IKER GAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

III. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2019ko maiatzaren 27, 28 eta 29
Baiona, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

GIZA ZIENTZIAK ETA ARTEA

EusTimeBank-TL corpusa: denbora-informaziodun testuetatik denbora-lerroetara

Begoña Altuna,
María Jesús Aranzabe eta
Arantza Díaz de Ilarraza

83-90 or.
https://dx.doi.org/10.26876/ikergazte.iii.01.11

EusTimeBank-TL corpusa: denbora-informaziodun testuetatik denbora-lerroetara

Altuna, Begoña[1]; Aranzabe, María Jesús[1] eta Díaz de Ilarraza, Arantza[1]

Ixa Taldea
Euskal Herriko Unibertsitatea

begona.altuna@ehu.eus

Laburpena

Ikerketa-lan honen helburua da denbora-informazioa etiketatuta duen EusTimeBank corpusa oinarri hartuta, denbora-lerroak eskuz etiketatuta dituen EusTimeBank-TL corpusa sortzea. KroniXa bezalako tresnek automatikoki sortzen dituzte denbora-lerroak eta horiek ebaluatzeko sortu da corpus hori. Lan honetan corpusa eraikitzeko egindako urratsen berri ematen da. Urrats horiek dira gertaerak kronologiako uneetan kokatzeko jarraitu beharreko irizpideak zein diren azaltzea, irizpide horiei jarraituz bi etiketatzailek egindako lana alderatzea eta ebaluatzea, eta, behin hori guztia eztabaidatuta, urre-patroia osatzea eta zein erabilera duen azaltzea.

Hitz gakoak: Denbora-informazioaren prozesamendua, denbora-lerroak, urre-patroia, etiketatzearen azterketa

Abstract

The goal of our work is creating the manually annotated EusTimeBank-TL timeline corpus based on the temporally annotated EusTimeBank corpus. Tools such as KroniXa automatically generate timelines and the corpus has been created for tehir evaluation. In this work we describe the process of the creation of the corpus. Those steps are defining the annotation guidelines for anchoring events to time points, comparing and assessing the annotation work two annotators have done following those guidelines and, after a discussion has been conducted, building the gold standard database and describing its uses.

Keywords: Temporal information processing, timelines, gold standard, annotation analysis

1. Sarrera eta motibazioa

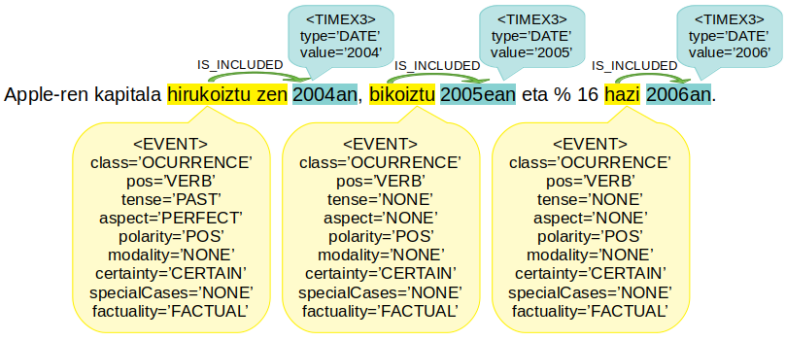
Hizkuntzaren Prozesamenduko (HP) ataza nagusietako bat testuetako informazioa automatikoki erauztea da. Horretarako, tresna automatikoak garatu behar dira eta, askotan, horiek garatzeko oinarrian ikasketa algoritmoak daude. Ikasketa algoritmoek testuetako informazio esanguratsua markaketa-lengoaien bidez etiketatuta duten testuak (urre-patroiak) behar dituzte. Horretan datza corpus etiketatuen interesa. Gainera, corpus horiek garatutako tresnak ebaluatzeko baliabide garrantzitsuak dira.

Denbora-informazioaren kasuan, *zer noiz* gertatzen den identifikatzen da. Esaterako, (1) adibideko esaldian Apple-ren kapitalak jasandako hiru hazkunde (*hirukoiztu*, *bikoiztu* eta *%16 hazi*) eta bakoitza noiz gertatu zen (*2004*an, *2005*ean eta *2006*an, hurrenez hurren) ageri dira.

(1) Apple-ren kapitala **hirukoiztu** egin zen *2004an*, **bikoiztu** *2005ean* eta % 16 **hazi** *2006an*.

Informazio hori HPko tresnekin baliatu ahal izateko, denbora-informazioa etiketatzeko markaketa-lengoaien bidez etiketatu behar da. 1. irudian, (1) adibideko esaldiaren etiketatzea irudikatu dugu EusTimeML markaketa-lengoaia (Altuna *et al.*, 2016) erabilita. Ikus daitekeenez, gertaerek `<EVENT>` etiketa hartzen dute eta denbora-adierazpenek, `<TIMEX3>`. Etiketa bakoitzak atributu zerrenda bat hartzen du eta horien bidez gertaeren edo denbora-adierazpenen atributuak (mota eta balio normalizatuak, besteak beste) esplizitu egiten dira. Halaber, gertaeren eta denbora-adierazpenen artean aldiberekotasuna (*IS_INCLUDED*) adierazten duten denbora-erlazioak etiketatu dira.

1. irudia. (1) adibideko esaldiaren etiketatzea EusTimeMLren bidez



Informazio hori baliatuta, 2. irudian ageri den denbora-lerroa sor daiteke. Denbora-lerro hori eraikitzeko, (1) adibideko esaldiko gertaerak gertatzen diren uneetara ainguratu edo lotu ditugu. Horretarako, zein gertaera zein unetan gertatu den kontuan izan dugu.

2. irudia. (1) adibideko esaldiko denbora-informazioan oinarrituta sortutako denbora-lerroa

Euskarazko denbora-informazioaren prozesamenduan, EusTimeML markaketa-lengoaia definitzeaz gain, horri jarraituta, EusTimeBank corpusa (Altuna, 2018) sortu dugu eta euskarazko denbora-informazioa automatikoki erauzten duten EusHeidelTime (Altuna *et al.*, 2017) eta bTime (Salaberri Izko, 2017) tresnak garatu ditugu. Zehazki, EusHeidelTimek denbora-adierazpenak identifikatzen eta sailkatzen ditu, eta ISO-8601 arauaren araberako balio normalizatua (ISO-TimeML working group, 2008) esleitzen die; bTimek, berriz, gertaerak eta denbora-erlazioak identifikatzen eta sailkatzen ditu.

EusHeidelTimek eta bTimek denbora-informazioa etiketatuta duten testuak itzultzen dituzte. KroniXa sistemak, etiketatuta dagoen informazio horretan oinarrituta, testuetako gertaerak ardatz kronologikoan kokatzen ditu. Lan honetan, KroniXa ebaluatzeko sortu dugun denbora-lerroak eskuz etiketatuta dituen EusTimeBank-TL corpusa nola garatu dugun deskribatuko dugu.

2. Arloaren egoera eta ikerketaren helburuak

Denbora-informazioa automatikoki tratatzeko sistemek informazio linguistikoa etiketatuta duten corpusak hartzen dituzte oinarritzat. Hain zuzen ere, mota horretako corpusen bitartez tresna horiek entrenatu eta ebaluatu egiten dira. KroniXa garatzeko eta ebaluatzeko baliabideak aukeratzeko, beste hizkuntzetan denbora-lerroen inguruan egin diren lanak aztertu ditugu .

2.1. Arloaren egoera

Denbora-lerroak automatikoki sortzeko saiakerak nabarmen ugaritu dira denbora-informazioa erauzteko sistemek denbora-informazio normalizatua erabilgarri egin dutenean. Esaterako, ikergaiaren gaurkotasunaren eredu da 2015ean SemEval ebaluazio-saioan egin zen denbora-lerroak sortzeko ataza (Minard *et al.*, 2015). Ataza horretan, hainbat sistema lehiatu ziren atazan definitutako denbora-lerroak eraikitzen eta, gaur egun, denbora-lerroak ebaluatzeko erabiltzen den metrika definitu zen. Ebaluazio-metrikak *Temporal Awareness Score* metrika (UzZaman *et al.*, 2012) eta gertaeren posizio erlatiboak hartzen ditu kontuan.

Gerora ere ugaritu dira denbora-lerroak automatikoki sortzeko sistemak. Reimers et al.-ek (2016) testuko gertaerak gertatzen diren egunetan ainguratzen edo kokatzen dituen sistema garatu dute. Cheng et al.-ek (2018) ere kronologiako uneen balio absolutuak hartzen dituzte denbora-lerroak sortzeko oinarri gisa. Leeuwenberg et al.-ek (2018), aldiz, denbora-lerroak sortzen dituzte gertaeren arteko posizio erlatiboetan oinarrituta.

Euskarari dagokionez, KroniXa sistemak testuko gertaerak kronologiako uneetan kokatzen ditu. Horretarako, oinarritzat hartzen ditu batetik, bTimek eta EusHeidelTimek lortutako denbora-informazioa eta, bestetik, euskararen prozesamendu-katetik (Otegi *et al.*, 2016) lortutako dependentzia sintaktikoen informazioa.

Testu bakarretik denbora-lerroak sortzen dituzten sistemak ebaluatzeko, ez da oraindik erabilera zabaleko urre-patroirik sortu. Ondorioz, tresnak ebaluatzeko unean uneko urre-patroiak sortu dira. Hain zuzen ere, urre-patroi funtzioa betetzen duen EusTimeBank-TL sortu dugu KroniXa ebaluatzeko.

2.2. Helburuak

Lan honen helburu nagusia da euskarazko denbora-lerroen EusTimeBank-TL corpusa sortzea. EusTimeBank corpusean (Altuna, 2018) etiketatuta dagoen denbora-informazioa (gertaerak, denbora-adierazpenak eta horien arteko denbora-erlazioak) oinarritzat hartuta, gertaerak kronologiako uneetan eskuz ainguratu dira. EusTimeBank-TL urre-patroitzat hartuko dugu eta denbora-lerroak automatikoki sortzen dituen KroniXa tresna ebaluatzeko erabiliko dugu. Helburu nagusi hori lortzeko, jarraian zerrendatu ditugun urrats hauek definitu ditugu:

- etiketatze-ataza eta irizpideak definitzea
- etiketatzaileen lanak ebaluatzea eta konparatzea
- denbora-lerroak sortzeko irizpide bateratuak definitzea eta urre-patroia sortzea

Egin ditugun urrats horiek hurrengo atalean deskribatuko ditugu.

3. Ikerketaren muina: testuetako denbora-lerroen eskuzko etiketatzea

Urre-patroitzat hartzen den EusTimeBank-TL corpusa sortzeko definitu ditugun etiketatze-ataza eta -irizpideak azalduko ditugu atal honetako 3.1 azpiatalean. Ondoren, bi etiketatzailek sortu dituzten denbora-lerroak konparatuko eta ebaluatuko ditugu 3.2 azpiatalean eta, jarraian, etiketatzaileen lanen azterketan identifikatutako zailtasunak kontuan hartuta, urre-patroia sortzeko prozesua deskribatuko dugu 3.3 azpiatalean.

3.1. Etiketatze-atazaren eta irizpideen definizioa

EusTimeBank corpusaren ebaluaziorako azpicorpusa osatzen duten albisteak erabili ditugu denbora-lerroak sortzeko. Corpus hori EusTimeMLri jarraituta etiketatuta dagoenez[1], markaketa-lengoaia horrek eskaintzen duen informaziotik gertaerak, denbora-uneak (egunak, orduak, etab.) adierazten dituzten denbora-adierazpenak eta horien artean sortzen diren denbora-erlazioak baino ez ditugu kontuan hartu.

Zehazki, hauek dira etiketatzaileei eman zaizkien irizpideak gertaera eta denbora-adierazpenak denbora-lerroetan koka ditzaten:

- Testuko informazioa baino ezin da erabili denbora-lerroan eta ezin da testuan ez dagoen informaziorik gehitu.
- Testu etiketatuan gertaera etiketa duen elementu oro agertuko da denbora-lerroan.
- Testuan uneak (datak eta orduak) adierazten dituzten denbora-adierazpenak baino ez dira denbora-lerroko aingurak edo ainguratze-puntuak izango.
- Testuan gertaera zein unetan gertatu(ko) den agertuz gero, gertaera une horretan ainguratuko da.
- Testuko gertaera testua idatzi den unean gertatzen bada edo egia bada, gertaera hori dokumentuaren sorreradatari (*DCT, Document Creation Time*) lotuko zaio.
- Gertaera ezin bazaio aingura zehatz bati lotu, XXXX-XX-XX aingurari lotuko zaio, betiere denbora-lerroan kokagune egokiena mantenduz.

Zer-nolako denbora-lerroak sortu definitu ondoren, bi etiketatzailek irizpide horiei jarraituta denbora-lerroak sortu dituzte. Horretarako, aurrez prestatutako kalkulu-orriak erabili dituzte, formatuan bat etortzea errazteko. Esperimenturako erabili ditugun testuak eta denbora-lerroak luzera ezberdinetakoak badira ere, denbora-lerro bakoitza definitutako formatuan sortzeak batez besteko ordubeteko lana eskatu du.

[1] Ikus 1. irudiko etiketak, atributuak eta horien balioak.

Eskuzko denbora-lerro baten eredua jarri dugu 3. irudian, eta urdin argiz markatu ditugu (1) adibideko aingura-gertaera erlazioak. Ikus daitekeenez, lehenengo zutabean unea-gertaera pareak denbora-lerroan hartzen duen posizioa (*Position ID*) definitu da[2]. *Timex Token*, *Timex ISO Value* eta *Timex Form* zutabeek denbora-uneari buruzko informazioa dute: denbora-unea testuan zein tokenek[3] adierazten duten, ISO-8601 ereduari jarraituta hartzen duen balio normalizatua eta denbora-adierazpenaren forma, hurrenez hurren. Jarraian, gertaera adierazten duen tokenaren identifikadorea (*Event Token*) eta testuan duen forma (*Event Form*) daude. Amaitzeko, eztabaidan baliagarriak izango diren iruzkinak gehitzeko tartea ageri da *Notes* zutabean.

3. irudia. Denbora-lerro baten eskuzko etiketatzea

Position ID	Timex Token	Timex ISO Value	Timex Form	Event Token	Event Form	Notes
1	w229	2004	2004an	w226	hirukoiztu	
2	w232	2005	2005ean	w231	bikoiztu	
3	w62,w63	FY2006	tarte berean	w64	irabazitako	
4	w91	2006	2006an	w89	igo	
4	w237	2006	2006an	w236	hazi	
5		XXXX-XX-XX		w123	iragarri	hemen?
5		XXXX-XX-XX		w155	iragarria	korref w123
5		XXXX-XX-XX		w156	zuen	
6	w52,w53,w54	FY2007	urte ekonomikoaren hasieran	w55	bikoiztu	ondorioa geroago dago
6		FY2007		w59	salmentak	
6		FY2007		w76	irabazita	
6		FY2007		w79	hautsi	
6		FY2007		w133	hobeak	
6		FY2007		w134	izan	
6		FY2007		w142	izan	
6		FY2007		w153	izango	
7	w158,w159,w	FY2007-Q1	urte ekonomikoko lehen hiruhilekoan	w179	saldu	berez F2007-Q1 eta 2006-Q4 berdinak dira
7	w158,w159,w	FY2007-Q1	urte ekonomikoko lehen hiruhilekoan	w182	horrek	
7	w158,w159,w	FY2007-Q1	urte ekonomikoko lehen hiruhilekoan	w186	salmentak	hemen?
7	w158,w159,w	FY2007-Q1	urte ekonomikoko lehen hiruhilekoan	w190	hazi	hemen?
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w44	izan	
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w42	eskari	
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w17	hazi	
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w29	igoera	korref w17
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w100	hazi	korref w17
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w103	ekoiztutako	hemen?
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w108	egon	
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w239	hazkunde	hemen?
8	w19,w20	2006-Q4	azkeneko hiruhilekoan	w244	salmenta	hemen?
9	w47	2006-12	Gabonetako	w48	salmenta	
10	w23,w24,w25	2006-12-30	2006ko abenduaren 30ean	w26	amaitu	

Formatu hori egokia da konparazioa errazteko, amaierako denbora-lerroan agertzen den informazioaz gain, informazio gehigarria ere baduelako. Hala, automatikoki sortutako denbora-lerroak ebaluatzeko, eskuzko etiketatzetik informazioa erauzi eta SemEval-2015ko atazan definitutako formatura moldatu dugu. Formatu horretan, aingura-gertaera erlazioak kronologikoki ordenatuta adierazten dira eta une berean gertatzen diren gertaerak multzokatzen dira. SemEvaleko formatu hori etiketatzaileen lana ebaluatzeko ere erabili dugu.

3.2. Etiketatzaileen lanen ebaluazioa eta konparaketa

Etiketatze-lana egin ondoren, sortutako denbora-lerroak aztertu, aspektu korapilatsuak komentatu eta ezberdintasunak eztabaidatu ditugu. Horretarako, etiketatzaileek ebaluazio-corpuseko 7 dokumentutan egindako lana konparatu dugu.

Etiketatzaileak bat etorri diren ikusteko, lehenik bi etiketatzaileen lanen arteko adostasuna neurtu dugu SemEval-2015eko atazako metrika erabilita. Horren bidez, dokumentu mailako gertaeren ordenaren doitasuna, estaldura eta horien arteko batez besteko harmonikoa (F neurria) neurtu ditugu. 1. taulan ikus daitekeenez, doitasuna % 61,83koa da, estaldura % 62,61koa eta F neurria % 62,22koa.

1. taula. Etiketatzaileen arteko adostasuna

Doitasuna	Estaldura	F neurria
61,83	62,61	62,22

Emaitza horiek erakusten dute testuetako informazioaren parte handi bat berdin antolatu dutela etiketatzaileek. Hala ere, agerian geratu da testuak interpretatzeko aukera bat baino gehiago izaten dela. Hain zuzen ere, ez da adostasun osorik egon denbora-lerro bat eraikitzean ere. Hala, desadostasunak zein izan diren ere aztertu dugu.

[2]Denbora-aingura bakoitzak posizio bat hartzen du eta aldi berean gertatzen diren gertaerek posizio bera hartzen dute.

[3]Tokenak esanahi osoa duten unitate minimoak dira; hau da, hitzak ez ezik puntuazio-markak, zenbakiak, laburtzapenak edo antzeko beste edozein karaktere.

Denbora-lerroak sortzean gertaerak denbora-aingurei lotu dizkiegunez, azterketaren lehen urratsean aingura berak identifikatu diren aztertu dugu. Horretarako, testuan esplizitu agertzen diren denbora-ainguretatik denbora-lerroetan zein kokatu diren identifikatu dugu. Orokorrean, aingura berak identifikatu dituzte bi etiketatzaileek eta kasu bakarrean baino ez dugu aurkitu unea adierazteko murriztapena betetzen ez duen denbora-aingura bat.

Hala eta guztiz ere, aingura egokiak denbora-lerroan kokatzean, anbiguotasuna dagoela ikusi ahal izan dugu. Arestian 3. irudian ikusi bezala, aingura batzuek ISO-8601 balio zehatza hartzen dute eta beste batzuek XXXX-XX-XX moduko balio zehaztugabeak. Balio esplizituei dagokienez, etiketatzaileek granularitatearen araberako ordena (unitate handienekoak aurretik) zaindu dute eta, ondorioz, balio esplizitua duten ainguruak beren artean ondo ordenatu dituzte denbora-lerroa orraztean egindako akats bat salbu.

Aingura zehatzak kokatzeko arazorik egon ez bada ere, erreferentzia lausoa adierazten duten aingurak ordenatzean desadostasunak ikusi ditugu. Esaterako, 4. irudian adierazi nahi izan dugunez, (2) adibideko denbora-adierazpenak ezberdin ordenatu dituzte. Etiketatzaile batek (A) *aspaldian* unea (2. posizioan) *uztailaren hondarrean* (4. posizioan) baino atzerago kokatu du eta beste etiketatzaileak (B) kontrako ordena interpretatu du. Izan ere, *aspaldian* moduko une zehaztugabeak irakurlearen interpretazioari lotuta daude.

(2) Kataluniako enpresan **aspaldian** da lanean (. . .) **uztailaren hondarrean** erabakitako adjudikazioa kontratuan ofizial egin behar da.

4. irudia. Balio zehaztugabeen eragina ainguren ordenan

Etik.	Position ID	Timex Token	Timex ISO value	Timex Form	Event Token	Event Form
A	2	w158	XXXX-XX-XX	aspaldian	w159	da
	4	w214,w215	2014-07	uztailaren hondarrean	w216	erabakitako
B	2	w214,w215	2014-07	uztailaren hondarrean	w216	erabakitako
	3	w158	XXXX-XX-XX	aspaldian	w159	da

Era berean, kontuan izan behar da XXXX-XX-XX bezalako aingura zehaztugabeen anbiguotasunak aingura-gertaera pareen posizio absolutuetan duen eragina. 5. irudian bi etiketatzaileen lanen laginak parean jarri ditugu. Ikus daitekeenez, biek ordenatu dituzte zuzen *ekainaren hasierako* eta *ekainaren 3an* ainguruak baita horiei gertaera berak ainguratu ere. Etiketatzaile batek (B), ordea, bi aingura horien artean zehaztugabeko *XXXX-XX-XX* ainguradun erlazioa gehitu du eta horrek ondorengoen posizio absolutuak baldintzatu ditu.

5. irudia. Balio zehaztugabeen eragina ainguren ordenan

Etik.	Position ID	Timex Token	Timex ISO value	Timex Form	Event Token	Event Form
	3	w29,w30	2009-06	ekainaren hasierako	w26	kendu
	3	w29,w30	2009-06	ekainaren hasierako	w11	kenduko
A	3	w29,w30	2009-06	ekainaren hasierako	w27	nahi
	4	w54,w55	2009-06-03	ekainaren 3an	w56	izango
	4	w54,w55	2009-06-03	ekainaren 3an	w58	entzunaldia
	3	w29,w30	2009-06	ekainaren hasierako	w26	kendu
	3	w29,w30	2009-06	ekainaren hasierako	w27	nahi
B	4		XXXX-XX-XX		w11	kenduko
	5	w54,w55	2009-06-03	ekainaren 3an	w56	izango
	5	w54,w55	2009-06-03	ekainaren 3an	w58	entzunaldia

Gertaeren ainguratzeari ere erreparatu diogu. Zehazki, gertaerak aingura berei lotu zaizkien aztertu dugu. Testuan esplizituki agertzen diren aingurak eta horiei esplizituki lotzen zaizkien gertaerak denbora-lerroan kokatzea ez da lan zaila eta era bertsuan identifikatu dituzte bi etiketatzaileek. Hala ere, hain lotura zuzena ez duten gertaerak ainguratzean ezberdintasun nabariak izan dira. Arazoa azaltzeko, (3) adibideko esaldia hartu dugu adibide gisa.

(3) ... *ostiralean* **jakinarazi** zuen bere kontzesionarioetakoa 1.100 inguru ixteko **asmoa zuela**.

(3) adibidean ikus daitekeenez, *ostiralean* izan zen *jakinarazpena*. *Asmoa izatea*, ordea, data berean (*ostiralean*) edo beste une zehaztugabe batean gertatzen dela interpretatu dute etiketatzaileek.

Aipatu bezala, oro har adostasun nabarmena egon da, baina hizkuntzaren beraren anbiguotasunak eragin handia izan du eta desadostasunerako patroiak aurkitzen saiatu gara. Etiketatzaileen lana aztertzean, hiru anbiguotasun-

fenomeno nagusi identifikatu ditugu: i) iraupen luzeko gertaeren ainguratzea, ii) korreferenteak diren gertaera-adierazpenak ainguratzea eta iii) diskurtso zuzen zein zeharkakoan agertzen diren gertaerak ainguratzea. Halaber, mundu ezagutzaz baino ezin desanbigua daitezkeen ainguraketak ere identifikatu ditugu.

Lehen kasuistikari dagokionez, denbora-lerroak eraikitzeko gidalerroetan adierazten da i) gertaera gertatzen den denbora-uneari ainguratuko zaiola eta ii) gertaera aingura bakarrari lotuko zaiola. Erregela horien arabera, iraupen luzeko gertaerak denbora-lerroko hainbat aingurari lot dakizkioke, baina bakarra aukeratu behar da.

(4) *Ekainaren 1a* bitarteko epea **du**.

Esaterako, (4) adibidean denboran luzatzen den *epea izatea* agertzen da. Izan dokumentua sortu zen datan zein *ekainaren 1ean*, gertaera horrek egia-balioa du eta zilegi da biei lotzea. Etiketatzaileek ere hala jokatu dute eta bakoitzak *du* gertaera aingura bati, sorrera-datari edo *ekainaren 1*ari, lotu dio.

Korreferenteak diren gertaerak ainguratzean ere aingura erabakitzeko irizpide desberdinak hartu dituzte etiketatzaileek. Gertaera bat aingura bakarrari lot dakiokeenez, gertaera horri erreferentzia egiten dioten adierazpen guztiak aingura berari lotu behar zaizkio. Kasu gehienetan ez da arazorik egon, baina korreferentzia ebazteko kasu zailak ere identifikatu ditugu. Esaterako, (5) adibideko *ofizialtzeko* eta *hori* gertaera beraren bi adierazpen dira, baina aingura banari (kurtsibaz) lotuta daude testuan. Gertaera-adierazpen horiek denbora-lerroan zein aingurari lotu erabakitzea zaila da, kontuan izanda balizko testuinguruetan daudela gertaera horiek.

(5) Bada aukeraren bat (...) *bihar bertan* **ofizialtzeko** (...), baina oso litekeena da **hori** *datorren astean* gertatzea.

Bestalde, bi gertaera-adierazpenen artean egiazko korreferentzia-erlaziorik dagoen ere aztertu dugu. Izan ere, gertaera bera den edo mota bereko beste gertaera bati erreferentzia egiten zaion erabakitzea ez da beti erraza. Adibidez, (6) adibidean *kontzesionarioak ixtea* aipatzen da. Batetik, etorkizunean egingo den gertaera da eta bestetik duela urte asko egin beharrekoa. Ondorioz, gertaera bera izanagatik ere, bi gauzatze dagoela interpreta daiteke (bat errealitatean egin ez bazen ere). Sinpletasunaren izenean, ordea, egokiena da gertaera bakartzat hartzea eta, ondorioz, aingura berari lotzea.

(6) ...*kontzesionarioetatik 1.100 ixteko* asmoa duela ... Jendeak pentsa dezake duela urte asko egin beharrekoa zela **hori**.

Testuan agertzen diren aipuetako gertaerak ainguratzean ere desadostasunak sortu dira. Diskurtsoa aztertzean, testuaren sorrera uneaz gain, iterazioaren unea ere kontuan hartu behar da. Horrek gertaeren ainguruak definitzean desadostasunak ekarri ditu, zailagoa egiten baita gertaerak zein uneren arabera ainguratu behar diren erabakitzea.

(7) Obamaren administrazioak *esan zuen* ez ziola diru-laguntza gehiagorik **emango** enpresari (...) betetzea adosten ez badute.

(7) adibideko esaldian *esan zuen* aditzak diskurtsoa sartzen du testuan; hain zuzen ere, diru-laguntzak ez emateari buruzko mezua. Etiketatzaileetako batek *emango ziola* lehenaldian (dokumentua sortu baino lehen) kokatu du eta besteak, etorkizunean (dokumentua sortu ondoren); aditz-denborak inplika dezakeena desberdin interpretatuta.

Amaitzeko, etiketatze-lanean azaleratutako zalantzak aztertu ditugu. Etiketatze-lanen azterketan ikusi bezala, hizkuntzaren beraren anbiguotasuna izan da atazaren zailtasuna justifikatzeko faktorerik garrantzitsuena. Beste faktore batzuek ere eragina izan dute, ordea. Zehazki, kazetaritza-testuen formak ere eragina izan du. Esaterako, albisteetan ohikoa izaten da, tituluaren ondoren, testu-gorputzeko informaziorik garrantzitsuena *lead* delakoan adieraztea. Gure corpuseko testuetan, *lead*eko aditzen denbora eta gorputzekoa ez dira beti bat etorri eta horrek gertaera bera bi aditz-denboraren bidez adieraztera eraman du.

(8) Akordio bat **egin zuen** Fiat autogilearekin. (...) Bitartean, akordio bat **egin du** Fiatekin.

Adibidez, (8) adibideko *egin zuen* eta *egin du* aditzek gertaera berari egiten diote erreferentzia. Aditz-denborari erreparatuz gero, ordea, badirudi lehena testuaren sorrera-unetik urrunago gertatu zela bigarrena baino, eta bi gertaera ezberdin direla pentsa daiteke.

Kazetaritza-testuen beste ezaugarri bat denbora-adierazpenen trataera da. Corpusean hainbat aldiz identifikatu dugu asteko eguna erabili dela testua publikatu den egunari erreferentzia egiteko. Hiztunok joera dugu asteko eguna eta testua ekoitzi den eguna bat datozenean, asteko egunak astebete lehenagoko edo ondorengo egunari erreferentzia egiten diotela pentsatzeko.

(9) General Motorsek **ostiralean** jakinarazi zuen kontzesionarioetako 1.100 inguru ixteko asmoa duela. (Sorrera data: 2009-05-15, ostirala)

Esaterako, (9) adibideko esaldia irakurtzean, *ostiralean* denbora-adierazpenari *2009-05-08* balioa esleituko genioke bai aditza lehenaldi burutuan dagoelako, bai iterazioaren egunari erreferentzia egiteko asteko eguna erabiltzea arrotz zaigulako. Testuingurua ezagututa,ordea, data zuzena *2009-05-15* dela egiazta dezakegu.

Hala, etiketatzeak konparatu, desadostasunak aztertu eta zailtasunak azaleratu ondoren, urre-patroia sortu dugu.

3.3. Etiketatze-irizpide bateratuen definizioa eta urre-patroiaren sorrera

Bi etiketatzaileen lana konparatu ondoren, urre-patroia bi fasetan sortu dugu. Lehenean, etiketatzearen azterketan landutako dokumentuak egokitu ditugu. Bi etiketatzaileen lanak kontuan hartuta, desadostasunak banan-banan aztertu eta horiek era bateratuan tratatzeko irizpideak finkatu ditugu gidalerroak eguneratzeko. Hala ere, etiketatzeen analisia egin dugunean ikusi bezala, batzuetan zenbait gertaeraren arteko ordena anbiguoa da eta kasu horietan bi etiketatzaileen artean adostutako ordena aukeratu dugu.

Bigarren fasean, gainontzeko denbora-lerroak izan ditugu aztergai. Etiketatzaile baten lana oinarritzat hartuta, gidalerro berrien araberako berreraikitzea egin dugu. Zehazki, berriz testua hartuta denbora-lerroko aingurak, gertaerak eta horien arteko ordena zuzena den egiaztatu dugu. Anbiguotasun-kasuetan bi etiketatzaileek adostutako ordena aukeratu dugu.

Denbora-lerroetako informazioa zuzena dela iritzi dugunean, ebaluazio-sistemak onartzen duen formatura bihurtu ditugu urre-patroiko denbora-lerroak KroniXak sortutakoak ebaluatu ahal izateko.

4. Ondorioak

Lan honen helburua automatikoki sortutako denbora-lerroak ebaluatzeko urre-patroia sortzea zen. Teknikoki zailtasunak gutxi izan badira ere, anbiguotasuna ebazteak edota testuetako informazio inplizitua interpretatzeak nabarmen oztopatu dute etiketatze-lana unean uneko erabakiak hartu behar izan baitira.

Testuak interpretatzeko anbiguotasunak agerian utzi ditu zailtasun batzuk. Denbora-lerroen azterketak erakusten duenez, testu beretik pertsona desberdinek denbora-lerro bera sortzea zaila da. Izan ere, gertaeren arteko segida logikoa (kausa-ondorio erlazioak, denbora-erlazio esplizituak, etab.) kontuan izanda ere, gertaera batzuen arteko ordena hainbat modutan interpreta daiteke.

Ondorioz, hizkuntzaren anbiguotasunagatik eta mundu errealeko informazioa eskuratu ezinagatik, gertaerak ordenatzean, bi interpretazio onargarri dagoela ikusi dugu. Gauzak hala, denbora-lerroen ebaluazioari dagokionez, ezin esan dezakegu eztabaidaren ondoren osatu dugun urre-patroia denik testuko denbora-informazioa adierazteko era zuzen bakarra.

Era berean, aukera zuzen bat baino gehiago egoteak orain arte baliatutako denbora-lerroen ebaluazio-erabakiak ezbaian jartzen ditu eta erronka berriak azaleratzen ditu, gaur egun arte automatikoki sortutako denbora-lerroak ordena bakarra onartzen duen urre-patroi bakarrarekin konparatu izan baitira. Etorkizunean, egoera anbiguoak okertzat hartzen ez dituzten ebaluazio-neurriak eta sistemak sortu behar izango dira eta urre-patroietan aukera aniztasuna nola islatu erabaki behar izango da.

5. Etorkizunerako planteatzen den norabidea

Denbora-lerroen urre-patroia KroniXaren funtzionamendua ebaluatzeko sortu dugu. Zehazki, urre-patroitzat hartu dugun corpusa tresnak automatikoki sortutako denbora-lerroekin erkatzeko erabiliko dugu. Hala ere, KroniXa hobetzen jarraituko badugu, tarteko ebaluazioak egiteko corpusa sortu behar izango dugu, tresna horrekin ebaluatzeko eta ebaluaziorako denbora-lerroekiko gehiegizko egokitzapena (*overfitting*) saihesteko.

Denbora-lerroen ebaluazioan aurrera egiteko, anbiguotasuna nola islatu da etorkizuneko erronkarik handiena.

Alde batetik, anbiguotasuna kontuan hartzen duen urre-patroia nola sortu erabaki behar izango dugu. Bestetik, anbiguotasun hori ontzat hartuko duen ebaluazio-sistema definitu behar izango dugu.

KroniXak testuetako denbora-lerroak sortzen dituenez, erabilgarria izan daiteke Humanitate Digitaletako (HD) hainbat alorretan. Esaterako, testuen laburpena egiteko, gai bati buruzko iturri desberdinetako informazioa kronologian kokatzeko edota infografiak sortzeko oinarria izan daiteke.

6. Erreferentziak

Altuna, Begoña, 2018. *Denbora-informazioaren azterketa eta corpusaren sorrera*. Donostia: Euskal Hizkuntza eta Komunikazioa saila, Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU) tesia.

——, María Jesús Aranzabe, eta Arantza Díaz de Ilarraza. 2016. Euskarazko denbora-egiturak etiketatzeko gidalerroak v2.0. Technical report, Lengoaia eta Sistema Informatikoak Saila, UPV/EHU. UPV/EHU/LSI/TR;01-2016. https://addi.ehu.es/handle/10810/17305.

——, María Jesús Aranzabe, eta Arantza Díaz de Ilarraza. 2017. EusHeidelTime: Time Expression Extraction and Normalisation for Basque. *Procesamiento del Lenguaje Natural* 59.15–22. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5488.

Cheng, Fei, eta Yusuke Miyao. 2018. Inducing Temporal Relations from Time Anchor Annotation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1833–1843. Association for Computational Linguistics.

ISO-TimeML working group. 2008. Language resource management – Semantic Annotation Framework (SemAF) – Part 1: Time and events. International Standard ISO/CD 24617-1(E), International Organization for Standardization, Switzerland. http://lirics.loria.fr/doc_pub/SemAFCD24617-1Rev12.pdf.

Leeuwenberg, Artuur, eta Marie-Francine Moens. 2018. Temporal Information Extraction by Predicting Relative Time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1237–1246, Stroudsburg, PA, USA. Association for Computational Linguistics. http://aclweb.org/anthology/D18-1155.

Minard, Anne-Lyse, Manuela Speranza, Eneko Agirre, Itziar Aldabe, Marieke van Erp, Bernardo Magnini, German Rigau, eta Ruben Urizar. 2015. SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 778–786, Denver, Colorado. Association for Computational Linguistics.

Otegi, Arantxa, Nerea Ezeiza, Iakes Goenaga, eta Gorka Labaka. 2016. A Modular Chain of NLP Tools for Basque. In *Proceedings of the 19th International Conference on Text, Speech and Dialogue — TSD 2016, Brno, Czech Republic*, ed. by Petr Sojka, Aleš Horák, Ivan Kopeček, eta Karel Pala, volume 9924 of *Lecture Notes in Artificial Intelligence*, 93–100. Springer International Publishing.

Reimers, Nils, Nazanin Dehghani, eta Iryna Gurevych. 2016. Temporal Anchoring of Events for the TimeBank Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2195–2204, Berlin, Germany. Association for Computational Linguistics.

Salaberri Izko, Haritz, 2017. *Rol semantikoen etiketatzeak testuetako espazio-denbora informazioaren prozesamenduan daukan eraginaz*. Donostia, Euskal Herria: Universidad del País Vasco/Euskal Herriko Unibertsitatea tesia.

UzZaman, Naushad, Hector Llorens, James F. Allen, Leon Derczynski, Marc Verhagen, eta James Pustejovsky. 2012. TempEval-3: Evaluating Events, Time Expressions, and Temporal Relations. *Computing Research Repository (CoRR)* abs/1206.5333. http://arxiv.org/abs/1206.5333.

7. Eskerrak eta oharrak

Lan hau Eusko Jaurlaritzaren Ekonomia, garapen eta azpiegitura sailak finantzatutako MODENA: Neurona-eredu aurreratua kalitate handiko itzulpengintza automatikorako proiektuari (KK-2018/00087), Espainiako Ekonomia eta Lehiakortasun Ministerioaren PRO-SAMED: PROcesamiento Semántico textual Avanzado para la detección de diagnósticos, procedimientos, otros conceptos y sus relaciones en informes MEDicos proiektuari (TIN2016-77820-C3-1-R) eta Europar Batasunaren LINGUATEC: mugakide diren hizkuntzen arteko lankidetza eta ezagutza-transferentzia hizkuntza-teknologian proiektuari (EFA227/16) esker egin da.