-entity-per-discourse hypothesis (Barrena *et al.* 2014).

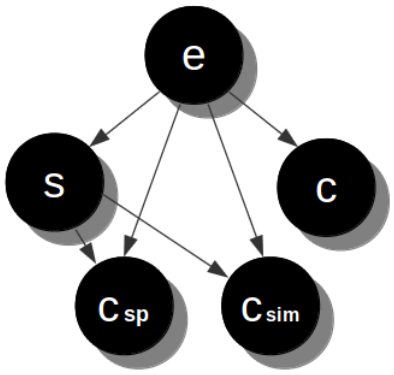
The Näive Bayes model is depicted in Figure 7.2. The candidate entity e of a given mention s, which occurs within a context c, is selected according to the following formula:

$$e = \arg\max_{e} P(s, c, c_{\text{sp}}, c_{\text{sim}}, e) =$$
$$\arg\max_{e} P(e)P(s|e)P(c|e)P(c_{\text{sp}}|e, s)P(c_{\text{sim}}|e, s)$$

The formula combines evidences taken from five different probabilities: the entity prior $p(e)$, the mention probability $p(s|e)$, the textual context $p(c|s)$, the selectional preferences $P(c_{\text{sp}}|e, s)$ and the distributional similarity $P(c_{\text{sim}}|e, s)$. This formula is also referred to as the "**Full model**", as we also report results of partial models which use different combinations of the five probability estimations.

Entity prior $P(e)$ represents the popularity of entity e, and is estimated as follows:

[6] We used a dump from 25-5-2011. This dump is close in time to annotations of the datasets used in the evaluation (c.f. Section 7.4)



7.2 Figure – Dependencies among variables in our Bayesian network.

$$P(e) = \frac{C(e) + 1}{|M| + N}$$

where $C(e)$ is the number of times the entity e is referenced within Wikipedia, $|M|$ is the total number of entity mentions and N is the number of distinct entities in Wikipedia. The estimation is smoothed using the *add-one* method.

Mention probability $P(s|e)$ represents the probability of generating the mention s given the entity e, and is estimated as follows:

$$P(s|e) = \theta \frac{C(e,s)}{C(e)} + (1-\theta)\frac{C(s)}{|M|}$$

where $C(e,s)$ is the number of times mention s is used to refer to entity e and $C(s)$ is the number of times mention s is used as anchor. We set the θ hyper-parameter to 0.9 according to developments experiments in the AIDA testa dataset (cf. Section 7.5.5).

Textual context $P(c|e)$ is the probability of entity e generating the context $c = \{w_1, \ldots, w_n\}$, and is expressed as:

$$P(c|e) = \prod_{w \in c} P(w|e)^{\frac{1}{n}}$$

where $\frac{1}{n}$ is a correcting factor that compensates the effect of larger contexts having smaller probabilities. $P(w|e)$, the probability of entity e generating word w, is estimated following a bag-of-words approach:

$$P(w|e) = \lambda \frac{C(e,w)}{\sum_w C(e,w)} + (1-\lambda)\frac{C(w)}{|M|}$$

where $C(e,w)$ is the number of times word w appears in the mention contexts of entity e, and $\sum_w C(e,w)$ is the total number of words in the mention contexts. The term in the right is a smoothing term, calculated as the likelihood of word w being used as an anchor in Wikipedia. λ is set to 0.9 according to development experiments done in AIDA testa.

Distributional Similarity $P(c_{\text{sim}}|e,s)$ is the probability of generating a set of similar entity mentions given an entity mention pair. This probability is calculated and estimated in exactly the same way as the textual context above, but replacing the mention context c with the mentions of the 30 most similar entities for s (cf. Section 7.2.1).

Selectional Preferences $P(c_{\text{sp}}|e,s)$ is the probability of generating a set of fillers c_{sp} given an entity and mention pair. The probability is again analogous to the previous ones, but using the filler entities of the selectional preferences of s instead of the context c (cf. Section 7.2.2). In our experiments, we select the 30 most frequent fillers for each selectional preferences, concatenating the filler list when more than one selectional preference is applied.

7.3.1 Ensemble model

In addition to the Full model, we created an ensemble system that combines the probabilities described above using a weighting schema, which we call "**Full weighted model**". In particular, we add an exponent coefficient to the probabilities, thus allowing to control the contribution of each model.

$$\arg\max_e P(e)^\alpha P(s|e)^\beta$$
$$P(c|e)^\gamma P(c_{\text{sp}}|e,s)^\delta P(c_{\text{sim}}|e,s)^\omega$$

We performed an exhaustive grid search in the interval $(0,1)$ for each of the weights, using a step size of 0.05, and discarding the combinations whose sum is not one. Evaluation of each combination was performed in the AIDA testa development set, and the best combination was applied in the test sets.[7]

[7]The best combination was $\alpha = 0.05$, $\beta = 0.1$, $\gamma = 0.55$ $\delta = 0.15$, $\omega = 0.15$

Dataset	Documents	Mentions
AIDA testa	216	4791
AIDA testb	231	4485
TAC14 DEL test	138	2817

7.1 Table – Document and linkable mention counts for AIDA and TAC14 DEL datasets.

7.4 Evaluation Datasets

The evaluation has been performed on one of the most popular datasets, the CoNLL 2003 named-entity disambiguation dataset, also know as the AIDA or CoNLL-Yago dataset (Hoffart *et al.* 2011). It is composed of 1393 news documents from Reuters Corpora where named entity mentions have been manually identified. It is divided in three main parts: *train*, *testa* and *testb*. We used *testa* for development experiments, and *testb* for the final results and comparison with the state-of-the-art. We ignored the training part.

In addition, we also report results in the Text Analysis Conference 2014 Diagnostic Entity Linking task dataset (TAC DEL 2014).[8] The gold standard for this task is very similar to the AIDA dataset, where target named entity mentions have been detected by hand. Through the beginning of the task (2009 to 2013) the TAC datasets were query-driven, that is, the input included a document and a challenging and sometimes partial target-mention to disambiguate. As this task also involved mention detection and our techniques are sensitive to mention detection errors, we preferred to factor out that variation and focus on the 2014.

The evaluation measure used in this paper is micro-accuracy, that is, the percentage of linkable mentions that the system disambiguates correctly, as widely used in the AIDA dataset. Note that TAC14 EDL included several evaluation measures, including the aforementioned micro-accuracy of linkable mentions, but the official evaluation measure was Bcubed+ F1 score, involving also detection and clustering of mentions which refer to entities not in the target knowledge base. We decided to use the same evaluation measure for both datasets, for easier comparison. Table 7.1 summarizes the statistics of the datasets used in this paper where document and mention counts are presented.

[8] `http://www.nist.gov/tac/2014/KBP/`

Method	m-acc
$P(e)P(s\|e)$	63.83
$P(e)P(s\|e)P(c_{\text{sim}}\|e,s)$	70.98

7.2 Table – Results on mentions with no context on the sports subset of testa, limited to 85% of the mentions (cf. Section 7.5.1).

7.5 Development experiments

We started to check the contribution of the acquired background information in the *testa* section of the AIDA dataset. In fact, we decided to focus first on a subset of *testa* about sports,[9] and also acquired background information from the sports sub-collection of the Reuters corpus.[10] The rationale was that we wanted to start in a controlled setting, and having assumed that the domain of the test documents and the source of the background information could play a role, we decided to start focusing on the sports domain first. Another motivation is that we noticed that the ambiguity between locations and sport clubs (e.g. football, cricket, rugby, etc.) is challenging, as shown in Figure 7.1.

7.5.1 Entity similarity with no context

In our first controlled experiment, we wanted to test whether the entity similarity resource provided any added value for the cases where the target mentions had to be disambiguated out of context. Our hypothesis was that the background information from the unannotated Reuters collection, entity similarity in this case, should provide improved performance. We thus simulated a corpus where mentions have no context, extracting the named entity mentions in the sports subset that had an entry in the entity similarity resource (cf. Section 7.2.1), totaling 85% of the 3319 mentions.

Table 7.2 shows that the entity similarity resource improves the results of the model combining the entity prior and mention probability, similar to the so-called most frequent sense baseline (MFS). Note that the combination of both entity prior and mention probability is a hard-to-beat baseline, as we will see in Section 7.6. This experiment confirms that entity similarity information is useful when no context is present.

[9]Including 102 out of the 216 documents in *testa*, totaling 3319 mentions.
[10]Including approx. 35K documents out of the 250K documents in Reuters

Method	**m-acc**
$P(e)P(s\|e)$	63.66
$P(e)P(s\|e)P(c\|e)$	66.18
$P(e)P(s\|e)P(c_{\text{sp}}\|e,s)$	67.33
$P(e)P(s\|e)P(c\|e)P(c_{\text{sp}}\|e,s)$	68.78

7.3 Table – Results on mentions with access to limited context on the sports subset of testa, limited to the 45% of mentions (cf. Section 7.5.2).

Method	**m-acc**
$P(e)P(s\|e)P(c\|e)$	69.54
$P(e)P(s\|e)P(c\|e)P(c_{\text{sp}}\|e,s)$	71.25
$P(e)P(s\|e)P(c\|e)P(c_{\text{sim}}\|e,s)$	72.64
Full	73.94

7.4 Table – Results on mentions with limited context on the sports subset of testa, limited to the 41% of the mentions (cf. Section 7.5.3)

7.5.2 Selectional preferences with short context

In our second controlled experiment, we wanted to test whether the selectional preferences provided any added value for the cases where the target mentions had limited context, that of the dependency template. Our hypothesis was that the background information from the unannotated Reuters collection, selectional preferences in this case, should provide improved performance with respect to the baseline generative model of context. We thus simulated a corpus where mentions have only short context, exactly the same as the dependency templates which apply to the example, constructed extracting the named entity mentions in the sports subset that contained matching templates in the selectional preference resource (cf. Section 7.2.2), totaling 45% of the 3319 mentions.

Table 7.3 shows that the selectional preference resource (third row) allows to improve the results with respect to the no-context baseline (first row) and, more importantly, with respect to the baseline generative model (second row). The last row shows that the context model and the selectional preference model are complementary, as they produce the best result in the table. This experiment confirms that selectional preference information is effective when limited context is present.

Models	Sport	Reuters
$P(e)P(s\|e)$	65.52	65.52
$P(e)P(s\|e)P(c\|e)$	72.81	72.81
$P(e)P(s\|e)P(c\|e)P(c_{\text{sp}}\|e,s)$	73.56	73.06
$P(e)P(s\|e)P(c\|e)P(c_{\text{sim}}\|e,s)$	75.73	76.62
Full	76.30	76.87

7.5 Table – Results on the entire sports subset of testa: middle column uses the sports subset of Reuters to acquire background information, right column uses the full Reuters (cf. Section 7.5.4).

7.5.3 Combinations

In our third controlled experiments, we combine all three context and background models and evaluate them in the subset of the sports mentions that have entries in the similarity resource, and also contain matching templates in the selectional preference resource (41% of the sports subset). Note that, in this case, the context model has access to the entire context. Table 7.4 shows that, effectively, the background information adds up, with best results for the full combined model (cf. Section 7.3), confirming that both sources of background information are complementary to the baseline context model and between themselves.

7.5.4 Sports subsection of AIDA testa

The previous experiments have been run on a controlled setting, limited to the subset where our constructed resources could be applied. In this section we report results for the entire sports subset of AIDA testa. The middle column in Table 7.5 shows the results for the two baselines, and the improvements when adding the two background models, separately, and in combination. The results show that the improvements reported in the controlled experiments carry over when evaluating to all mentions in the Sport subsection, with an accumulated improvement of 3.5 absolute points over the standard NED system (second row).

The experiments so far have tried to factor out domain variation, and thus the results have been produced using the background information acquired from the sports subset of the Reuters collection. In order to check whether this control of the target domain is necessary, reproduced the same

System	testa
$P(e)P(s\|e)$	73.76
$P(e)P(s\|e)P(c\|e)$	78.98
$P(e)P(s\|e)P(c\|e)P(c_{\text{sp}}\|e,s)$	79.32
$P(e)P(s\|e)P(c\|e)P(c_{\text{sim}}\|e,s)$	81.76
Full	81.90
$P(e)^{\alpha}P(s\|e)^{\beta}P(c\|e)^{\gamma}$	85.20
Full weighted	**86.62**

7.6 Table – Results on the full testa dataset (cf. Section 7.5.5).

experiment using the full Reuters collection to build the background information, as reported in the rightmost column in Table 7.5. The results are very similar,[11] with a small decrease for selectional preferences, a small increase for the similarity resource, and a small increase for the full system. In view of these results, we decided to use the full Reuters collection to acquire the background knowledge for the rest of the experiments, and did not perform further domain-related experiments.

7.5.5 Results on AIDA testa

Finally, Table 7.6 reports the results on the full development dataset. The results show that the good results in the sports subsection carry over to the full dataset. The table reports results for the baseline systems (two top rows) and the addition of the background models, including the Full model, which yields the best results.

In addition, the two rows in the bottom report the results of the ensemble methods (cf. Section 7.3.1) which learn the weights on the same development dataset. These results are reported for completeness, as they are an overestimation, and are over-fit. Note that all hyper-parameters have been tuned on this development dataset, including the ensemble weights, smoothing parameters λ and θ (cf. Section 7.3), as well as the number of similar entities and the number of fillers in the selectional preferences. The next section will show that the good results are confirmed in unseen test datasets.

[11] The two first rows do not use background information, and are thus the same.

System	AIDA	TAC14
$P(e)P(s\|e)$	73.07	78.31
$P(e)P(s\|e)P(c\|e)$	79.98	82.11
$P(e)P(s\|e)P(c\|e)P(c_{\text{sp}}\|e,s)$	81.31	82.61
$P(e)P(s\|e)P(c\|e)P(c_{\text{sim}}\|e,s)$	82.72	83.24
Full	82.85	83.21
$P(e)^{\alpha}P(s\|e)^{\beta}P(c\|e)^{\gamma}$	86.44	81.61
Full weighted	**88.32**	**83.46**

7.7 Table – Overall micro accuracy results on the AIDA testb and TAC 2014 DEL datasets.

7.6 Overall Results

In the previous sections we have seen that the background information is effective improving the results on development. In this section we report the result of our model in the popular AIDA testb and TAC14 DEL datasets, which allow to compare to the state-of-the-art in NED.

Table 7.7 reports our results, confirming that both background information resources improve the results over the standard NED generative system, separately, and in combination, for both datasets (Full row). All differences with respect to the standard generative system are statistically significant according to the Wilcoxon test (p-value < 0.05).

In addition, we checked the contribution of learning the ensemble weights on the development dataset (testa). Both the generative system with and without background information improve considerably.

The error reduction between the weighted model using background information (Full weighted row) and the generative system without background information (previous row) exceeds 10% in both datasets, providing very strong results, and confirming that the improvement due to background information is consistent across both datasets, even when applied on a very strong system. The difference is statistically significant in both datasets.

7.7 Related Work

Our generative model is based on (Han and Sun 2011), which is basically the core method used in later work (Barrena *et al.* 2015; Lazic *et al.* 2015)

System	AIDA	TAC14
Full weighted	88.32	83.46
Barrena *et al.* 2015	83.61	80.69
Lazic *et al.* 2015	86.40	—
(Alhelbawy & Gaizauskas,14)	*87.60	—
Chisholm and Hachey 2015	88.70	—
Pershina *et al.* 2015	*91.77	—
TAC14 best Ji *et al.* 2014	—	82.70

7.8 Table – Overall micro accuracy results on the AIDA testb and TAC 2014 DEL datasets, including the current state-of-the-art. Starred results are not comparable, see text.

with good results. Although the first do not report results on our datasets the other two do. (Barrena *et al.* 2015) combines the generative model with a graph-based system yielding strong results in both datasets. (Lazic *et al.* 2015) adds a parameter estimation method which improved the results using unannotated data. Our work is complementary to those, as we could also introduce additional disambiguation probabilities (Barrena *et al.* 2015), or apply more sophisticated parameter estimation methods (Lazic *et al.* 2015).

Table 7.8 includes other high performing or well-known systems, which usually use complex methods to combine features coming from different sources, where our results are only second to those of (Chisholm and Hachey 2015) in the AIDA dataset and best in TAC 2014 DEL. The goal of this paper is not to provide the best performing system, but yet, the results show that our use of background information allows to obtain very good results.

(Alhelbawy and Gaizauskas 2014) combines local and coherence features by means of a graph ranking scheme, obtaining very good results on the AIDA dataset. They evaluate on the full dataset, i.e. they test on train, testa and testb (20K, 4.8K and 4.4K mentions respectively). Our results on the same dataset are 84.25 (Full) and 88.07 (Full weighted), but note that we do tune the parameters on testa, so this might be slighly over-estimated. Our system does not use global coherence, and therefore their method is complementary to our NED system. In principle, our proposal for enriching context should improve the results of their system.

(Pershina *et al.* 2015) propose a system closely resembling (Alhelbawy and Gaizauskas 2014). They report the best known results on CONNL 2003 so far, but unfortunately, their results are not directly comparable to the rest

of the state-of-the-art, as they artificially insert the gold standard entity in the candidate list.[12]

In (Chisholm and Hachey 2015) the authors explore the use of links gathered from the web as an additional source of information for NED. They present a complex two-staged supervised system that incorporates global coherence features, with large amount of noisy training. Again, using additional training data seems an interesting future direction complementary to ours.

We are not aware of other works which try to use additional sources of context or background information as we do. (Cheng and Roth 2013) use relational information from Wikipedia to add constraints to the coherence model, and is somehow reminiscent of our use dependency templates, although they focus on recognizing a fixed set of relations between entities (as in information extraction) and do not model selectional preferences. (Barrena *et al.* 2014) explored the use of syntactic collocations to ensure coherence, but did not model any selectional preferences.

Previous work on word sense disambiguation using selectional preference includes (McCarthy and Carroll 2003) among others, but they report low results. (Brown *et al.* 2011) applied wordNet hypernyms for disambiguating verbs, but they did not test the improvement of this feature. (Taghipour and Ng 2015) use embeddings as features which are fed into a supervised classifier, but our method is different, as we use embeddings to find similar words to be fed as additional context. None of the state-of-the-art systems, e.g. (Zhong and Ng 2010), uses any model of selectional preferences.

7.8 Discussion

We performed an analysis of the cases where our background models worsened the disambiguation performance. Both distributional similarity and selectional preferences rely on correct mention detection in the background corpus. We detected that mentions where missed, which caused some coverage issues. In addition, the small size of the background corpus sometimes produces arbitrary contexts. For instance, subject position fillers of *score* include mostly basketball players like `Michael_Jordan` or `Karl_Malone`. A similar issue was detected in the distributional similarity resource. A larger corpus would produce a broader range of entities, and thus use of larger background corpora (e.g. Gigaword) should alleviate those issues.

[12] `https://github.com/masha-p/PPRforNED/readme.txt`

Another issue was that some dependencies do not provide any focused context, as for instance arguments of *say* or *tell*. We think that a more sophisticated combination model should be able to detect which selectional preferences and similarity lists provide a focused set of instances.

7.9 Conclusions and Future Work

In this article we introduced two novel kinds of background information induced from corpora to the usual context of occurrence in NED: (1) given a mention we used distributionally similar entities as additional context; (2) given a mention and the syntactic dependencies in the context sentence, we used the selectional preferences of those syntactic dependencies as additional context. We showed that similar entities are specially useful when no textual context is present, and that selectional preferences are useful when limited context is present.

We integrated them in a Bayesian generative NED model which provides very strong results. In fact, when integrating all knowledge resources we yield the state-of-the-art in the TAC KBP DEL 2014 dataset and get the third best results in the AIDA dataset. Both resources are freely available for reproducibility.[13]

The analysis of the acquired information and the error analysis show several avenues for future work. First larger corpora should allow to increase the applicability of the similarity resource, and specially, that of the dependency templates, and also provide better quality resources.

[13] `http://ixa2.si.ehu.es/anderbarrena/2016ACL_files.zip`

8

Ondorioak eta etorkizuneko ildoak

EID atazaren helburu nagusia testuetako izen-aipamenak ezagutza-basean dagokion entitatearekin lotzea da. Ataza honek gaur egungo interneteko bilatzaileen artean garrantzi handia dauka, eta bilaketa semantikoen zein ezagutza-baseen erabilera egokirako ezinbesteko urratsa da.

Tesiaren helburu eta motibazio nagusia **EID sistemen azterketa eta metodo berrien proposamena da**. Horretarako, arloaren egoera bideratzen duten bi korronte nagusien azterketa egin da. Batetik, algoritmo globalak erabiliz, eta bestetik, algoritmo lokalak aplikatuz. Gainera, bi ereduak modu egonkor eta eraginkorrean konbinatzeko metodoak aztertu dira. Jarraian, motibazio honi lotutako ondorio eta ekarpenak azalduko dira. Gai hauek 4. eta 5. kapituluetan jorratu dira.

- **Algoritmo globalak, ausazko ibilbideak Wikipedia grafoan.**

  Eneko Agirre, Ander Barrena and Aitor Soroa. ***Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation.*** *arXiv.org CoRR 2015.*

  Algoritmo globalek testuetako izen-aipamenen entitate-hautagaien artean, ezagutza-base batekiko koherenteenak direnak aukeratzen dituzte. Horretarako, HAD atazan arrakastaz aplikatu den ausazko ibilbideetan oinarritutako sistema egokitu da. Bide batez, EID atazan ausazko ibilbideak aplikatzeko Wikipediako hiperesteken egitura optimoa zein den aztertu da.

Hiperesteken azterketatik artikuluen arteko esteka zuzenak erabili ordez, grafo osoan daudenak erabiltzea hobesten da. Gainera, grafoa sortzean adabegiak elkarrenganako estekak dituztenean soilik lotuz, desanbiguazioaren emaitza hobea dela frogatu da. Azkenik, info-tauletatik eta kategoria egituratik datozen estekak baztertzea gomendatzen da.

Grafoa eraikitzeko argibideak jarraituz, eta ausazko ibilbideak hainbat datu-multzotan aplikatuz, arloaren egoerako emaitzak gainditu edo pareko balioak lortzen dira (ikus 4.9 taula).

4. kapituluko ekarpen nagusia: EID atazan ausazko ibilbideak aplikatzeko Wikipediaren grafo egitura berritzailea da. Horretaz gain, sistema eta Wikipediatik erauzitako baliabideak (hiztegia eta grafoa) libre eta eskuragarri daude.[1]

- **Algoritmo globalak eta lokalak konbinatzen.**

> Ander Barrena, Aitor Soroa and Eneko Agirre. ***Combining Mention Context and Hyperlinks from Wikipedia for Named Entity Disambiguation.*** *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics *SEM 2015.* Denver, Colorado, USA. 2015.

Algoritmo lokalak izen-aipamena inguratzen duten hitzetan oinarritzen dira desanbiguazioa burutzeko. Horretarako, izen-aipamenaren entitate-hautagaien testuinguruko hitzekin konparatzen dituzte. Tesi honetan, konparaketa hau egiteko hitz multzoak erabili dira. Eredu Bayesiar sortzaileetan oinarritutako eredu lokala, eta 4. kapituluan garatu den eredu globala konbinatu dira. Bi ereduak modu osagarrian konbinatzen dira. Izan ere, bakoitzak bere aldetik lortzen dituen emaitzak konbinatzean, aurrekoak hobetzen dira kasu guztietan, batean izan ezik (ikus 5.2 taula). Sistema zortzi datu-multzotan ebaluatu da, eta zortzitik bostetan arloaren egoerako emaitza onenak eskuratu dira.

Halaber, eredu lokalak eta globalak sistemari egiten dion ekarpena pisuen bidez zehaztu daiteke (ikus 5.5 atala). Garapeneko AIDA-testa datu-multzoan pisu egokienak kalkulatu, eta AIDA-testb datu-multzoan aplikatuz arloaren egoerako balioak lortzen dira (ikus 5.1 taula).

[1] `http://ixa2.si.ehu.es/ukb/`

5. kapituluko ekarpen nagusia: Batetik, EID atazan algoritmo lokalak aplikatzeko hitz multzoen erabilera. Bestetik, eredu sortzaileen bidezko hitz multzoen eta ausazko ibilbideen konbinaketa berritzailea.

Tesiko bigarren helburu eta motibazioa **EID atazarako teknika eta ezaugarri gehigarriak ikertzea da**. EID sistemek dituzten gabeziak gainditzeko, eta orokorrean edozein sistemaren emaitzak hobetzeko, bi ikerketa lerro aztertu dira. Hasteko, HAD atazan arrakastaz aplikatu den hipotesi batean oinarrituta, eredu globalaren, eredu lokalaren eta *Spotlight* izeneko sistemaren emaitzak hobetu dira. Ondoren, eredu lokala oinarritzat hartuta kanpotiko corpusetatik eskuratutako informazioaz emaitzak hobetzea lortu da. Proposatu diren hobekuntzak sistemetan aldaketarik egin gabe aplikatu daitezke, beraz, beste edozein sistemak barneratu ditzakeen propietateak dira. Jarraian, motibazio hauei loturiko ondorioak azalduko dira. Gai hauek 6. eta 7. kapituluetan jorratu dira.

- **Entitate bakarra diskurtsoan eta agerkidetzan**

Ander Barrena, Eneko Agirre, Bernardo Cabaleiro, Anselmo Peñas and Aitor Soroa. ***"One Entity per Discourse" and "One Entity per Collocation" Improve Named-Entity Disambiguation.*** *Proceedings of the 25th International Conference on Computational Linguistics COLING 2014.* Dublin, Ireland. 2014.

Ikerketa lerro honetan HAD atazan adierekin betetzen den "Adiera bakarra diskurtsoan eta agerkidetzan" hipotesia entitateekin betetzen dela frogatu da. Esate baterako, aipamen berdina behin baino gehiagotan azaldu baldin bada dokumentu berean "Entitate bakarra diskurtsoan" %96-98an betetzen da (AIDA eta TAC09 azpimultzoetan). Gainera, aipamena behin baino gehiagotan azaldu bada dokumentuen bilduman, "Entitate bakarra bilduman" %86-75 artean betetzen da. Bestalde, agerkidetza sintaktiko berdinean aipamena behin baino gehiagotan azaltzen bada, "Entitate bakarra agerkidetzan" %91n beteko da. Azkenik, agerkidetza konplexuagoak edo proposizioak aztertuz hipotesia %98an betetzen da.

Azterketaren ondoren, eredu globalean, lokalean eta *Spotlight* izeneko EID sisteman hipotesia aplikatu da. Kasu guztietan hobekuntzak egon

dira (ikus 6.5 eta 6.6 taulak), eta kasurik onenean F-neurria 8 puntutan hobetzea lortu da.

6. kapituluko ekarpen nagusia: edozein EID sistemak barneratu dezakeen "Entitate bakarra" propietatea hein handi batean betetzen dela. Ikerketa hau aurrera eramateko erabili diren baliabideak eskuragarri daude, beste edozein sistema hipotesi hauetaz aberastu daitekeela frogatu nahi bada.[2]

- **EID sistemak kanpo-ezagutzaz elikatzen**

> Ander Barrena, Aitor Soroa and Eneko Agirre. ***Alleviating Poor Context with Background Knowledge for Named Entity Disambiguation.*** *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics ACL 2016.* Berlin, Germany. 2016

EID sistemek arazoak izaten dituzte testuinguru urria duten aipamenak desanbiguatzeko orduan. 7. kapituluan arazo hau leuntzeko eta desanbiguazioan laguntzeko testuinguru gehigarria eskuratzea proposatzen da. Horretarako, etiketatu gabeko kanpotiko corpusetatik testuinguru gehigarriak eskuratzen dira bi aldaera ezberdinetan:

1. Alde batetik, desanbiguatu nahi den aipamenaren antzerakoak diren, edo espazio-bektorialean gertu dauden izen-aipamenekin sortzen dena (aipamen antzekoenak aurrerantzean).
2. Bestetik, izen-aipamenaren dependentzia-sintaktikoaren hautapen-murriztapenekin sortzen dena (hautapen-murriztapenak aurrerantzean).

Bi testuinguru hauek desanbiguazioan egin dezaketen ekarpena neurtzeko eredu Bayesiar lokala erabili da. Horretarako, ereduak testuinguru berrien sorkuntza ere zenbatesten du. Eragiketa sinple honek hiru informazio iturriak osagarriak direla erakusten du. Hori gutxi balitz, TAC14 datu-multzoan arloaren egoerako emaitzak lortzen ditu, eta AIDA datu-multzoan hirugarren emaitzarik onenak.

[2] `http://ixa2.si.ehu.es/OEPDC`

Horretaz gain, testuinguru urria duten izen-aipamenen kasuetan azterketa sakonagoa egin da. Honekin, testuinguru berritzaileen ekarpena frogatu nahi da. Horretarako, kontrolpeko esperimentuak egin dira garapeneko azpimultzo ezberdinetan eta egoera ezberdinak planteatuz:

- Aipamen antzekoenen testuinguruaren ekarpena ebaluatzeko, izen-aipamenaren testuinguru normala kontuan izan gabe ebaluatu da. Baliabide honekin bakarrik 7 puntuko hobekuntza lor daitekeela erakusten da (ikus 7.2. taula).
- Hautapen-murriztapenen testuingurua ebaluatzeko, testuinguru urria duten izen-aipamenen azpimultzoa sortu da. Hautapen-murriztapenekin testuinguru urriarekin ebaluatuz baino emaitza hobeak lortzen dira (ikus 7.3. taula). Hautapen-murriztapenen testuingurua eta izen-aipamenaren testuinguru urria konbinatzean osagarriak direla ikusten da, konbinaketak oinarri-lerroa 5 puntutan hobetzen baitu.
- Azkenik baliabide guztien ekarpenak ebaluatu dira. Testuinguru guztiak osagarriak direla ikusi da, izan ere, konbinaketa bakoitzak aurrekoaren emaitzak hobetzen ditu (ikus 7.4. taula).

7. kapituluko ekarpen nagusia: kanpotiko corpusetatik eskuratu den ezagutza da. Bi aldaera berritzaileetan planteatu da: aipamen antzekoenen eta hautapen-murriztapenen testuinguruetan errepresentatua. Edozein EID sistemarentzat aipamen antzekoenen eta hautapen-murriztapenen testuinguruak sortu eta aplikatzeko baliabide guztiak eskuragarri daude.[3]

Etorkizuneko ildoei dagokienean, EID atazako arloaren egoerak etengabe eguneratzen dihardu. Argitalpen kopuruak gora egiten duen heinean, emaitzek ere gora egiten dute. Honek, gero eta konpetentzia maila altuagoak ezartzen ditu, horren adibide AIDA datu-multzoan arloaren egoeraren bilakaera da. Kapituluaren arabera %82tik %85era pasa da, eta azkenik %92ra igo da (4. 5. eta 7. kapituluak hurrenez hurren). Jarraian, lehiakortasun honek bultzatuta etorkizunerako ideia nagusiak azalduko dira.

Algoritmo globalen inguruan, 4. kapituluan Wikipediako hiperesteken azterketak esteka zuzenen aurrean grafo osoaren erabilera hobesten du. Hala

[3] `http://ixa2.si.ehu.es/anderbarrena/2016ACL_files.zip`

ere, beranduago argitaratu diren artikuluetan (Alhelbawy and Gaizauskas 2014, Pershina *et al.* 2015) azpigrafo edo esteka zuzenak erabiliz kontrakoa erakutsi dute. Artikuluetako autoreekin hitz egin ondoren, ebaluatzerako garaian hainbat ezberdintasun daudela ikusi da. Beraz, artikuluetako emaitzak ez dira arloaren egoerarekin konparagarriak. Hala ere, etorkizunerako, (Alhelbawy and Gaizauskas 2014) eta (Pershina *et al.* 2015) artikuluetako algoritmoak ber-inplementatzea pentsatu da. Bide batez, ebaluazio irizpide berdinak erabiliz esteka zuzenak eta grafo osoa erabiltzearen arteko eztabaida argituz.

Algoritmo lokalei buruz, testuinguruaren modelatzea hitz multzoetatik haratago hedatu nahi da. Adibidez, (Lazic *et al.* 2015) artikuluan aurkezten den bidea jarraituz. Horretarako, bi ideia nagusi proposatzen dira:

- Batetik, testuinguruen modelatzean hitzen errepresentazio bektoriala erabiltzea (Mikolov *et al.* 2013). Demagun, hitz bakoitza dimentsio jakin bateko bektorean errepresentatzen dela. Testuinguruko hitzen bektoreen batez-bestekoak edo baturak, testuingurua dimentsio horretan errepresentzeko aukera eskaintzen du. Errepresentazio honek eredu Bayesiarretatik haratago ikasketa teknika eraginkorragoak erabiltzea ahalbidetzen du. Horien artean, gaur egun hainbeste arrakasta duten sare neuronalak.

- Bestetik, hitz multzoen ikuspegi berritzailea proposatu nahi da. Horretarako, hitz multzoak kategoria multzoetara hedatzeko lehen urratsak eman dira. Hiztegia erabiliz testuinguruko hitz bakoitza entitate multzotara hedatzen da (hitz bakoitzari hiztegian lotuak dauden entitate guztiak esleituz). Ondoren, Wikipediako kategoria egituran oinarrituta entitateak kategoria multzotara hedatzen dira. Esate baterako, *page* hitza kategoria hauetara hedatuko litzateke:

  - *Living_ people*
  - *World_ Wide_ Web*
  - *Handley_ Page_ aircraft*
  - ...

  7. kapituluko eredu Bayesiarrean testuinguru ezberdinak konbinatu diren gisa, hitz multzoak eta kategoria multzoak konbinatu nahi dira. Lehenengo esperimentu batzuek informazio osagarriak direla erakutsi dute.

Ildo orokorrei dagokienean tesi-lan honetan proposatu diren EID sistemetatik, ausazko ibilbideetan oinarritzen dena libre eta edozeinentzat eskuragarri dagoen bakarra da. Sistemaz gain, Wikipediako grafoak eta hiztegiak atzigarri daude.[4] Etorkizunean eredu lokala libre eta atzigarri jartzeko lehen pausuak eman dira. Izan ere, Wikipediako testuinguruak erauzteko sistema eta algoritmo lokala iada berrinplementazio prozesuan daude. Berrinplementazioan memoria eta efizientzia kontuetan arreta berezia jarri da. Izan ere, hau da eredu lokalaren ahulgunea. Libre jartzearekin batera *ixapipes*[5] hizkuntzaren analisi katean barneratzeko asmoa dago.

Azkenik, tesi honetan landu diren ereduak ingurune eleaniztunean probatu nahiko lirateke. Izan ere, esperimentu guztiak ingelesezko datu-multzoetan egin dira arloaren egoerarekin konparatu ahal izateko. Horregatik, sistemen eraginkortasuna beste hizkuntzetan probatu nahi da (Euskara edo Gazteleran egin den gisa (Pérez de Viñaspre 2015) lanean). Hala ere, euskarazko Wikipedia ingelesezkoa baino askoz txikiagoa da, eta honek, sistemen eraginkortasunean nabarmen eragiten du. Edonola ere, euskarazko Wikipediako grafoa ingelesezko grafoko estekekin aberastuz desanbiguazioa hobetuko dela espero da.

[4] `http://ixa2.si.ehu.es/ukb/`

[5] `http://ixa2.si.ehu.es/ixa-pipes/`

Bibliografia

Agirre E., Cuadros M., Rigau G., and Soroa A. Exploring Knowledge Bases for Similarity. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

Agirre E., de Lacalle O.L., and Soroa A. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–88, 2014.

Agirre E. and Soroa A. Personalizing PageRank for Word Sense Disambiguation. *Proceedings of 14th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, 2009.

Agirre E., Soroa A., Alfonseca E., Hall K., Kravalova J., and Pasca M. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. *Proceedings of annual meeting of the North American Chapter of the Association of Computational Linguistics (NAAC)*, Boulder, USA, June 2009.

Agirre E., Barrena A., and Soroa A. Studying the wikipedia hyperlink graph for relatedness and disambiguation. *CoRR*, abs/1503.01655, 2015.

Agirre E. and Edmonds P. *Word Sense Disambiguation: Algorithms and Applications*. Springer Publishing Company, Incorporated, 1st edition, 2007. ISBN 1402068700, 9781402068706.

Alhelbawy A. and Gaizauskas R. Graph ranking for collective named entity disambiguation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 75–80, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Amigó E., Gonzalo J., Artiles J., and Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.*, 12 (4):461–486, 2009.

Baroni M., Dinu G., and Kruszewski G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of The 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, USA, 2014.

Barrena A., Agirre E., Cabaleiro B., Peñas A., and Soroa A. "one entity per discourse"and "one entity per collocation"improve named-entity disambiguation. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2260–2269, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

Barrena A., Soroa A., and Agirre E. Combining mention context and hyperlinks from wikipedia for named entity disambiguation. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, 101–105, Denver, Colorado, June 2015. Association for Computational Linguistics.

Bizer C., Lehmann J., Kobilarov G., Auer S., Becker C., Cyganiak R., and Hellmann S. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3): 154–165, September 2009. ISSN 1570-8268.

Bohnet B. Very high accuracy and fast dependency parsing is not a contradiction. *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 89–97, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

Bollacker K., Evans C., Paritosh P., Sturge T., and Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge.

Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, SIGMOD '08, 1247–1250, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-102-6.

Brin S. and Page L. The Anatomy of a Large-scale Hypertextual Web Search Engine. *Proceedings of the seventh international conference on World Wide Web 7*, WWW7, 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

Brown S.W., Dligach D., and Palmer M. Verbnet class assignment as a wsd task. *Proceedings of the Ninth International Conference on Computational Semantics*, IWCS '11, 85–94, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Bunescu R.C. and Pasca M. Using encyclopedic knowledge for named entity disambiguation. *Proceesings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 9–16, Trento, Italy, 2006. The Association for Computer Linguistics. ISBN 1-932432-59-0.

Carmel D., Chang M.W., Gabrilovich E., Hsu B.J.P., and Wang K. Erd'14: entity recognition and disambiguation challenge. *ACM SIGIR Forum*, 48 lib., 63–77. ACM, 2014.

Carpuat M. One translation per discourse. *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, 19–27, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-31-2.

Chang A.X., Spitkovsky V.I., Yeh E., Agirre E., and Manning C.D. Stanford-UBC Entity Linking at TAC-KBP. *Proceedings of TAC 2010*, 758 lib., Gaithersburg, Maryland, USA, 2010.

Cheng X. and Roth D. Relational inference for wikification. *EMNLP*, 2013.

Chisholm A. and Hachey B. Entity disambiguation with web links. *Transactions of the Association for Computational Linguistics*, 3:145–156, 2015. ISSN 2307-387X.

Cornolti M., Ferragina M., and Ciaramita M. A framework for benchmarking entity-annotation systems. *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, page 249âĂŞ260, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-2035-1.

Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 708–716, Prague, Czech Republic, 2007.

Cucerzan S. and Sil A. The msr systems for entity linking and temporal slot filling at tac 2013. *Proceedings of the Sixth Text Analysis Conference (TAC 2013)*, page 10, Gaithersburg, Maryland USA, 2013. National Institute of Standards and Technology (NIST).

Daiber J., Jakob M., Hokamp C., and Mendes P.N. Improving efficiency and accuracy in multilingual entity extraction. *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, 121–124, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1972-0.

De Marneffe M.C. and Manning C.D. Stanford typed dependencies manual. *URL http://nlp. stanford. edu/software/dependencies manual. pdf*, 2008.

Erk K. A simple, similarity-based model for selectional preferences. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 216–223, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Fernandez I. *Euskarazko Entitate-Izenak: identifikazioa, sailkapena, itzulpena eta desanbiguazioa.* Doktoretza-tesia, Euskal Herriko Unibertsitatea UPV/EHU, 2012.

Ferragina P. and Scaiella U. Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software*, 29(1):70–75, July 2012. ISSN 0740-7459.

Fokkens A., van Erp M., Postma M., Pedersen T., Vossen P., and Freire N. Offspring from reproduction problems: What replication failure teaches us. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1691–1701, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Gabrilovich E. and Markovitch S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proc of IJCAI*, 6–12, 2007.

Gale W.A., Church K.W., and Yarowsky D. One sense per discourse. *Proceedings of the workshop on Speech and Natural Language*, HLT '91, page 233âĂŞ237, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. ISBN 1-55860-272-0.

García N.F., Arias-Fisteus J., and Fernández L.S. Comparative evaluation of link-based approaches for candidate ranking in link-to-wikipedia systems. *J. Artif. Intell. Res. (JAIR)*, 49:733–773, 2014.

Guo Y., Che W., Liu T., and Li S. A graph-based method for entity linking. *Proceedings of 5th International Joint Conference on Natural Language Processing*, page 1010âĂŞ1018, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

Hachey B., Radford W., and Curran J. Graph-based Named Entity Linking with Wikipedia. *Proceedings of the 12th international conference on Web information system engineering*, WISE'11, 213–226, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 978-3-642-24433-9.

Hachey B., Radford W., Nothman J., Honnibal M., and Curran J.R. Evaluating Entity Linking with Wikipedia. *Artif. Intell.*, 194:130–150, January 2012. ISSN 0004-3702.

Han X. and Sun L. A generative entity-mention model for linking entities with knowledge base. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, 945–954, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-87-9.

Han X. and Sun L. An entity-topic model for entity linking. *EMNLP-CoNLL*, 105–115, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Haveliwala T. Topic-sensitive PageRank. *Proceedings of the 11th international conference on World Wide Web (WWW'02)*, 517–526, New York, NY, USA, 2002. ISBN 1-58113-449-5.

He Z., Liu S., Li M., Zhou M., Zhang L., and Wang H. Learning entity representation for entity disambiguation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 30–34, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Hoffart J., Seufert S., Nguyen D.B., Theobald M., and Weikum G. Kore: Keyphrase overlap relatedness for entity disambiguation. *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, 545–554, New York, NY, USA, 2012a. ACM. ISBN 978-1-4503-1156-4.

Hoffart J., Suchanek F., Berberich K., and Weikum G. YAGO2: a Spatially and Temporally Enhanced Knowledge Base from Wikipedia. *Artificial Intelligence*, 194:28–61, 2012b.

Hoffart J., Yosef M., Bordino I., Fürstenau H., Pinkal M., Spaniol M., Taneva B., Thater S., and Weikum G. Robust Disambiguation of Named Entities in Text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-937284-11-4.

Houlsby N. and Ciaramita M. A scalable gibbs sampler for probabilistic entity linking. In de Rijke M., Kenter T., de Vries A., Zhai C., de Jong F., Radinsky K., and Hofmann K., editors, *Advances in Information Retrieval*, 8416 lib. of *Lecture Notes in Computer Science*, 335–346. Springer International Publishing, 2014. ISBN 978-3-319-06027-9.

Hovy E., Navigli R., and Ponzetto S. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27, January 2013. ISSN 0004-3702.

Hughes T. and Ramage D. Lexical Semantic Relatedness with Random Graph Walks. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

Islam A. and Inkpen D. Second order co-occurrence pmi for determining the semantic similarity of words. *Proceedings of the International Conference*

on Language Resources and Evaluation (LREC 2006), 1033–1038, Genoa, Italy, 2006.

Jelinek F. and Mercer R. Interpolated Estimation of Markov Source Parameters from Sparse Data. *Pattern recognition in practice*, 381–397, 1980.

Ji H., Grishman R., Dang H.T., Griffitt K., and Ellis J. Overview of the tac 2010 knowledge base population track. *Third Text Analysis Conference (TAC 2010)*, 2010.

Ji H., Nothman J., and Hachey B. Overview of tac-kbp2014 entity discovery and linking tasks. *Proc. Text Analysis Conference (TAC2014)*, 2014.

Klein D. and Manning C.D. Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, 423–430. Association for Computational Linguistics, 2003.

Krovetz R. More than one sense per discourse. *NEC Princeton NJ Labs., Research Memorandum*, 1998.

Lazic N., Subramanya A., Ringgaard M., and Pereira F. Plato: A selective context model for entity resolution. *Transactions of the Association for Computational Linguistics*, 3:503–515, 2015. ISSN 2307-387X.

Lee H., Chang A., Peirsman Y., Chambers N., Surdeanu M., and Jurafsky D. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4), 2013.

Lehmann J., Monahan S., Nezda L., Jung A., and Shi Y. LCC Approaches to Knowledge Base Population at TAC 2010. *Proceedings of the Text Analysis Conference*, Gaithersburg, Maryland, USA, 2010.

Lewis D.D., Yang Y., Rose T.G., and Li F. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

Lopez de Lacalle O. *Domain-Specific Word Sense Disambiguation.* Doktoretza-tesia, Euskal Herriko Unibertsitatea UPV/EHU, 2009.

Martinez D. *Supervised Word Sense Disambiguation: Facing Current Challenges.* Doktoretza-tesia, Euskal Herriko Unibertsitatea UPV/EHU, 2004.

Martinez D. and Agirre E. One sense per collocation and genre/topic variations. *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, page 207âĂŞ215, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

McCarthy D. and Carroll J. Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences. *Comput. Linguist.*, 29(4):639–654, December 2003. ISSN 0891-2017.

McNamee P., Dang H., Simpson H., Schone P., and Strassel S. An Evaluation of Technologies for Knowledge Base Population. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC10)*, 369–âĂŞ372., Valletta, Malta, 2010.

McNamee P. and Dang H. Overview of the TAC 2009 Knowledge Base Population track. *Proceedings of the Second Text Analysis Conference*, 2009.

Mihalcea R. and Csomai A. Wikify!: linking documents to encyclopedic knowledge. *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, 233–242, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-803-9.

Mikolov T., Chen K., Corrado G., and Dean J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Miller G.A. and Charles W.G. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

Milne D. and Witten I.H. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, January 2013. ISSN 0004-3702.

Milne D. and Witten I. An Effective, Low-Cost Measure of Semantic Relatedness Obtained from Wikipedia Links. *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence*, Chicago, Illinois, USA, 2008a.

Milne D. and Witten I. Learning to Link with Wikipedia. *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM*

'08, page 509, New York, New York, USA, 2008b. ACM Press. ISBN 9781595939913.

Moro A., Raganato A., and Navigli R. Entity linking meets word sense disambiguation: a uniiĭňĄed approach. *Transactions of the Association of Computational Linguistics*, 2:231–244, May 2014.

Nastase V. and Strube M. Transforming Wikipedia into a large Scale Multilingual Concept Network. *Artificial Intelligence*, 194:62–85, 2013.

Navigli R. and Ponzetto S. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250, 2012a.

Navigli R. and Ponzetto S. BabelRelate! A Joint Multilingual Approach to Computing Semantic Relatedness. In Hoffmann J. and Selman B., editors, *AAAI*. AAAI Press, 2012b.

Navigli R. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, February 2009. ISSN 0360-0300.

Noreen E.W. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons, 1989.

Pedersen T., Patwardhan S., and Michelizzi J. Wordnet::similarity: Measuring the relatedness of concepts. *Demonstration Papers at HLT-NAACL 2004*, HLT-NAACL–Demonstrations '04, 38–41, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

Peñas A. and Hovy E. Filling knowledge gaps in text for machine reading. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 979–987. Association for Computational Linguistics, 2010.

Pérez de Viñaspre J. Wikipedia eta anbiguetate lexikala, 2015.

Pershina M., He Y., and Grishman R. Personalized page rank for named entity disambiguation. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 238–243, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Pilehvar M.T., Jurgens D., and Navigli R. Align, Disambiguate and Walk: a Unified Approach for Measuring Semantic Similarity. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 1341–1351, Sofia, Bulgaria, 2013.

Ponzetto S. and Strube M. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30:181–212, 2007.

Ponzetto S. and Strube M. Taxonomy Induction based on a Collaboratively built Knowledge Repository. *Artificial Intelligence*, 175:1737–1756, 2011.

Press W., Teukolsky S., Vetterling W., and Flannery B. *Numerical Recipes: The Art of Scientific Computing V 2.10 With Linux Or Single-Screen License*. Cambridge University Press, 2002. ISBN 9780521750363.

Radinsky K., Agichtein E., Gabrilovich E., and Markovitch S. A word at a time: computing word relatedness using temporal semantic analysis. *Proceedings of the 20th international conference on World wide web*, WWW '11, 337–346, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4.

Ratinov L., Roth D., Downey D., and Anderson M. Local and Global Algorithms for Disambiguation to Wikipedia. *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, 1375–1384. The Association for Computer Linguistics, 2011. ISBN 978-1-93243.

Resnik P. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159, November 1996.

Rubenstein H. and Goodenough J. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

Spitkovsky V.I. and Chang A.X. A Cross-lingual Dictionary for English Wikipedia Concepts. *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, 2012.

Strube M. and Ponzetto S.P. Wikirelate! computing semantic relatedness using wikipedia. *Proceedings of the National Conference on Artificial Intelligence*, 21 lib., 1419–1424. Menlo Park, CA; Cambridge, MA; London; AAAI Press, 2006.

Suchanek F.M., Kasneci G., and Weikum G. YAGO: A large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 2008.

Taghipour K. and Ng H.T. Semi-supervised word sense disambiguation using word embeddings in general and specific domains. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 314–323, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Tsatsaronis G., Varlamis I., and Vazirgiannis M. Text Relatedness Based on a Word Thesaurus. *J. Artif. Intell. Res. (JAIR)*, 37:1–39, 2010.

Varma V., Bharat V., Kovelamudi S., Bysani P., GSK S., N K.K., Reddy K., Kumar K., and Maganti N. Iiit hyderabad at tac 2009. Barne-txostena, NIST, 2009.

Yarowsky D. One sense per collocation. *Proceedings of the workshop on Human Language Technology*, HLT '93, page 266âĂŞ271, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. ISBN 1-55860-324-7.

Yazdani M. and Popescu-Belis A. Computing text semantic relatedness using the contents and links of a hypertext encyclopedia. *Artificial Intelligence*, 194:176–202, January 2013. ISSN 0004-3702.

Yeh E., Ramage D., Manning C., Agirre E., and Soroa A. WikiWalk: Random walks on Wikipedia for Semantic Relatedness. *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, 41–49, Suntec, Singapore, August 2009. Association for Computational Linguistics.

Zhong Z. and Ng H.T. It makes sense: A wide-coverage word sense disambiguation system for free text. *ACL: System Demonstrations*, 78–83, Uppsala, Sweden, 2010.

Glosategia

agerkidetza (*co-ocurrence*)

Dokumentuan bi termino edo gehiago elkarren ondoan izateari deritzo, zoriz lortutakoa baino handiagoa den maiztasunarekin.

antzekotasun edo antzekotasun semantiko (*semantic similarity*)

Hitzen arteko sinonimia (auto-berebil) eta hiperonimia/hiponimia (taxi-auto) erlazioak bere baitan hartzen dituen kontzeptua.

ausazko ibilbideak (*random walks*)

Ezagutza-baseen egiturak duen informazioa ustiatzeko metodo globala. Ezagutza-basea grafo bezala errepresentatuz, adabegiek grafoaren egituran duten garrantzia erlatiboa kuantifikatzen dute.

bilatzaile (*search engine*)

Konputagailuetan informazioa bilatzeko garatutako informazioaren berreskurapenerako sistema. Erabiltzaileak kontsulta bidez sistemari zer bilatu nahi duen adierazten dio, eta sistemak kontsulta horren araberako elementuak itzuliko dizkio.

datu-multzo (*dataset*)

EID sistemen ebaluazioa egiteko erabiltzen den datu-bilduma, hainbat dokumentuz osatua egongo dena.

entitatea (*entity*)

Entitateak existitzen diren (edo ziren) pertsona, leku edo erakundeak dira. Adibidez, `Jeff_Beck` edo `Donostia`.

entitate-hautagaiak (*candidate entities*)

Izen-aipamen batek ezagutza-basean erreferentziatu ditzakeen entitateen zerrenda dira. Adibidez, *Beck* aipamenak, beste batzuen artean, `Jeff_Beck` edo `Beck_Weathers` entitateak erreferentziatu ditzake.

entitate izenduna (*named entity*)

Izen propioa duen entitatea.

entitate izendunen desanbiguazioa EID (***named entity disambiguation*, NED**)

Ataza honen helburua izen-aipamenak dagokion entitate-hautagaiarekin lotzea da.

errendimendu (*efficiency*)

EID sistemen desanbiguazioaren azkartasuna adierazten duen neurria.

ezagutza-basea (*knowledge-base*)

Entitate eta kontzeptuei buruzko informazioa duen biltegi edo lexikoia.

hedapen (*expansion*)

Testu zati bati hitz berriak gehitzeko teknika, beti ere, hitz horiek testuko hitzekin nolabaiteko erlazio edo ahaidetasun semantikoa dutelarik.

hitz multzoak (*bag of words*)

EID atazan erabiltzen den desanbiguaziorako eredua. Eredu honetan dokumentuak hitzez betetako zaku bezala ikus daitezke; hau da, dokumentuan hitzek duten segida erabat galtzen da, eta, hortaz, ordena ez da kontuan hartzen.

hitzen adiera-desanbiguazio, HAD (***word sense disambiguation, WSD***)

Konputazio-metodoak erabiliz hitzen agerpenei adiera egokia esleitzen dien prozesua.

hizkuntzaren prozesamendu, HP (***natural language processing*, NLP**)

Hizkuntzaren tratamendu automatikoaren inguruko ikerrarloa.

izen-aipamena (*mention*)

Entitate jakin bat testuetan aipatzeko erabili den hitz-katea. Adibidez, `Jeff_Beck` entitatea *Beck* edo *Jeff* izen-aipamenekin izendatu daiteke.

kontzeptua (*concept*)

Kontzeptuak objektuen irudikapen abstraktuak dira. Adibidez, musikariak, mendiak edo kotxeak.

leuntze (*smoothing*)

Besteak beste, dokumentuan agertzen ez diren terminoei zero probabilitatea esleitu beharrean, probabilitate-masa txiki bat esleitzeko teknika. Hitz gutxitan esanda, gertaera ezagunentzat estimatutako probabilitatea txikiagotu eta gertaera ezezagunentzat estimatutako probabilitatea handiagotzen du teknika honek.

oinarri-lerroko sistema (*baseline system*)

Lantzen ari den arazoaren soluzio sinplea, oinarritzat hartu ohi dena emaitzen konparaketak egiterakoan. Sistema honek lortzen duen emaitza hobetzea izango da egiten diren esperimentuen helburua.

ontologia (*ontology*)

Mundu errealaren eskema kontzeptuala, non hitzekin izendatzen ditugun kontzeptuak modu hierarkikoan antolatuta dauden.

stopword

EID sistemen eraginkortasunean ekarpenik egiten ez duten eta, ondorioz, testuetatik kanpo uzten diren hitzak. Horien adibide dira, esaterako, artikuluak, preposizioak eta juntagailuak, edo bilduman oso ohikoak diren beste hainbat hitz.

WordNet

Ingeleseko hitz eta adierei buruzko informazioa duen ezagutza-base lexikala. Izen, aditz, adjektibo eta adberbioak aurkitzen dira bertan

synset delakoen arabera antolatuta eta hainbat erlazio semantikorekin lotuta.