Lurrikarei buruzko informazioa eskuratzen Twitter bidez.

Ander Intxaurrondo, Eneko Agirre eta **Oier Lopez de Lacalle**

Ixa Taldea. Euskal Herriko Unibertsitatea.

Laburpena

Lan honetan, mikroblogetatik gertaera konplexuak erauzten dituen sistema bat aurkezten dugu, urruneko gainbegiraketa erabiliz. Denbora errealeko datu iturri hauetako testuak laburrak, sintaxi zaratatsukoak eta anbiguoak dira; baina informazio kantitate handiak topatu ditzakegu. Gure ekarpena lurrikaren domeinuan ebaluatzen dugu, 20 argumentutik gorako gertaerekin. Ezagutza basea eta txio garrantzitsuak dituen datu multzoa publikoki dago eskuragarri, biak ingelesez daude.

Hitz gakoak: Hizkuntzaren prozesamentua, gertaeren erauzketa, urruneko gainbegiraketa, ezagutza baseak

Abstract

In this work, we introduce an event extraction approach that extracts complex event templates from microblogs, using distant supervision. These near real-time data source texts are short, ambiguous and contain dirty syntax; but we can find lots of information. We evaluate our contribution on the domain of earthquakes, with events with up to 20 arguments. The dataset containing the knowledge base and relevant tweets is publicly available, both in English.

Keywords: Language processing, event extraction, distant supervision, knowledge bases

1 Sarrera eta motibazioa

Twitter baliabide ona bilakatu da denbora errealean gertaera desberdinei buruko datuak lortzeko era azkarrean, informazio erauzketa (IE) ohiko egunkari artikuluak ez diren beste informazio iturrietan aplikatzera motibatuz. Era askotako informazioa eskuratu dezakegu, hala nola artista baten emanaldi bati buruzkoa, hegazkin istripuak, eta abar. Txioek hizkera kolokiala, sintaxi eta diskurtso zaratatsua, eta informazio anbiguoa izateko joera dute; hala ere, informazio kantitate handiak aurki ditzakegu.

Lan honetan gertaera erauzketa (GE) sistema bat garatu dugu. GE sistemak, testuetako gertaerak identifikatzen saiatzen dira, eta testuinguruko elementu desberdinek jokatzen duten rolak identifikatzen saiatzen dira. Aukeratu dugun domeinua lurrikarena da, eta lurrikara bakoitzeko 20 argumentu desberdini buruzko informazioa eskuratzen dugu, automatikoki aukeratutako txio sorta batekin.

Informazio erauzketa sistema onenetako corpusak eskuz etiketatzen dira, oso emaitza onak ematen dituzte, baina etiketatze prozesu honen kostua oso garestia da. Lan honetan, eskuzko etiketazioaren kostu garestia alde batera uzten dugu, eta entrenamenduko corpusak automatikoki eskuratzeko algoritmo bat erabili: urruneko gainbegiraketa (UG).

Lan honetarako egindako ekarpenak ondorengoak dira:

1. Hau da urruneko gainbegiraketaren bidez mikroblogetatik argumentu askotako gertaera konplexuak erauzten dituen lehen lana.
2. Lurrikarei buruzko ezagutza base bat jarri dugu publikoki eskuragarri, baita lurrikara bakoitzari dagozkion txioak ere. Txioak eta ezagutza basea ingelesez daude.

Hasteko, lan honetan ezagutza baseak eta urruneko gainbegiraketa zer diren azalduko dugu. Jarraian lurrikarei buruzko ezagutza basea nola sortu dugun azaldu, eta lurrikara bakoitzari buruzko txioak nola eskuratu ditugun komentatuko dugu. Ondoren esperimentuak eta emaitzak erakutsiko ditugu. Amaitzeko, lan honi buruzko ondorioak eta etorkizuneko lanak aurkeztuko ditugu.

1 Irudia: Bi infotaulen adibideak.

(a) Bernardo Atxagaren infotaula.

Datu pertsonalak	
Izen osoa	Jose Irazu Garmendia
Ezizena	*Bernardo Atxaga*
Jaio	1951ko uztailaren 27a Asteasu, Gipuzkoa (Euskal Herria)
Bikotekidea(k)	Asun Garikano
Webgunea	http://www.atxaga.org/

(b) Lurrikara baten infotaula.

Date	15:40 PDT, April 4, 2010
Duration	89 seconds
Magnitude	7.2 M_w
Depth	10 kilometers (6 mi)
Epicenter	32.128°N 115.303°W
Countries or regions	Mexico United States
Max. intensity	IX[1]
Tsunami	No
Landslides	Yes
Aftershocks	Yes
Casualties	4 killed, at least 100 injured in the vicinity of Mexicali.[2]

2 Arloko egoera eta ikerketaren helburuak

Ezagutza base bat (EB) ezagutza kudeatzeko datu base berezi bat da. Ezagutzaren bilketa, antolaketa eta berreskurapena konputazionalki egiteko baliabideak hornitzen ditu. Azken urteetan informazio erauzketan eta lengoaia naturalaren prozesamenduan geroz eta gehiago erabiltzen dira. Gehien erabiltzen direnak DBpedia[1] eta Freebase[2] dira.

Ezagutza baseak hainbat kontzeptu eta entitateren multzoak dira, eta entitate hauei buruzko informazioa era eskematikoan eta ulergarrian irudikatzen dute. Entitate bakoitzak erlazio batzuk ditu, erlazio bakoitzak izen bat jasotzen du eta beste entitate, kontzeptu edo balio batekin erlazionatuta dago. Ezagutza baseetan aurki ditzakegun entitateak pertsonak, erakundeak, lekuak, denbora adierazpenak eta beste hainbat motatakoak izan daitezke, bai entitate nagusia, baita erlazioan parte hartzen duen bigarren entitatea ere.

Wikipediako infotaulak oso baliagarriak dira ezagutza baseak sortzeko. Infotaulak Wikipediako artikulu baten eskubialdean aurki ditzakegu, artikuluko informazioaren laburpen bat emanez. 1a irudian Wikipediako Bernardo Atxaga idazlearen artikuluko[3] infotaula dugu, bertatik ondorengo erlazioak eskuratu ditzakegu, besteak beste:

- Bernardo Atxaga - *jaioteguna* - 1951ko uztailaren 27a
- Bernardo Atxaga - *jaioterria* - Asteasu
- Bernardo Atxaga - *izen_osoa* - Jose Irazu Garmendia

Urruneko gainbegiraketa (UG) (Mintz *et al.*, 2009) lanean erlazio erauzketarako proposatutako paradigma bat da. Hurbilketa honek automatikoki etiketatzen ditu corpusak. UGren motibazio nagusia eskuzko lanak sahiestea da, hala nola corpusen eskuzko etiketatzea.

UGren arabera, ezagutza base batek bi elementuren artean erlazio bat dagoela zehazten badu, eta bi elementu hauek esaldi berean agertzen badira, esaldi horrek erlazio hori adieraziko du nola edo hala.

Corpus batetik Bernardo Atxagari buruzko esaldiak berreskuratu ondoren, esaldi hauetan dauden entitate desberdinak detektatu behar ditugu. 1 taulan Bernardo Atxagari buruzko esaldi desberdinak ditugu, aurretik aipatutako erlazioak adieraziz.

Urruneko gainbegiraketa ia ez da erabili gertaerei buruzko informazioa erauzteko. (Benson *et al.*, 2011) da GE eta UG batu dituen lehen lana, Twitterreko txioak erabiliz. Esperimentu hauetan, astista des-

[1] `http://dbpedia.org/About` . Euskaraz `http://eu.dbpedia.org/index.php?title=Azala`
[2] `https://www.freebase.com/`
[3] `http://eu.wikipedia.org/wiki/Bernardo_Atxaga`

1 Taula: Bernardo Atxagari buruzko esaldi desberdinak, eskuineko zutabeak idazlea eta letra lodiz jarritako elementuen arteko erlazioa adierazten du.

Esaldia	**Erlazioa**
Bernardo Atxaga 1951ko uztailaren 27an jaio zen Asteasu herrian.	*jaioteguna*
Bernardo Atxaga, **1951eko uztailaren 27an** jaioa, idazle ospetsu bat da.	*jaioteguna*
Bernardo Atxaga 1951ko uztailaren 27an jaio zen **Asteasu** herrian.	*jaioterria*
Asteasu da **Bernardo Atxaga** jaio zen herria.	*jaioterria*
Bernardo Atxaga izengoitiz, agiri ofizialetako izen deiturez **Jose Irazu Garmendia** (...)	*izen_osoa*
Bernardo Atxaga da **Jose Irazu Garmendiaren** goitizena.	*izen_osoa*

berdinek New York hirian egindako emanaldiei buruzko informazioa lortzen saiatzen dira, baina bakarrik emanaldi bat non egin duten jakin nahi dute, emanaldiari buruzko informazio sakonagoa (ordua, ikusle kopurua,...) alde batera utzita. Gure gertaera erauzketa esperimentuetan berriz, lurrikarei buruzko informazio asko erauzten dugu.

(Reschke *et al.*, 2014) lanean ere UG erabiltzen dute gertaerak erauzteko. Lan honetan, hegazkin istripuei buruzko hainbat informazio erauzten dute: eguna, istripu mota, istripua gertatu den lekua, hegaldi zenbakia, hildakoak eta abar. Gure lana eta hau oso parekoak dira, baina beraiek berri agentzien dokumentuak aztertuz eskuratzen dute informazio hori, guk ordea, Twitter erabiltzen dugu.

UG gertaera erauzketan aplikatzeko, aurretik testuetako gertaerak identifikatzea komeni da. UGren algoritmoa ezin dugu zuzenean aplikatu GErako, gertaeraren izena ez delako esaldietan esplizituki aipatzen. UG gertaeren erauzketara moldatzeko, ondorengo heuristikoa proposatzen dugu: esaldi bat gertaera konkretu bati buruzkoa bada, esaldian dagoen aipamen batek batek ezagutza baseko argumentu baten balio berdina badu, aipamen horrek argumentu mota hori adieraziko du nola edo hala.

3 Ikerketaren muina

Atal honetan, lurrikarei buruzko EBa nola sortu dugun azalduko dugu, txioak eskuratzeko jarraitutako irizpideekin batera. Sistemak txio bakoitza nola prozesatu duen azalduko dugu, eta amaitzeko emaitzak erakutsi.

3.1 Lurrikarei buruzko ezagutza basearen sorkuntza

Lan honetarako, ingelesezko Wikipediako infotauletan oinarritutako ezagutza base bat sortu dugu. Ezagutza base honetan 2009ko hasiera eta 2013ko uztailaren arteko lurrikarak aurki ditzakegu. EB honetan, lurrikarari buruzko hainbat informazio dago bilduta, hala nola eguna, lekua, magnitudea eta abar. Guztira 108 lurrikara desberdinei buruzko informazioa bildu dugu.

1b irudian, ingeleseko Wikipediako infotaula bat dugu. Infotaula hau Mexikoko Baja Californian[4] gertatutako lurrikara batena da.

Ezagutza basea 20 argumentu desberdinez osatuta dago. 2 taulak argumentu horiek biltzen ditu, argumentu bakoitzaren datu mota zein den adieraziz, honen esanahia euskaraz, eta aurretik adibide bezala erabili dugun lurrikararen datuekin. Argumentu motak ondorengoak dira: E eguna, D denbora, L lekua, Z zenbakizkoa eta B boolearra (bai ala ez). Asteriskoa (*) duten argumentuek balio bat baino gehiago onartzen dute. 4. zutabeak argumentu bakoitzeko zenbat informazio dugun ezagutza basean adierazten du; ikus dezakegunez, lurrikara guztiek dute eguna, ordua, estatua, magnitudea eta koordenatu geografikoei buruzko informazioa; aurrelurrikarak, iraupena, desagertu kopurua eta beste argumentu batzuei buruzko informazioa 10 lurrikara baino gutxiagotan aurki dezakegu. Azken zutabea hurrengo azpiatalean azalduko dugu.

[4] `http://en.wikipedia.org/wiki/2010_Baja_California_earthquake`

2 Taula: Lurrikarei buruzko ezagutza basearen argumentuak, hauen esanahia euskaraz, argumentu motak, adibide bat, argumentu bakoitzaren balio kopurua EBan, eta urruneko gainbegiraketaren bidez zenbat aldiz etiketatu dugun argumentu bakoitza datu multzoan.

Argumentua	**Euskaraz**	**Mota**	**Adibidea**	**# EB**	**# UG**
date	Eguna	E	2010-4-4	108	291
time	Ordua	D	T22:40:00	108	378
country	Estatua	L	Mexico	108	6294
region	Herrialdea	L	Baja California	77	2598
city	Hiria	L	-	77	1426
latitude	Latitudea	Z	32.128	108	2
longitude	Longitudea	Z	-115.303	108	4
dead	Hilkakoak	Z	4	71	143
injured	Zaurituak	Z	100	39	22
missing	Desagertuak	Z	-	8	-
magnitude	Magnitudea	Z	7.2	108	933
depth (km)	Sakonera (km)	Z	10	99	27
affected-country(*)	Estatu kaltetua	L	United States	37	436
affected-region(*)	Herrialde kaltetua	L	-	4	-
landslides	Lubiziak	B	yes	8	7
tsunami	Tsunamiak	B	-	10	408
aftershocks	Erreplikak	Z	-	20	5
foreshocks	Aurrelurrikarak	Z	-	3	6
duration	Iraupena	D	00:01:29	7	-
peak-acceleration	Azelerazio sismikoa	Z	-	8	-
			Guztira	**1116**	**13562**

3.2 Txioak Twitterretik eskuratzen

Lurrikarei buruzko txioak eskuratzeko, Topsy Labs[5] enpresaren baliabideak[6] erabili ditugu. Enpresa hau baliabide sozialen edukien bilaketa eta analisira jarduten da.

Lurrikara bakoitzeko bilaketak egitean, *earthquake* hitz gakoa erabili dugu, ezagutza basean agertzen zen kokapenarekin (hiriak, herrialdeak eta estatuak) batera zehaztuz. Lurrikara gertatu baino egun bat lehenago eta hortik 7 egun geroago idatzitako txioak eskuratzen ditugu bakarrik.

Lurrikara gertatu baino egun bat lehenagoko tweetak eskuratzeak arrazoi bat dauka: ordu eremuak[7]. Txioak ez daude geolokalizatuta, kontuan hartu behar da txiolariak ez direla profesionalak eta txiokatzean lurrikara gertatutako unea aipatzeko beren bizilekuko ordua erabiltzen dutela denbora estandarraren ordez.

Lortutako txio asko lurrikaren erreplikei buruzkoak dira. Erreplikak lurrikara nagusiaren ondoren gertatutako beste lurrikara batzuk dira, hauek epizentrotik gertu daude eta normalean nagusiak baino magnitude txikiagoa dute. Erreplikak direla eta, txioetako informazioan eta EBko informazioan kontraesanak egongo dira, sistemaren ikasketa prozesua nahastuz eta ebaluatzean emaitza okerrak itzuliz.

3.2.1 Erreplikak antzematen

Erreplikak gertaera desberdin bezala tratatu ditugu, eta ezagutza basean zeuden lurrikarei buruzko txioak bakarrik edukitzearren, metodo oldarkor bat aplikatu dugu erreplikei buruzko txioak baztertzeko. Heuristiko hau aplikatzeko, txioak kronologikoki ordenatu ditugu. Heuristiko hau txio desberdinetan aipatzen diren denbora adierazpenetan oinarritzen da:

[5] `http://topsy.com`
[6] `http://api.topsy.com/doc`
[7] `http://eu.wikipedia.org/wiki/Ordu-eremu`

1. Txioetan lurrikara bakoitzeko lortutako lehen denbora adierazpena gordetzen dugu. *ordua:minutua* patroia erabili da denbora adierazpen hauek antzemateko. Segunduak ez ditugu kontuan hartzen.
2. Geroagoko txio batean agertzen den denbora adierazpena lehenengoarekiko desberdina bada, bai ordua bai minutua, txio hau erreplika bati buruz ari dela ulertzen dugu. Txio hau eta ondoren datozen beste guztiak kentzen ditugu, hemendik aurrera jasoko ditugun txioak erreplika horri edo beste batzuei buruzkoak izango direlakoan. Ordua lurrikara nagusiko orduarekiko desberdina bada baina minutua berdina, orduan lurrikara nagusitzat hartzen dugu txio hau, denbora eremu desberdin batean dagoen pertsona batek txiokatu duelakoan.

Bukaerako datu multzoak 108 lurrikara desberdin ditu eta guztira 7841 txio desberdin. Batazbesteko 72 txio ditugu lurrikara bakoitzeko, gehienez 654 txio eta gutxienez 2 edukiz. 19 lurrikarek 10 txio baino gutxiago dituzte.

3.3 Aipamenen etiketatzea txioetan

Urruneko gainbegiraketaren algoritmoa aplikatuz, lurrikara bakoitzaren txioak bildu eta EBko argumenturen baten balioarekin bat egiten duten aipamenak etiketatu ditugu. Adibide bezala, Baja Californiako lurrikarari buruzko txio bat hartuko dugu:

- Update : Earthquake in Baja California, Mexico upgraded to 7.2 magnitude, from 6.9 - USGS
  (Eguneraketa: Baja California, Mexikoko lurrikararen magnitudea 6.9tik 7.2ra eguneratuta - USGS)

Ezagutza basea aztertu ondoren (2 taula), honela etiketatzen da urruneko gainbegiraketaren bidez:

- Update : Earthquake in <region>Baja California< /region> , <country>Mexico< /country> upgraded to <magnitude>7.2< /magnitude> magnitude , from 6.9 - USGS

Guztira 13562 aipamen etiketatu ditu UG sistemak. 2 taularen azken zutabean aurki dezakegu argumentu bakoitza zenbat aldiz etiketatu den txioen datu multzoan.

3.4 Argumentuen kategorizazioa ikasketa automatikoarekin

Gure sistemak, nolabait esateko, *burmuin* bat dauka integratuta, **sailkatzaile** deiturikoa. Sailkatzailearen bidez informazioa kudeatzeko teknikari **ikasketa automatikoa** deitzen zaio. Ikasketa automatikoa bi fasetan dago banatuta:

- **Entrenamendu fasea:** sailkatzailearen eginbeharra etiketatutako txio guztien egitura ikastea da, txioetako elementuen ezaugarri linguistikoak aztertuz, informazio horren eredu bat sortzeko.
- **Iragarpen fasea:** sailkatzaileak beste lurrikara batzuei buruzko txioak jasotzen ditu, etiketatu gabe. Honek ikasitakoa praktikan jarri eta txioetatik informazio garrantzitsua eskuratzen du.

Sailkatzailea entrenatzeko, txio bakoitzaren ezaugarri linguistikoak behar ditugu, sailkatzaileak hauetatik ikas dezan. Horretarako, txioak tokenizatu ditugu, beste era batera esanda, hitzen banaketa bat egin, eta hitz bakoitzaren lema, kategoria gramatikala eta entitate izen mota eskuratu. Ezaugarri linguistikoen sorkuntza Stanfordeko CoreNLP tresnaren[8] bidez egin dugu.

3 taulak aurretik jarri dugun txioaren ezaugarriak irudikatzen ditu, lehenengo zutabeak txioko hitza adierazten du, eta beste zutabeetan hitz bakoitzaren lema, kategoria gramatikala, entitate izen mota eta kategoria ageri dira. Kategoria ezagutza baseko argumentua da.

Ikasketa automatikorako hainbat sailkatzaile desberdin aurki ditzakegu. Bakoitzak ikasketarako bere teknika dauka. Gure esperimentuetarako erabilitako sailkatzailea "Baldintzazko hausazko eremua" da (BHE, ingelesez, *Conditional Random Field*[9]). Sailkatzaile hau etiketatze sekuentzialean oinarritzen da, eta hitz bakoitzaren inguruko hitzak aztertzen ditu datu multzoa entrenatzean, baita hitz baten etiketa iragartzean ere. Aukeratutako BHE sailkatzailea Stanfordeko CoreNLP tresnarena da.

[8] `http://nlp.stanford.edu/software/corenlp.shtml`
[9] Wikipedian: `http://en.wikipedia.org/wiki/Conditional_random_field`

3 Taula: Txio baten aurreprozesaketa. Erabilitako txioa taularen gainean dago. Txioan hitzak banandu dira eta bakoitzarentzat bere lema, kategoria gramatikala eta entitate izen motaren balioak lortu. Azken zutabean, hitzari dagokion kategoria dago, urruneko gainbegiraketaren bidez sistemak etiketatutakoa.

Update : Earthquake in <region>Baja California</region> , <country>Mexico</country> upgraded to <magnitude>7.2</magnitude> magnitude , from 6.9 - USGS

Hitza	**Lema**	**Kat. gram.**	**Entitate izena**	**Kategoria**
Update	Update	NNP	O	O
:	:	:	O	O
Earthquake	earthquake	NN	O	O
in	in	IN	O	O
Baja	Baja	NNP	LOCATION	region
California	California	NNP	LOCATION	region
,	,	,	O	O
Mexico	Mexico	NNP	LOCATION	country
upgraded	upgrade	VBN	O	O
to	to	TO	O	O
7.2	7.2	CD	NUMBER	magnitude
magnitude	magnitude	NN	O	O
,	,	,	O	O
from	from	IN	O	O
6.9	6.9	CD	NUMBER	O
-	-	:	O	O
USGS	usg	NN	ORGANIZATION	O

Sailkatzailea Txinako lurrikara bati buruzko adibide honetatik ahalik eta informazio gehien lortzen saiatzen da:

- Earthquake in western China kills more than 60 - The 7.1 quake struck around 33 km below the surface in Yushu county ... `http://ow.ly/173YIJ`
  (Txinako mendebaldeko lurrikarak 60 pertsona baino gehiago hil ditu - 7.1eko dardara gainazaletik 33 km-ko sakoneran talka Yushu udalerrian ... `http://ow.ly/173YIJ` *)*

Txio honen ezaugarri linguistikoak aztertu ondoren, gai izan beharko litzateke ondorengo informazioa eraUzteko:

Argumentua	**Estatua**	**Herrialdea**	**Hildakoak**	**Magnitudea**	**Sakonera**
Balioa	Txina	Yushu	60 baino gehiago	7.1	33 km

Sailkatzaleak lurrikara konkretu baten txio berri guztiak aztertu ondoren, argumentu bakoitzarentzat iragarpen desberdinak egiten ditu, baina argumentu gehienek balio bakarra onartzen da. Iragarpen egokiena aukeratzeko, *NoisyOR* metodoa erabili dugu. *NoisyOr* egokia da kategorizaziorako, ereduen konfidantza (zenbat eta probabilitate altuagoa, orduan eta puntuazio altuagoa) eta jarioa (zenbat eta aipamen gehiago etiketa baterako iragarri, orduan eta altuagoa izango da etiketaren puntuazioa) ondo orekatzen dituelako:

$$NoisyOr(a, i) = 1 - \prod_{p \in P} (1 - p) \tag{1}$$

a argumentuaren izena da eta i argumentuaren iragarpen potentziala. Txio bakoitzean, sailkatzaileak iragarpen probabilitate bat (p) ematen dio hitz bakoitzari argumentu bakoitzeko. P aldagaiak iragarpen probabilitate guztiak multzokatzen ditu, a argumenturako. Formula hau (Surdeanu *et al.*, 2012) lanetik hartu dugu.

4 Taula: Ebaluazioaren emaitzak.

Sistema	Doitasuna	Estaldura	F1 neurria
UG	50.60	17.79	24.07
Eskuzkoa	47.65	26.69	34.21

3.5 Emaitzak

Entrenamendurako, ezagutza baseko lurrikaren %75a erabili dugu, gainentzekoa ebaluaziorako.

Gure sistemaren emaitzak ebaluatzeko erabili ditugun ebaluazio metrikak doitasuna, estaldura, eta F1 neurria dira.

Doitasunak sistemak itzulitako emaitza zuzenen kopurua itzulitako guztiekin konparatzen du:

$$Doitasuna = \frac{\#(Emaitza\_zuzenak)}{\#(Sistemak\_itzulitako\_emaitzak)} \tag{2}$$

Estaldurak sistemak itzulitako emaitza zuzenen kopurua EBan dauden balio guztiekin konparatzen du:

$$Estaldura = \frac{\#(Emaitza\_zuzenak)}{\#(Asmatu\_behar\_direnak)} \tag{3}$$

Eta **F1 neurria** doitasuna eta estalduraren arteko batazbesteko harmonikoa da:

$$F1Neurria = 2 * \frac{doitasuna * estaldura}{doitasuna + estaldura} \tag{4}$$

Gure sistemaren eraginkortasuna ondo neurtzeko, txio guztiak eskuz etiketatu ditugu, eta ikasketa automatikoa aplikatu, aurretik aipatu dugun metodologia erabiliz. Eskuzko etiketatzearen bidez, gure sistemak lortuko lukeen emaitza onena kalkulatu dezakegu, eta UG sistemak lortutakoarekin konparatu.

4 taulan ikus ditzakegu UG algoritmoaren eta eskuzko etiketatzearen ebaluazioen emaitzak. Gure sistemak estaldura txikiagoa dauka, baina eskuzkoak baino doitasun hobea. F1 neurritik ez gaude urruti.

Ikusten den bezala, urruneko gainbegiraketak potentzial handia dauka gertaeren erauzketarako. Mikroblogak erabiltzea informazioa lortzeko eraginkorra dela frogatu dugu ere.

4 Ondorioak

Artikulu honetan Twitterreko txioetatik lurrikarei buruzko informazioa eskuratzeko sistema bat aurkeztu dugu. Horretarako, urruneko gainbegiraketaren (UG) algoritmoa gertaera erauzketarako moldatu dugu. UG algoritmoak corpusak automatikoki etiketatzen ditu, eskuzko lan garestia ekidituz. Lan honetan frogatzen dugu posible dela UG gertaera erauzketarako ere aplikatzea. Esperimentuetarako aukeratutako domeinua lurrikarena da.

Lan honetan mikroblogetatik gertaerei buruzko informazioa eskuratzea posible dela frogatzen dugu, gertaera erauzketan hauen potentziala argi utziz, nahiz eta hauek hizkera kolokiala, sintaxi zaratatsua eta informazio anbiguoa eduki.

Gure esperimentuetarako lurrikarei buruzko ezagutza base bat sortu dugu. Horrez gain, ezagutza basean dauden lurrikara desberdinei buruzko txioak eskuratu ditugu Twitterretik, gure esperimentuetan erabiltzeko. Ezagutza basea eta txioen datu multzoa publikoki eskuragarri daude [10].

Gure sistemaren eraginkortasuna neurtzeko, eta lortuko lukeen emaitza maximoa jakiteko, txioak eskuz etiketatu genituen. Txio hauen entrenamenduak UGrekin erabilitako txioen urrats berdinak jarraitzen ditu. Guk lortutako emaitzak eskuzkoaren emaitzetatik gertu daude, UG algoritmoak mikroblogetatik gertaerak eraztzeko duen gaitasuna frogatuz.

[10] `https://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1427727946/publikoak/earthquake-kb-dataset.zip`

5 Etorkizuneko lanak

Txioak eskuz etiketatzean, hauetan aurki ditzakegun datuak oso dinamikoak direla konturatu gara. Horrez gain, sistemak iragarritako balio asko ezagutza basekoen oso hurbilak zirela ere. Txioetako informazioa EBkoekiko antzekoa denean, hauek ere etiketatzen egingo ditugu esperimentuak, emaitzak hobetzeko. Ebaluazioan antzeko balioak partzialki ontzat hartzea ere lan polita litzakete.

Sarreran aipatu bezala, txioetan topatu dezakegun informazioa oso anbiguoa da. Anbiguotasun horri aurre egiteko asmoa dugu, lurrikara bereko txioen artean hauen testuingurua elkarbanatuz.

Lan honetan esperimentuak domeinu bakar baterako bakarrik egin ditugu, eta komeni zaigu beste domeinutan frogak egitea. (Reschke *et al.*, 2014) lanerako, hegazkin istripuei buruzko ezagutza base bat sortu zuten; EB hau aprobetxatu dezakegu Twitterretik istripu hauei buruzko txio desberdinak eskuratu eta gure esperimentuak errepikatzeko.

Amaitzeko, Interesgarria litzateke gure sistema moldatzea informazioa Twitterretik denbora errealean eskuratzeko.

Erreferentziak

Benson, Edward, Aria Haghighi, eta Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Intxaurrondo, Ander, 2015. *Ezagutza baseen aberasketa urruneko gainbegiraketaren bidez: analisiak eta hobekuntzak*. Euskal Herriko Unibertsitatea tesia.

Mintz, Mike, Steven Bills, Rion Snow, eta Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

Reschke, Kevin, Martin Jankowiak, Mihai Surdeanu, Christopher D. Manning, eta Daniel Jurafsky. 2014. Event extraction using distant supervision. In *Proceedings of LREC*.

Surdeanu, Mihai, Julie Tibshirani, Ramesh Nallapati, eta Christopher D. Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (EMNLP-CoNLL)*.

6 Eskerrak eta oharrak

Arizonako Unibertsitateko Mihai Surdeanu ikerlariari eskerrak eman nahi dizkiogu, lan honetan eman digun laguntzagatik.

Esker instituzionalak Eusko Jaurlaritzako Hezkuntza, Unibertsitate eta Ikerketa Sailari, ikerketa lan hau egiteko emandako ikertzaileak prestatzeko bekarengatik.

Lan hau egile nagusiaren tesiaren eratorria da (Intxaurrondo, 2015).

Irudiaren iturria: `http://zthiztegiberria.elhuyar.org/artikuluak/Lurrikara`