I. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2015eko maiatzaren 13, 14 eta 15
Durango, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Izen-aipamenak desanbiguatu eta Wikipediara lotzen

Ander Barrena, Eneko Agirre, Jokin Perez de Viñaspre eta *Aitor Soroa*

669-675 or.
https://dx.doi.org/10.26876/ikergazte.i.92

ANTOLATZAILEA: udako euskal unibertsitatea
BABESLEAK: Eusko Jaurlaritza Gobierno Vasco; BFA DFB Bizkaiko Foru Aldundia Diputación Foral de Bizkaia; eman ta zabal zazu UPV EHU
LAGUNTZAILEAK: Deusto Universidad de Deusto Deustuko Unibertsitatea; Euskal Herria Udalbiltza; Mondragon Unibertsitatea; upna Universidad Pública de Navarra Nafarroako Unibertsitate Publikoa

Izen-aipamenak desanbiguatu eta Wikipediara lotzen

Ander Barrena eta **Eneko Agirre** eta **Jokin Perez de Viñaspre** eta **Aitor Soroa**

IXA Taldea - Euskal Herriko Unibertsitatea UPV/EHU

Laburpena

Izen-aipamenen desanbiguazio atazaren helburu nagusia testu batean agertu diren izen-aipamenak identifikatu eta Wikipediako entitateekin lotzea da. Lotura hau egitean, dokumentuaren ulermena errazteaz gain, irakurleak behar duen informazio guztia eskura jartzen zaio. Kontuan izan behar da, izen bat entitate ezberdin asko izendatzeko erabili daitekeela, hau da, anbiguoa dela. Artikulu honetan Wikipediatik ikasten duen ezagutzaz baliatuz, eredu Bayesiar bat azalduko da izen-aipamenak automatikoki desanbiguatu eta dagokion entitateari lotuko dituena.

Hitz gakoak: Izen-aipamenen Desanbiguazioa, Hizkuntzaren Prozesamendua, Wikipedia

Abstract

Linking entities with a knowledge base is a key issue in named entity disambiguation. This task enriches documents and makes them more understandable for readers. Due to name variation and ambiguity, name mentions should be linked to the correct Wikipedia entity. In this paper, we propose a Bayesian model which uses the knowledge extracted from Wikipedia to automatically disambiguate mentions.

Keywords: Named Entity Disambiguation, Natural Language Processing, Wikipedia

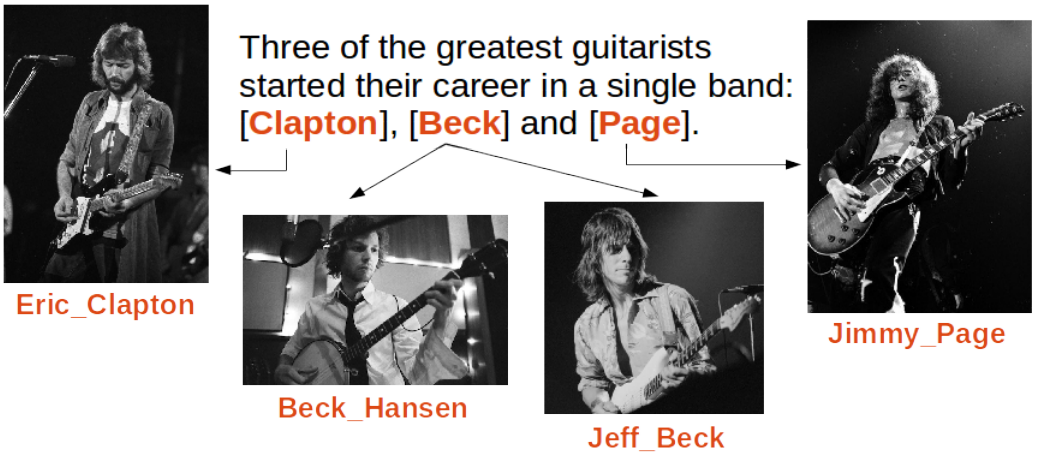
1 Sarrera eta motibazioa

Azken urteetan Wikipedia[1] bezalako ezagutza-trukerako web orriei esker, entitate-izenen ezagutza base erraldoiak edozeinen eskura daude. Interneten gaur egun ohikoak diren berrien orrialdeetan edo foro askotako elkar-ekintzetan, erabiltzailearentzat ezezagunak diren izen asko agertzen dira. Erabiltzailearen joera arruntena, pertsona, erakunde edo munduko toki horren informazioaren bila Wikipediara jotzea da. Ekintza honek, irakurgai den testuaren haria galtzea ekartzen du. Gainera, klik bakarrean informazioa eskura jar daiteke testua informazio askorekin aberastuz. Lan honen motibazio nagusia hori da: erabiltzaileari testua ulergarriagoa egiteko eta bere jakin-mina asetzeko, beharrezkoak diren artikuluak automatikoki eskura jartzea.

Motibazio hau borobiltzen duen adibide bat 1. irudian ikus daiteke. Clapton, Beck eta Page musikarien web orriak zuzenean atzigarri jarrita, erabiltzaileak Wikipediako informazio guztia eskura dauka, testua ulergarriagoa bihurtuz. Ataza honetan bi arazo nagusi aurki daitezke. Gerta liteke izen batek Wikipediako entitate bati baino gehiagori erreferentzia egitea. 1. irudian ikus daiteke *Beck* izenak *Jeff_Beck* edo *Beck_Hanson* erreferentziatu ditzakeela era berean. Gainera entitate bera izen ezberdinekin azaldu daiteke dokumentuetan, *Jeff_Beck* entitatea, *Jeff*, *Beck* edo *Geoffrey Arnold Beck* bezala azaldu daiteke. Beraz sistemak erabaki bat hartu beharko du horrelako kasuen aurrean, izen-aipamena desanbiguatuz.

[1] `http://www.wikipedia.org`

1 Irudia: Edozein egunkari edo interneteko Web orri batean aurki daitekeen testua. Bertan, izen-aipamenak Wikipedia orrietara lotuta azaltzen dira, baina *Beck*-en kasuan ez da argitzen zein den. Hots, *Beck* izen anbiguoa azaltzen dela, testua *Beck_Hansen* edo *Jeff Beck*-i buruz mintzo da?



2 Arloaren egoera eta ikerketaren helburuak

Izenen desanbiguazioaren arloan, ikerketaren zati handiena ingelesez egiten da. Gainera ikerketarako baliabideak, eta gainontzeko artearen egoera ere hizkuntza honetan daude. Horregatik artikulu honetako adibideak eta esperimentuak ingelesezko dokumentuetan egin dira, batez ere, beste ikerketekin alderatu ahal izateko.

Izenen desanbiguazioari dagokionez, arloaren egoeran oso ohikoak dira ikasketa automatikoan oinarritutako algoritmoak (ikus Bunescu eta Pasca (2006); Cucerzan eta Sil (2013); Mihalcea eta Csomai (2007); Milne eta Witten (2008); Hoffart *et al.* (2011)). Sistema hauek ehunka ezaugarri konbinatzen dituzte Bektore Euskarridun Makina edota Ausazko Basoak bezalako algoritmoekin. Ikasketa automatikoan oinarritutako sistemak izatean, dokumentuen eskuzko etiketatua beharrezkoa da algoritmo hauek bertatik ikas dezaten.

Ikerketa honen helburua, izen-aipamenak desanbiguatzeko, zuzenean Wikipediatik ikasiko duen sistema bat garatzea izango da. Horretarako, Wikipedian entitateen portaera aztertuko da eta honekin testu berri baten aurrean eredu Bayesiar baten bitartez izen-aipamenak desanbiguatuko dira.

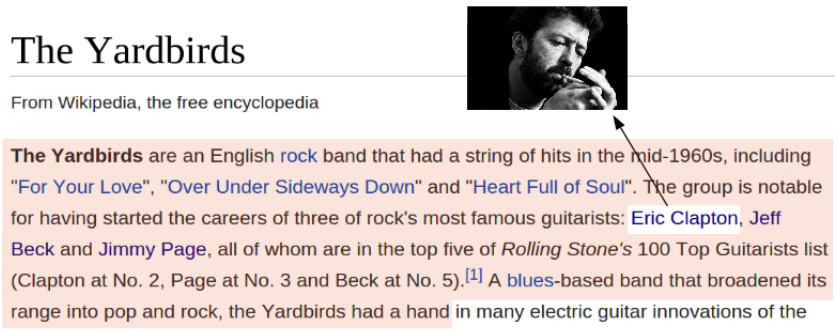
3 Wikipediatik ikasten

Wikipedia entziklopedia eleanitza artikulu sorta erraldoia bezala deskriba daiteke. Bere egituran, artikuluen barnean beste artikuluetara doazen estekak aingura bitartez egiten dira, eta artikuluan izen batez identifikatzen dira. Aingura izenak, sarrerak ez bezala, errepikatuak egon daitezke artikulu ezberdinetan. Ezaugarri hauek, informazio-iturri gisa oso interesgarriak dira, entitate-izenen desanbiguaziorako baliabide ezin hobea bihurtuz. Wikipedian entitateen aipamenak aztertuko dira, horretarako hasieratik bukaerara prozesatuko da artikuluz artikulu.

2. irudian Wikipedia artikulu baten lagina azaltzen da. Ikasketa prozesuan, aingura bat azaltzen den aldi bakoitza entitate baten aipamen bat bezala ikus daiteke. Gertaera bakoitzeko jarraian datozen ezaugarriak erauziko dira (adibidean irudiko *Eric Clapton* aingura erabiliko da):

- Lehenik "Hiztegia" deituko den fitxategi batean, *Eric Clapton* aingura, *Eric_Clapton* artikulu edo entitatera lotu dela zerrendaratuko da. Aingura eta artikuluen arteko erlazioa, aingura hori, artikulu batera erreferentzia moduan agertu den kopuruaren kontaketaz osatzen da. Hiztegia sortzeko oinarriak Chang *et al.* (2010) artikulutik hartu dira.
- Bigarrenik, aingura horren testuingurua gordeko da. Hau, 2. irudian azpimarratua dagoen tes-

2 Irudia: Wikipediako *The_Yardbirds* artikuluaren lagina.



tua litzateke. 50 hitzetako leihoa baino urrunago dauden terminoak izen-aipamenarekin loturarik ez dutela suposatzen da. Entitate bakoitzeko fitxategi bat sortuko da Wikipedian aingura bezala azaldu den testuinguru guztiekin. Hemendik aurrera Wikipediatik erauzi den testuingurua, entitate-testuingurua deituko da.

- Azkenik, aingura horrekin azaldu diren beste aingurak kontuan hartuta, entitateen arteko loturak egiten dira. Horrela, *Eric_Clapton*, *Jeff_Beck* eta *Jimmy_Page* entitateak elkarren artean erlazionatuko dituen grafoa sortuko da. Grafoko adabegiak entitateak izango dira eta ertzak hauen arteko erlazioak (artikulutik artikulura dauden ainguren arabera).

Laginaren adibidea ikusi ondoren, Wikipedia osoan pausu berdinak aplikatuz, ezagutza osoa kodetuko da hainbat baliabide ezberdinetan:

1. Hiztegi bat, aingura eta entitateak erlazionatzen dituena.
2. Entitate bakoitzeko, hau azaldu den testuinguruen bilduma.
3. Entitateen arteko erlazioak kodetzen dituen grafoa.

4 Entitate izenak Wikipediara lotzen

Hasieran aipatu den bezala, artikulu honek aurkezten duen metodoaren helburua betetzeko bi pausu eman beharko dira. Lehenik, testu batean azaldu den izen-aipamen bakoitzeko, izen honi lotuta dauden entitate hautagaiak zerrendaratuko dira. Bigarrenik, hautagaien artean bat aukeratuko da, hau da, izen-aipamena desanbiguatuko da Wikipediako entitate bakar bati lotuz. Jarraian datozen ataletan pausu bakoitza banan-banan azalduko da.

4.1 Izen-aipamenen entitate hautagaiak zerrendaratzen

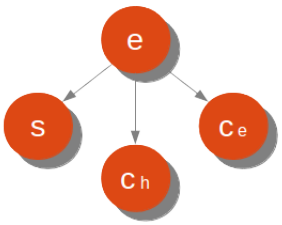
Testu batean identifikatu diren aipamenentzat hautagaiak izango diren entitateak sortzeko, 3. atalean aipatu den hiztegia erabiliko da. Testuan azaldu den izen-aipamena hiztegiko aingura balitz bezala tratatuz, aingurari lotuak azaldu diren entitateak, hautagaiak izatera pasako dira. Adibidez, 1. irudian, *Beck* hiztegian bilatu eta entitate hautagaiak hartuko lirateke: *Beck_Hanson* eta *Jeff_Beck*.

4.2 Entitate hautagaiak sailkatzen: Eredu Bayesiarra

Pauso honetan, eredu Bayesiarrar baten bitartez hautagaien artean bat aukeratuko da. Han eta Sun (2011) artikuluak finkatzen dituen oinarriak erabilita, c aipamen-testuinguru batean agertzen den s izen-aipamenetik, hautagai izan daitekeen $e1, e2, e3...$ entitateak sortzen ditu. Entitate bakoitzeko, lau probabilitate biderkatzen dira. Probabilitate hauek, 3. irudian ikus daitekeen diagraman, Bayesen Teoremaren oinarriak[2] aplikatuz lortzen dira.

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

3 Irudia: Izen-aipamen anbiguoak sailkatzeko eredua. e entitate bat emanda, s izen batekin gauzatuko da c testuinguru batean. Bayesen teorema aplikatuz azpian dagoen formula ebazten da.



$$P(s, c_h, c_e, e) = P(e)P(s|e)P(c_h|e)P(c_e|e)$$

Demagun *"Three of the greatest guitarrists started their career in a single band: Clapton, Beck and Page"* adibidean *Beck* desanbiguatu nahi dela. 3 irudian dagoen ereduko elementu bakoitza horrela gauzatuko zen:

- s edo aipamena, dokumentuan azaldu den *Beck* hitz katea izango da. Desanbiguatu nahi den izena hain zuzen.
- e edo entitateak, *Beck* aipamenaren hautagaiak diren *Beck_Hansen* eta *Jeff_Beck* izango dira. Kontuan hartu hiztegiak 83 artikulu zerrendatzen dituela, baina adibidea azaltzeko 2 erabiliko dira. Errealitatean desanbiguazioa 83 horien artean egiten da.
- c aipamen-testuingurutik bi probabilitate ezberdin zenbatetsiko direnez, 2 informazio iturri ezberdin erauziko dira:
  - c_h edo aipamen-testuinguruko hitzak. Kasu honetan *Three*, *of*, *the*, *greatest*, *guitarrists*, *started*, *their*, *career*, *in*, *a*, *single*, *band*, *Clapton*, *and* eta *Page* hitzek osatuko dute. Algoritmoak ikasteko erabili duen patroiari jarraituz, s-ren inguruan 50 hitz ezkerrera eta beste hainbeste eskuinera.
  - c_e edo aipamen-testuinguruko gainontzeko entitateen aipamenak, adibidean *Clapton* eta *Page*-k osatuko dute. Hauek 3 atalean azaldu den grafoarekin zerikusia izango dute.

Lehen esan bezala lau probabilitateen biderketa bidez $P(s, c_h, c_e, e) = P(e)P(s|e)P(c_h|e)P(c_e|e)$ hautagaiak diren entitateak sailkatuko dira. Probabilitate altuena lortzen duena izango da s izenarekin c testuan agertu den aipamenari dagokion e entitatea. Beraz entitate irabazlea aurreko formularen maximoa lortzen duena izango da.

$$e = \arg\max_e P(s, c_h, c_e, e) = \arg\max_e P(e)P(s|e)P(c_h|e)P(c_e|e)$$

Jarraian formula hau ebazteko, probabilitate bakoitzak zein ezagutza deskribatzen duen eta nola zenbatetsi den azalduko da.

4.2.1 Entitatearen probabilitatea

$P(e)$ izendatu den banaketak, e entitatea Wikipedian zein ospetsua den adieraziko du egiantza handieneko zenbatezketaz. Horretarako jarraian dagoen formula erabiliko da probabilitate hau kalkulatzeko.

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

M Wikipediako entitate agerpen guztien kontaketa da eta N entitate ezberdin kopurua. $Count(e)$-k e entitate horri dagokion agerpen kopurua izango da. $Count(e)$ balioari +1 leunketa aplikatzen zaio 0 probabilitatearen arazoa saihesteko[3]. Kontaketa guzti hauek hiztegitik ateratzen dira.

[3]Arazo honekin biderkagaietako bat 0 izanda probabilitate osoa 0 litzateke eta hau saihestu nahi da.

4.2.2 Entitatearen izenaren probabilitatea

$P(s|e)$ banaketak, e entitatea s izenaz agertzeko duen probabilitatea erakusten du. Horretarako egiantza handieneko zenbatezketaz baliatuz formula hau erabili da.

$$P(s|e) = \frac{Count(e,s)+1}{Count(e)+S}$$

$Count(e,s)$-k e entitatea s izenarekin agertu den kontaketa balioa erakusten du. $Count(e)$-k aldiz, e entitatea s izen ezberdin guztiekin zenbat aldiz agertu den kalkulatzen du, edo lehen esan bezala agerpen kopuru totala. Banaketa honetako kontaketak ere 3 atalean azaldu den hiztegitik ateratzen dira. Aurreko biderkagaiari aplikatu zaion leunketa berdina aplikatuz, S-k e entitatea zenbat aingura ezberdinekin azaldu den adierazten du.

4.2.3 Entitatearen testuinguruaren probabilitatea

$P(c_h|e)$ banaketak e-ren entitate-testuinguruan, c aipamen-testuinguruko t termino edo hitz bakoitzaren probabilitatea $P'_e(t)$ banaketa zenbatetsiko du. Horretarako termino horren agerpen kopurua zati termino guztien kontaketa kalkulatzen da (kontaketak entitate-testuingurukoak dira, Wikipediatik ikasi direnak). Formula jarraian datorrena da.

$$P'_e(t) = \frac{Count_e(t)}{\sum_t Count_e(t)}$$

$P'_e(t)$ banaketari leunketa metodo bat aplikatu zaio 0 probabilitate arazoari aurre egiteko. Horretarako, probabilitateari termino horrek Wikipedian orokorrean agertzeko duen probabilitatea $P_g(t)$ gehitu zaio Jelinek eta Mercer (1980) artikuluak aipatzen dituen pausoak jarraituz. Jarraian $P_e(t)$-ren zenbatespena kalkulatuko duen azken formula $P_g(t)$ leunketa probabilitatea barne duela.

$$P_e(t) = \lambda P'_e(t) + (1-\lambda)P_g(t)$$

λ parametroari 0.95 balioa ezarri zaio garapen esperimentuetan oinarrituz. Beraz, s izen-aipamenaren c aipamen-testuinguruak n termino baditu, e entitate batentzat $P(c|e)$-ren zenbatespena hau litzateke:

$$P(c_h|e) = P_e(t_1)P_e(t_2)...P_e(t_n)$$

Terminoen artean dependentziak daudela badakigu, izan ere termino askok probabilitate handia dute bata bestearen ondoren agertzeko. Baina banaketa honetan terminoen arteko independentzia asumitzen da, zenbatespena errazten baitu.

4.2.4 Entitatearen testuinguruaren probabilitatea grafoaren arabera

Laugarren biderkagaiak e entitateak c aipamen-testuinguruan azaldu diren beste entitateekin azaltzeko duen probabilitatea neurtzen du. Kasu honetan, labur esanda, Wikipediako entitateekin osatu den grafoan duten garrantziaren arabera egiten da. Probabilitate hau zenbatesteko Agirre eta Soroa (2009) eta Agirre *et al.* (2014) artikuluetan azaltzen diren algoritmoak erabili dira, *Personalized Page Rank* algoritmoa hain zuzen. Ezaugarri hau $P(c_e|e)$ probabilitate banaketa bezala izendatuko da. Hau kalkulatzeko UKB[4] softwarea erabilida.

5 Esperimentuak

Sistemaren oinarriak azaldu ondoren, bere eraginkortasuna ebaluatuko da. Ingelesezko entitate izenen desanbiguazioa ebaluatzeko, hainbat datu-multzo eskuragarri daude. Horien artean 2009. urtetik hona, urtero ospatzen den TAC-KBP[5] txapelketako Entity-Linking atazakoak. Helburua ingelesezko testu

[4] http://ixa2.si.ehu.es/ukb/
[5] http://www.nist.gov/tac/2014/KBP/

ezberdinetako izen-aipamenak Wikipediako azpimultzo batetik sortutako ezagutza-baseko entitateekin lotzea da. Horretarako dokumentu berrietatik, foroetatik edo interneteko web orrietatik biltzen dituzte. Ondoren bertan azaltzen diren aipamenak dagokion entitatearekin eskuz etiketatzen dituzte. Urtero ospatzen den konferentzia honetako 6 datu-multzoak erabiliko dira ebaluazioan (TAC09, TAC10, TAC11, TAC12, TAC13 eta TAC14). Esan beharra dago txapelketako antolatzaileek, izen-aipamen anbiguoak dituzten dokumentuak hautatzen dituztela. Bestalde, AIDA eta KORE[6] deituriko beste bi datu-multzoetan ere ebaluatuko da. Lehenak berrietatik eskuratu diren dokumentuak biltzen ditu, bigarrenak, aldiz, testu oso motzak eta izen oso anbiguoz osaturiko dokumentuak. Eredu Bayesiarraren parametro eta ezaugarriak finkatzeko, garapen esperimentuak AIDA datu-multzoren beste atal batean egin dira.

AIDA, KORE, TAC09 eta TAC10-en ebaluaziorako erabiltzen den metrika zehaztasuna da. Zehaztasunak entitate egokira zenbat aipamen lotu diren, zati datu-multzoko aipamen kopuru totala kalkulatzen du. TAC11-tik TAC14 bitartean metrika aldatu eta bCubed+ (ikus Amigó *et al.* (2009)) metrika erabiltzera pasa ziren. Datu multzo bakoitzarentzat, sistemaren emaitza eta orain arte argitaratu den emaitzarik onena alderatuko dira.

1 Taula: Datu-mutzo ezberdinetan, eredu Bayesiarraren emaitza eta arloaren egoeran argitaratu den emaitza onena konparatzen dira. Beltzez nabarmenduta artearen egoera gainditzen duten emaitzak ageri dira.

	AIDA	KORE	TAC09	TAC10	TAC11	TAC12	TAC13	TAC14
$P(e)P(s\|e)$	67.51	36.11	67.46	76.76	68.27	46.60	68.29	62.55
$P(e)P(s\|e)P(c_h\|e)$	76.01	61.11	78.45	**85.29**	76.42	57.74	**76.47**	71.92
$P(e)P(s\|e)P(c_h\|e)P(c_e\|e)$	82.59	70.14	**82.15**	**85.39**	**81.64**	**72.20**	**74.56**	75.31
arloaren egoera	84.89	71.50	79.00	80.60	80.10	68.50	71.80	79.60

1. taulan sistemaren emaitzak ikus daitezke. Biderkagai ezberdinak konbinatuz, eredu Bayesiarraren konbinazio posibleak azaltzen dira. Lehen lerroan, testuinguruaren informazioa alde batera utzita, bi ezaugarrien konbinazioak ematen dituen emaitzak ikus daitezke. Ondoren, bigarren lerroan, testuinguruko hitzen probabilitatea biderkatuz emaitzak gora egiten dutela ikus daiteke, arloaren egoerako emaitzetara gerturatuz. Azkenik, testuinguruko entitateek grafoan lortzen duten probabilitateari esker, algoritmoak arloaren egoerako emaitza parekoak lortzen dituela ikus daiteke.

Aipatzekoa da TAC10, TAC12 eta TAC13 datu multzoetan, arloaren egoerarekiko dagoen aldea. Gainera AIDA eta KORE datu-multzoetan emaitza onena hobetzea lortu ez den arren, onenarekiko distantzia ez da inoiz 2,5 puntu baino handiagoa. TAC14-en ordea, emaitzak hobetzeko aldea badago.

Kontuan hartu behar da arloaren egoerako emaitzak sistema eta sortzaile ezberdinen eskuetatik datozela, hau da, ez dela sistema berdina datu-multzo guztietan exekutatu. Beraz datu multzoaren arabera sistema optimizatu daiteke egoera horretara egokituz. Aurkeztu den eredu Bayesiarrak aldiz, portaera egonkorra azaldu du datu-multzo guztietan emaitza onak lortuz.

6 Euskarazko testuak Wikipediara lotzen

Ingelesezko testuetan ereduak duen eraginkortasuna ikusi ondoren, euskarazko testuetako izen-aipamenak Wikipediara lotzeko saiakera bat egin da. Honetarako Fernandez (2012) tesian erabili ziren CorpusB bildumako euskarazko testuak erabili dira. Hala ere, baliabide faltagatik, eredu Bayesiarra ez da bere osotasunean probatu. Testuinguruko hitzen probabilitatea alde batera utziz, jarraian dagoen formula erabili da: $P(s, c_e, e) = P(e)P(s|e)P(c_e|e)$. Fernandez (2012) tesian izen-aipamenak Wikipediara lotzean %76.20 zehaztasuna lortzen da. Eredu Bayesiarrak, dokumentu berdinen gainean %87.84-ko zehaztasuna lortzen du.

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

7 Ondorioak

Artikulu honetan, testuak ulergarriagoak egiteko dokumentu bateko izen aipamenak automatikoki Wikipediara lotzen dituen eredu Bayesiar bat aurkeztu da. Sistemak behar duen ezagutza guztia Wikipediatik ikasten du. Hori gutxi balitz, izen anbiguoei aurre egiteko gai dela frogatu da. Gainera, arloaren egoerarekin konparatuz eta datu-multzoaren arabera, emaitzak oso gertu edo hauen gainetik aurkitzen direla ikusi da. Sistema berdinak datu-multzo ezberdinetan lortu dituen emaitza egonkorrek aipamena merezi dute.

8 Etorkizunerako planteatzen den norabidea

Etorkizunean, eredu Bayesiarretatik haratago beste eredu probabilistiko batzuk erabiltzea pentsatzen da. Horretarako Markov-en kateetan oinarritutako algoritmoak erabiliz iadanik zenbatetsi diren probabilitateak barneratuko dira. Modu honetan ereduari ahalik eta etekin gehien ateratzea espero da.

Erreferentziak

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

Agirre, Eneko, Oier Lopez de Lacalle, eta Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics* 40.57–88.

Amigó, Enrique, Julio Gonzalo, Javier Artiles, eta Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Inf. Retr.* 12.461–486.

Bunescu, Razvan C., eta Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*. The Association for Computer Linguistics.

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

Cucerzan, Silviu, eta Avirup Sil. 2013. The msr systems for entity linking and temporal slot filling at tac 2013. In *Proceedings of the Sixth Text Analysis Conference (TAC 2013)*, p. 10. National Institute of Standards and Technology (NIST).

Fernandez, Izaskun. 2012. Entitate-izenak euskaraz: Identifikazioa, sailkapena, itzulpena eta desanbiguazioa.

Han, X., eta L. Sun. 2011. A Generative Entity-mention Model for Linking Entities with Knowledge Base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, 945–954.

Hoffart, J., M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, eta G. Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, United Kingdom 2011*, 782–792.

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

Mihalcea, Rada, eta Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 233–242. ACM.

Milne, D., eta I.H. Witten. 2008. Learning to Link with Wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, p. 509, New York, New York, USA. ACM Press.