eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Informatika Ingeniaritzako Gradua

Konputazio espezialitatea

Gradu Amaierako Proiektua

Code-Switching fenomenoaren detekzio automatikoa Twitter-eko txioetan

---

Egilea

Ander Corral Naves

informatika fakultatea
facultad de informática

2017

Aipamenak

Eskerrak eman nahi dizkiet:

- **Iñaki Alegria Loinaz**-i, proiektuan egindako zuzendaritza eta kontseilu lan guztiarengatik.
- **Larraitz Uria Garin**-i, IXA taldeko kidea, *corpus*-aren anotazioan igarotako ordu guztiengatik eta egindako zuzenketa lanengatik.

Laburpena

Proiektu honen helburua Twitter-eko txioetan euskara-gaztelaniazko *code-switching* delako fenomeno linguistikoa aztertzea da. Horretarako, errealitatearen corpus adierazgarri bat sortu da, euskarazko, gaztelaniazko eta txio elebidunak bilduz, hainbat pertsonaia publiko eta bestelako erabiltzaileengandik. CRF sekuentzia etiketatzailea erabili da sistemari ezagutza emateko, algoritmo honek datu sekuentziak tratatzeko duen izaeragatik. CRF algoritmoaren doiketa izan da proiektuaren atal nagusia, asmatze-tasen hobekuntza lortzeko asmoz. Emaitza gisa, denbora errealean, *streaming* bidez, lortutako txioen *code-switching*-a analizatuko duen aplikazioa sortu da, non bilaketak erabiltzaile konkretu baten edo hitz konkretu baten arabera egiteko aukera dagoen. Gainera, EUS-ES *code-switching* fenomenoaren detekzio automatikoan lehen urratsak ematea bilatu da, etorkizuneko proiektuentzat abiapuntua finkatuko duen proiektua burutuz.

Gaien aurkibidea

Irudien aurkibidea

Taulen aurkibidea

1. KAPITULUA

Sarrera

Azken urteotan, sare sozialen erabilerak gorakada nabarmena jasan du eta Facebook, Twitter, Instagram edo LinkedIn motako sare sozialen sorrerari bidea eman die. Hauen artean, Twitter plataformak, munduko edozein lekutako erabiltzaileek edozein gairi buruzko iritziak, gustuak, aipamenak, etab., 140 karaktereetan idazteko aukera ematen du.

Hori dela eta, Twitter testuen datu-base ikaragarria bihurtu da eta ondorioz, informazio iturri baliagarria enpresa, gobernu eta erakundeentzat. Hala ere, datu gordinetatik informazio esanguratsua lortzeko prozesu baten beharra dago. Hori dela eta, *data mining* edo *machine learning* bezalako kontzeptu eta teknikek sekulako arrakasta eduki dute konputazioaren munduan, datu gordinetatik informazio esanguratsua eskuratzeko duten gaitasunarengatik. Asko dira sare sozialen ustiatzea helburu duten aplikazioak, erabiltzaile eta tendentzia konkretuei buruzko datu esanguratsuak lortzeko aukera hobeezina baita.

Aplikazio hauen funtsezko helburua gehienetan ekonomikoa izan arren, badira beste motako helburuak, hala nola, linguistikoa. Azken mota honetakoa izango da proiektu honen hurbilpena, lan honen helburu nagusia Twitter plataforma baliatuz Euskal Herriko zonaldean agertzen den *code-switching* fenomeno linguistikoa analizatzea izanik. Fenomeno hau gehienetan ahozko hizkera informalean agertzen den arren, Twitter-eko txio askok daukaten izaera informal eta ahozkoaren parekoa dela eta, informazio iturri garrantzitsua da, eta fenomenoaren analisian lehen urratsak emateko aukera hobeezina. Gainera, ahozko fenomenoek zailtasun gehigarri bat daukate, hauek analizatzeko pertsonen grabazioak behar dira eta hauen hitz egiteko moduak naturaltasuna galtzen du grabatzailearen

aurrean. Twitter-eko txioek, aldiz, naturaltasun hori mantendu egiten dute eta ondorioz errealitatea modu zehatzagoan errepresentatzea espero da.

Linguistikaren eremuan, arau eta ohitura finkatuak daude testuak analizatzerako orduan. Hala eta guztiz ere, Twitter-eko txioetako testuek ezaugarri oso bereziak dituzte orokorrean, testu formal eta egituratuen aldean. Izan ere, txioek itxura informala, ahozkoaren parekoa, luzera motza, eta egitura lasaia daukate. Gainera, beste testuetan agertzen ez diren elementuen agerpenak ohikoak dira, esate baterako, traolak, erabiltzaile identifikatzaileak, estekak, emotikonoak, etab. Hori dela eta, testu hauek analizatzeak erronka handiagoa suposatzen du, eta ohiko arau eta ohituretatik urruntzea komenigarria da testuak analizatzerako orduan.

Hego Euskal Herri mailan, bi hizkuntza ko-ofizial existitzeak, Gaztelania eta Euskara, *code-switching* edo kode aldaketa delako fenomeno linguistikoa agertzea laguntzen du. Bi hizkuntzek erabilera zabala daukate eta ondorioz pertsonen arteko komunikazio moduan eragin handia daukate, bi hizkuntzak batera agertzen diren adibideak asko izanik. Beraz, proiektuaren irismena hego Euskal Herriko lurraldera mugatuta egongo da. *Code-switching* fenomenoa bi hizkuntza edo gehiago modu trukagarrian erabiltzean datza. Normalean ahozko erregistroan gertatzen den arren, idatzizko forman ere aurki daiteke. Lehen aipatu moduan, Twitter plataformako txioen izaera dela eta, erabiltzaileak bi hizkuntzak modu naturalean eta trukagarrian erabiltzea espero da.

Proiektuaren helburua Twitter-eko txioetan EUS-ES *code-switching* fenomeno linguistikoa aztertzea denez, fenomeno hau modu automatikoan detektatuko duen aplikazioa garatu da. Horretarako, errealitatearen corpus adierazgarri bat sortu da, euskarazko, gaztelaniazko eta txio elebidunak bilduz, hainbat pertsonaia publiko eta bestelako erabiltzaile desberdinetatik. Sistemari ezagutza emateko CRF sekuentzia etiketatzailea erabiltzea erabaki da. Izan ere, algoritmo honek datu sekuentziak tratatzeko duen izaerak aproposa egiten du algoritmoaren erabilera txioen testuak analizatzeko. CRF etiketatzailearen doiketa prozesua aurrera eraman da, asmatze-tasen hobekuntza lortzeko asmoz. Emaitza gisa garatutako aplikazioak, denbora errealean, *streaming* bidez, lortutako txioen *code-switching*-a analizatuko du, non bilaketak erabiltzaile konkretu baten edo hitz konkretu baten arabera egiteko aukera dagoen. Aplikazioaren kodea GitHub bidez atzigarri dago honako helbidean: https://github.com/anderleich/CodeSwitchingDetection

Gainera, gradu amaierako lan honetan EUS-ES *code-switching* fenomenoaren detekzio

automatikoan lehen urratsak ematea bilatu da, etorkizuneko proiektuentzat abiapuntutzat baliagarria izango dena. Fenomeno honen detekzio automatikoak abantaila handiak ekar ditzake hizkuntza teknologia desberdinetan. Izan ere, sarri agertzen den fenomenoa izanik, ezinbestekoa da honen ulermena eta antzematea testuak tratatzerako orduan. Honen adibidea izan daiteke itzulpen automatikoa, non beharrezkoa den *code-switching*-aren detekzioa, hizkuntza desberdinak bereizteko eta itzulpen egokia burutzeko.

2. KAPITULUA

Proiektuaren Kudeaketa

Atal honetan proiektuaren garapenarekin hasi aurretik egin beharreko kudeaketaren inguruko aspektu nagusiak azaltzen dira; alde batetik, proiektuaren helburu nagusiak azalduz eta bestetik, proiektuaren irismena azalduz. Irismenaren atalean, betekizunak, mugarriak, lanaren deskonposaketa, arrisku kudeaketa eta emangarri nagusiak aztertuko dira.

Helburuak

Atal honetan, proiektuak bete beharreko helburu nagusiak azaltzen dira. Helburu hauek proiektua hasterako orduan finkatu egin ziren eta diseinu eta garapen fasean kontutan hartzea ezinbestekoa izango da lan arrakastatsu bat lortzeko. Gradu amaierako proiektua izanik, honek suposatzen dituen aparteko helburuak ere zehazten dira.

Proiektuaren helburu nagusia Twitter sare sozialeko txioetan gertatzen den EUS-ES *code-switching* fenomeno linguistikoa automatikoki detektatzea izango da. Helburu nagusia finkatuta, bitarteko beste azpi-helburu batzuk finkatu dira:

- *Corpus* anotatu bat sortzea *code-switching* adibideekin.
- *Corpus* anotatu batean oinarrituta *code-switching* fenomenoaren detekzio automatikoa egitea, ikasketa automatikoko CRF sekuentzia etiketatzailea erabiliz.
- Denbora errealean *code-switching* fenomenoa detektatu eta etiketatuko duen aplikazioa garatzea.

- EUS-ES *code-switching* fenomenoaren detekzio automatikoan lehen urratsak ematea, etorkizuneko proiektuei begira.

Proiektuaren gaiarekin zerikusia duten helburuez gain, Gradu Amaierako Proiektua izanik, proiektuaren nondik norakoak jasoko duen memoria idaztea eta defentsarako aurkezpena prestatzea ezinbestekoa izango da. Gainera, graduan zehar, eta batez ere konputazio adarrean, ikasitakoa eta eskuratutako gaitasunak praktikan jartzea bilatuko da.

Irismena

Proiektuaren irismena analizatuko da atal honetan. Garrantzitsua da proiektuaren irismena Gradu Amaierako Proiektuari estimatutako garapen denborara egokitzea proiektua bideragarria izan dadin. Proposatu daitezkeen hobekuntzak etorkizun baterako kontutan hartuko lirateke edo proiektuaren garapenean zehar, irismenean zehaztutako betebeharrak bete direla ikusiko balitz, irismena eguneratzea pentsa daiteke.

Betekizunak

Proiektuaren helburuak betetzeko garapenean zehar bete beharreko betekizunak zehazten dira atal honetan. Betekizun hauek argi edukitzea ezinbestekoa izango da proiektu arrakastatsu bat lortzeko.

1. Ezagutza automatikoaren muina izango den *corpus* gordinaren bilketa eta osaera.

    - Twitter aplikazioan ikerketa lana burutu beharko da *code-switching*-a daukaten txioak eskuratzeko. Horretarako, erabiltzaile konkretuak eta momentuko traola nagusiak aztertuko dira txio elebidunen adibideen bila.
    - Twitter API-aren bidez txioak eskuratzeko balioko duen programa garatzea. Heuristiko sinple bat aplikatuko da euskarazko eta gaztelaniazko hiztegien laguntzaz.
    - Lortutako txioak eskuz gainbegiratu eta iragaziko dira, aurreko pausuko heuristikoak izan ditzakeen arazoak konpontzeko.
    - Lortutako txio guztiekin bukaerako *corpus* gordin orokorra sortu beharko da.

Code-switching-a daukaten txioetaz aparte, zenbait txio elebakar ere lortu beharko dira, sistemaren ezagutza maila handitu eta osatzeko. Horrela, entrenatutako algoritmoak txio elebakarrak eta txio elebidunak bereizteko informazio gehigarria izango du. Twitter-ek gaztelaniazko txioak zuzenean bilatzeko aukera ematen du bere bilaketa aurreratuan, baina euskarazko txioak lortzeko ikerketa lana burutzea ezinbestekoa izango da.

2. Corpus gordinaren anotazioa. Txioen anotazioa hitzez hitz burutuko da.

   - Anotatzerako orduan jarraitu beharreko gomendio eta arauak azalduko dituen gidalerroa osatu beharko da.
   - Etiketatzeko lanean lagunduko duen interfaze grafikoa garatu beharko da. Interfaze honek txioak banan-banan etiketatzeko aukera eman beharko du, honako funtzionalitate nagusiak eskainiz:
     – Hurrengo txiora joateko botoi bat eskainiko du.
     – Aurreko txiora joateko botoi bat eskainiko du.
     – Nabigazio librea ahalbidetuko duen eremua eskainiko du, joan nahi den txio zenbakira joateko aukera emanez.
     – Tokenizatzaile baten laguntzaz txioaren testuko tokenak bananduko dira eta interfazeak token hauek banaka anotatzeko aukera emango du.

3. CRF sekuentzia etiketatzailearen inplementazioa eta esperimentazioa.

   - CRF algoritmoa Python programazio lengoaian inplementatzeko liburutegi egoki bat bilatuko beharko da.
   - CRF etiketatzailearen entrenamendu, ebaluazio eta test faseak burutuko dituen programa garatu beharko da.
   - Programa honek sekuentzia elementuen ezaugarri funtzioak doitzeko eta aukera desberdinak probatzeko aukera eman beharko du.
   - CRF algoritmoaren ebaluazio eta test faseen emaitzak gorde beharko dira esperimentazio desberdinak konparatzeko eta ondorioak ateratzeko.
   - *Code-switching* detekzioaren *baseline*-aren kalkulua egin beharko da problemaren zailtasun mailaren erreferentzia puntu bat edukitzeko.

4. Denbora errealean txioak eskuratuko dituen eta *code-switching*-a detektatuko duen aplikazioa burutu beharko da.

- *Corpus* guztia erabiliko da ikasketa faserako.
- Txioak *streaming* bidez lortzeko Twitter API-ak eskaitzen dituen erremintak aztertu beharko dira.
- Behin txioak analizatuta, JSON formatuan itzuliko dira beste aplikazioek formatu hau irakurtzeko daukaten erraztasuna dela eta, adibidez, Javascript lengoaiak.

5. Proiektuaren nondik-norakoak azalduko dituen memoria osatzen joan beharko da proiektuaren bizi-ziklo osoan zehar.

Mugarriak

Proiektuan zehar hainbat mugarri zehaztu dira egin beharreko atazen erreferentzia epeak edukitzeko eta proiektuaren garapen fasea kontrolatzeko. Mugarri hauek orientagarriak izango dira eta proiektuan zehar sortutako aldaketa edo arazoen arabera moldatu beharko dira.
Mugarriak bi multzotan banatu dira: barne mugarriak eta kanpo mugarriak. Barne mugarriak ez dira derrigorrezkoak izango, baina proiektuaren garapena kontrolpean eramateko beharrezkoak izango diren erreferentziak edukitzeko lagungarriak izango dira. Kanpo mugarrietan, aldiz, derrigorrezko entregatze data ofizialak kontutan izan dira.

Barne mugarriak

- **2017-III-31**: Proiektuaren plangintza eta betekizun nagusiak zehaztuko dituen dokumentuarekin amaitu.
- **2017-IV-24**: *Corpus* gordinaren osaera amaitu eta hau anotatzeko programaren garapena amaitu.
- **2017-V-1**: *Corpus* anotatua amaitu.
- **2017-V-21**: CRF algoritmoaren entrenatze eta esperimentazio fasea amaitu.
- **2017-VI-4**: Txioak denbora errealean *streaming* bidez eskuratu eta *code-switching*-a detektatzeko programaren garapena amaitu.
- **2017-VI-18**: Proiektuaren memoria amaitu.

Kanpo mugarriak

- **2017-VI-23**: Unibertsitateko ADDI plataformaren bitartez memoriaren PDF entrega orokorra eta GitHub plataforma bidez, garatutako kodea publiko jarri.

- **2017-VII-(10-13)**: Epaimahaiaren aurrean proiektuaren aurkezpena eta defentsa burutu.

Lanaren deskonposaketa egitura, LDE

LDE diagramak proiektuan zehar egin beharreko lana atazetan banatzeko helburua dauka. Horrela, burutu beharreko lanaren kontrola eramatea bilatzen da, suerta daitezkeen aldaketa edo arazoak gutxitzeko. 2.1 irudian ikus daiteke LDE diagramaren grafikoa.

Kudeaketa

Atal honen barnean kontutan edukiko dira proiektuaren kudeaketarekin zerikusia duten atazak. Hala nola, proiektua hasterakoan egin beharreko plangintza, proiektuaren garapenean zehar egindako kontrol eta jarraipen bilerak, etab.

Ikerketa

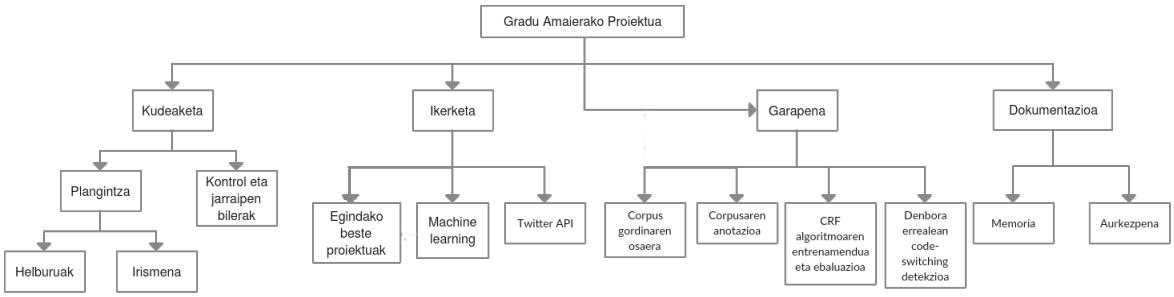
Ikerketa atalean proiektuari hasiera emateko burutu beharreko ikerketa lanaren parte diren atazak bilduko dira. Adibidez, *code-switching* fenomenoaren inguruan egindako beste proiektuak, *machine-learning* algoritmo egokiaren analisia, Twitter API-aren analisia, etab.

Garapena

Atal honetan proiektuaren garapen fasean egin beharreko ataza desberdinak hartuko dira kontutan. Besteak beste, corpus gordinaren osatzea, honen anotazio fasea, CRF entrenamendu eta balidazio fasea eta denbora errealean *code-switching*-a detektatuko duen aplikazioaren garapena.

Dokumentazioa

Proiektua amaitutzat emateko proiektuaren nondik norakoak bilduko dituen memoria eta aurkezpena burutuko dira atal honetan.



2.1 Irudia: LDE diagramaren grafikoa.

Arrisku kudeaketa

Atal honen helburua proiekturen garapenean suerta daitezkeen arriskuak aurreikustea da, arrisku horiek kontutan hartzeko eta proiektuaren arrakastan eragina gutxitzeko.

- Ikasketa automatikoaren muina izango den *corpus*a aldatzea edo galtzea. Garapenean zehar *corpus*arekin probak egingo dira, hala nola, txioak gehitu, anotazioa zuzendu, *corpus*-a analizatu, etab. Prozesu hauek guztietan, *corpus*a aldatzeko edo galtzeko arriskua dago. Hori dela eta, *corpus*aren bertsio ezberdinak gordetzea eta segurtasun kopiak edukitzea ezinbestekoa izango da.

- Garatutako kodearen segurtasun kopia bat edukitzea ezinbestekoa izango da. Horretarako periodikoki Google-ek eskaitzen duen Drive plataforma erabiliko da. Beste aukera GitHub plataforma erabiltzea izango da.

- Aurreikusitako epeak ez betetzea. Gerta daiteke, proiektuaren lan karga dela eta zehaztutako epeak betetzeko denborarik ez izatea. Proeiktu arrakastatsu entregatzeko epemuga batzuk daudenez, hauek betetzea derrigorrezkoa da. Horretarako, jarraipen eta kontrol bilerak burutuko dira, prozesua egiaztatu eta kontrolpean mantentzeko.

Emangarri nagusiak

Proiektuaren emangarri nagusiak proiektua modu arrakastatsuan aurkezteko ofizialki entregatu behar diren dokumentuak izango dira.

- EUS-ES *code-switching* fenomenoa denbora errealean detektatuko duen aplikazioaren kodea, GitHub plataforma bidez.
- Corpusa etiketatzeko interfaze grafikoaren kodea, GitHub plataforma bidez.
- Proiektuaren nondik norakoak jasoko dituen memoria, unibertsitateko ADDI plataformaren bidez.

3. KAPITULUA

Oinarri teorikoak

Motibazioa

Code-switching fenomeno linguistikoa oso zabalduta dago bi hizkuntza edo gehiago erabiltzen diren zonaldeetan. Hori dela eta, zonalde horietako erabiltzaileek hizkuntzak modu trukagarrian erabiltzen dituzte ahozko erregistroan zein idazkera informalean. Fenomenoaren eragina zabala da kasu hauetan, eta honek hainbat erronka suposatzen ditu modan dauden zenbait hizkuntza teknologientzako. Analisi sintaktikoa, itzulpen automatikoa, hizketa-ezagutza automatikoa (*Automatic Speech Recognition, ASR*), informazio erauzketa edo prozesaketa semantikoa dira hizkuntza teknologia horien adibide batzuk. Lengoaia naturalaren prozesamendurako metodo tradizionalak eraginkorrak dira kasu elebakarretan, baina hauen eraginkortasuna gutxitu egiten da hizkuntza bat baino gehiago nahasita dauden kasuetan.

Hori dela eta, azken urteetan *code-switching* fenomenoaren detekzioa funtsezkoa bihurtu da hizkuntza teknologientzat. Helburu honen harira, 2016an burutu zen *Overview for the Second Shared Task on Language Identification in Code-Switched Data* delako bilkura SIGCAT elkartearen eskutik [4]. Bilkura hau EMNPL (*Empirical Methods in Natural Language Processing*) konferentziaren parte izan zen eta lehiaketa moduan bideratu zen, hizkuntzaren prozesaketako 9 taldek parte hartuz. Lehiaketaren helburua Twitter-eko txioetan *code-switching*-a detektatzea zen, horretarako bi adibide hartuz: Arabiar estandarraren eta Arabiar dialektalaren artekoa, MSA-DA, eta gaztelania eta ingelesaren artekoa,

SPA-ENG.

Eskatutako eginkizuna betetzeko, talde gehienek CRF algoritmoa hartzea erabaki zuten ikasketa automatikoko metodo gisa. Aukera hau berretsita geratu zen, taldeen emaitzak ikusterakoan. Izan ere, CRFen sekuentziak automatikoki etiketatzeko izaera dela eta, emaitza egokiak ematen ditu testu-token sekuentziaz osatutako problema baten aurrean. Emaitzek erakutsi zuten bezala, talde hoberenek %89 inguruko asmatze-tasa lortu zuten SPA-ENG kasurako eta %75 inguruko asmatze tasa MAS-DA kasurako.

Ondorioz, SIGCAT elkarteak burututako bilkura aurrekaritzat hartuko da proiektu honen garapenareko. CRF sekuentzia etiketatzaileek lortutako emaitzak azpimagarriak direla eta, proiektu honetan erabiltzea pentsatu da. Atal honetan, CRF-ei inguruko atal teorikoak azalduko dira, bai eta hauen aurrekariak diren HMM eta MEMM ereduei buruzkoak ere.

Iragarpen egituratua (Structured prediction)

Klasifikazio-ereduek irteera diskretuak aurresateko metodo ahaltsu eta errotuak eskaintzen dituzte. Hala ere, badira hainbat aplikazio non objektu konplexuagoak aurresan nahi diren, hala nola, lengoaia naturalerako analisi zuhaitzak.

Iragarpen egituratu kontzeptua klasifikatzaileen eta eredu grafikoen arteko konbinazioa dira. Datu aldagai anitzak modelatzeko ahalmena eta sarrerako datu sekuentziak aurresateko ahalmena konbinatzen dituzte.

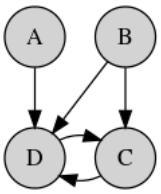
Elkarren arteko dependentzia daukaten aldagai multzo baten iragarpena aplikazio askoren funtsezko gaitasuna da. Aplikazio hauetan ausazko aldagaiez osatutako $y = \{y_0, y_1, ..., y_T\}$ irteerako bektorea iragarri nahi da, gainbegiratutako x ezaugarrri bektorea emanik. Honen adibide ezagun bat lengoaia naturalaren *PoS tagging* etiketatze prozesua[1] da, non y_s s hitzari dagokion kategoria gramatikala den eta sarrerako x bektorea ezaugarri bektoreetan banatuta dagoen $\{x_0, x_1, ..., x_T\}$. x_s bakoitzak s posizioan dagoen hitzari buruzko informazioa edukiko du, hala nola, identitatea, aurrizkiak, atzizkiak, lexikoietatik lortutako datuak, etab.

[1] *PoS tagging* etiketatze prozesu bat da, non hitz bakoitzari dagokion kategoria gramatikala edota bestelako marka lexikalen bat esleitzen zaion.

Aldagai anitzak aurresateko problema baten aurrean, lehen hurbilpena aldagaiak independenteki tratatuko dituen klasifikatzaile bat erabiltzea izan daiteke, non s posiziorako $x \rightarrow y_s$ ematen den. Horrela, sekuentzia elementuetan banatzen da eta independenteki kontsideratzen dira.

3.1 Irudia: Sekuentzia elementuen banatzea.

Hala ere, hurbilpen honek ez ditu aldagaien arteko dependentziak kontutan hartzen eta ondorioz informazio asko galdu egiten da bidean. *PoS tagging* adibidean, euskararen kasuan, izen bati askotan adjektibo batek jarraitzen dio. Ondorioz, irteerako aldagaien dependentziak modu egokian adierazteko grafo ereduak[2] (*Graphical models*) erabiltzen dira. 3.2 irudian ikus daiteke eredu grafikoen errepresentazioa. Aldagaien arteko erlazioak gorde-



3.2 Irudia: Eredu grafiko baten adibidea, non gezi bakoitzak aldagaien arteko dependentzia adierazten duen.

ko eta kontutan hartuko dituen ohiko eredu grafikoa Markov-en Eredu Ezkutua (*Hidden Markov Model*, HMM) da. Erabiltzen den beste eredu bat Markov-en Entropia Maximoaren Eredua (*Maximum Entropy Markov Model*, MEMM) da.

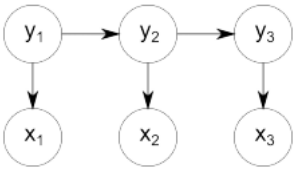
[2] *Graphical models*, eredu probabilistikoa da non ausazko aldagaien arteko dependentziak adierazteko grafo bat erabiltzen den, erpin bakoitza aldagai bat izanik.

HMM, Markov-en eredu ezkutua

HMM ereduek probabilitate bateratua esleitzen diete ezaugarri-bektore eta ausazko aldagai bektore bikote bati, $p(x,y)$, non x gainbegiratutako ezaugarri bektorea den eta y iragarri nahi den ausazko aldagaien bektorea. Parametroak probabilitate bateratuaren fidagarritasuna maximizatzeko entrenatzen dira entrenamendu datuekiko. Ondoren, egoera-kate baten probabilitatea lortzeko baldintzapeko probabilitatearen erregela erabiltzen da.

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

Ekuazio honek suposatzen du x ezaugarri bektore guztien probabilitateak edukitzea eta hau bakarrik posiblea da obserbazio katearen elementuak elkarren artean independenteak badira, hau da, x_s ezaugarri bektorea y_s aldagaiarekiko bakarrik menpekoa bada. Adierazpen honek sekuentzia-elementuen arteko independentzia handia suposatzen du. Gainera, HMMak Markov propietateak betetzen dituzte, hau da, ausazko aldagaiak independenteak dira elkarrekiko, aurrekoarekiko izan ezik.



3.3 Irudia: HMM-en dependentzia adibidea.

HMM dependentziak:

- Obserbazio bakoitza bere egoerarekiko menpekoa da bakarrik.
- Egoera bakoitza aurreko egoerarekiko menpekoa da bakarrik.

Independentzia suposizio hauen ondorioz, posible da ezaugarri-bektore eta ausazko aldagai bektore bikotearen probabilitate bateratuaren faktorizazioa lortzea.

$$p(x,y) = \prod_{s=1}^{T} p(y_s|y_{s-1})p(x_s|y_s) \qquad \text{non, } p(y_1|y_0) = p(y_1)$$

Ezagutza aditua (Expert knowledge)

Bilatzen den ereduak sarrera datuetatik aparte informazio gehiago kontutan hartzea ahalbidetzen du ezagutza adituak. Adibidez, izen entitate aztertzaile batek maiuskulaz hasten diren hitzak kontutan hartzea. Bestalde, desiragarria da erabilitako ereduak indukzio bidez datuetatik informazioa eskuratu ahal izatea. Hau da, hain nabariak ez diren patroi eta erlazioak eskuratzea.

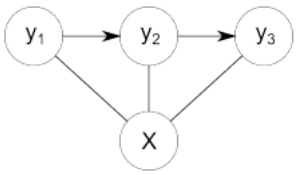
MEMM, Markov-en entropia maximoko eredua

MEMM ereduak baldintzapeko ereduak dira. HMM ereduen ideiak konbinatzen dituzte datu-sekuentziak entropia maximoko ereduekin tratatuz. Entropia maximoaren helburua sarrerako datuak modu egokienean irudikatuko duen eredua aurkitzea da, beti ere, probabilitate distribuzioa uniforme mantenduz.

HMM-kin konparatuz, MEMM ereduek ezagutza aditua erabiltzen dute. MEMM eredu baten sarrera ezaugarri-funtzioak dira. Funtzio hauek datu-sekuentzi batetik, x, uneko indizetik, s, uneko etiketatik, y_s, eta aurreko etiketatik, y_{s-1}, kalkulatzen dira.

$$f(y_s, y_{s-1}, x, s)$$

MEMM ereduetan, HMM ereduek ausazko aldagaien eta ezaugarri bektoreen artean suposatzen zuten dependentzia desagertu egiten da. Gainera, ereduari informazio gehigarria emateko aukera eskaintzen dute ezaugarri funtzioen bitartez (ezagutza aditua).



3.4 Irudia: MEMM-en dependentzia adibidea.

MEMM dependentziak:

- Egoera bakoitza aurreko egoerarekiko (Markov) eta datu sekuentziatik lortutako ezaugarrietatik dependentea da.

MEMMak baldintzapeko ereduak direnez, baldintzapeko probabilitatea kalkulatzen dute, $p(y|x)$, hau da, ausazko aldagaien probabilitatea datu-sekuentzia jakinda.

$$p(y|x) = \prod s = 1^T p(y_s|y_{s-1}, x)$$

CRF, Conditional Random Fields

Aurreko bi ereduek, soberan dauden suposizio asko egiten dituzte (HMM) edo zenbait problematan erabilgarriak ez diren arazoak dituzte (MEMM). Bilatzen den ereduak ezagutza aditua erabiltzea beharrezkoa da. Gainera Makov-en legea malguagoa bihurtzea edo kentzea bilatzen da, dependentzia norabidea (Markov, aurrekoarekiko sekuentzia elementuarekiko menpekotasuna) desagertzeko eta datu konplexuagoekin lan egiteko, adibidez, irudiak.

Aurreko ereduak gogoratuz, pentsa dezakegu datu-sekuentzia etiketatu batean, $y = (y_s)_{s \in 1..T}$, eta gainbegiratutako objektu ezagun batean honi esleituta, x_s HMM ereduen kasuan eta ezaugarri funtzioa MEMM kasuan. Etiketa-kate hau grafo bat da, non erpinak etiketak, y, diren eta ertzak hauen arteko konexioak. Sekuentzien kasuan, kate bat, (y_s, y_{s+1}) motako konexioekin.

Etiketatutako datu sekuentzia batentzat, CRF ereduaren definizioaren arabera, objektu eta etiketa parea (x,y) CRF bat izango da baldin eta $p(y_t|X, y_1, ...y_{s-1}, y_{s+1}, ..., y_T) = p(y_t|X, y_{t-1}, y_{t+1})$ bada. Hau da, y_s etiketaren probabilitatea obserbazioekiko, x eta bere bizilagunekiko, y_{t-1} eta y_{t+1}, menpekoa da bakarrik eta ondorioz, independentea beste etiketa guztiekiko.

CRF-ek informazioa irudikatzeko ezaugarri funtzioak erabiltzen dituzte. Funtzio hauek, ereduak x obserbazioak interpretatzeko modua dira eta problemaren dependentziak definitzen dituzte, hau da, irteerako grafoa. CRFek inposatzen duten murrizketa bakarra bizilagunak ez diren Y multzoko elementuen independentzia da eta ondorioz, malgutasun handiagoa eskaintzen dute ezaugarri-funtzioetarako. Funtzio horietan hainbat datu barneratu daitezke:

- Edozein informazio x multzoko elementu bat edo gehiagotik.
- Edozein informazio y multzoko elementu batetik.

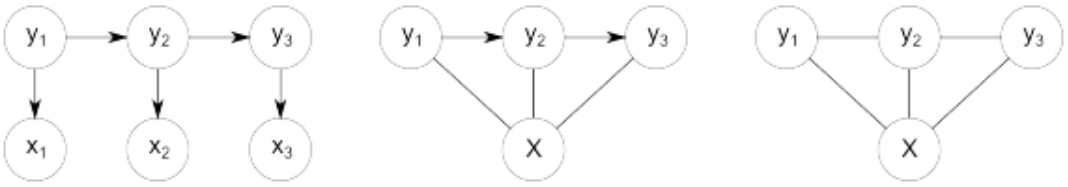
- Edozein informazio y multzoan bizilagunak diren elementuetatik.

- Aurrekoen konbinazioak.

Normalean ezaugarri-funtzioek propietate edo egitate bat adierazten dute balio bitarrak erabiliz (0 edo 1). Horrela, MEMM ereduen kasuan bezala, CRF ereduan ezagutza aditua sartzeko aukera ematen da, sarrerako x elementu guztietatik informazioa deribatuz edota y irteerako elementu bizilagunen dependentzien bidez.

CRF linealak

CRF linealak CRF ereduen formarik sinpleena dira. HMM eta MEMM ereduen egitura berdina jarraitzen dute eta ondorioz erabilpen kasu berdinak dituzte. Hauen desberdintasunak erabilitako dependentzietan daude.



3.5 Irudia: (1) HMM ereduen dependentziak, (2) MEMM ereduen dependentziak eta (3) CRF linealen dependentziak.

- HMM ereduek, Markov-en propietatea betetzen dute, y_s etiketa y_{s-1} etiketarekiko dependentea da bakarrik honako probabilitatearekin: $p(y_s|y_{s-1})$. Gainera, x_s obserbazioa y_s etiketarekiko dependentea da, $p(x_s|y_s)$ izanik.

- MEMM ereduen kasuan, Markov-en propietatea ere betetzen da, baina kasu honetan y_s etiketak aurreko etiketarekiko, y_{s-1}, eta x bektorearekiko dependentzia dauka, probalilitatea honakoa izanik: $p(y_s|y_{s-1},x)$.

- CRF ereduek, ordea, y_s etiketa bakoitzak ondoko etiketa bizilagunekiko (y_{s-1} eta y_{s+1}) eta obserbazio bektorearekiko, x, dependentzia daukala suposatzen dute.

Erabilpen kasuak

Kate linealeko CRF ereduek hainbat erabilpen kasu dituzte lengoaia naturalaren prozesamenduan. Erabilgarriak dira sekuentzia segmentazioa edota sekuentzien etiketatzea behar deneko kasuetan. Erabilpen kasu hauetako batzuk honakoak dira:

- Testuen hitz-gakoen erauzketa burutzeko.
- Izen entitateen erazagupena burutzeko, (NER, *Named Entitity Recognition*).
- Testuen sentimendu analisia burutzeko.
- Hitzen kategoria gramatikalen azterketa burutzeko, (*Part-of-Speech Tagging*).
- Hizkera erazagupena burutzeko.

4. KAPITULUA

Erabilitako teknologiak

Atal honetan proiektuaren garapenean zehar erabilitako teknologia nagusiak aipatuko dira. Erabilitako teknologia hauek, egin izan beharreko lanaren eskakizunak betetzeko erabilgarri suertatu dira. Hiru multzo nagusi bereiz daitezke: Twitter plataformako txioen atzipena ahalbidetzen dituztenak, Twitter REST APIa eta Tweepy, CRF sekuentzia etiketatzailearen erabilpena ahalbidetzen dituztenak, *sklearn-crfsuite* Python liburutegia, eta inplementazio faserako programazio-lengoai desberdinak, hala nola, Python, Java, Html+CSS eta Javascript.

Twitter REST API

Twitter plataformak REST API bat eskaintzen du sare sozialaren datuak atzitzeko. API-ak abstrakzio maila bat sortzen du, hainbat funtzio eta prozedura eskainiz, hainbat programazio lengoaietara moldatzeko. Dokumentazio zabala eskaintzen da funtzionalitate guztien inguruan, APIa erabili nahi duten garatzaileentzat [14]. Hala eta guztiz ere, atal honetan proiektuan erabilitakoari bakarrik erreferentzia egingo zaio.

Dokumentazioaren arabera, programazio hizkuntza gehienetarako liburutegiak daude, esate baterako, Java, Python, Perl edo C++. Proiektu honetan Python lengoaiarako *Tweepy* liburutegia erabili da [15].

API-a erabili baino lehen erabiltzailearen edo aplikazioaren identifikazioa beharrezkoa

da. Identifikazio prozesua *OAuth* bidez egiten da. Dokumentazioaren arabera, bi identifikazio posible existitzen dira:

- **Aplikazio mailako identifikazioa.** Metodo honen bidez, egindako eskaerek aplikazioa bakarrik identifikatu egiten dute. Desgaituta egongo dira erabiltzailearekin zerikusia duten ekintzak.
- **Erabiltzaile mailako identifikazioa.** Metodo honen bidez, egindako eskaerek bai erabiltzailea eta bai aplikazioa identifikatu egiten dituzte. Identifikazio honek erabiltzaileak bere profila ikusteko, txioak idazteko, beste erabiltzaile batzuk jarraitzeko, etab., egiteko aukera dauka. Gainera aplikazio mailako identifikazioaren ekintzak ere burutu daitezke.

Proiektu honetan erabilitako aplikazioak erabiltzaile identifikazioa erabiltzen du. Bai aplikazio mailako identifikazio egiaztagiria eta bai erabiltzaile mailako identifikazio egiaztagaria web bidez lortu daitezke ondoko estekan https://apps.twitter.com/.

Bi identifikazio motek eskaera kopuruan murrizketak dituzte, zerbitzarien asetzea ekiditeko, baina banatuta kontsideratzen dira bakoitzarentzat. Murrizketen inguruko informazio guztia API-ak eskaintzen duen dokumentazioan aurki daiteke. Murrizpen balioak kontutan eduki behar dira aplikazioa garatzerako orduan eskaerekin arazorik ez edukitzeko. Izan ere, eskaera kopurua handia bada deskonexioa gertatuko da. Twitter-ek gomendatzen du eskaera kopuru handiko aplikazioek, *cache* modukoa erabiltzea eskaerak murrizteko.

API-ari egindako eskaeren emaitzak JSON formatuan jasotzen dira. API-aren dokumentazioaren ikus daiteke erantzunei buruzko informazio guztia. Proiektuaren irismenari begira, erabili diren JSON eremuak bakarrik aipatuko dira.
Txio baten eremuak:

- **id_str**: txioaren identifikatzailea *string* formatuan.
- **text**: txioaren testua.
- **user**: txioa igo duen erabiltzaile objektua.

Erabiltzaile baten eremuak:

- **id_str**: erabiltzailearen identifikatzailea *string* formatuan.

- **name**: erabiltzailearen izena.
- **screen_name**: erabiltzailearen Twitter-eko erabiltzaile izena.
- **profile_image_url**: erabiltzailearen irudiaren esteka.

API-ak *streaming* funtzio bat eskaintzen du, txioak denbora errealean eskuratzeko. Txioen eskaera bi modutara egin daiteke:

- **follow**: metodo honen bidez erabiltzaile konkretu baten txioak eskuratu daitezke denbora errealean. Erabiltzaileak sortutako txioak, erabiltzaileak *retweet* egindakoak, erabiltzaileak sortutako txioen erantzunak eta erabiltzaileak sortutako txioen *retweet*-ak eskuratuko dira.
- **track**: metodo honen bidez hitz konkretuak dituzten txioak eskuratu daitezke. Horrela, *hashtag* bidezko bilaketa egin daiteke.

Tweepy

Twitter REST API-a Python programazio lengoaian erabiltzeko eskuragarri dagoen liburutegia da. Proiektu honetan liburutegi hau erabili da, bai *corpus*a osatzeko txioak eskuratzeko eta bai Twitter API-aren *streaming* funtzioa erabiltzeko. *Streaming* funtzioaren laguntzaz, txioak denbora errealean lortzeko aukera izan da. Liburutegia erabiltzeko tutoriala, funtzioak eta gainontzeko informazioa liburutegiak eskaintzen duen dokumentazioan daude atzigarri [15].

sklearn-crfsuite

sklearn-crfsuite CRFSuite inplementazioa Python programazio lengoaian erabiltzeko liburutegia, *wrapper*, da. CRFSuite CRF algoritmoaren inplementazioa da C lengoaian. Online dokumentazio argia eskaintzen du eta instalatzeko azkarra eta zuzena da.

Hunspell

Hunspell ortografia egiaztatzailea eta morfologia aztertzailea da. Hainbat hizkuntzatarako eskuragarri dago eta software librea da. Proiektu honetan euskarazko eta gaztelaniazko hiztegiak erabili dira. Python-en erabiltzeko eskuragarri dago.

Bestelakoak

Atal honetan proiektuaren garapenean erabilitako bestelako teknologiak aipatzen dira. Teknologia hauek graduan zehar ezagutzat ematen dira eta ondorioz ez dira azalduko. Hala ere, teknologia hauek funtsezkoak izan dira garapen fasean eta ondorioz, hauen ezagutza beharrezkoa da proiektua aurrera eramateko.

- **Python** programazio lengoaia.
- **Java** programazio lengoaia.
- **Html + CSS**, web egitura.
- **Javascript**, web programazio lengoaia.
- **TeXstudio**, testu zientifikoak idazteko interfazea.

5. KAPITULUA

Proiektuaren garapena

Proiektuaren garapenaren atalean bi azpiatal nagusi bereizi dira, batetik diseinu fasea, non inplementazioa hasi aurretik egin beharrekoa nola egingo den zehaztu den, eta bestetik inplementazio fasea, non egin beharrekoa aurrera eramateko urratsak burutu diren. Azken fase honetan garatu diren bost urrats nagusiak aipatuko dira: corpus gordinaren osatzea, corpusaren anotatzea, CRF etiketatzailearen esperimentazioa, denbora errealean *code-switching*-a detektatuko duen aplikazioaren garatzea eta emaitzak modu dotoreagoan adieraziko dituen web interfazearen garatzea.

Diseinu fasea

Proiektuaren garapen fasea proiektuaren atal garrantzitsuena izango da, bertan proiektuaren produktua sortuko delako. Atal hau aurretik planifikatzea eta diseinatzea beharrezkoa izango da egin beharrekoa argi edukitzeko eta gerta daitezkeen arazo edo atzerapenak gutxitzeko. Proiektuaren garapen fasean, 2.1 eta 2.2 ataletan aipatutako helburu eta betebeharrak burutuko dira.

Horretarako, garapen fasea lau fase desberdinetan banatu da.

- ***Corpus* gordinaren osatzea**. Fase honetan Twitter plataformatik txio elebakarrak eta *code-switching* adibideak atzituko dira. Gainera, ikerketa lana burutu beharko da *code-switching* egiten duten erabiltzaileak aurkitzeko.

- ***Corpus* gordinaren anotazioa**. Fase honetan aurreko fasean lortutako *corpus*a eskuz anotatuko da. Lan hori errazteko interfaze grafiko bat garatu beharko da. Txioen testuak hitzez hitz anotatuko dira, sortuko den gidalerroko arauei jarraituz.
- **CRF etiketatzailearen esperimentazioa**. Fase honetan, CRF sekuentzia etiketatzailearen algoritmoaren entrenatze eta ebaluazio faseak inplementatuko dira. Horretarako hainbat esperimentu egingo dira, CRF-en ezaugarri funtzioak doituz asmatzetasak hobetzeko asmoz. Fase honen garrantzia aipagarria da, honek zehaztuko duelako garatutako produktuaren fidagarritasuna.
- **Denbora errealean *code-switching* fenomenoa detektatuko duen aplikazioaren garapena**. Sortutako *corpus* anotatua eta CRF sekuentzia etiketatzailea erabilita, txioetan *code-switching* fenomenoa automatikoki detektatuko duen aplikazioa garatuko da fase honetan. Horretarako, Twitter API-aren *streaming* funtzioa erabiliko da.

Fase hauek ez dira ordena kronologikoan burutuko, baizik eta iteratiboki garatzen eta hobetzen joango dira. Horrela, iterazio bakoitzaren bukaerako bertsioarekin probak eta ebaluazioak egingo dira. Proba horien ondoren, ikusitako arazo nagusiak konponduko dira eta hobekuntzak burutuko dira. Garapena iteratiboki egiteak, fase honen kontrola eta jarraipena erraztuko du, eta sor daitezkeen arazoak gutxitu eta lehenago konpondu.

Corpus gordinaren osatzea

Sortutako *corpus*a errealitatetik hurbilen egotea ezinbestekoa izango da burutu beharreko *code-switching*-aren detekzio automatikoa modu fidagarri eta eraginkor batean lortzeko. *Corpus* adierazgarri eta anitz batek emaitza hobeagoak lortuko ditu, kasu desberdin gehiago kontutan hartuko dituelako. Horretarako, txioak eskuratu aurretik ikerketa lana burutu da. *Code-switching* adibideak lortzeko, txioak eskuratzerako orduan, iragazketa heuristiko bidez eta manualki egin da.

Ikerketa lana

Code-switching-a izan dezaketen txioak lortzeko, lehenik eta behin, Twiter-en normalean bai euskaraz eta bai gaztelaniaz idazten duten erabiltzaileak bilatu dira. *Corpus*a errepresentagarria eta anitza izateko, erabiltzaile desberdinetatik txioak lortu dira. Izan ere,

erabiltzaile bakoitzak bere estilo propioa dauka idazterako orduan eta hauen idazteko modua patroi ez bihurtzeko aniztasuna ezinbestekoa da. Pertsonaia publikoen txioak, euskal telebista saioei buruzko erabiltzaileen txioak eta momentuko euskarazko *trending topic* nagusiak analizatu dira horretarako.

Lan honetarako erabilgarria izan da *umap.eus* webgunea, https://umap.eus/. Webgune honek Twitter plataformako euskarazko jarduna monitorizatzen du eta *ranking* desberdinak bistaratzen ditu. Adibidez, eskuragarri daude euskaraz gehien idazten duten erabiltzaileen zerrenda, euskarazko traola edo joera nagusiak, etab.

Txioak eskuratzea

Txioak eskuratzeko Tweepy liburutegia erabili da. Honen bidez, Python programazio lengoaiaz heuristiko sinple bat garatu da *code-switching*-a detektatzeko, euskarazko eta gaztelaniazko *Hunspell* hiztegiak erabiliz. Tweepy liburutegiak erabiltzaile konkretu baten txioak eskuratzeko funtzioak eskaintzen ditu. Aipatu bezala, API-ak murrizpenak ditu zerbitzatu ditzakeen txio kopuruan eta beraz, erabilitako metodoarekin gehienez 3.200 txio eskuratu daitezke erabiltzaile batetik. Hala ere, kopuru hau nahikoa izan da proiektuarentzat erabiltzaile aniztasuna lortu nahi izan baita.

Code-switching-a detektatzeko erabilitako heuristikoaren deskribapena:

- Txioaren testua token desberdinetan banatu espresio erregularren laguntzaz.
- Uneko tokena euskarazko hiztegian baldin badago, euskarazko token kopuruari bat gehitu.
- Uneko tokena gaztelaniazko hiztegian baldin badago, gaztelaniazko token kopuruari bat gehitu.
- Bukaeran euskarazko eta gaztelaniazko tokenak badaude, orduan *code-switching* fenomenoa gertatu da eta txioa gordeko da.

Emaitza moduan CSV fitxategi bat sortzen da eskuratutako *code-switching* txioekin. Heuristikoak, *hunspell* erremintak eskaintzen dituen euskarazko eta gaztelaniazko hiztegiak

erabiltzen ditu. Hau da, hain zuzen ere, heuristikoaren arazo bat, bi hizkuntzetako hiztegietan dauden hitzak ager daitezkeelako eta ondorioz, *code-switching*-a detektatzen delako. Arazo honi aurre egiteko, txioak eskuratu eta gero, manualki iragazi behar izan dira, txioak kenduz.

Code-switching adibideez gain, txio elebakar batzuk ere eskuratu dira, sistemaren ezagutza iturria osatu eta hobetzeko. Txio hauek eskuratzeko ere Tweepy erabili da. Twitter-eko API-ak gaztelaniazko txioak bakarrik bilatzeko aukera dauka, hizkuntza parametroa erabiliz. Euskarazko txio elebakarren kasurako, aldiz, ez dago horrelako aukerarik eta eskuz bilatu behar izan dira *umap.eus* webgunearen laguntzaz.

Lortutako txio guztiak batu egin dira *corpus*aren azken bertsio lortzeko. 5.1 taulan *corpus*aren ezaugarri nagusiak ikus daitezke, bai txioen hizkuntzaren araberako kopuruak[1] eta bai erabiltzaile araberako kopuruak. Ikus daitekeenez, *corpus* adierazgarri eta anitz sortzea lortu da.

Txioak guztira	1765
Gaztelanizko txioak	467
Euskarazko txioak	337
Code-switching txioak	959
Bestelako txioak	2
Erabiltzaileak guztira	424
Gaztelaniazko erabiltzaileak	249
Euskarazko erabiltzaileak	172
Code-switching erabiltzaileak	85

5.1 Taula: *Corpus* gordinaren ezaugarri orokorrak.

Corpusaren anotatzea

Corpusaren anotatzea aditu batek manualki burututako prozesua izan da. Hasierako anotazioa aditu batek manualki egitea beharrezkoa da, zehaztasuna eta fidagarritasuna lortzeko proiektuaren abiapuntuan. Aurretik EUS-ES *code-switching* analizatzen duen beste proiekturen bat egotekotan, prozesu hau proiektu horrek erabilitako etiketatzaile automatikoarekin egin zitekeen. Hala ere, kasu horretan ere, eskuzko gainbegiratzea beharrezkoa litzateke *corpus*aren fidagarritasuna bermatzeko.

[1] Bestelako txio bezala sailkatu dira euskarazko eta gaztelaniazko tokenik ez dauzkaten txioak.

Prozesu honetan hiru pausu eman dira: alde batetik, anotatzeko arauak dituen gidalerroa zehaztu da, bestetik, etiketatze fasean lagunduko duen interfaze grafikoa garatu da, eta bukatzeko, aditu batek *corpus*aren eskuzko anotazioa egin du.

Anotatzeko gidalerroak

*Corpus*a anotatzeko arauak finkatzea beharrezkoa da anotazioaren osotasuna bermatzeko. Gidalerro honek, anotazio prozesua burutuko duen pertsonari erreferentziatzat balioko dio.

Txioak banan-banan analizatuko dira eta token mailako anotazioa erabiliko da. Horretarako, txioak aurreprozesatu beharko dira tokenizatzaile baten laguntzaz. Tokenizatzaileak, txioen egitura eta berezitasunak kontutan hartu behar izango ditu. Izan ere, txioek egitura eta token bereziak dituzte, hala nola, traolak, erabiltzaile identifikatzaileak, URL-ak, emotikonoak, zenbakiak, puntuazio-markak, etab. Token desberdinak detektatzeko, espresio erregularrak erabiliko dira.

```
regex_str = [
        r'(?:@\w+)', #Twitter erabiltzaile izenak
        r"(?:\#\w+)", #Traolak
        r'(?:https?\S*)', # URLak
        r'(?:\d+(?:[.,]\d+)?)', #Zenbakiak
        r'(?:\w+)', #Hitzak
        r'(?:[?¿.,;:!¡()/\-"]+)', #Puntuazio-markak
        r'(?:\S)' #Bestelakoak, egongo balitz
]
```

Etiketak

Anotatzeko erabiliko diren etiketak baldintzatzen dituzte etiketatzaile automatikoaren emaitzak eta ebatzi nahi den problemaren araberakoak izango dira. Jarraian, *corpus*a anotatzerakoan erabiliko diren etiketen espezifikazioa agertzen da.

- **ES**, gaztelaniazko hitza adierazten du.

- **EUS**, euskarazko hitza adierazten du.
- **ID**, Twitter-eko erabiltzaile baten identifikatzailea adierazten du.
- **URL**, esteka bat adierazten du.
- **IE**, izen entitate bat adierazten du.
- **NH**, bi hizkuntzak token berdinean nahasita daudeneko kasua adierazten du.
- **ANB**, testuinguruaren arabera hitz anbiguoa deneko kasua adierazten du.
- **EG**, etiketa gabeko tokena adierazten du.

ES etiketa

ES etiketak gaztelaniazko hitza adierazten du. 5.2 taulan ikus daiteke ES etiketaren erabilpen kasu baten adibidea. Bi kasu desberdin bereizten dira:

1. Hitza gaztelaniazko testuinguruan bakarrik erabil daiteke eta inolako kasutan ez euskarazko testuinguruan.
2. Hitza bai gaztelaniazko eta bai euskarazko testuinguruan erabil daiteke, baina emandako testuinguruaren arabera gaztelaniazko hitza da.

Tokena	**Etiketa**
Oso	EUS
ona	EUS
!	EG
Así	ES
me	ES
gusta	ES
!	EG
3	EG
veces	ES
txapeldun	EUS
Gorka	IE

5.2 Taula: ES etiketaren adibidea

EUS etiketa

EUS etiketak euskarazko hitza adierazten du. 5.3 taulan ikus daiteke ES etiketaren erabilpen kasu baten adibidea. Bi kasu desberdin bereizten dira:

1. Hitza euskarazko testuinguruan bakarrik erabil daiteke eta inolako kasutan ez gaztelaniazko testuinguruan.

2. Hitza bai gaztelaniazko eta bai euskarazko testuinguruan erabil daiteke, baina emandako testuinguruaren arabera euskarazko hitza da.

Tokena	Etiketa
Oso	EUS
ona	EUS
!	EG
Así	ES
me	ES
gusta	ES
!	EG
3	EG
veces	ES
txapeldun	EUS
Gorka	IE

5.3 Taula: EUS etiketaren adibidea

ID etiketa

ID etiketak Twitter-eko erabiltzaile identzifikatzailearen formatua daukan tokena adierazten du, aurretik @ karakterea daukaten tokenak, hain zuzen ere. 5.4 taulan ikus daiteke ID etiketaren erabilpen kasu baten adibidea.

Tokena	Etiketa
Segi	EUS
horrela	EUS
@Mikel	ID
!!!!	EG
Tu	ES
puedes	ES
!	EG

5.4 Taula: ID etiketaren adibidea

URL etiketa

URL etiketak esteka formatua daukan tokena adierazten du. 5.5 taulan ikus daiteke URL etiketaren erabilpen kasu baten adibidea.

Tokena	Etiketa
Sistemari	EUS
ez	EUS
Di	ES
no	ES
al	ES
sistema	ES
https://t.co/HfV45F0ero	URL

5.5 Taula: URL etiketaren adibidea

IE etiketa

IE etiketak izen entitateak adierazten ditu. 5.6 taulan ikus daiteke IE etiketaren erabilpen kasu baten adibidea. Izen entitateak izen propio edo bereziak dira. Izen hauek pertsonak, lekuak, erakundeak, pelikulak, abestiak, etab., adierazteko erabiltzen dira. Normalean hizki larriz hasiko dira, baina ez kasu guztietan, testuaren izaera informala dela eta. Izen entitateak hitz bat baino gehiagokoak izan daitezke, eta gaztelaniazko kasuan artikuluak ere egon daitezke. Ager daitezkeen siglak ere IE etiketarekin bidez anotatuko dira.

Tokena	Etiketa
Gran	ES
bar	ES
La	IE
Reina	IE
del	IE
Sur	IE
!	EG
Bueltatuko	EUS
gea	EUS

5.6 Taula: IE etiketaren adibidea

NH etiketa

NH etiketak nahasitako hitza adierazten du, hau da, bi hizkuntzak nahasita daudeneko kasua, non lehenengo zatia gaztelaniazkoa den eta bigarren zatia, aldiz, euskarazkoa, edo alderantziz. 5.7 taulan ikus daiteke NH etiketaren erabilpen kasu baten adibidea.

Tokena	Etiketa
Voy	ES
a	ES
egituratuar	NH
mi	ES
trabajo	ES

5.7 Taula: NH etiketaren adibidea

ANB etiketa

ANB etiketak testuinguruaren arabera anbiguoak diren hitzak adierazten ditu. Hitz anbiguoak bi hizkuntzetakoak izan daitezke, adibidez: *familia, sistema, goza*, etab. Hitz hauek egokiak dira bi hizkuntzetan eta testuinguruaren arabera ezin badaiteke zehaztu zein hizkuntzatakoa den ANB etiketa erabiliko da. 5.8 taulan ikus daiteke ANB etiketaren erabilpen kasu baten adibidea.

Tokena	Etiketa
Sistema	EUS
honek	EUS
ez	EUS
du	EUS
iraungo	EUS
,	EG
claro	ES
que	ES
no	ES
!	EG
Familia	ANB

5.8 Taula: ANB etiketaren adibidea

EG etiketa

EG etiketak agertzen diren bestelako tokenak adierazten ditu. Kasu honetan sartuko dira bestelako hizkuntzak, hitz ulertezinak, emotikonoak, letrarik ez daukaten tokenak (puntuazio markak, zenbakiak...), RT tokena, etab. 5.9 taulan ikus daiteke EG etiketaren erabilpen kasu baten adibidea.

Tokena	Etiketa
Lol	EG
ikaragarria	EUS
!	EG
Je	EG
suis	EG
Charlie	IE
Hhjagsdfjak	EG

5.9 Taula: EG etiketaren adibidea

Bestelako aipamenak

karaktereaz hasten diren traolak token bakar baten moduan tratatuko dira eta aurreko etiketen espezifikazioa erabiliz etiketatuko dira, hitz arrunt bat izango balitz bezala. Ahal den kasuetan, dagokien hizkuntza zehaztuko da. Ezinezkoa den kasuetan, hitz arrarotzat hartuko dira eta EG etiketa erabiliko da. 5.10 taulan ikus daiteke traolen anotatze adibidea.

Tokena	Etiketa
Gran	ES
actuación	ES
Goazen	EUS
#TuCaraMeSuena	ES

5.10 Taula: Traolen anotatze adibidea

Gaizki idatzitako hitzak dagokien hizkuntzan etiketatuko dira, beti ere ulergarriak badira. Ulergarriak ez badira, EG etiketarekin anotatuko dira, hitz arrarotzat hartuz. 5.11 taulan ikus daiteke gaizki idatzitako hitzak anotatzeko adibidea.

Tokena	Etiketa
K	ES
grande	ES
eres	ES
!!!	EG
Leohiak	EUS
beti	EUS

5.11 Taula: Gaizki idatzitako hitzen anotatze adibidea

Luzatutako hizkiak normalizatu egingo dira lehenik tokenizazio fasean, Adibidez *goazeeeeeen!* hitza normalizazioa ondoren *goazen!* izango da.

Laburduren kasuan, dagokien hitz osoaren hizkuntzaren etiketa erabiliko da, beti ere ulergarriak badira. Ezinekoa den kasuetan EG etiketa erabiliko da. 5.12 taulan ikus daiteke hitzen laburdurak anotatzeko adibidea.

Tokena	Etiketa
Mz (Maite zaitut)	EUS
Mikel	IE
Q (que)	ES
ganas	ES
de	ES
verte	ES

5.12 Taula: Laburtutako hitzen anotatze adibidea

Anotatzeko interfaze grafikoa

Anotatze fasea azkartzeko eta erosoagoa bihurtzeko Java programazio lengoaiaz interfaze grafiko bat garatu da. Honek CSV formatuan dagoen *corpus*aren fitxategia irakurtzen du eta interfaze grafiko bidez txioak banan-banan erakusten ditu, tokenizatuta.

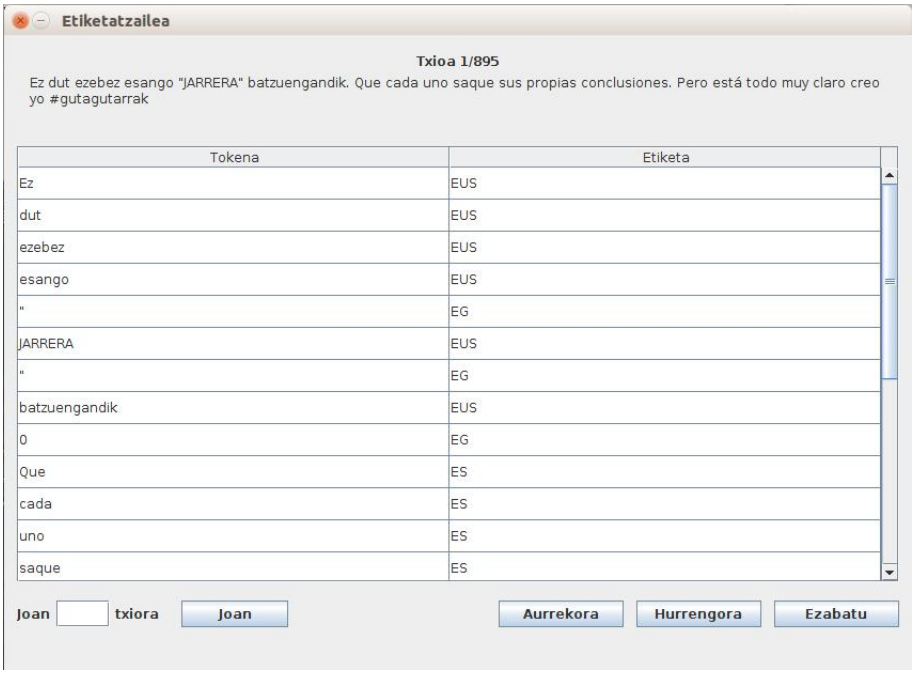
Interfaze grafikoak txioaren testu osoa erakusten du testuingurua ulertzeko. Izan ere, testuinguruaren arabera erabikiko da token bakoitzaren anotazioa. Horrez gain, testuko token bakoitzeko, gidalerroan aipatutako etiketen artean aukeratzeko zerrenda hedagarri bat eskaintzen da. 5.1 irudian ikus daiteke interfaze grafikoaren pantaila-irudi bat.

Interfazeak txioetan zehar nabigatzeko aukera ematen du, *Aurrekora* eta *Hurrengora* botoien bidez. Gainera, *Joan* botoiaren bidez, nahi den txioa atzitu daiteke, txioaren zenbakia sartuta. Txioak *corpus*etik ezabatzeko aukera dago, *Ezabatu* botoiaren bitartez. Lau botoi hauek *corpus*a gorde egiten dute sakatzerakoan, sarrerako CSV fitxategia eguneratuz.

Interfaze grafikoa erabiltzeko exekutagarri sortu da. Linux-en erabiltzeko adibididez, terminaletik honako komandoa erabiliko da:

```
java -jar Etiketatzailea.jar corpus_CSV_fitxategia
```

Ezabatutako txioak *DeletedTweets.csv* fitxategian gordeko dira, geroago erabili nahi badira. Hala ere, ez dira anotatuta gordeko. Interfaze grafikoaren kodea eta exekutagarria GitHub plataformaren bitartez eskuragarri daude, honako helbidean: https://github.com/anderleich/CodeSwitchingDetection



5.1 Irudia: Anotatzeko interfaze grafikoa.

Eskuzko anotatzea

*Corpus*aren eskuzko anotatzea aditu batek burutu du, garatutako interfaze grafikoaren laguntzaz. Prozesu luzea denez, garapen fasearekin jarraitu ahal izateko bertsio desberdinak erabili dira periodikoki probak egiteko. Lortutako bukaerako *corpus* anotatua erabili da CRF etiketatzailearen ezagutza iturritzat eta lortutako emaitzak CRF ebaluazio eta test faseetan ikus daitezke, 5.4 atalean. 5.13 taulan ikus daiteke anotazioaren ondoren *corpus*aren etiketen distribuzioa eta token kopuru totala.

Token kopurua	34771
ES token kopurua	13891
EUS token kopurua	7234
ID token kopurua	1399
URL token kopurua	740
EG token kopurua	7939
ANB token kopurua	112
NH token kopurua	82
IE token kopurua	3353

5.13 Taula: *Corpus* anotatuaren etiketen distribuzioa.

Etiketatzailearen esperimentazioa

Atal honetan CRF algoritmoa ebaluatzerako orduan egindako esperimentazio desberdinak azaltzen dira. CRF algoritmoak sekuentzia-elementu bakoitzeko ezaugarriak (*features*) onartzen ditu, hau da, elementu horri buruzko informazioa ematen duten hainbat aldagai. Ezaugarri hauek ebatzi nahi den problemaren araberakoak izango dira, kasu honetan hitzen eta testuaren analisian lagunduko dutenak. Ezaugarri hauek doituz algoritmoaren asmatze-tasak alda daitezke. Honen harira, zazpi esperimentu desberdin burutu dira. Jarraian, esperimentu horiek zeintzuk izan diren ikus daitezke eta hurrengo ataletan banan banan azalduko dira.

- Uneko sekuentzia elementua bakarrik kontutan hartu edo aurreko eta hurrengo *n* bizilagunak kontutan hartu.
- Uneko sekuentzia tokenaren azken *n* hizkiak kontutan hartu.
- Uneko sekuentzia tokenaren hasierako *n* hizkiak kontutan hartu.
- Uneko sekuentzia tokena titulua den kontutan hartu, hau da, lehenengo hizkia hizki larria da.
- Uneko sekuentzia tokena puntuazio ikurra da.
- Uneko sekuentzia tokenak azentua dauka edozein posiziotan.
- Uneko sekuentzia tokena hizki larriz idatzia dago.

Baseline metrikaren kalkulua

Ikasketa automatikoaren sailean, *baseline* deritzo datu multzo baten iragarpena egiteko metodoari. Honen bitartez problemaren zailtasun maila ikus daiteke lortutako metrikaren arabera. Metrika hau ikasketa automatikoko algoritmoaren errendimenduaren erreferentzia puntutzat, *baseline*, hartzen da eta algoritmoaren esperimentazio desberdinek lortutako asmatze-tasak konparatzeko erabilgarria da. Erreferentzia puntuaren azpitik lortutako emaitzek algoritmoaren aukeraketa edo entrenamendu maila desegokia adierazten dute.

*Code-switching corpus*aren *baseline* metrika lortzeko, honako heuristikoa erabili da, zehaztutako ordena zehatzean:

- Aurreko tokena ez bada puntua, harridura-ikurra edo galdera-ikurra eta uneko tokena hizki larriz hasten bada, orduan IE etiketa esleituko da.
- Uneko tokena hizki larriz osatuta badago, orduan IE etiketa esleituko da.
- Uneko tokena gaztelaniazko hiztegian baldin badago, orduan ES etiketa esleituko da.
- Uneko tokena euskarazko hiztegian baldin badago, orduan EUS etiketa esleituko da.
- Uneko tokena espresio erregular bidez esteka dela detektatzen bada, orduan URL etiketa esleituko da.
- Uneko tokena espresio erregular bidez Twitter-eko erabiltzaile identifikatzailea dela detektatzen bada, orduan ID etiketa esleituko da.
- Bestelako kasuetan, EG etiketa esleituko da.

Baseline kalkuluaren emaitzak, ondorengo ataletan egindako esperimentuen ontasuna neurtzeko erreferentzi puntutzat hartuko dira. 5.14 taulan ikus daitezke *baseline* kalkuluaren emaitzak.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza amatzea
Baseline	0.77759	0.03853	0.73484

5.14 Taula: *Baseline* asmatze-tasak ehunekotan.

1. Esperimentua: Kontutan hartutako bizilagun kopurua

Lehenengo esperimentu honetan, sekuentziaren dependentziak aztertu dira, hau da, aurreko eta hurrengo bizilagunek, tokenek, uneko tokenaren etiketa zehazteko daukaten eragina. Horretarako, lau proba kasu aztertu dira:

- Uneko tokenak bakarrik osatzen du sekuentzia elementuaren ezaugarri multzoa.
- Aurreko, uneko eta hurrengo tokenek osatzen dute sekuentzia elementuaren ezaugarri multzoa.

- Aurreko bi tokenek, uneko tokenak eta hurrengo bi tokenek osatzen dute sekuentzia elementuaren ezaugarri multzoa.
- Aurreko hiru tokenek, uneko tokenak eta hurrengo hiru tokenek osatzen dute sekuentzia elementuaren ezaugarri multzoa.

5.15 taulan ikus daitezke CRF algoritmoa entrenatu eta ebaluatu ondoren lortutako asmatze-tasak. Ikus daitekeenez, emaitza hoberenak ondoz ondoko lehen bizilagunak bakarrik kontutan hartzerakoan lortu dira.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Uneko tokena bakarrik	0.84967	0.18182	0.85227
Aurrekoa, unekoa, hurrengoa	0.85356	0.19318	0.84091
2 aurreko, unekoa, 2 hurrengo	0.84883	0.16185	0.83815
3 aurreko, unekoa, 3 hurrengo	0.84153	0.13095	0.82738

5.15 Taula: 1. esperimentazioaren asmatze-tasak.

2. Esperimentua: Uneko tokenaren azken n hizkiak

Esperimentu honen helburua tokenen arteko patroiak finkatzea da. Jakina da hizkuntza bakoitzak sarri agertzen diren atzizki edo hitz bukaerak dituela. Gaztelaniazko kasuan, adibidez, *ión, mente, ar, er, ir*, etab. Euskarazko kasuan, aldiz, *ekin, aren, ki, tzen*, etab. Hori dela eta, $n = 1,2,3,4$ kasuak aztertu dira.

5.16 taulan ikus daitezke CRF algoritmoa entrenatu eta ebaluatu ondoren, kasu bakoitzean lortutako asmatze-tasak. Ikus daitekeenez, emaitza hoberenak $n = 1,2,3,4$ deneko kasuan lortu dira.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
n=1	0.88494	0.24432	0.85227
n=2	0.89032	0.26136	0.90909
n=3	0.87806	0.23864	0.85795
n=4	0.86820	0.22159	0.85795
n=1,2	0.89420	0.26705	0.89205
n=1,2,3	0.89480	0.26705	0.87500
n=1,2,3,4	0.89779	0.26705	0.86364

5.16 Taula: 2. esperimentazioaren asmatze-tasak.

3. Esperimentua: Uneko tokenaren hasierako n hizkiak

Esperimentu honen helburua tokenen arteko patroiak finkatzea da. Aurreko kasuan bezala, hizkuntza bakoitzak sarri agertzen diren aurrizki edo hitz hasierak dituela. Gaztelaniazko kasuan, adibidez, *des, re, pre*, etab. Euskarazko kasuan, aldiz, *des, erre, aurre*, etab. Kasu batzuetan komunean izango dute aurrizkia, baina gehienetan hizkuntza antzemateko ezaugarri esanguratsua izango da. Honetaz aparte, Twitter erabiltzaile izenek beti @ karakterea edukiko dute aurretik eta estekek *http* karaktere segida. Ondorioz, hauek atzemateko modua ere izango da. Hori dela eta, $n = 1,2,3,4$ kasuak aztertu dira.

5.17 taulan, CRF algoritmoa entrenatu eta ebaluatu ondoren kasu bakoitzean lortutako emaitzak ikus daitezke. Ikus daitekeenez, emaitza hoberenak, $n = 1,2,3$ deneko kasuan lortu dira. Aurreko kasuan ez bezala, kasu honetan $n = 3$ arteko atzizkiak bakarrik kontutan hartzea nahikoa izan da.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
n=1	0.92050	0.36932	0.90341
n=2	0.91990	0.35795	0.89773
n=3	0.91632	0.35227	0.89773
n=4	0.91213	0.33523	0.89773
n=1,2	0.92050	0.36932	0.91477
n=1,2,3	0.92409	0.38636	0.90909
n=1,2,3,4	0.92319	0.39773	0.90909

5.17 Taula: 3. esperimentazioaren asmatze tasak.

4. Esperimentua: Uneko tokena titulua da

Esperimentu honen helburua sistemari izen entitateak etiketatzeko ezagutza ematea da. Horretarako, uneko tokena hizki larriz hasten den ala ez adieraziko da bere ezaugarrietan.

5.18 taulan ikus daitezke CRF algoritmoa entrenatu eta ebaluatu ondoren lortutako emaitzak. Ikus daitekeenez, emaitza hobeagoak lortu dira ezaugarri honen bidez, aurretik lortutakoekin alderatuz.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Aurrekoa	0.92409	0.38636	0.90909
Titulua da?	0.94142	0.47727	0.92045

5.18 Taula: 4. esperimentazioaren asmatze-tasak.

5. Esperimentua: Uneko tokena puntuazio ikurra da

Esperimentu honen helburua sistemari puntuazio ikurrak detektatzeko ezagutza ematea da. Horretarako, uneko tokena puntuazio ikurra den ala ez adieraziko da bere ezaugarrietan.

5.19 taulan ikus daitezke CRF algoritmoa entrenatu eta ebaluatu ondoren lortutako emaitzak. Ikus daitekeenez, emaitzak okertu egin dira puntuazio-marka den ezaugarria kontutan hartzerakoan. Ondorioz, ezaugarri hau ez da kontutan hartuko eta aurreko esperimentuko asmatze-tasekin jarraituko da.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Aurrekoa	0.94142	0.47727	0.92045
Puntuazio ikurra da?	0.93933	0.46591	0.90341

5.19 Taula: 5. esperimentazioaren asmatze tasak.

6. Esperimentua: Uneko tokenak azentua dauka

Esperimentu honen helburua sistemari gaztelaniazko hitzak detektatzeko ezagutza gehiago ematea da. Horretarako, uneko tokenak hizki azentudun bat daukan ala ez adieraziko

da bere ezaugarrietan.

5.20 taulan ikus daitezke CRF algoritmoa entreantu eta ebaluatu ondoren lortutako asmatze-tasak. Ikus daitekeenez, kasu honetan ere azentudun tokena den adierazten duen ezaugarriak emaitzak okertu ditu. Ondorioz, ez da ezaugarri hau kontutan hartuko eta 4. esperimentuan lortutako asmatze-tasak kontutan hartuko dira.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Aurrekoa	0.94142	0.47727	0.92045
Azentuduna da?	0.94142	0.48864	0.91477

5.20 Taula: 6. esperimentazioaren asmatze tasak.

7. Esperimentua: Uneko tokena hizki larriz dago

Esperimentu honen helburua sistemari izen entitateak detektatzeko ezagutza gehigarria ematea da. Izan ere, kasu gehienetan hizki larriz idatzita dagoen tokenak siglak direla adieraziko du. Kontutan eduki behar da, hizki larriz idatzita dauden txioetan ez dela horrela izango.

5.21 taula ikus daitezke CRF algoritmoa entrenatu eta ebaluatu ondoren lortutako emaitzak. Ikus daitekeenez, kasu honetan ere ez dira emaitzak hobetu. Ondorioz, tokena hizki larriz idatzita dagoen ez da ezaugarritzat hartuko eta 4. esperimentuan lortutako asmatze tasak kontutan hartuko dira.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Aurrekoa	0.94142	0.47727	0.92045
Hizki larriz dago?	0.93903	0.47159	0.92045

5.21 Taula: 7. esperimentazioaren asmatze tasak.

Test fasea

Test fase honen helburua ebaluazio fasean gerta daitekeen *overfitting*-a ekiditea da. Gerta daiteke ebaluazio fasean erabilitako entrenamendu multzoa gehiegi doitu izana ebaluazio

multzoak asmatze-tasa nabargarriak emateko. Ondorioz, etiketatzailearen erabilgarritasuna murriztu egiten da, ebaluazio kasuetarako bakarrik emaitza onak lortuz. Test fasearen bitartez, ebaluazioan zehar ezaugarri-funtzioei egindako doiketak balioztatzen dira, bai eta etiketatzailearen ontasuna ebaluatu.

Test fase honetan metrika desberdinak kalkulatu dira etiketatzailearen inguruan. Metrika hauen bitartez sailkatzailearen ontasunaren inguruko informazioa lortu daiteke.

5.22 taulan ebaluazio eta test faseetan lortutako asmatze tasen arteko konparazioa ikus daiteke. Hiru asmatze-tasa desberdin konparatzen dira, etiketa mailako asmatzea, sekuentzia mailako asmatzea eta hizkuntza asmatzea (EUS, ES edo *Code-switching*).

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Ebaluazio fasea	0.94142	0.47727	0.92045
Test fasea	0.93072	0.41477	0.86932

5.22 Taula: Ebaluazio eta test faseen asmatze tasen konparazioa.

5.23 taulan etiketa posible bakoitzeko *precision, recall, F1-score* eta *support* balioak ikus daitezke. *Precission* balioak etiketa bat zuzen iragarri deneko kasuen ehunekoa adierazten du. *Recall* balioak etiketa baten test multzoko agerpen guztietatik zuzen etiketatutako ehunekoa adierazten du. *F1-score* balioak *precission* eta *recall* balioen arteko erlazioa adierazten du. *Support* balioak etiketa bakoitzeko test multzoko agerpen kopuru totala adierazten du.

Etiketa	Precision	Recall	F1-score	Support
ID	1.000	1.000	1.000	145
IE	0.813	0.703	0.754	353
EG	0.984	0.984	0.966	789
NH	0.000	0.000	0.000	10
ANB	0.714	0.556	0.625	9
URL	1.000	1.000	1.000	69
ES	0.926	0.969	0.947	1276
EUS	0.916	0.953	0.934	769
batazbestekoa/ guztira	0.927	0.931	0.928	3420

5.23 Taula: Test multzoko etiketa bakoitzeko *precission, recall, F1-score* eta *support* emaitzak.

Azkenik, 5.24 taulan CRF etiketatzaileak *code-switching* txioen eta txio elebakarren *precision, recall, F1-score* eta *support* balioak ikus daitezke.

Mota	Precision	Recall	F1-score	Support
Code_switching	0.85149	0.92473	0.88660	101
Elebakarrak	0.89333	0.81707	0.85350	75

5.24 Taula: Test multzoko *code-switching* txioen eta txio elebakarren *precission, recall, F1-score* eta *support* balioak.

Etiketatzailearen ontasun maila modu zehatzagoan kalkulatu ahal izateko eta bukaerako balidazioa burutzeko *cross-validation* metodoa erabili da. Horrela, algoritmoaren asmatze tasen batezbestekoa eta desbideratze tipikoa kalkulatu dira. 5.25 taulan 10 iterazio ondoren lortutako emaitzak ikus daitezke.

	Etiketa asmatzea	Sekuentzia asmatzea	Hizkuntza asmatzea
Batazbestekoa	0.93822	0.44926	0.87477
Desbideratze tipikoa	0.00519	0.02773	0.02365

5.25 Taula: *Cross-validation* emaitzak.

Denbora errealean *code-switching*-a detektatu

Atal honetan proiektuaren produktu eta helburu nagusia den aplikazioa garatu da. Aplikazioaren betekizuna txioak denbora errealean atzitzea da eta hauen ondorengo *code-switching* fenomenoaren detekzioa burutzea. Txioak denbora errealean lortzeko, Twitter API-aren *streaming* funtzioa erabili da. Funtzio honek txioak erabiltzaile konkretu baten arabera edo hitz konkretu baten arabera txioak iragazteko aukera ematen du. Aplikazioa Python programazio lengoaia erabiliz garatu da eta kodea GitHub bidez atzigarri dago honako helbidean: `https://github.com/anderleich/CodeSwitchingDetection`

Sarrerako parametro moduan *corpus* anotatuaren fitxategi izena, JSON formatuan emaitzak itzultzeko fitxategiaren izena, hitz edo erabiltzaile konkretu baten araberako bilaketa den zehaztuko duen aukera, eta aurreko aukeran zehaztutako hitz edo erabiltzaile zerrenda. Komandu kontsolatik exekutatu daiteke honako komanduaren bidez:

```
usage: TweepyStreaming.py [-h] corpus output {word,user} query [query ...]

Performs code switching analysis in Twitter streaming

positional arguments:
corpus       Name of the corpus file, with extension
output       Name of the JSON output file, no extension
{word,user}  Search by word or by user
query        List of queries (words or users)

optional arguments:
-h, --help   show this help message and exit
```

Aplikazioak emaitza moduan analizatutako txio guztiak JSON formatuan idazten ditu zehaztutako fitxategian. Txio guztiek egitura bera jarraitzen dute eta jarraian JSON fitxategi horren eremuak azaltzen dira:

```
{ tweets : [
        text : txioaren testua,
        tweet_id : txioaren identfikatzaile zenbakia,
        user_id : erabiltzailearen identifikatziale zenbakia ,
        user_image : erabiltzailearen profilaren irudia,
        user_name : erabiltzailearen izena,
        user_screenname : erabiltzailearen Twitter erabiltzaile izena,
        tags : [
                token: tokenaren karaktere katea,
                tag: tokenari dagokion etiketa
                ],
        language : ES|EUS|CS, code-switching detekzioa
        ]
}
```

Emaitzen errepresentazioa

Atal hau ez zegoen hasierako irismenaren barruan aurreikusita. Aurreko fasearen garapenean ordea, emaitza gisa itzultzen zen JSON fitxategia modu ulergarriagoan errepresentatzeko beharra ikusi zen. Hori dela eta, emaitzen errepresentazio dotoreagoa eta ulergarriagoa garatu da web interfaze sinple batez. 5.2 irudian ikus daiteke interfazearen pantaila-irudi bat.

Web interfaze honek, JSON fitxategia irakurtzen du Javascript bidez. Javascript lengoaia erabiltzea erabaki da, programazio lengoai honek dagoeneko liburutegiak dituelako JSON fitxategiak modu erraz eta azkarrean tratatzeko.

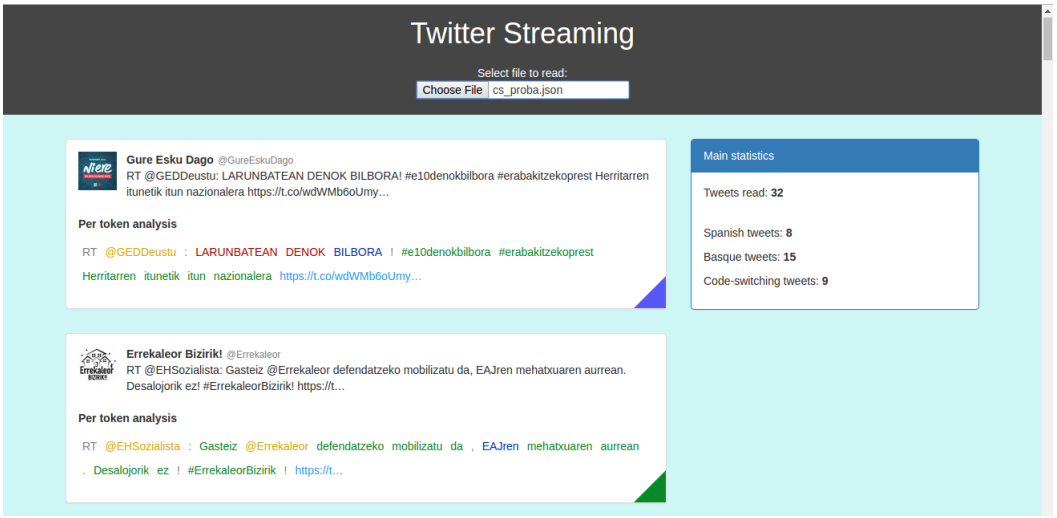
Interfaze grafikoa edozein nabigatzailetan exekutatu daiteke. Irakurri nahi den JSON fitxategia zehaztu behar da. Ondoren, interfazeak fitxategia irakurtzen du eta txioak banan banan erakusten ditu Twitter-eko itxura jarraituz. Gainera, etiketatzaile automatikoak sortutako etiketak erabiltzen dira testuko token bakoitza kolorez adierazteko. Honakoa da kolore banaketa:

- **Gorria**: ES etiketa duen tokena.
- **Berdea**: EUS etiketa duen tokena.
- **Horia**: ID etiketa duen tokena.
- **Urdin argia**: URL etiketa due tokena.
- **Grisa**: EG etiketa duen tokena.
- **Urdin iluna**: IE etiketa duen tokena.
- **Beltza**: ANB eta NH etiketak dituzten tokenak.

Gainera, txio bakoitzari *code-switching* analisiari dagokion etiketa orokorra esleitzen zaio. Hau ere kolorez adierazten da:

- **Gorria**: gaztelaniazko txioa.
- **Berdea**: euskarazko txioa.
- **Morea**: *code-switching* adibidea den txioa.

Aplikazioak sarrerako fitxategia irakurtzen duenean, zenbait datu orokor ere ematen ditu. Hala nola, irakurritako txio kopurua, euskarazko txio kopurua, gaztelaniazko txio kopurua eta *code-switching*-aren adibideak diren txioak.



5.2 Irudia: Emaitzak modu dotorean adierazteko web interfazea.

6. KAPITULUA

Ondorioak

Code-switching fenomenoa bi hizkuntza edo gehiago modu trukagarrian erabiltzean gertatzen den fenomeno linguistikoa da. Hizkuntza bat baino gehiago elkarbizitzan dauden zonaldeetan gertatzen da eta Euskal Herria zonalde interesgarria da fenomeno honen ikerketarako. Izan ere, euskara eta gaztelania hizkuntza ofizialak dira eta erabilera zabaldua dute herritarren artean. Hori dela eta, ahozko nahiz idatzizko erregistro informalean *code-switching* fenomenoaren adibide ugari aurki daitezke.

Fenomeno linguistiko honen detekzioak aurrerapauso handiak suposatzen ditu hizkuntzaren teknologien eremuan. Izan ere, *code-switching*-a detektatzea beharrezkoa da lengoaia naturalaren prozesaketarako. Analizatzaile sintaktikoak, itzultzaile automatikoak edota hizkera detektatzaileak dira hizkuntza teknologien adibide batzuk. Aplikazio hauek funtzionaltasun egokia eduki dezaten, beharrezkoa da *code-switching*-a detektatzea eta ulertzea. Proiektu honek bide horretan lehen pausuak ematea izan du helburu eta lortutako emaitzak etorkizun batean lengoaia naturalaren detekziorako hainbat aplikazioren abiapuntutzat har litezke.

Proiektu honetan, Twitter sare soziala erabiltzea erabaki da, plataformaren izaera informala dela eta. Izan ere, Twitter-eko txioek egitura informala eta lasaia daukate, ahozko erregistroaren parekoa. Honek abantailak ditu *code-switching* fenomenoa aztertu behar denean, kasu natural eta errepresentagarri asko aurki daitezkeelako. Ahozko grabaketekin konparatuta, eguneroko gaien inguruko adibide naturalagoak lor daitezke Twitter bidez, grabaketetan naturaltasun edo espontaneotasun hori galdu egiten baita. Hori dela eta,

code-switching fenomenoaren azterketan lehen pausuak emateko aukera egokia da Twitter plataforma.

Proiektu honen garapen fasean, 2 atalean planifikatutako helburu eta betebehar guztiak burutu dira. Horretarako, Twitter plataformatik erauzitako txioekin *corpus* anotatu errepresentagarria osatu da, euskarazko eta gaztelaniazko txio elebakarrak eta *code-switching* adibideak lortuz. Anotazioa eskuz garatu du aditu batek, garatu den interfaze grafikoaren laguntzaz. Anotazio fase horretan, token bakoitzaren etiketa zehaztu da. Ondoren, *corpus*a ezagutza iturritzat hartuz, CRF sekuentzia etiketatzailea erabiltzea erabaki da, sekuentziak tratamendurako daukan izaera dela eta. CRF-aren ebaluazio eta test faseek aurretik kalkulatutako *baseline*-aren asmatze tasak hobetu dituzte %93 inguruko asmatze tasa lortuz sekuentzia elementu etiketazioan. Gainera, %86 inguruko asmatze tasak lortu dira hikuntzaren detekziorako, hau da, gaztelania, euskaraz edo *code-switching* etiketentzat. Ondorioz, CRF etiketatzailearen aukeraketa egokia suertatu da eta 3.1 puntuan azaldutako ikerketan lortutako emaitzekin bat dator.

Hizkuntza parea	Entrenamendu fasea	Ebaluaketa fasea	Test fasea	Guztira
MSA-DA	8862	1117	1258	11237
SPA-ENG	8733	1857	10716	21306
EUS-ES	1412	176	176	1764

6.1 Taula: Beste hizkuntzetan egindako *code-switching* ikerketetan erabilitako *corpus*en ezaugarrien konparaketa.

Mota	Taldea	Elebakar F1	Code-switching F1
SPA-ENG	(Al-Badrashiny and Diab, 2016)	0.83	0.69
	(Xia, 2016)	0.86	0.79
	(Samih et al., 2016)	0.92	0.88
	(Shrestha, 2016)	0.90	0.86
	(Sikdar and Gambck, 2016)	0.91	0.87
MSA-DA	(Shrestha, 2016)	0.72	0.34
	(Al-Badrashiny and Diab, 2016)	0.83	0.37
	(Samih et al., 2016)	0.89	0.50
EUS-ES	-	0.85	0.89

6.2 Taula: Beste hizkuntzetan egindako *code-switching* ikerketetan emaitzen *F1-score* konparaketa.

Proiektuaren helburu nagusia eta produktua zen denbora errealeako EUS-ES *code-switching*-aren detekzio automatikoa burutzea lortu da. Horretarako, aplikazio bat garatu da den-

bora errealean atzitutako txioak automatikoki etiketatzen dituena. Aplikazioak emaitzak JSON formatuan itzultzen dituenez, beharrezkoa ikusi da emaitza hauek modu ulergarriagoan azalduko dituen web interfazea garatzea. Aplikazioa zein web interfazea atzigarri daude GitHub plataforman honako helbidean: https://github.com/anderleich/CodeSwitchingDetection

Proiektuaren garapen iteratiboak abantaila aipagarriak ekarri ditu gertatutako atzerapen eta arazoen aurrean. Izan ere, iterazioen bidez garapenaren jarraipen eta kontrola eraman da, iterazio bakoitzean azaldutako arazo eta hobekuntzei soluzioa aurkituz. Gainera, proiektuaren tutorearekin eta *corpus* anotatzailearekin egindako jarraipen bileren bidez, besteen ikuspuntua ulertzea lortu da eta hauen helburuak modu zehatzagoan betetzea lortu da.

Etorkizuneko lanak

Proiektu honen irismenean zehaztutako guztia egitea lortu da. Gainera hasierako irismenean aurreikusita ez zegoen web interfazearen garapena gehitu da. Hala ere, zehaztutako irismenak EUS-ES *code-switching*-aren detekzioan lehen urratsak ematea helburu du eta beraz, oraindik garapen urrats gehiago egitea posible da.

Horregatik, jarraian etorkizun batean egin daitezkeen hobekuntza edota funtzionalitate gehigarri batzuk aipatzen dira:

- *Corpus* anotatuaren adibide kopurua handitzea, sistemaren ezagutza iturria osatzeko eta ondorioz ikasketa prozesua hobetzeko.
- Anotazioan erabili diren etiketen aukera kopurua handitzea, beharretara egokitzeko. Proiektu honetan etiketa bakar batzuk aukeratu dira bakarrik, proiektuaren helburuetarako nahikoak zirelako. Hala ere, etiketa aukera kopurua handitzeak doitasun handiagoa emango dio sistemari eta detektatu nahi diren tokenetara egokitu daiteke.
- Bestelako *machine learning* algoritmo etiketatzaileak erabiltzea, adibidez, SVM.
- Web aplikazio bat garatzea *code-switching* detekzioa *online* burutzeko.

Ikasitako lezioak

Proiektua aurrera eramateko eta zehaztutako helburuak betetzeko graduan zehar eskuratutako gaitasunak ezinbestekoak izan dira. Gaitasun hauek ez dira soilik graduko ikasgaietan eskuratutako gaitasun teorikoak izan, baizik eta metodologiarekin lotutakoak ere. Horregatik, programazio gaitasunak eta aspektu teorikoak bezainbeste garrantzitsu izan dira proiektu baten diseinu eta planifikazio gaitasunak, arazoen aurrean jokatzeko modua, eta orokorrean, jarraitutako metodologia.

Garapen faseko urratsak modu iteratiboan burutu izanak sekulako abantailak izan ditu eta bukaerako emaitzan eragina izan du. Izan ere, iterazio bakoitzean egindakoa ebaluatu da lortutakoaren kalitatea bermatzeko. Horrela, ikusitako arazoak konpontzea eta hobekuntzak aurrera eramatea lortu da. Proiektuaren jarraipena eta kontrola eramateko modu fidagarria da eta epemugak betetzeko beharrezkoa. Honetaz aparte, proiektuko zuzendariarekin edukitako jarraipen eta kontrol bilerak ezinbestekoak izan dira lanaren biderapenerako.

Ikasitako lezio guztietatik nabarmentzekoak dira:

- **Kodearen berrerabilpena**. Garapenean zehar dagoeneko inplementatuta zeuden liburutegi batzuk erabili dira. Proiektu baten aurrean, gurpila berriz ez asmatzeak denbora eta lan asko aurreztu dezake. Beraz, besteek egindakoa berrerabiltzea oso lagungarria suerta daiteke. Gaur egun, sarean informazio eta liburutegi asko aurki daitezke eta gomendagarria da hauek kontutan hartzea proiektuaren diseinu fasean. Proiektu honetan erabilitako liburutegi batzuk *Tweepy, Hunspell* eta *sklearn-crfsuite* izan dira.
- **Konputazio adarreko oinarri teorikoak praktikan ikustea**. Graduan zehar bakoitzaren gustuen arabera adar desberdinak aukeratzeko aukera dago. Aukeratutako adarrean aspektu teorikoak ikasten dira, zenbait lan praktiko bidez indartuz. Hala ere, gradu amaierako proiektuaren garapenak, aspektu teoriko horiek modu sakonago eta pertsonal batean praktikan jartzeko aukera ezin hobea eskaintzen du. Horrela, adarraren inguruan motibazioa sustatu egiten da, mundu errealean aplikagarri izan daitezkeen aplikazioak burutuz. Egindako proiektu honetan linguistikaren munduan EUS-ES *code-switching* detekzio automatikoan lehen urratsak ematea lortu da.
- **Etorkizunerako gaitasunak lortzea**. Gradu amaierako proiektua aurrera eramate-

rakoan, etorkizun batean erabilgarri suerta daitezkeen gaitasunak eskuratzen dira. Izan ere, lan zein inbestigazio munduan burutu beharreko lanekin oso erlazionatuta dago, non bezeroak produktu bat garatzeko eskatzen duen. Proiektuan zehar zuzendariak bezero paper hori betetzen du. Lan autonomoa sustatu egiten da. Gainera, epemugen garrantzia eta ondorioz proiektu planifikatu eta ondo diseinatu baten beharra ikusten da.

Bibliografia

[1] Giovanni Molina, Nicolas Rey-Villamizar, Thamar Solorio, *Dept. of Computer Science, University of Houston, Houston TX, 77004*, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Mona Diab, *Dept. of Computer Science, George Washington University, Washington DC, 20052 Overview for the Second Shared Task on Language Identification in Code-Switched Data* 2016.

[2] *Annotation guidelines for the Second Shared Task on Language Identification in Code-Switched Data* 2016. http://care4lang1.seas.gwu.edu/cs2/guidlines/Annotation_Guidelines_2016.pdf

[3] Prajwol Shrestha, Verscend Technologies Pvt. Ltd. *Codeswitching Detection via Lexical Features using Conditional Random Fields*, pages 121-126. 2016

[4] EMNLP 2016 *Second Workshop on Computational Approaches to Code Switching* 2016: http://care4lang1.seas.gwu.edu/cs2/call.html.

[5] Charles Sutton and Andrew McCallum *An Introduction to Conditional Random Fields* 2012.

[6] Sergio-Nabil Khayyat Arranz *Campos Aleatorios Condicionales* 2011: Universidad Autónoma de Madrid, Escuela Politécnica Superior, Departamento de Ingeniería Informática

[7] Data Science Central, The Online Resource For Big Data Practitioners *Conditional Random Fields (CRF): Short Survey*, posted by Nikitinsky Nikita, 2017. http://www.datasciencecentral.com/profiles/blogs/conditional-random-fields-crf-short-survey

[8] Wikipedia *F1 score*: 2017

[9] TutorialsPoint *Python - Tutorial*: 2017. https://www.tutorialspoint.com/python/

[10] W3Schools *Java tutorials*: 2017. http://www.w3schools.in/java-tutorial/intro/

[11] W3Schools *JavaScript Tutorial*: 2017. https://www.w3schools.com/Js/

[12] W3Schools *HTML5 Tutorial*: 2017. https://www.w3schools.com/html/

[13] W3Schools *CSS Tutorial*: 2017. https://www.w3schools.com/css/default.asp

[14] Twitter Developers *Twitter Developer Documentation, REST APIs*: 2017. https://dev.twitter.com/rest/public

[15] Tweepy *Tweepy: An easy-to-use Python library for accessing the Twitter API*: 2017. http://www.tweepy.org/

[16] Mikhail Korobov *sklearn-crfsuite*: 2017. http://sklearn-crfsuite.readthedocs.io/en/latest/

[17] Naoaki Okazaki *CRFSuite, A fast implementation of Conditional Random Fields (CRFs)*: 2017. http://www.chokkan.org/software/crfsuite/

[18] Hunspell: 2017. http://hunspell.github.io/

Eranskinak

A. ERANSKINA

Bilera aktak

1. Bilera akta

Data: 2017/03/24
Kideak: Ander Corral, Iñaki Alegria
Lekua: Informatika fakultatea, Donostia

Proiektuari hasiera emateko bilera. Bileran proiektuaren ideai nagusiak eta nondik norakoak zehaztu dira. Proiektuaren egutegia eta epe-mugak finkatu dira.
Erabili daitezkeen zenbait baliabide ikusi dira, bai eta bestelako hizkuntzetan egindako *code-switching* lanak aztertu dira.

Egitekoa:

- *Code-switching* inguruan egindako beste lanak gainbegiratu.
- Corpus osatzeko Twitter-eko erabiltzaileak bilatzen joan.
- Corpusa osatzeko txioak eskuratzeko programa garatzen hasi.
- Memoriarekin hasi, helburuak eta irismena bukatu.

2. Bilera akta

Data: 2017/04/25
Kideak: Ander Corral, Iñaki Alegria
Lekua: Informatika fakultatea, Donostia

Proiektuaren inguruko jarraipen eta kontrol bilera. Bileran orain arte egindakoaren garapena gainbegiratu da eta burutu beharreko hurrengo pausuak zehaztu dira.

Egitekoa

- Corpus gordina osatu
- Anotatze fasean lagunduko duen interfaze grafikoa garatu
- CRFen ingurua informazioa bilatu

3. Bilera akta

Data: 2017/05/16
Kideak: Ander Corral, Iñaki Alegria, Larraitz Uria
Lekua: Informatika fakultatea, Donostia

Corpusaren anotazioaren inguruko bilera. Larraitz adituarekin bilera egin da, corpus fitxategia entregatzeko eta garatutako interfaze grafikoa nola dabilen erakusteko. Anotazioan erabili beharreko gidalerroa ere azaldu da eta hobekuntza proposamenak zehaztu.

Egitekoa

- CRF liburutegia Python programazio lengoaiarentzat aukeratu.
- CRFen inplementazioa garatu.
- Anotatzeko gidalerroak hobekuntzekin eguneratu

4. Bilera akta

Data: 2017/05/17
Kideak: Ander Corral, Larraitz Uria
Lekua: Informatika fakultatea, Donostia

Korpusaren anotazioaren inguruko bilera. Larraitzekin bilera egin da anotatzeko gidalerroen inguruko zalantzak argitzeko.

5. Bilera akta

Data: 2017/05/31
Kideak: Ander Corral, Iñaki Alegria, Larraitz Uria
Lekua: Informatika fakultatea, Donostia

Proiektuaren jarraipen eta kontrol bilera. Corpus anotatuaren inguruko zalantzak argitu eta anotazioan egin beharreko aldaketak finkatu dira. Gainera orain arte egindako garapena gainbegiratu da, hala nola, CRFen inplementazioa, denbora errealean txioak eskuratzeko aplikazioa eta web interfazea.

Egitekoa

- Txio elebakarrak lortu eta anotatu
- CRFen inplementazioa hobetu, emaitza gehiago bistaratuz eta kodea txukunduz.
- Denbora errealean txioak eskuratzeko aplikazio hobetu.
- Web interfazea bukatu.
- Memorian aspektu teoriakoak idazten hasi.

6. Bilera akta

Data: 2017/06/06
Kideak: Ander Corral, Iñaki Alegria
Lekua: Informatika fakultatea, Donostia

Proiektuaren jarraipen eta kontrol bilera. Orain arte egindako garapena gainbegiratu da, hala nola, CRFen inplementazioa, denbora errealean txioak eskuratzeko aplikazioa eta web interfazea.

Egitekoa

- CRFen inplementazioa hobetu, emaitza gehiago bistaratuz eta kodea txukunduz.
- Denbora errealean txioak eskuratzeko aplikazioari erabiltzaile edo hitz konkretuen arabera bilaketak egiteko funtzionalitatea gehitu.
- Memoria osatze joan egindako garapenarekin.

7. Bilera akta

Data: 2017/06/13
Kideak: Ander Corral, Iñaki Alegria
Lekua: Informatika fakultatea, Donostia

Proiektuaren jarraipen eta kontrol bilera. Orain arte egindako garapena gainbegiratu da eta memoriaren lehen bertsioa gainbegiratu eta hobekuntzak proposatu dira.

Egitekoa

- Garatutako kodea GitHub plataformara igo
- Memoria osatu zehaztutako proposamenekin.

8. Bilera akta

Data: 2017/06/21
Kideak: Ander Corral, Iñaki Alegria, Larraitz Uria
Lekua: Informatika fakultatea, Donostia

Proiektuaren jarraipen eta kontrol bilera. Azken bilera izan da entrega aurretik. Memoriaren akatsak gainbegiratu dira eta azken proposamenak finkatu.

Egitekoa

- Garatutako azken kodea GitHub plataformara igo
- Memoria zuzendu eta bukatu zehaztutako proposamenekin.