Universidad del País Vasco Euskal Herriko Unibertsitatea

Muturretik muturrerako informazio erauzketa eleaniztuna eta hizkuntzen arteko terminoen lerrokatze neuronala

Egilea: Mari Susperregi Indakoetxea
Tutorea: Olatz Perez de Viñaspre Garralda, Ander Barrena Madinabeitia

HAP/LAP

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko proiektua

2020ko iraila

Sailak: Lengoaia eta Sistema Informatikoak, Konputagailuen Arkitektura eta Teknologia, Konputazio Zientziak eta Adimen Artifiziala, Euskal Filologia, Elektronika eta Telekomunikazioak.

Laburpena

Lan honen helburua testu-bikote elebidunetatik terminoak lortzea da, hau da, euskarazko eta gaztelaniazko esaldi parekatuak emanda, esaldi horietan agertzen diren termino-bikote esanguratsuak lortzea: glosarioak egiteko, hiztegiak aberasteko, NLP atazetan erabili ahal izateko... Helburu hori lortzeko *Erauzterm* eta *Itzulterm* (antzeko helburua lortzeko aurretik egindako tresnak) tresnekin lortutako corpusa erabili da; corpus horiek eta sare neuronalen teknikak erabiliz, *Itzulterm* tresna gaur egungo teknologietara eguneratzea izan da helburua. Lan hau garatzeko sekuentziatik sekuentziarako hurbilpenean oinarritua dagoen *transformer* teknologia erabili da. Orain arte erabili diren metodo linguistiko eta estatistikoak erabili ordez sare neuronalen teknikak erabili dira ataza hau garatzeko; hau da, hizkuntza bakoitzeko terminoak erauzi eta terminoak lerrokatzea bata bestearen ondoren egin ordez, ekintza guztiak aldi-berean egingo dira, muturretik muturrerako ataza bihurtuz eta erroreen propagazioa gutxiagotuz. Sistema ebaluatzeko BLEU metrika erabiltzeaz gain, lan honetarako berariaz sortu den TEB (*Termino-Erauzle Balidazioa*) metrika ere erabili da. Metrika horrek BLEUk kontuan hartzen ez dituen eta terminologia-erauzketarako garrantzitsuak diren ezaugarri batzuk hartzen ditu kontuan ebaluazioa egiteko.

Master amaierako lan honetarako garatutako sistemak BLEU metrikan 0,78 puntuko balioa lortu du. Eta ebaluaziorako erabili den oinarri lerroarekiko BLEU metrikan 50 puntuko hobekuntza lortu du. Ondorioz, terminologia-erauzketa ataza sare neuronalen teknologiak erabiliz garatu daitekeela frogatu da.

Abstract

The aim of this work is to obtain terms of bilingual text pairs. In other words, using matched phrases in Basque and Spanish, obtain pairs of significant terms that appear in the phrases for multiple purposes: making glossaries, enriching dictionaries to use them in NLP tasks, etc. For this objective, the corpus used was obtained with the tools *Erauzterm* and *Itzulterm* (previous tools to achieve a similar objective). For the development of this work transformer technology based on sequence to sequence architecture has been used. Instead of using the linguistic and statistical methods as in previous works, neural network techniques have been used to develop this task. In other words, instead of extracting the terms of each language and performing the alignment of the terms in a pipeline, all actions are carried out simultaneously, becoming end-to-end tasks and reducing the spread of errors.

In addition to the system evaluation BLEU metric, the TEB metric has been used, which was created specially for this work. This metric takes into account some features that BLEU does not contemplate and which are important for terminological extraction. The system developed for this master's thesis has obtained a value of 0.78 points in BLEU metric. And it has achieved an improvement of 50 points in the BLEU metric compared to the baseline used for evaluation. Consequently, it has been shown that the terminological extraction task can be developed using neuronal network technologies.

Gaien aurkibidea

Irudien zerrenda

Taulen zerrenda

Glosarioa

Aipamen
Testuan zehar entitate bati erreferentzia egiten dion espresio testuala.

Atentzio-mekanismoa
Token batek, sekuentzia baten baitako tokenei jarritako arreta.

Esaldi parekatu
Bata bestearen itzulpena diren hizkuntza desberdineko bi esaldi

Gold standard
Sistema baten baliagarritasuna ebaluatzeko erabiltzen den erreferentzia corpusa.

Helburu hizkuntza
Itzulpen automatikoan hizkuntza batetik beste hizkuntza batera egiten dira itzulpenak, helburu hizkuntza bigarren hizkuntzari esaten zaio

Hizkuntzaren prozesamendua
Hizkuntzaren tratamendu automatikoaren inguruko ikerketa-lerroa.

Hitz anitzeko termino
Hitz bat baino gehiago dituen termino bat.

IE
Ingelesezko *Information Extraction* terminoaren laburtzapena, euskaraz *Informazio erauzketa* (IE) terminoa erabiltzen da.

Informazio erauzketa
Testu batetik informazioa lortzeko hizkuntzaren tratamendu automatikoaren ataza bat.

Itzulpen-memoria
Bi hizkuntzetako esaldi parekatuak dituen corpusa, normalean tmx formatuan egoten da.

Jatorri hizkuntza
Hizkuntza batetik bestera egiten den itzulpenaren lehenengo hizkuntza.

Lerrokatzea
Bi esaldi parekatuetako unitateak parekatzea.

NERC

HAP masterra

Ingelesezko *Name Entity Recognition and Classification* terminoaren laburtzapena, euskaraz *Izen entitateen ezagutza eta klasifikatzea*

Offset
Hitz batek testu batean duen posizioa (karaktereak kontatuz).

Seq2seq
Ingelesezko *sequence to sequence* terminoaren laburtzapena, euskaraz *sekuentziatik sekuentziara.*

Termino-bikote / termino-pare
Lan honetan termino hau horrela erabiliko da: Hizkuntza desberdinetan bata bestearen itzulpena diren terminoak

Terminologia-erauzketa
testu batetik terminologia nagusia lortzea.

Token
Hizkuntzaren prozesamenduaren arloan testu zatituaren unitate bat da. Orokorrean hitzaren baliokidea da, baina hitz-zatia edo karakterea ere adieraz dezake.

Sare neuronal / neurona sare
(Neural Network, NN), informazioa prozesatzeko adimen artifizialaren arloan erabiltzen den eredu matematiko bat.

Segmentazioa
Testu bateko hitzak bereiztea, tokenizazioa ere esaten zaio.

Segmentua
Itzulpen-memoria bateko esaldi parekatu bat.

Tokenizatzailea
Testu bateko hitzak bereizten dituen tresna edo aplikazioa.

Transformer
Auto-arretan oinarritutako sekuentziatik sekuentziarako arkitektura neuronala.

1 Sarrera

Hizkuntzaren Prozesamenduaren alorraren barruan hainbat ataza biltzen dira, horien artean, lan honetan landu den terminologia-erauzketa izeneko ataza. Labur esanda, terminoerauzketa, testu multzo bat emanda, testu horretan agertzen diren termino nagusiak erauzteari esaten zaio. Kasu honetan gainera, terminologia erauzteaz gain bi hizkuntza desberdinetan baliokideak diren termino-bikoteak erauzteko sistema bat eraiki nahi da.

1.1 Motibazioa

Testu zientifikoetan agertzen den terminologia eta hiztegietan dagoen informazio lexikal eleaniztuna askotan ez datoz bat; terminologia etengabe sortzen eta garatzen da, garapen hori testuetan egiten da eta antolatutako glosario edo hiztegietara iristen bere denbora behar du. Ondorioz, testu espezializatuetan erabiltzen den terminologia askotan ez da hiztegietan katalogatua egoten, baina interesgarria da testu horietan oinarritutako glosario edo hiztegiak osatzea. Bai kontsulta material gisa erabiltzeko, baina baita hizkuntzen prozesamendu automatikoaren beste ataza batzuetan erabiltzeko ere. Adibidez: informazio erauzketarako (Wang et al. (2018); Singh (2018)), testuen laburpen automatikorako (Allahyari et al. (2017)), itzulpen automatikorako (Etchegoyhen et al. (2018)), ontologietarako (Pociello et al. (2008))...

Testuetatik terminologia erauzteko ataza ez da berria, euskararentzat aurretik garatutako tresnak badaude; hala ere, euskararentzat orain arte sortutako tresnak teknika-linguistikoetan eta teknika estatistikoetan oinarritu izan dira. Teknika linguistikoek hizkuntzen egitura morfosintaktiko edo ereduak hartzen dituzte oinarritzat; teknika estatistikoek aldiz esaldietako osagaiek elkarrekin agertzeko duten 'joera' neurtzen dute:

ErauzTerm (Gurrutxaga et al. (2004)): teknika linguistiko eta teknika estatistikoak baliatuz euskarazko testuetatik terminologia erauzteko teknologia. Bi urrats egiten ditu, lehenengoan teknika linguistikoak erabiliz termino hautagaiak erauzten dira; bigarrenean aldiz teknika estatistikoak erabiliz termino hautagai egokienak detektatzen dira.

ELexBI (Gurrutxaga et al. (2006)): Itzulpen-memorietatik euskara-gaztelania hizkuntzetako termino-bikoteak erauzteko tresna. Lehenengo pausuan bi hizkuntzetako terminoak erauzten dira, bigarren pausuan aldiz bi hizkuntzetako terminoak parekatzen dira. Euskarazko terminoak erauzteko *Erauzterm* tresna erabiltzen du; gaztelaniazkoak erauzteko, aldiz, *Freeling 2.1* (Padró et al. (2004)). *Freeling* tresnak, teknika linguistiko eta estatistikoak erabiliz, hizkuntza askotako esaldiak aztertu eta zuhaitz egitura itzultzen du, *Elexbi*k zuhaitz egitura horietako izen-sintagmak hartzen ditu termino hautagaitzat. Ondoren, metodo estatistikoak erabilita itzulpen-unitate bereko termino hautagaien artean, bata bestearen baliokideak (edo itzulpenak) direnak hautatzen ditu.

Orain arte erabili diren metodo linguistiko eta estatistikoak erabili ordez sare neuronalen teknikak erabiliko dira ataza hau garatzeko; gainera, muturretik muturrerako ataza bihurtuko da, hau da, azpiatazak bata bestearen ondoren exekutatu ordez denak batera gauzatuko dira, errore propagazioa gutxiagotuz.

1.2 Atazaren ezaugarriak

1.2.1 Informazio erauzketa eleaniztuna

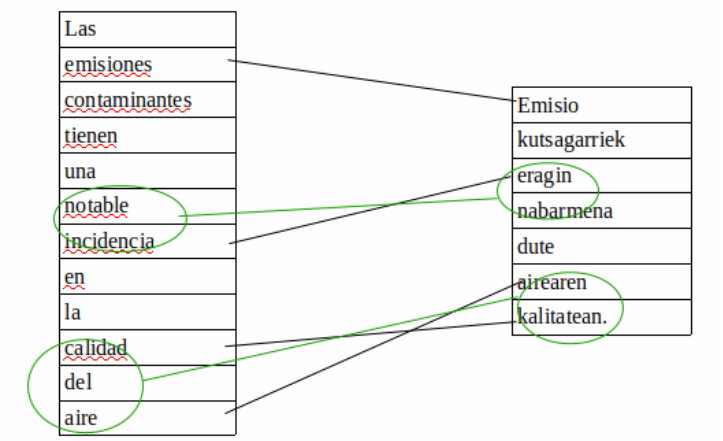
Terminologia-erauzketa (ingelesez, *terminology extraction*) informazio erauzketaren barne dagoen arlo bat da. Terminologia-erauzketaren helburua corpus bat oinarri izanik, automatikoki termino esanguratsuak eskuratzean datza.

Komunitate birtual bateko gai baten ezagutza irudikatzeko lehen pausoetako bat gai horretarako termino garrantzitsuekin glosario bat osatzea da. Dokumentu sorta batean oinarrituz gai baten inguruko termino garrantzitsuak eskuratzeko metodo ezberdinak definitu izan dira.

Orokorrean, terminologia automatikoki erauzteko, termino hautagaiak ateratzen dituzten prozesagailu linguistikoak erabili izan dira. Terminologia-erauztea arlo askotarako izan daiteke garrantzitsua: "antzekotasun semantikoak" (*semantic similarity*) bilatzeko eta "informazioa eta ezagutzaren kudeaketa" (*information and knowledge management*) eta antzekoetan oinarri bezala erabili ahal izateko, "testu-multzoen alorra identifikatzeko" (*Topic modeling*)...

Ataza honetan gainera, hizkuntza bakarreko terminologia erauzteaz haratago, bi hizkuntzetako (*es* eta *eu*) terminologia erauzi nahi da, terminologia-erauzketa ataza bera konplikatuz.

Adibidez, 1. irudian energia alorreko itzulpen-memoria bateko bi esaldi daude; eta, parekatutako bi esaldi horietan ondoko termino-bikoteak ikus daitezke:



Irudia 1: Terminologia-erauzketaren adibidea

- emisiones / emisio
- notable incidencia / eragin nabarmena
- calidad del aire / airearen kalitate
- incidencia / eragin
- calidad / kalitate

- aire / aire

Adibide honetan erraz ikus daitezke termino erauzketa atazaren ezaugarri batzuk:

- Normalean terminoak izen kategoriako sintagmak izan ohi dira: izena, izena + adjektiboa, izena + izena eta abar.
- Jatorri eta helburu hizkuntzetan ez dira berdin ordenatuta egoten.
- Hitz anitzeko terminoak detektatzea zailagoa da hitz bakarrekoak detektatzea baino.

Lan honetan hitz bakarreko terminoen erauzketa hartuko da oinarritzat. Hizkuntzei dagokionez aldiz, gaztelania eta euskararen arteko terminologia-erauzketa landuko da.

1.2.2 Lerrokatzea

Bi hizkuntzetako terminologia erauzteaz gain, hizkuntza bakoitzeko terminoak beste hizkuntzako terminoekin parekatzea ere bada ataza honen egitekoetako bat. Azpiataza hori ez da erraza, 1. irudian ikusten den bezala, helburuko esaldian terminoen ordena ez baita sarrerako esaldian termino baliokideek duten ordena bera. Lan honetan aukeratu diren hizkuntzak (*es* eta *eu*) gramatikalki desberdinak direnez, terminoen ordena ere desberdina da, beste hizkuntza batzuk aukeratuz gero terminoen arteko ordena antzekoagoa izango litzateke.

Orain arteko sistemek metodo estatistikoak erabiltzen zituzten lerrokatzea burutzeko, orain aldiz sare neuronalekin egingo da saiakera. Sistema estatistikoek esaldi parekatuetako jatorri hizkuntzako termino bat, helburu hizkuntzako terminoekin agertzeko joera neurtzen dute. Orain arteko sistema guztietan metodo estatistikoak erabili dira azpiataza hori burutzeko:

- *Itzulterm*, (Elhuyar Fundazioa et al. (2009))
- *Sketch Engine*, (Baisa et al. (2015))
- *Mutual Bilingual Terminology Extraction*, (Ha et al. (2008))
- *A Model for Multilingual Terminology Extraction via a Medical Social Network* (AYADIa et al. (2017))
- *Multilingual open relation extraction using cross-lingual projection* (Faruqui eta Kumar (2013))

Lan honetan aurkeztutako atazan, aldiz, aldi berean ikasiko du sistemak bi hizkuntzetako terminoak erauzi eta parekatzen.

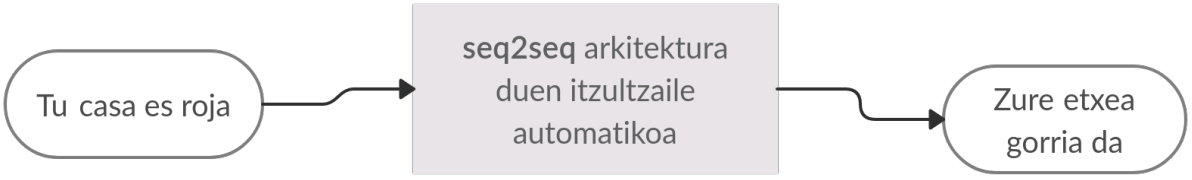
1.2.3 Sekuentziatik sekuentziarako arkitektura

Terminologia erauzteko sistemetan, orain arte, egitekoak pausuka edo bata bestearen ondoren egin dira. Adibidez, ondorengo hau izan daiteke sekuentzia bat:

- 1. pausua: jatorrizko hizkuntzako corpusetik terminologia erauzi.
- 2. pausua: helburuko hizkuntzako corpusetik terminologia erauzi.
- 3. pausua: aurreko bi pausuetako terminologia lerrokatu.

Kasu honetan, sare neuronalen teknikak erabiliz ekintza guztiak aldi berean egingo badira ere, *Named-entity Recognition* bezalako atazetan ohikoa izan da sailkatzaileen hurbilpena erabiltzea (Agerri eta Rigau (2016)). Sailkatzaileak hitz bakoitzari etiketa bat esleitzen dio: terminoa den edo ez, terminoaren hasiera den, termino baten barruko hitza den, termino baten amaiera den eta abar. Sailkatzaileetan ekintzak bata-bestearen ondoren exekutatzen dira, sekuentziatik sekuentziarako arkitekturetan aldiz ekintza guztiak aldi-berean exekutatzen dira, erroreen propagazioa saihestuz.

Sekuentziatik sekuentziarako (*sequence to sequence*, seq2seq) hurbilpena, hizkuntzaren prozesamenduko hainbat atazatan arrakastaz erabiltzen den hurbilpena da. Hizkuntzaren prozesamenduko edozein ataza testutik testurako ataza batean bihurtu daiteke "*Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*"(Raffel et al. (2019)) artikuluan proposatu zen bezala. Seq2seq hurbilpena darabilen atazarik ezagunena itzulpen automatikoa da, baina testuen laburpen automatikorako edo elkarrizketa eta galdera-erantzun sistemetarako ere erabili izan da. Itzulpen automatikoko adibide bat 2. irudian ikusi daiteke:

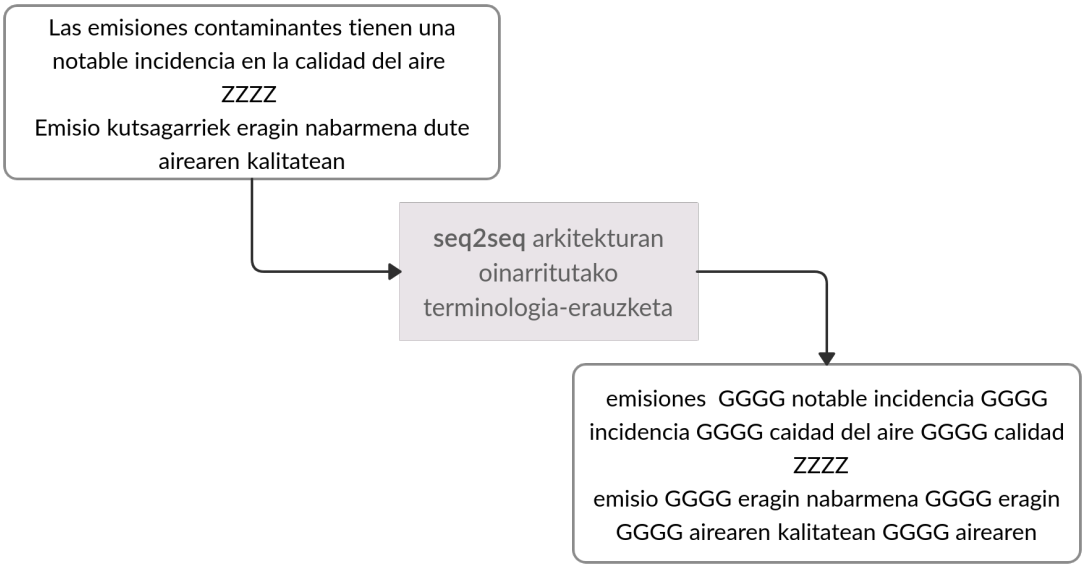


Irudia 2: Seq2seq arkitekturan oinarritutako itzulpen automatikoa.

Sekuentziatik sekuentziarako hurbilpenean, sare neuronalari sarreran sekuentzia bat ematzen zaio, eta irteeran beste sekuentzia bat itzuliko du; tarteko lan guztiak aldi-berean egiten dituelarik. Itzulpen automatikoaren kasuan, sarrerako sekuentzia jatorri-hizkuntzako testua izaten da, eta irteerako sekuentzia itzulpenari dagokion helburu-hizkuntzako testua. Terminologia-erauzketa atazarentzat seq2seq hurbilpenean erabiliko den sarrera eta irteera sekuentzien formatua 3. irudian ikus daiteke.

Lan honetan, 3. irudian azaltzen den formatuarekin osatu da corpusa. Sarreran 'zzzz' hitz katearekin bereizitako bi esaldi egongo dira, bi esaldiok bata bestearen itzulpena izanik. Irteeran, aldiz, 'zzzz' katearekin bereizitako bi termino segida egongo dira, termino

segidako termino bakoitza 'gggg' katearekin bereizita dagoelarik. Gainera, termino segida bakoitzeko termino baliokideak segidan duen posizioaren arabera ordenatuta daude, hau da, lehenengo segidako 3. terminoa bigarren segidako 3. terminoaren baliokidea da.



Irudia 3: Seq2seq arkitekturan oinarritutako terminologia-erauzketa.

Gaur egun seq2seq arkitektura erabiltzen duen teknologia asko dago. Lan honetan *Fairseq* ("*Facebook AI Research Sequence-to-Sequence Toolkit written in Python*", Ott et al. (2019)) erabili da. Fairseq python programazio lengoaian oinarritua dago, kode irekikoa da eta gitHub-etik eskura daiteke[1]. Sekuentziekin lan egiteko tresna-multzo handia duen liburutegia da: edozein ikerlari edo garatzailek erabili dezake bere modeloa entrenatu eta probatzeko. Itzulpen automatikoa, testuen laburpena, lengoaia modelizazioa eta antzeko testu sorkuntza atazetan erabiltzeko definitua dago. *Fairseq* kodearen definizioan aipatzen ez bada ere, lan honetan termino-erauzketarako erabiliko da.

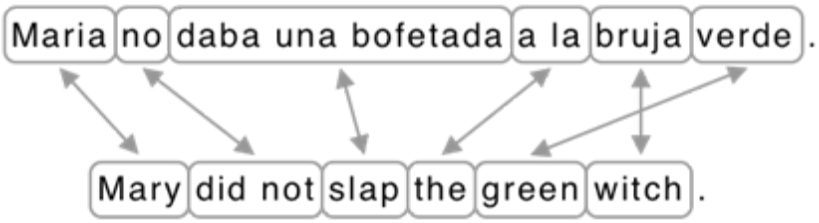
1.3 Proiektuaren helburuak

Master amaierako proiektu hau garatzeko ideia Elhuyarren sortu zen; Elhuyarren garatutako *Itzulterm* (2009) teknologia zaharkitua geratu da eta sare neuronalen bidez teknologia hori berritzeko saiakera da hau. *ItzulTerm* tresnak aurretik aipatutako *ElexBI* du oinarritzat, erabiltzaileari termino-erauzketa prozesu osoa eskaintzen dio: itzulpen-memorietan oinarrituta automatikoki terminologia erauzi, eskuz balidatu eta termino-zerrenda deskargatzeko aukera ematen duen web-ingurunea da. Automatikoki erauzitako terminologia hori ondoko atazetarako erabiltzen da: hiztegiak elikatzeko, itzulpen-memorien sistema hobetzeko, itzulpen automatikoan erabiltzeko, etab.

[1] https://github.com/pytorch/fairseq

Lan honetan gaztelania eta euskara hizkuntzak erabili diren arren, beste hizkuntzetako corpusak edukita hizkuntza-bikote gehiagotara zabal daiteke termino-erauzketa.

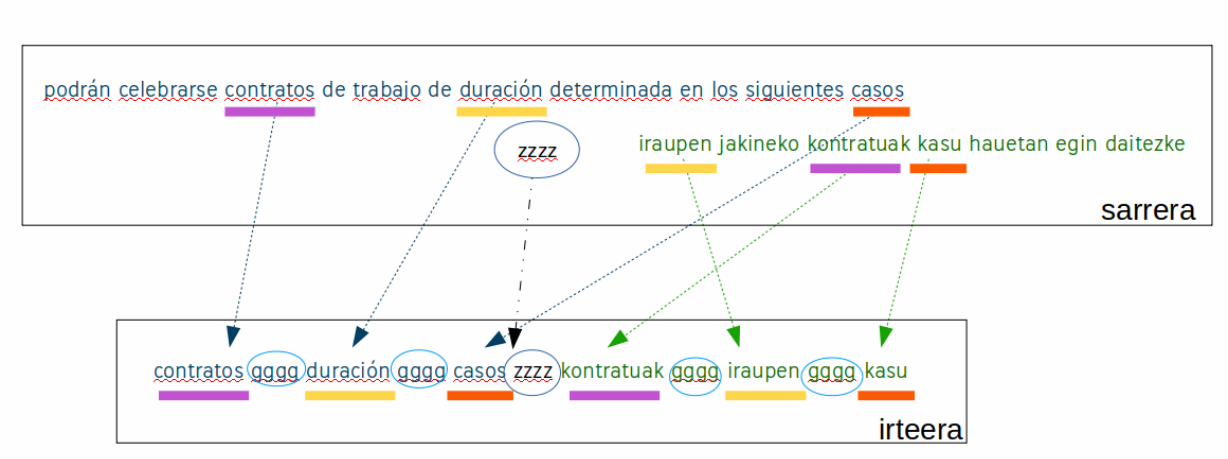
Proiektua garatzeko erabili den teknologia sekuentziatik sekuentziarako (*sec2sec*) eredu bat da, *fairseq* (Ott et al. (2019)). Itzulpen automatikorako erabiltzen da normalean seq2seq arkitektura, 4. irudian (Iturria: *Real-World Natural Language Processing* liburua[2]) ikus daiteke adibide bat. Kasu honetan, gaztelania eta ingelesaren arteko itzulpena da; sekuentzia batetik hasi (*Maria no daba una bofetada a la bruja verde* (*es*)) eta beste sekuentzia bat (*Mary did not slap the green witch* (*en*)) itzultzen duen sistema.



Irudia 4: Esaldi baten itzulpena (es-en).

Iturria: *Real-World Natural Language Processing* liburua.

Itzulpen automatikoarentzat ikusi den adibidearen antzeko egitura bat sortu beharko da terminologia-erauzketa atazarentzat ere, hau da, sekuentzia batetik hasi eta beste sekuentzia bat itzuliko duen sistema bat sortzeko sarrera eta irteera sekuentziak definitu behar dira. Ikusi 5. irudia.



Irudia 5: Terminologia-erauzlea seq2seq ikuspuntua.

Sarrera sekuentzia: *podrán celebrarse contratos de trabajo de duración determinada en*

[2]https://livebook.manning.com/book/real-world-natural-language-processing/chapter-6/v-4/19

los siguientes casos zzzz iraupen jakineko kontratuak kasu hauetan egin daitezke. 'zzzz' katearekin bereizitako bi esaldi, bi hizkuntza desberdinetan (*es-eu*).

Irteera sekuentzia: *contratos gggg duración gggg casos zzzz kontratuak gggg iraupen gggg kasu.* Termino-bikoteen zerrendak: terminoak 'gggg' katearekin bereizita daude; 'zzzz' katearekin aldiz bi hizkuntzetako termino-segidak bereizten dira. Hizkuntza bakoitzeko terminoak bikoteka lotuta daude, zerrendako posizioaren arabera.

Hauek dira terminologia-erauzketa atazaren ezaugarriak:

- *es* hizkuntzako termino hautagaiak esaldian duten orden berean daude.
- *eu* hizkuntzako terminoak, *es* hizkuntzako terminoen ordenaren arabera daude.
- 'zzzz' katea bi hizkuntzetako esaldiak eta terminoak bereizteko erabiltzen da.
- Irteerako sekuentzian 'zzzz'ren bi aldeetan dauden termino kopurua bera da.
- Itzulpen automatikoko atazetan ez bezala, kasu honetan, bai sarreran eta bai irteeran, bietan, bi hizkuntzak (*es-eu*) daude.
- Lortutako terminoak dagokion hizkuntzako esaldian bere horretan agertzen dira, hau da, sistemak itzultzen dituen terminoak ez daude lematizatuta.

Master amaierako lan hau honela dago egituratua: hasteko, terminologia-erauzketa behar bezala garatzeko behar diren atazen aurkezpena egingo da **??**. atalean: Informazioaren erauzketa, lerrokatzea, sekuentziatik sekuentziarako arkitekturak, segmentazioa eta ebaluazio-metrikak. Ataza horien jatorria eta gaur egungo egoeraren berri emango da. Atal mamitsuena 3. atala da, bertan ataza lantzeko erabili den metodologia aurkeztuko da: corpusaren jatorria (3.1) eta egin zaizkion eraldaketak (3.3); erabili den sare neuronalaren ezaugarriak (3.6); eta, sistema ebaluatzeko garatu den algoritmoaren berri emango da (3.8). Hurrengo atalean, 4. atalean, ataza garatzeko egindako esperimentuak, sistemaren emaitzak eta adibide erreal batzuk ikusiko dira. Amaitzeko, 5. atalean, master amaierako lan honen ondorioak eta sistemari egiteko hobekuntza posible batzuk aurkeztuko dira.

2 Aurrekariak

Atal honetan, sare neuronalen aurreko terminologia-erauzketaren oinarriak izan diren atazak azalduko dira: tokenizazioa, informazioaren erauzketa eta lerrokatzea. Ondoren, sare neuronalen agerpenaren ondorioz, muturretik muturrerako sistemen arkitekturaren oinarriak azalduko dira. Eta bukatzeko, sistemen ebaluazio metriken nondik norakoak azalduko dira.

2.1 Informazioaren erauzketa

Gaur egun hizkuntza naturalaren prozesamenduan erabiltzen den *"entitateen ezagutza"* terminoa 1996an sortu zen, MUC-6 (*Message Understanding Conference*) konferentzian (R. Grishman & Sundheim 1996). Konferentzia, berez, informazio erauzketa (IE, *Information Extraction*) atazari buruzkoa zen: egiturarik gabeko testuetatik (egunkarietako artikuluak...) informazioa erauzteko moduak aztertzeko egin zen. Eztabaida haietan jabetu ziren testuetako pertsona-izen, erakunde-izen eta gai garrantzitsuak identifikatu eta klasifikatzearen garrantziaz. Definitu zuten atazari NERC (*Name Entity Recognition and Classification*) izena eman zitzaion eta geroztik hizkuntza naturalaren prozesamenduaren eta zehazki informazioaren erauzketa atazaren azpi-ataza garrantzitsuenetakoa da.

Euskarari dagokionez ere entitateen ezagutza ataza asko erabili da:

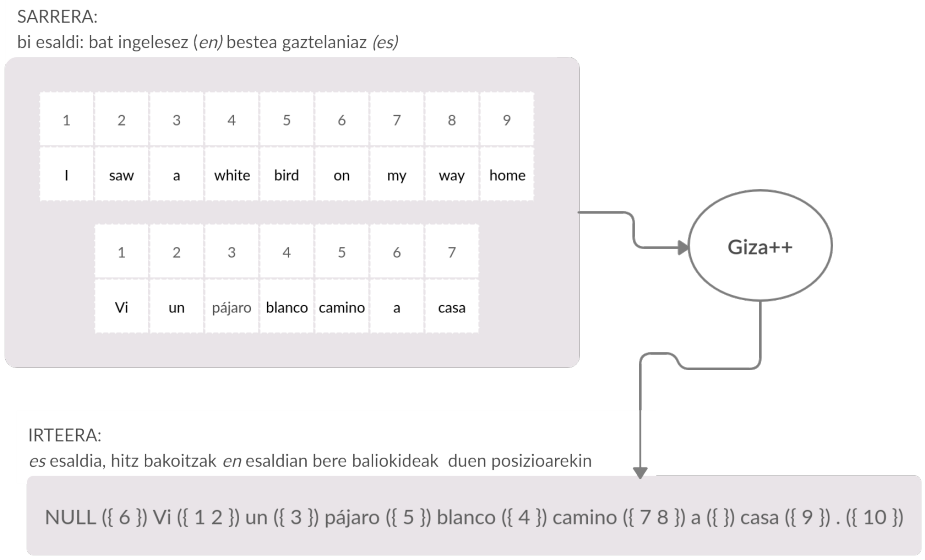
- ***Erauzterm***: ikusi 1.1. atala: euskara hizkuntzarentzat termino erauzlea
- ***EleXBi***: ikusi 1.1. atala: euskara-gaztelania corpus paraleloetatik termino baliokideen erauzlea
- ***ixa-pipes-nerc*** (Agerri eta Rigau (2016)): hizkuntza desberdinetako testuetan izen entitateak markatzen dituen sistema, ondoko hizkuntzetarako balio du: euskara, gaztelania, ingelesa, alemana, neerlandera eta italiera.
- ***BERTeus*** (Agerri et al. (2020)): Gaur egungo azken teknologiak erabiliz (Conneau et al. (2020), Devlin et al. (2018)), euskararentzat aurre-entrenatutako hizkuntza-eredu erraldoia sortu eta besteak beste NERC atazan emaitza oso onak ematen dituen sistema; euskararen kasuan punta-puntako teknologia.

2.2 Lerrokatzea

Esan bezala, ataza honek, bi hizkuntzetako esaldietan terminoak identifikatzeaz gain, termino horiek parekatu edo lerrokatu ere egin behar ditu; hau da, jatorrizko hizkuntzako esaldian dauden terminoak helburuko hizkuntzako esaldiko terminoekin lotu behar ditu. Lotze honi lerrokatzea esaten zaio. Orain arte itzulpen automatikoan asko erabiltzen zen lerrokatzea; orain, sare neuronalen ondorioz itzulpen sistemarako ez da lerrokatzerik behar. Bai ordea itzulpenaren ondorengo tratamendurako: estiloa mantentzeko, glosarioak sortzeko, erreferentziak markatzeko...

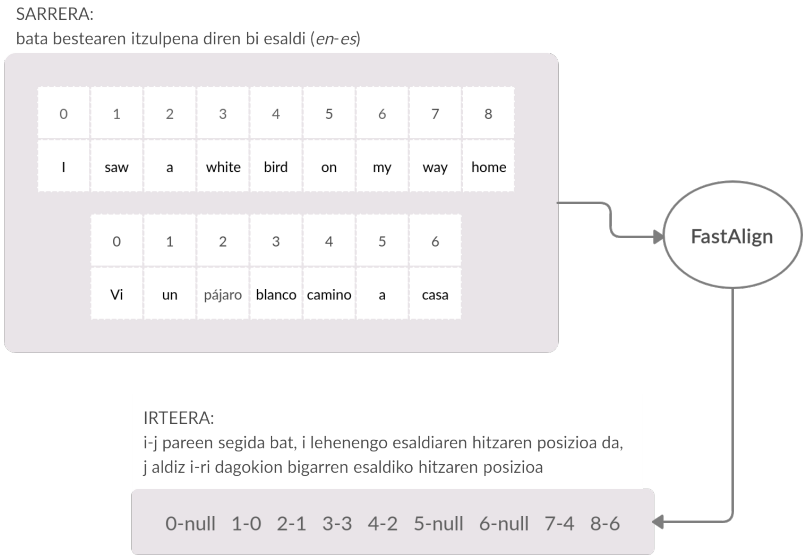
Azken 20 urteetan erreferente izan den lerrokatze sistemetako bat *Giza++* da (Och eta Ney (2003)). 1.999an sortu bazen ere, 2.003an egin zitzaizkion ekarpenen ondoren egin zen ezaguna sistema hau. *Giza++* sistemak IBMk garatutako hitzen lerrokatze estatistikoa (Brown et al. (1993)) du oinarritzat. Gaur egun, sare neuronalen erabilerarekin zaharkitua dagoen arren, oraindik asko erabiltzen den sistema da.

Giza++ aplikazioak, bata bestearen itzulpena diren bi esaldi emanda, bi esaldi horietako elementuak lotzen ditu metodo estatistikoetan oinarrituta. Sistema honen irteera bigarren esaldia da, baina hitz bakoitzaren ondoan lehenengo esaldian hitz horren baliokideak duen posizioa agertzen da (ikusi 6. irudia).



Irudia 6: *Giza++* lerrokatzailearen adibide bat.

Lerrokatze prozesuan ezaguna den beste sistema bat *FastAlign* (Dyer et al. (2013)) da. Hau ere IBMk garatutako hitzen lerrokatze estatistikoan (Brown et al. (1993)) oinarritua dago. Sarreran, *Giza++* sisteman bezala, elkarren itzulpena diren bi esaldi hartzen ditu; irteeran, sarrerako esaldi bat itzuli ordez, sarrerako esaldietako hitzen posizio-bikote segida bat itzultzen du.



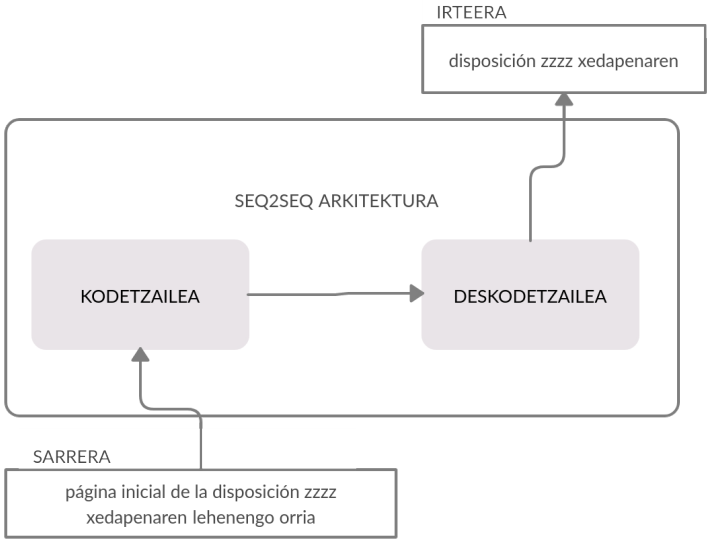
Irudia 7: *FastAlign* lerrokatzailearen adibide bat.

Giza++ eta *FastAlign* sistemak azken 20 urteetan erreferente izan baldin badira ere, aurten sare neuronaletan oinarritutako lehenengo lerrokatze sistema bat aurkeztu da; *Giza++* sistemaren emaitzak hobetzen dituen bat: *End-to-End Neural Word Alignment Outperforms GIZA++* (Zenkel et al. (2020)). Oso berria da oraindik, ikusi beharko da datozen hilabeteetan honen erabilerak ematen dituen emaitzak.

2.3 Sekuentziatik sekuentziarako arkitektura eta *transformer*ak

Hizkuntzaren prozesamenduko atazetan, *seq2seq* hurbilpenerako sare neuronalen arkitektura bat baino gehiago dago. Hala ere, gehienek oinarrian kodetzaile-deskodetzaile (*encoder-decoder*) arkitektura dute. Arkitektura horiek bi pausutan funtzionatzen dute: lehenengo pausuan, sarrerako esaldia emanda kodetzaileak kodetu egiten du; bigarren pausuan, aldiz, deskodetzaileak kodetutako esaldia deskodetu egiten du, irteerako esaldia sortuz (Ikusi 8. irudia).

HAP masterra



Irudia 8: seq2seq code-decode arkitektura, terminologia-erauzketa atazako adibide batekin.

Kodetzaileak sarrera sekuentziako hitzak $(x_1, x_2, ..., x_n)$ $(z_1, z_2, ..., z_n)$ errepresentazio segida batean kodetzen ditu. Deskodetzaileak, kodetzaileak itzulitako errepresentazioa irteera sekuentzia $(y_1, y_2, ..., y_n)$ batean bihurtzen du, buelta bakoitzean unitate bat sortuz. Irteera sekuentziako y_i unitatea sortzeko aurretik sortutako sekuentzia $(y_1, y_2...y_{i-1})$ hartzen du kontuan.

Kodetzaile-deskodetzaile arkitekturan oinarrituta dauden hiru sistema nagusiak ondokoak dira: neurona-sare errepikakorrak (*Recurrent Neural Network*, RNN) (Sutskever et al. (2014)), neurona-sare konboluzionalak (*Convolutional neural network*, CNN) (Gehring et al. (2017)) eta *transformer*ak (Vaswani et al. (2017)). Sekuentziatik sekuentziarako sistema eredurik onenak sakonera handiko kodetzaile-deskodetzaileak erabiltzen dituzten sare neuronal konplexuetan oinarritzen dira. Errendimendu handiena kodetzailea eta deskodetzailea arreta-mekanismo baten bidez konektatzen dituzten sistemek ematen dute. *Transformer*ak, gainera, kodetzaile eta deskodetzailea arreta-mekanismo batekin osatzen diren sareak dira. *Attention is all you need* artikuluan (Vaswani et al. (2017)) esaten denez, itzulpen automatikoko atazetan egindako esperimentuetan *transformer*-ereduek kalitate hobea lortzen dute eta entrenatzeko denbora gutxiago behar dute. *Transformer*en arkitektura nolakoa den 9. irudian ikus daiteke.

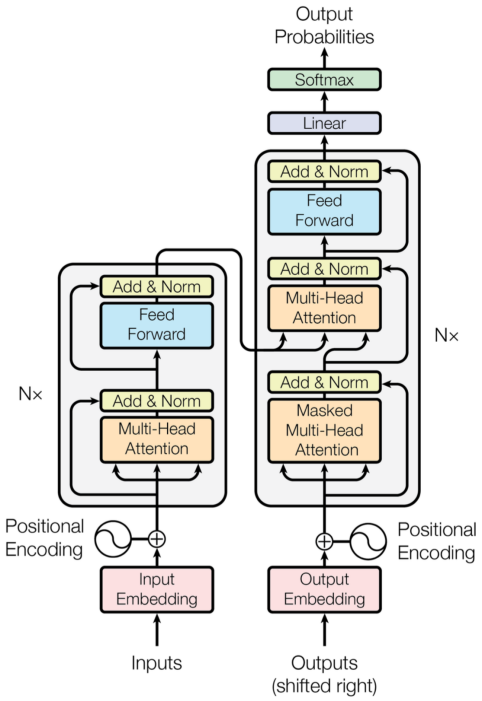


Figure 1: The Transformer - model architecture.

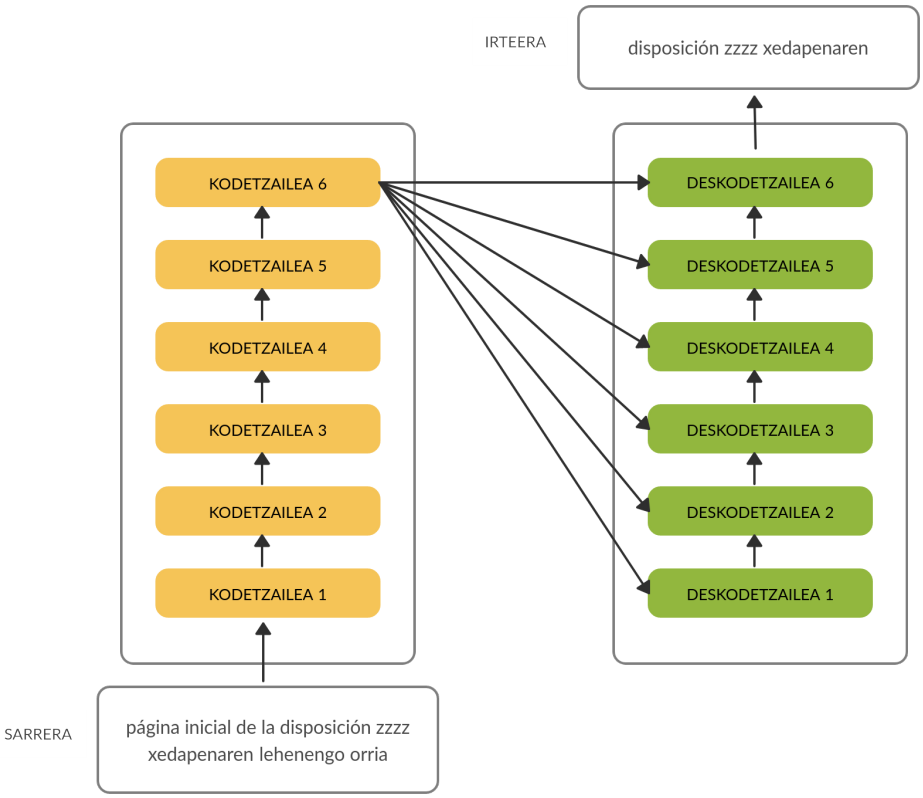
Irudia 9: *Transformer*en arkitektura.

Iturria: *Attention is all you need*, (Vaswani et al. (2017))

Attention is all you need (Vaswani et al. (2017)) artikuluaren xedea ongi azalduta dago *The Illustrated Transformer*[3] izeneko web-sarreran (Alammar (2018)).

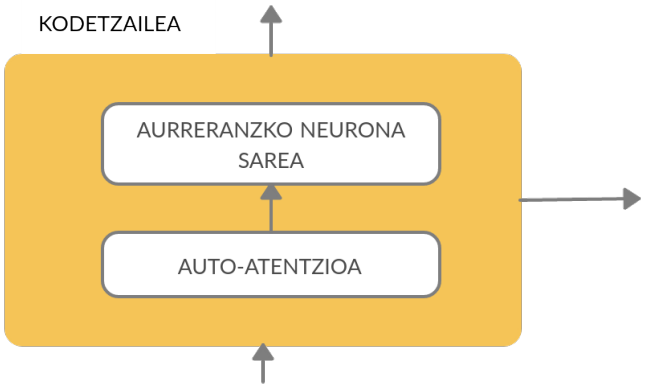
Kodetzailearen zatia bata bestearen gainean dauden kodetzaile pila batekin osatua dago; deskodetzailea ere bata bestearen ondoren exekutatzen den deskodetzaile pila bat da. Bi piletan geruza kopuru bera egotea gomendagarria da. Ikusi 10. irudia:

[3]http://jalammar.github.io/illustrated-transformer/



Irudia 10: *Transformer*aren kodetzaile-deskodetzaile geruzak.

Kodeketa blokeko kodetzaile guztiek egitura bera dute eta kode bakoitza bi azpigeruzatan bereizita dago. Ikusi 11. irudia:



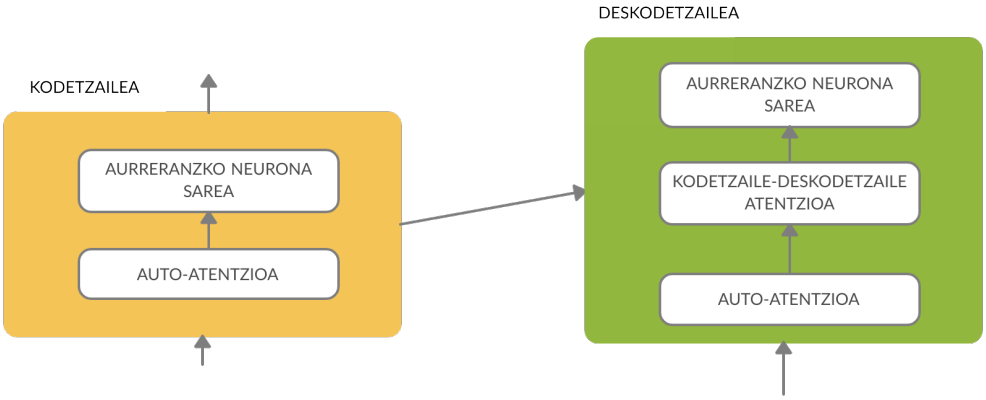
Irudia 11: Kodetzaile pilako geruza bat auto-atentzioarekin.

Atentzio-mekanismoa (*attention-mechanism*) sekuentzia batean token jakinetan arreta

jartzeari esaten zaio. Atentzio-mekanismoa sare neuronalen bidez ikasten den atentzio-funtzio bat da, eta sekuentziako hitz garrantzitsuenei pisu handiagoa esleitzean datza. Auto-atentzioa (*self-attention*), edo barne-atentzioa, sekuentzia batek bere buruaren gainean arreta jartzeari esaten zaio. Sekuentzia baten errepresentazioa sortzeko, sekuentziako token bakoitzarentzat sekuentziaren posizio bakoitzari atentzio-mekanismo bat aplikatzen dio.

Atentzio-funtzio bakarra ikasi ordez, hainbat atentzio-funtzio edo buru aldi berean ikasteak emaitza hobeak ematen ditu. Horri buru anitzeko auto-atentzioa (*multi-head self-attention*) esaten zaio. Atentzio buru asko izateak buru bakoitzak posizio bakoitzeko errepresentazioan arreta jartzea ahalbidetzen du, buru bakoitzaren dimentsioa txikituz. Ondorioz, sare neuronalaren kostu konputazionala ere ez da asko handituko. *Transformer* arkitekturan, buru-anitzeko atentzioa, kodetzaileko eta deskodetzaileko auto-atentzioan eta kodetzailetik deskodetzailera doan atentzioan aplikatzen da.

Kodetzaileko lehenengo geruzako auto-atentzioak hitz bakoitza kodetzeko sekuentziako gainontzeko hitzei erreparatzen die. Auto-atentzioaren irteera, kodetzaileko bigarren azpi-geruzan dagoen aurreranzko neurona-sare trinkoaren sarrera izango da (ikusi 11. irudia). Aurreranzko sare neuronal trinkoa (*fully-connected feed-forward network*) sekuentziako posizio bakoitzari berdin-berdin aplikatzen zaio, inguruko hitzei erreparatu gabe. Deskodetzaileak ere bi geruza horiek baditu, baina bi geruzen artean sarrera sekuentziako zati garrantzitsuenak kontuan hartzen dituen atentzio geruza bat ere badago.

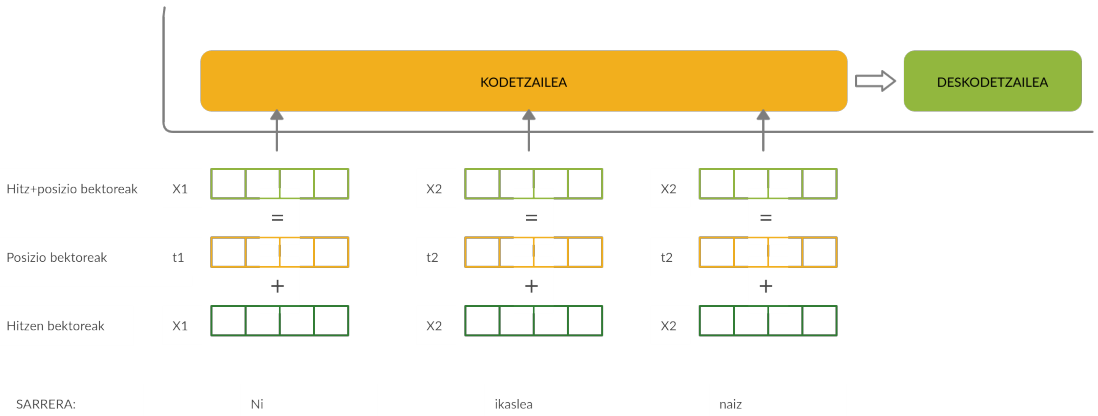


Irudia 12: Kodetzaile eta deskodetzailearen barneko arkitekturak.

*Transformer*aren deskodetzaile geruzak (*decoder layer)* ere baditu buru-anitzeko auto-atentzioa eta aurreranzko neurona-sare trinkoaren geruzak. Baina bi azpi-geruzen artean, kodetzaileko irteeraren gaineko buru-anitzeko atentzio mekanismoa duen azpi-geruza bat du (12. irudia). Deskodetzailearen auto-atentzioak aurretik tratatutako posizioak bakarrik hartzen ditu kontuan, momentuko posiziotik eskuinera dauden tokenak baztertuz.

*Transformer*aren funtzionamenduaz gain, *transformer*ak hartzen dituen datuen errepresentazioa ere garrantzitsua da. *Transformer*ak corpusetik ikasteko prozesuan, besteak

beste, hiztegi bat sortzen du, corpus horretan aurkitu dituen hitz desberdin guztiekin edo aurretik esandako tamaina maximo bat gainditu arte.



Irudia 13: Sarrerako esaldiko hitzen eta hauen posizioen errepresentazioa.

*Transformer*ak sarrera bezala jasotzen duen esaldiko hitz bakoitzeko bektore bat sortuko da $(x_1, x_2, ...x_n)$; eta hitz bakoitzak esaldian duen posizioarekin beste bektore bat sortuko da $(t_1, t_2, ...t_n)$. Sortutako bi bektoreen batuketa eginez sortuko da kodetzaileari bidaliko zaion lehenengo bektore multzoa $(x_1, x_2, ...x_n)$; lehen esan bezala, esaldiko hitz bakoitzeko bektore bat. Hasieran osatutako bektore horiek kodetzaileko pausu bakoitzean transformatu egingo dira, deskodetzailearentzat sarrera izango den bektore segida osatu arte. Deskodetzailean ere antzeko prozesu bat egingo da, azkenean bektore horiek berriro hitz bezala sailkatu arte.

2.4 Tokenizazioa/segmentazioa

Corpusa tokenizatzeko erabili den teknika BPE (*Byte Pair Encoding*) (Sennrich et al. (2016)) teknika da. BPE segmentazio teknika bat da. Tokenizatzaile batek testuko hitzak banatzen ditu; segmentatzaile batek, aldiz, hitzen barruko osagaiak banatzen ditu. Informazio linguistikorik erabili gabe, karaktere-segiden maiztasunetan oinarrituta, BPEk hitzetatik atzizki, aurrizki eta antzekoak bereizten ditu. Hitzak xehetzeaz gain, hitz bakoitzeko osagaiak hiztegian aparte sartzen ditu, bakoitza bere maiztasunarekin. Beraz, hitz deklinatuak ez dira hiztegian gordeko eta hiztegiaren tamaina txikiagoa izango da. Horren ondorioz, gainera, hiztegiaren tamaina maximora iristea gehiago kostako zaio, eta hiztegian sartuko ez diren hitz ezezagunak gutxiago izango dira.

BPEk egiten dituen segmentazio adibide batzuk ondokoak dira:

berezi@@ ren

hau@@ etatik

HAP masterra

eba@@ keta

jaurlari@@ tza

instituzio@@ etako

kalte@@ garri

eskola@@ tik

eskola@@ raino

denda@@ tik

Testuko hitzak horrela bereizita, testuan dauden elementu guztien hiztegi bat osatzen da. Hiztegi sarrera bakoitzari corpusean duen agerpen kopurua esleitzen zaio. Ondorioz, hiztegian ez ditu "eskolatik" eta "eskolaraino" hitzak sartuko; "eskola" eta "tik" osagaiak sartuko ditu, bakoitza bere agerpen kopuruarekin. Segmentazioaren ondoren egin daitezkeen atazetan lemen maiztasuna ziurragoa izango da eta emaitza hobeak emango ditu, kasu gehienetan.

Morfologikoki konplexuak diren hizkuntzentzat oso baliagarria da BPE teknika. Lematizatzea beste bide bat izan zitekeen, baina informazioa galtzea ekartzen duenez eta lematizazio erroreak ekiditeagatik segmentazioa erabili da.

2.5 Ebaluazio metrikak

Itzulpen automatikoa egiten duen ataza baten egokitasuna neurtzeko, egokiena pertsonek egindako ebaluazioa da, eskuzkoa. Egokia bai, baina garestia ere bai. Eskuzko lan hori ekiditeko sortu zen 2002an BLEU metrika.

BLEU (*Bilingual Evaluation Understudy*) izeneko metrika, itzulpen automatikoko sistemen kalitatea neurtzen duen metodo automatiko bat da (Papineni et al. (2002)). Itzultzaile automatiko batek itzulitakoa eta itzultzaile profesional batek itzulitakoa konparatzen ditu. Itzulpen automatiko bat itzulpen profesional batekiko zenbat eta hurbilago egon, orduan eta hobea da. BLEU sistema azkarra, merkea eta hizkuntzekiko independentea da; ondorioz, arrakasta handia du.

Itzultzaile profesionalek egindako itzulpenak eta itzulpen horien jatorrizko esaldiak konparatuz parametro batzuk "ikasten" ditu; beraz, BLEU ona lortzeko kalitate oneko itzulpenak dituen corpus handi bat beharko da. BLEUk ondoko ezaugarriak hartzen ditu kontuan itzulpenaren kalitatea neurtzeko:

- **Egokitzapena** (*adequacy*): jatorrizko esaldiaren esanahia helburuko esaldiak nola barne hartzen duen.
- **Fidagarritasuna** (*fidelity*): helburuko esaldiaren esanahia zenbateraino gerturatzen den jatorrizko esaldiaren esanahira.
- **Hitz-jarioa** (*fluency*): gramatika aldetik esaldiak nola osatuta dauden eta esaldien interpretaziorako erraztasuna neurtzen du.

Ezaugarri horiez gain beste parametro batzuk (bi esaldien luzera, hitzen ordena, sintagmen osaera...) kontuan hartuta formula matematiko baten bidez puntuazio bat ematen dio itzulpenari. BLEU metrikak ez du sistema baten baliagarritasuna neurtzen; garapen fasean dagoen sistema bat hobetzen doala frogatzeko edo bi sistema konparatzeko balio duen metrika bat da.

Lan honetako ataza ez da itzulpen automatiko bat. Baina bi esaldi dituen testu batetik abiatuta, termino zerrenda bat duen testu batera iritsiko den "itzulpen"bat bezala uler daitekeenez, BLEU metrika baliogarria izango da ebaluaziorako.

3 Metodologia

Atal honetan, esaldi-bikoteetatik termino-bikoteak lortzeko egin den bidea azalduko da. Hasteko, sistema entrenatzeko erabiliko den corpusa lortzeko (3.1); eta corpus hori sistema entrenatzeko egokia den formatura (3.2) bihurtzeko egin diren pausuak (3.3) azalduko dira. Corpusaren ezaugarri nagusiak 3.4 atalean aurkeztuko dira. Ondoren, corpusa entrenatu ahal izateko nola multzokatu den erakutsiko da (3.5); eta sortutako corpusarekin sare neuronala entrenatzeko erabili diren hiperparametroak azalduko dira (3.6). Amaitzeko, ebaluaziorako sortu den metrika (3.8) eta ebaluaziorako kontuan hartzeko heuristikoei (3.7) buruzko informazioa emango da.

3.1 Corpusaren eraketa

Ataza hau garatzeko behar den corpusak ondoko ezaugarriak izan behar ditu: alde batetik, sarrera sekuentzia bezala erabiliko diren itzulpen-memoriak, edo bata bestearen itzulpenak diren esaldi pareak; bestalde, irteera sekuentzia bezala erabiliko diren termino-bikoteak.

3.1.1 *Itzulterm* tresnarekin sortutako corpusa

Aurretik, 1.3. atalean azaldu den bezala, *Itzulterm* inguruneak erabiltzaileari ondokoak egiteko aukera ematen dio: itzulpen-memoria bat aukeratu, bertako terminologia automatikoki erauzi (*EleXBi* erabiliz), terminologia hau landu eta glosarioa esportatu. *Itzulterm* tresnak itzulpen-memoriak eta hauetan landutako terminologia mysql datu-baseetan gordetzen du; lan hau egiteko behar den corpusa osatzeko datu-base horietako informazioa erabili da.

*Itzulterm*ek erabiltzaileak sarrera bezala ematen dizkion itzulpen-memoriako esaldi-bikoteak eta esaldi hauetatik lortutako termino-bikoteak datu-base batean gordetzen ditu. Erabiltzaileak sortzen duen erauzketa bakoitzeko datu-base bat sortzen duelarik. Corpus bakoitzari buruz gordetzen duen informazioa ondokoa da: sarrera hizkuntza, helburu hizkuntza, segmentu kopurua, sarrera hizkuntzako hitz kopurua eta helburu hizkuntzako hitz kopurua (ikusi 1. taula).

Corpus kodea	Sarrera hizkuntza	Helburu hizkuntza	Segmentu kopurua	Sarrera hizkuntzako hitz kopurua	helburu hizkuntzako hitz kopurua
IRvQQD	es	eu	2.761	56.399	40.609
7__Hy	es	eu	5.166	100.220	71.290
hST1Y	es	eu	2.761	56.399	40.609

Taula 1: *Itzulterm*ek gordetzen dituen corpusen datuak.

Datu-base bakoitzean, aldiz, itzulpen-memoria baten erauzketa datuak gordetzen dira:

HAP masterra

- Jatorrizko esaldia.
- Jatorrizko esaldiaren itzulpena.
- Esaldi-pare bakoitzean dauden termino-bikoteak. Termino-bikote bakoitzari buruz gordetzen den informazioa ondokoa da:
  - Jatorrizko hizkuntzan dagoen terminoa: agerpena, lema, esaldian duen posizioa (*offset*), corpuseko agerpen kopurua...
  - Helburu hizkuntzan dagoen terminoa: agerpena, lema, esaldian duen posizioa (*offset*), corpuseko agerpen kopurua...
  - *Bal* eremua.

Bal eremu horrek, termino-bikote bakoitzaren lantze egoera gordetzen du. Izan ere, lehenago esan den bezala, *Itzulterm*ek *ElexBI* bidez termino-bikoteak automatikoki erauzteaz gain, erabiltzaileari termino-bikote horiek lantzeko aukera ematen dio, hau da, termino-bikote hori glosariorako hautatu/ezeztatu/moldatzeko aukera. *Bal* eremuan ondoko balioak har ditzake:

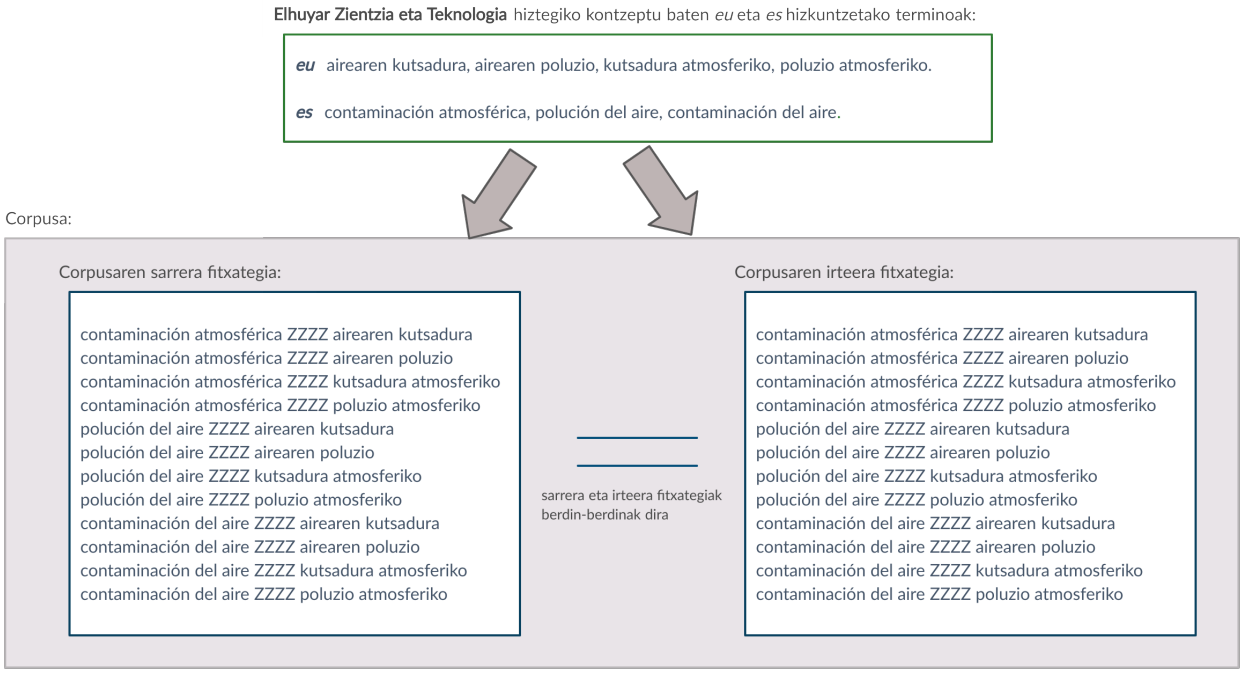
- 1: Erabiltzaileak eskuz onartu du termino-bikotea.
- 2: Erabiltzaileak eskuz baztertu du termino-bikotea.
- 3: Beste erauzketaren batean onartu da termino-bikote hau.
- 5: Termino-bikote hau hiztegi orokor batean badago.
- 7: *ElexBI*k proposatutako termino-bikotea editatu egin du erabiltzaileak.
- 13: Termino-bikote hau 3 eta 5 glosarioetan dago.
- 17: Termino-bikote hau erabiltzaileak ez du ukitu. *ElexBI*k automatikoki proposatutakoa dela besterik ezin da jakin, termino-bikote esanguratsua izan daiteke, edo ez.

Bal eremuko informazio hori aztertu ondoren, corpuserako *gold standard* (GS) moduan aukeratuko diren termino-bikoteak 1, 3, 7 eta 13 balioak dituztenak dira. Datu-basean multzo horietan dauden termino-bikoteak benetan termino-bikote esanguratsuak direla esan daiteke. Baina horrek ez du esan nahi 17 edo 5 zenbakiak dituzten termino-bikoteak esanguratsuak ez direnik, litekeena baita erabiltzaileak glosario lanketa ahaztu edo baztertu izana.

HAP masterra

3.1.2 Elhuyar Zientzia eta Teknologia hiztegia

Corpusari terminologia zientifikoa gehitzeko *Elhuyar Zientzia eta Teknologia* hiztegiko (ZTH) terminoak corpusean gehitzea erabaki da. Horretarako, ZTH hiztegiko kontzeptu bereko *es* eta *eu* termino-bikoteen konbinazioak bere horretan gehitu zaizkio corpusari, hau da, kasu honetan, sarrerako esaldi-parea eta irteerako termino-bikotea berdin-berdinak izango dira. Ikusi 14. irudia:



Irudia 14: Elhuyar Zientzia eta Teknologia hiztegiko terminoak corpuseko formatuan.

3.2 Atazarako Formatua

Esan bezala, datu-baseetatik lortutako informazioa ondokoa da: batetik, itzulpen-unitateak, hau da, bata bestearen itzulpena diren esaldi-bikoteak; bestalde, eta esaldi horietatik erauzitako termino-bikoteak. Aukera desberdinak aztertu ondoren, ondokoa da corpuserako erabaki den formatua:

- *Sarrera* fitxategia (lerro bat): *esaldia_es ZZZZ esaldia_eu.*
- *Irteera* fitxategia (lerro bat): *terminoa1_es GGGG terminoa2_es GGGG ... terminoaX_es ZZZZterminoa1_eu GGGG terminoa2_eu GGGG ... terminoaX_eu.*

Bi fitxategietako lerroak parekatuta daude, hau da, sarrera fitxategiko Y lerrotik lortutako terminoak irteera fitxategiko Y lerroan daude. Adibidez ikusi 2. taula.

Sarrera fitxategia (276.196 lerroa)	**Irteera fitxategia (276.196 lerroa)**
Se produce una **banda** deprimida semicircular en la zona antero lateral (externa) del **muslo**, producida por una **atrofia** del tejido adiposo subcutáneo, derivada de un **proceso** inflamatorio. 2, 3, 4, 5	banda gggg muslo gggg atrofia gggg proceso
zzzz	zzzz
Banda deprimitu erdizirkularra gertatzen da **izterraren** (kanpoko) aurre-alboko aldean, azalpeko ehun adiposoaren **atrofiak** eragina eta **prozesu** inflamatorio batetik eratorria.2, 3, 4, 5	banda gggg izterraren gggg atrofiak gggg prozesu

Taula 2: Ataza garatzeko sortutako corpusaren formatua, adibidea. Adibide honetan zuriuneak eta lerro saltoak gehitu dira irakurketa errazteko, baina corpus errealean lerro bakoitza jarraian, eta garbi agertuko da.

3.3 Corpusaren aurreprozesaketa

Atal honen hasieran (3.1.) azaldutako corpusa aberasteko egin diren ekintzak azalduko dira atal honetan.

3.3.1 Kodeketa arazoak

Itzulterm aplikaziora igo diren itzulpen-memoriak iturri askotako jatorria dute; gainera, itzulpen-memoria berean jatorri eta formatu desberdinetako dokumentuak itzultzen dira, ondorioz, kodeketa arazo handiak topatu dira: *utf-8*, *latin-1*, *windows-1521*, ezkutuko karaktereak... Automatikoki egiteko bide egokirik topatu ez denez, corpusa garbitzen eskuzko lan handia egin da. Ondorioz, lan honetarako behar den tamainako corpus zatia bakarrik garbitu da.

3.3.2 Termino-bikoteen hedapena

*ItzulTerm*en sortutako glosarioak erabili dira corpusak osatzeko; glosario horiek momentuaren beharren arabera sortutako glosarioak dira. Hau da, erabiltzaileak termino bat glosarioan sartuko du momentuan lantzen ari den glosarioaren ezaugarrien (tamaina, alorra, teknikoa, etab.) arabera. Tamalez, *Itzulterm*ek ezaugarri horiek ez ditu esplizituki gordetzen; ondorioz, *Itzulterm*en gordetako memoria guztiak berdin tratatuta sortzen den memoria homogeneoa ez izatea eragiten du. Adibidez, ikusi 3. taula:

HAP masterra

Esaldia es:	Esaldia eu:
La aparición en 1995 en unas oficinas bancarias en Bruselas, de una acumulación de casos (llegaron hasta 900) y la aparición posterior de algunos fenómenos parecidos en edificios de oficinas nuevos (en Cataluña en 2007) ha hecho que se hayan puesto en marcha otras hipótesis causales relacionadas con las condiciones laborales para la aparición de esta enfermedad.	1995an, gaixotasunaren kasu-pilaketa bat agertu zen Bruselako banku-bulego batzuetan (900raino iritsi ziren); ondoren, antzeko beste gertakari batzuk agertu ziren bulego-eraikin berri batzuetan (Katalunian 2007an). Horiek guztiak direla eta, gaixotasun hori agertzeko lan-baldintzekin erlazionaturiko beste kausa-hipotesi batzuk jarri dira abian.

Taula 3: Esaldi-bikote batetik alorraren arabera lor daitezkeen termino-bikoteak.

Adibidean (3. taulan) medikuntza alorreko esaldi-bikote bat dago; bertan, berdez markatuta dauden termino-bikoteak glosariorako hautatuko lituzke erabiltzaileak. Baina testu bera administrazioko glosario bat lantzeko erabiliko balitz, horiz markatuta daudenak hautatuko lirateke. Ondorioz, corpusean testua *ItzulTerm*en landu zeneko terminologia bakarrik gordeko balitz, berdeak bakarrik gordeko lirateke; eta sistemak horiak termino bezala ez sailkatzen ikasiko luke. Baina hori ez da zuzena, horiak ere terminoak baitira, beste alor batekoak, baina terminoak azken finean.

Arazo hori ekiditeko, hasierako corpusean ondoko lanketa automatikoa egin da:

- Corpusean agertzen diren termino-bikote guztiekin zerrenda bat osatu.
- Sortutako zerrenda horretan agertzen diren termino-bikote guztiak sarrera corpuseko segmentu guztietan markatu hautatutako termino-bikote bezala.

Lanketa honekin, hasierako corpusa homogeneoagoa bihurtu da, sistemari ikasteko aukerak zabalduz.

3.3.3 Segmentu errepikatuak kendu

Corpusa sortzeko erabilitako metodoa *Itzulterm* aplikazioak urteetan sortu dituen erauzketetan oinarritu denez, segmentu errepikatu asko daude. Segmentu horiek guztiak kendu egin dira hasierako corpusetik.

3.4 Corpusaren ezaugarriak

Proiektu honen lehenengo fasean *Itzulterm* tresnarekin urteetan egindako erauzketa guztiak bildu ziren, ondoko kopuruak lortuz:

- Esaldi-bikote (*es-eu*) kopurua: 2.380.527.

- Jatorrizko hizkuntzako (*es*) hitz-kopurua: 42.994.367.
- Helburu-hizkuntzako (*eu*) hitz-kopurua: 30.306.525.

Hala ere, lan hau egiteko corpus osoaren lagin bat bakarrik erabili da. Dokumentu honetan, hemendik aurrera corpusa aipatzen denean laginari egiten dio erreferentzia, ez hasieran bildutako corpus osoari. Erabili den laginaren ezaugarriak ondokoak dira:

- Corpuseko sarreraren ezaugarriak:
  - Esaldi-bikote (*es-eu*) kopurua: 105.999.
  - Sarrerako jatorri-hizkuntzako (*es*) hitz-kopurua: 2.001.659.
  - Sarrerako helburu-hizkuntzako (*eu*) hitz-kopurua: 1.453.809.
- Corpuseko irteeraren ezaugarriak:
  - Irteerako jatorri-hizkuntzako (*es*) hitz-kopurua: 769.114.
  - Irteerako helburu-hizkuntzako (*eu*) hitz-kopurua: 769.209.
  - Termino-bikote kopurua (*es-eu*): 429.896.

3.5 Entrenamendu multzoak

Sare neuronalak entrenatzeko corpusa zatitu egin da. Zati bat entrenatzeko (*train*), beste bat garapenerako (*dev*) eta beste bat ebaluaziorako (*test*).

- **train**: sarea entrenatzeko erabiltzen den datu-multzoa.
- **dev**: garapen fasean entrenatutako modeloa probatzeko eta hiperparametroak optimizatzeko erabiltzen den datu-multzoa.
- **test**: modelorik onena lortutakoan, modeloa probatzeko erabiltzen den datu-multzoa.

Gainera, modelo hoberena lortzeko bidean, entrenamenduetan corpus desberdinak erabili dira: 10.000 lerro dituen entrenamendu-corpusa, 25.000 lerrokoa, 50.000 lerrokoa, 75.000 lerrokoa eta 100.000 lerrokoa. Honela lerro-kopurua handitu ahala sistema nola hobetzen den aztertu ahal izango da. Entrenamendu corpusen ezaugarriak 4. taulan ikus daitezke:

HAP masterra

Fitxategia	Lerro kopurua	Sarrera		Irteera		
		Hitz kopurua (*es*)	**Hitz kopurua (*eu*)**	**Hitz kopurua (*es*)**	**Hitz kopurua (*eu*)**	**Termino -bikote kopurua**
dev	1.999	35.084	25.302	12.108	12.115	6.789
test	4.000	71.228	51.720	25.758	25.771	14478
train_10K	10.000	200.926	140.129	72.582	72.571	40.577
train_25K	25.000	497.104	347.931	187.260	187.258	104.334
train_50K	50.000	891.134	626.400	340.112	340.148	190.715
train_75K	75.000	1.361.506	969.403	526.952	527.009	295.299
train_100K	100.000	1.895.347	1.376.787	731.248	731.323	408.629

Taula 4: Entrenamendurako erabili diren corpusen ezaugarriak.

3.6 Sekuentziatik sekuentziarako hurbilpena

Seq2seq hurbilpenerako erabili den tresna *fairseq* da. Lehenago azaldu den bezala, *Fairseq* python programazio lengoaian idatzita dago eta sekuentziekin lan egiteko tresna-multzo handia du. *Fairseq* tresnak ondoko funtzio nagusiak ditu:

- *fairseq-preprocess*: datuen aurreprozesaketa egiteko agindua: besteak beste, sare neuronalak erabiliko dituen hiztegiak sortu eta entrenamendurako corpuseko testua minuskuletara pasatzen du.
- *fairseq-train*: modelo berri bat entrenatzeko agindua.
- *fairseq-generate*: entrenatutako modeloari aurreprozesatutako corpus bat emanda, modeloak itzuliko duen irteera erakusten duen agindua.
- *fairseq-interactive*: testu formatuan emandako sarrerarekin modeloak itzultzen duen irteera erakusten duen agindua.
- *fairseq-score*: *fairseq-generate* aginduarekin sortutako irteera bat erreferentziazko itzulpenekin konparatu eta BLEU balioa itzultzen duen agindua.
- *fairseq-eval-lm*: hizkuntza modeloaren ebaluazioa.

Lan honetan, behar diren modeloak sortzeko, *fairseq*en lehenengo hiru aginduak erabili dira. Aginduetan erabili diren hiperparametroen balioak *fairseq* sistemak defektuz dituenak dira. *Fairseq*en dokumentazioan defektuzko hiperparametro hauek egokiak direla aipatzen da eta aldaketak egiteko beharrik ez da ikusi. Erabilitako hiperparametroen zerrenda 5. taulan ikus daiteke:

HAP masterra

Hiperparametroa	**Balioa**
Architecture	*transformer_iwslt_de_en*
Number of layers	6
Number of units	512
Vocabulary-size	90.000
Preprocess-workers	8
Embedding-size	512
Atention-head	4
Optimization	*adam*
adam-betas	(0,9, 0,98)
Criterion	*label_smoothed_cross_entropy*
Smothing	0,1
Batch-size	30
max-tokens	4.096
Learning-rate	0,0005
lr-scheduler	*inverse_sqrt*
warmup-updates	4.000
Drop-out	0,3
Beam-width	5
Best checkpoint metric	*loss*
keep-last-epochs	3
nBest	3

Taula 5: Entrenamenduetan erabilitako hiperparametroen balioak.

3.7 NBest algoritmoa eta heuristikoak

Itzulpen automatikoan sarrerako esaldi baten itzulpen zuzenak bat baino gehiago izan daitezke. Modelo batetik itzulpen bat baino gehiago lortzeko algoritmoetako bat NBest (Chow eta Schwartz (1989)) da. NBest algoritmoa denborarekiko sinkronoa den Viterbi motako bilatzaile bat da eta itzulpen multzo batetik egokienak diren n itzulpen aukeratzen ditu. Horrela, ondorengo pausu batean eta horretarako definitu daitezkeen heuristikoen bidez, n horietatik egokiena erabakitzeko aukera ematen du. Seq2seq arkitekturan NBest hiperparametroa gehituta esaldi baten n itzulpen lortzen dira (ikusi 15. irudia).

HAP masterra

Irudia 15: NBest parametroari 3 balioa emanda sistemak itzultzen dituen aukerak.

NBest-ek lortutako n emaitzetatik egokiena erabakitzeko aukera ematen du, atazaren araberako heuristiko batzuk definituz. Lan honetan ondoko heuristikoak definitu dira:

- Sistemak GS-eko irteeran dauden termino-bikoteetatik ahalik eta gehien itzultzea. Adibidez, 15. irudiko 1-Best proposamenean ez dago "*incapacidad-ezgaitasun*" termino-bikotea.
- Sistemak itzultzen duen termino bakoitza GS-eko dagokion hizkuntzako esaldian egotea. Adibidez, 15. irudiko 3-Best proposameneko "*parada*" terminoa ez dago GS-eko sarrera esaldian.
- Sistemak itzultzen dituen termino-bikoteak GS-eko irteeran egotea. Adibidez, 15. irudiko 2-Best proposameneko "*incapacidad-langile*" termino-bikotea ez dago GS-aren irteeran.

3.8 TEB: Termino-Erauzle Balidazioa. Lan honetarako egindako ebaluazio algoritmoa

Sortutako sistema baloratu eta BLEU metrikarekin lortutako emaitza osatzeko, beste metrika bat sortu da, ondoko arrazoiengatik:

HAP masterra

- BLEUk esaldien luzera kontuan hartzen du, itzulpen automatikoan oinarritua dagoelako. *Itzulterm*ek *gold standard*ean (GS) ez dauden termino-bikote dezente itzultzen ditu; ondorioz, irteeran itzultzen duen testuaren luzera asko aldatzen da. Honek sistemaren balorazioan penalizazio handia sortuz.

- Termino erauzketan GS-eko irteerako termino-bikoteetatik ahalik eta gehien lortzea ongi baloratu beharko litzateke. BLEUk antzekotasuna bakarrik neurtzen du. Ondorioz, termino-bikote antzekoak itzultzen dituen sistema bat oso ongi baloratuko du. Baina termino-bikote berak orden desberdin batean edo beste termino-bikoteen artean baldin badatoz ebaluazioaren emaitza kaxkarra izango da.

Arrazoi horiek ikusita sistema berria (eta *Itzulterm*) baloratzeko beste metrika bat eraiki da. Algoritmo horrek ondoko formularen emaitza itzultzen du:

- Sarrera bezala ondoko esaldi-bikotea izanda,

  $esaldia_{es}\ zzzz\ esaldia_{eu}$

- Sarrera horrentzat *gold standard*-eko irteeran ondoko termino-bikote multzoa izanda,

  $T = [T_{es_1} - T_{eu_1},\ T_{es_2} - T_{eu_2},\ ..., T_{es_n} - T_{eu_n}]$
  non $T_{es_1},\ T_{es_2},\ ..., T_{es_n} \in esaldia_{es}$
  non $T_{eu_1},\ T_{eu_2},\ ..., T_{eu_n} \in esaldia_{eu}$

- Eta, baloratu nahi den sistemak hasierako esaldiarentzat ondoko termino-bikote multzoa erauzita,

  $X = [X_{es_1} - X_{eu_1},\ X_{es_2} - X_{eu_2},\ ..., X_{es_m} - X_{eu_m}]$

- X irteera T irteerarekin konparatu eta egokitasuna itzuliko duen metrika honela definituko litzateke:

$$\{X \equiv T \rightarrow 6\}\ \vee \tag{1}$$
$$\{(\exists X_{es_j} \in esaldia_{es} \rightarrow 2) + (\exists X_{eu_j} \in esaldia_{eu} \rightarrow 2) + \tag{2}$$
$$(\exists X_{es_j} \notin esaldia_{es} \rightarrow -1) + (\exists X_{eu_j} \notin esaldia_{eu} \rightarrow -1) + \tag{3}$$
$$\frac{2 * \sum_{j=1}^{m}([X_{es_j} - X_{eu_j}] \in T)}{n+m} + \tag{4}$$
$$\frac{\sum_{j=1}^{m}([X_{es_j}] \in T_{es})}{2 * (n+m)} + \frac{\sum_{j=1}^{m}([X_{eu_j}] \in T_{eu})}{2 * (n+m)}\} \tag{5}$$

  – (1) GSaren emaitza eta balidatu nahi den sistemaren emaitza berdina baldin bada, puntuazio maximoa (6) itzuli.

- (2) Sistemak itzuli dituen terminoetatik, terminoren bat jatorrizko esaldian eta dagokion hizkuntzan baldin badago, 2 puntu batu.
- (3) Sistemak itzuli dituen terminoetatik, terminoren bat jatorrizko esaldian eta dagokion hizkuntzan ez baldin badago, puntu bat kendu.
- (4) Sistemak berriak GSaren emaitzan dauden termino-bikoteetatik ahalik eta gehien erauztea baloratu nahi da formula horrekin.
- (5) Sistema berriak GSaren emaitzako terminoetatik ahalik eta gehien erauztea baloratu nahi da formula horrekin.

Formula honek 0.0-tik 6.0rako balio bat itzuliko du esaldi bakoitzeko; erauzketa bat balidatu nahi denean lerro bakoitzean itzuli duen balioaren batezbestekoa egingo da, eta emaitzari *100/6 aplikatuko zaio, 0 eta 1 arteko balioa lortzeko.

Formularen funtzionamendua adibide batzuen bidez ikus daiteke 6. eta 7. tauletan:

Irteera nondik	**Itzulitako termino zerrenda**	**TEB**	**Puntuazioaren arrazoiak**
GSaren sarrera	en caso de **duda** se someterá al arbitraje del **director** y a la **legislación** que regule estos **aspectos** de la **actividad** profesional zzzz **zalantzarik** izanez gero **zuzendariak** ebatziko du lanbide **jardueraren alderdiak** arautzen dituen **legediaren** arabera		
GSaren irteera	duda gggg director gggg legislación gggg aspectos gggg actividad zzzz zalantzarik gggg zuzendariak gggg legediaren gggg alderdiak gggg jardueraren		
1-Best	duda gggg director gggg legislación gggg aspectos gggg actividad zzzz zalantzarik gggg zuzendariak gggg legediaren gggg alderdiak gggg jardueraren	6	GSaren emaitzaren berdina da, puntuazio maximoa
2-Best	duda gggg directora gggg legislación gggg aspectos gggg actividad zzzz zalantzarik gggg zuzendariak gggg legediaren gggg alderdiak gggg jardueraren	3.6	Termino bat ez dago sarrerako testuan: *directora*. Termino-bikote bat ez dago GSaren irteeran: (*directora-zuzendariak*).
3-Best	duda gggg director gggg legislación gggg aspecto gggg actividad zzzz zalantzarik gggg zuzendariak gggg legediaren gggg alderdiak gggg jardueraren	3.6	Termino bat ez dago sarrerako testuan: *aspecto*. Termino-bikote bat ez dago GSaren irteeran: (*aspecto-alderdiak*).

Taula 6: TEB metrikaren adibide bat.

Irteera nondik	Itzulitako termino zerrenda	TEB	Puntuazioaren arrazoiak
GSaren sarrera	**preacuerdo** sobre el **convenio** de **colectivos** laborales al **servicio** la **administración** de la **comunidad** autónoma de euskadi para 1996 1997 zzzz euskal autonomi **elkarteko administrazioaren zerbitzuan** diharduten lan itunpeko **taldeen** 1996 1997rako **hitzarmenari** buruzko **aurreakordioa**		
GSaren irteera	preacuerdo gggg convenio gggg colectivos gggg servicio gggg administración gggg comunidad zzzz aurreakordioa gggg hitzarmenari gggg taldeen gggg zerbitzuan gggg administrazioaren gggg elkarteko		Urdinez markatutako termino-bikoteak sistemak ez ditu erauzi: (*preacuerdo-aurreakordioa*), (*comunidad-elkarteko*).
1-Best	convenio gggg colectivos gggg administración gggg administración zzzz hitzarmenari gggg taldeen gggg zerbitzuan gggg administrazioaren	3.25	-. GS-eko termino-bikoteak falta dira:(*preacuerdo-aurreakordioa*), (*comunidad-elkarteko*). -. Itzulitako bikote batzuk ez daude GSean: (*administración-zerbitzuan*). -. GS-ko irteerako 3 termino-bikote ditu.
2-Best	convenio gggg colectivos gggg administración zzzz hitzarmenari gggg taldeen gggg administrazioaren	4.667	-. GS-eko termino-bikoteak falta dira: (*preacuerdo-aurreakordioa*), (*comunidad-elkarteko*). -. GS-ko irteerako 3 termino-bikote ditu.
3-Best	convenio gggg colectivos gggg servicio gggg administración zzzz hitzarmenari gggg taldeen gggg zerbitzuan gggg administrazioaren	5.5	-. GS-eko termino-bikoteak falta dira: (*preacuerdo-aurreakordioa*), (*comunidad-elkarteko*). -. GS-ko irteerako 4 termino-bikote ditu.

Taula 7: TEB metrikaren beste adibide bat.

Metrika honek baditu hobetzeko puntuak, baina BLEUren emaitzak osatzeko baliogarria izango da. Algoritmo honi, TEB (*Termino Erauzle Balidazioa*) izena jarri zaio.

4 Esperimentazioa eta emaitzak

Atal honetan, lehenengo, sortuko den sistema zerekin konparatu nahi den definituko da (4.1.). Ondoren, datu-multzo desberdinak erabiliz sistema optimizatu eta lan honetako sistema sortuko da (4.2.1.). Ondoren, sistemaren emaitzak kuantitatiboki (4.2.2.) eta kualitatiboki (4.2.3. eta 4.2.4.) azalduko dira.

4.1 Oinarri lerroa

Sistema honen eraginkortasuna neurtu ahal izateko, sistemaren emaitzak beste sistema baten emaitzekin konparatu behar dira. Euskara-gaztelania hizkuntza-bikoterako dagoen sistema bakarra *Itzulterm* da; beraz, garapeneko eta ebaluazioko corpusekin *Itzulterm*ek itzultzen duen emaitza *gold standard*aren emaitzarekin konparatuko da. Horretarako, garapeneko sarrera fitxategia *Itzulterm* aplikazioan exekutatu eta lortzen den emaitza izango da gero konparaziorako erabiliko den oinarria.

Itzulterm exekutatuta lortu den emaitza *gold standard*aren irteera fitxategiarekin konparatu ondoren lortutako BLEU eta TEB metriken balioak 8. eta 9. tauletan ikus daitezke, hurrenez hurren.

BLEU metrikaren taulan (8. taula) ikusten den bezala, garapeneko eta ebaluaziorako corpusetan bina BLEU kalkulatu dira: lehenengoa corpusaren irteera fitxategia bere horretan hartuta kalkulatu da; bigarrena aldiz, corpusaren irteera fitxategitik 'zzzz' eta 'gggg' tokenak kenduta kalkulatu da. BLEU metrikak lerroen antzekotasuna neurtzen du eta lerro guztietan 'zzzz' token bana eta 'gggg' n token daudenez, lerroen antzekotasuna handitu egiten da. 'zzzz' eta 'gggg' tokenak ez dira terminoak; terminoak bereizteko kateak, baizik. Ondorioz, token horiek garbitu ondorengo BLEU zenbakia errealagoa da.

Corpusa	**BLEU garbitu gabea**	**BLEU garbia**
Dev	0,412	0,267
Test	0,392	**0,248**

Taula 8: *gold standard*a *Itzulterm*en exekutatuta lortutako emaitzaren BLEU balioa.

TEB metrika kalkulatzeko 'zzzz' eta 'gggg' tokenak ez dira kontuan hartzen, ondorioz, kalkulu bakarrarekin nahikoa da kasu honetan (ikusi 9. taula).

Corpusa	**TEB metrikaren emaitza**	$TEB * 100/6$
Dev	4,634	0,7698
Test	4,619	**0,7699**

Taula 9: *gold standard*a *Itzulterm*en exekutatu ondoren lortutako emaitzaren TEB balioa.

Bi taulak aztertuz, *Itzulterm*ek *gold standard*arekiko lortutako puntuazioa **0.248** da BLEU metrikan eta **0.7699** TEB metrikan. Hemendik aurrera, lan honetan entrenatutako modeloen egokitasuna konparatzeko erreferentzia moduan erabiliko dira bi balio horiek.

4.2 Garapena

Garapena *Google Colaboratory* (*Colab*)zerbitzuan egin da. Nabigatzaile batekin konekta daitekeenez sistema eragilearekiko independentea da eta ordenagailu normal batekin egin daiteke. *Colab*ek ondoko abantailak ditu: GPUarekin lana egiteko aukera, kodea partekatzeko erraztasuna, *Google drive*kin bat egiteko aukera, corpusak *Drive*n edukitzeko aukera... *Colab*en bertsio libreak desabantailak ere baditu: 15Gko muga du, gainera exekuzio denbora ere kontrolatu egiten du, 10 bat ordu pasatutakoan deskonektatu egiten da eta ordu batzuk pasa behar dira berriro konektatu ahal izateko eta abar.

4.2.1 Corpusaren tamaina egokiena definitzen

Sistemarik egokiena aukeratzeko, 3.5. atalean azaldu den bezala, lerro-kopuru desberdinetako corpusekin egin da entrenamendua; horrela, termino-erauzketa atazarentzat entrenamendurako corpusaren tamaina optimoa definitzeko asmotan. Gainera, corpus bakoitza bi aldiz entrenatu da, BPE segmentazioa erabiliz eta BPE erabili gabe.

Egindako entrenamendu guztien emaitzak (ikusi 10. taula) aztertu ondoren aukeratuko da modelo egokiena.

Lerro kopurua	**BPE: bai-ez**	***Epoch* kop**	***Best epoch***	***Loss***	**Sarrera hiztegia**	**Irteera hiztegia**	***Dev* BLEUa**	***Dev* BLEU garbia**
10.000	ez	292	132	3,204	22.888	7.824	0,596	0,446
10.000	bai	248	93	3,477	20.568	7.824	0,687	0,519
25.000	ez	109	22	3,261	44.520	15.456	0,335	0,157
25.000	bai	392	52	2,498	38.880	15.472	0,724	0,564
50.000	ez	242	72	2,588	77.936	25.912	0,698	0,542
50.000	bai	123	67	2,411	65.384	25.992	0,803	**0,678**
75.000	ez	101	71	2,859	100.024	32.504	0,863	**0,778**
75.000	bai	83	48	2,999	81.432	32.648	0,749	0,603
100.000	ez	94	92	3,7	119.304	38.064	0,236	0,135
100.000	bai	97	42	3,334	83.920	36.704	0.672	0.498

Taula 10: Lan honetarako egindako sistemaren modeloa lortzeko egin diren probak.

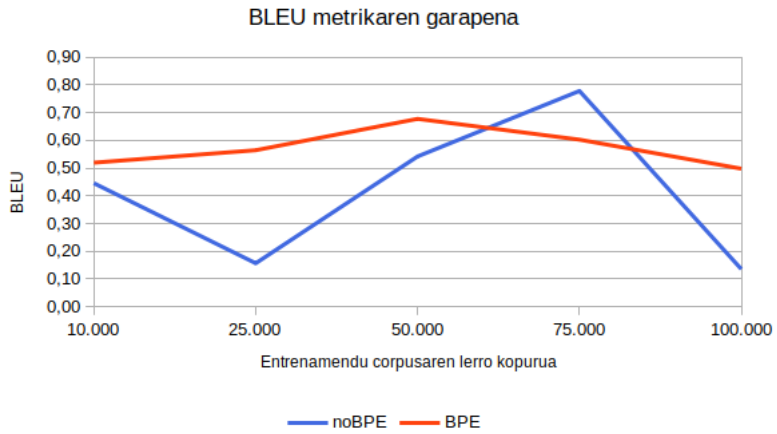
Entrenamendu guztien taulako (ikusi 10. taula) zutabeek ondokoa adierazten dute:

- **Lerro kopurua**: Modeloa entrenatzeko erabili den lerro kopurua.
- **BPE bai-ez**: Modeloa entrenatzeko BPE segmentatzailea erabili den edo ez.

- ***Epoch* kop**: Entrenamendu garaian datu-multzoari eman zaion buelta kopurua.
- ***Best epoch***: Entrenamendu garaian garapeneko datu-multzoari eman zaizkion buelta guztietan *loss* balio txikiena duena.
- ***Loss***: Sare neuronalaren errore-tasa.
- **Sarrera hiztegia**: Corpuseko sarreratik lortu den hiztegiaren hitz kopurua.
- **Irteera hiztegia**: Corpuseko irteeratik lortu den hiztegiaren hitz kopurua.
- ***Dev* BLEU**: Garapeneko corpusetik lortutako emaitzetan BLEU metrikak eman duen balioa.
- ***Dev* BLEU garbia**: Garapeneko corpusetik modeloak itzuli duen emaitzari 'zzzz' eta 'gggg' tokenak kenduta lortzen den BLEU balioa, aurrekoa baino errealagoa da.

Egindako entrenamendu guztietatik emaitza hoberenak ematen dituen modeloa 75.000 lerroko corpusa eta BPE erabili gabeko entrenamendutik lortu da. Esandako modeloarekin BLEU metrikan 0.78ko balioa lortu da. Modelo honi *Termino Erauzle Modeloa* (TEM) izena jarri zaio, hemendik aurrera dokumentu honetan horrela agertuko da.

Modelo egokiena zein den erabakitzeko, emaitzen taulako (10. taula) 9. zutabearen emaitzak 16. irudiko grafikoan errepresentatu dira. Hasiera batean corpusa handitu ahala modeloak emaitza hobea lortzen badu ere, 75.000 lerrotik aurrera emaitzak okertu egiten direla ikus daiteke, bai BPE aplikatuta, baita gabe ere.



Irudia 16: Garapeneko corpusa handitu ahala BLEU metrikaren eboluzioa.

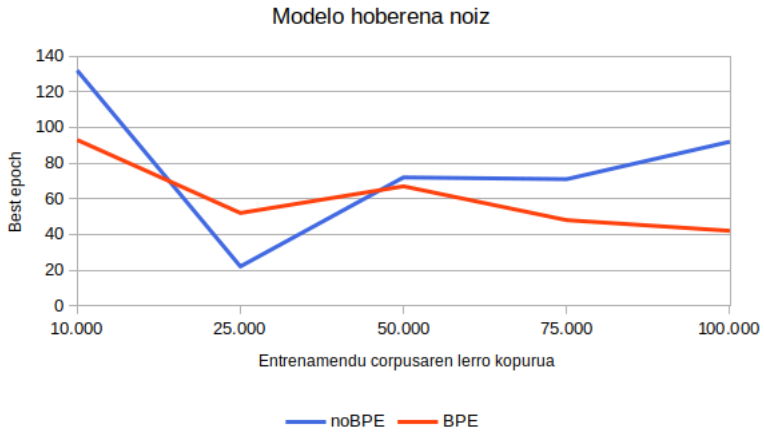
BLEU balioaren eboluzioaren interpretazio bat ondokoa da:

- BPE gabeko entrenamenduan hiztegiaren tamaina maximoa gainditzearen ondorioz, maiztasun txikiko hitzak hiztegian ez sartzea. Lehenago esan den bezala BPE erabili

gabe hitz deklinatu bakoitza sarrera bezala sartzen da hiztegian. Ondorioz, corpusa handitu ahala, hiztegia gehiegi handitzen da eta ikaste errendimendua jaitsi egiten da.

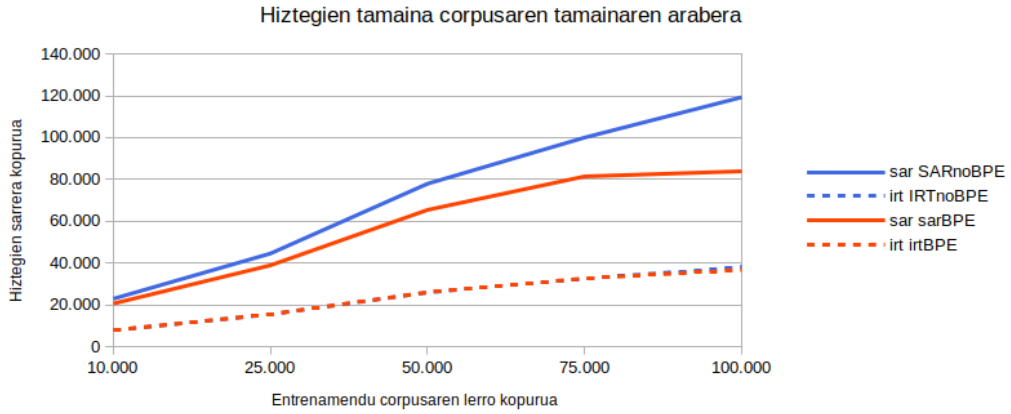
- Hizkuntzaren prozesamenduko atazetan, BPE erabiltzearen onurak aipatu dira 2.4 atalean. Hala ere, kasu honetan, BPE erabilita ere, 60.000 lerrotik aurrera emaitzak asko okertzen dira. Logikaren arabera, datu gehiago izanda eta hiztegian hitz gehiago sartzeko aukera izanda (BPEk ez ditu deklinatutako hitz guztiak gordetzen) emaitza hobeak lortu beharko lirateke. Baina ez da hori gertatu.
- *Transformer*ak bi hizkuntzen arteko itzulpenak egiteko oso eraginkorrak dira; baina ataza honetako sarreran bi hizkuntza daude eta irteeran ere bai. Hori izan daiteke emaitzak okertzearen arrazoietako bat.
- Corpusa txikiegia izatea, hau da, 75.000 lerrotik 100.000 lerrotara pasatzean litekeena da oso gai desberdineko testuak sartu izana eta horrek emaitzetan eragina izatea. Baliteke, corpus osoarekin eta BPE aplikatuz entrenatuta emaitza hobeak lortzea. Baina lan honetan *Colab*en mugak direla eta ezin izan da proba hori egin.

Emaitzen taulako (10. taula) hirugarren eta laugarren zutabeei erreparatuta sortu den grafikoan (ikusi 17. irudia), nabaria da corpusa handitu ahala entrenamenduak datu-multzoan eman behar dituen buelta kopurua txikiagoa dela. 25.000 lerroko corpusean gertatzen dena salbuespena da.



Irudia 17: Entrenamenduan *epoch* hoberena zenbatgarrena den.

Hiztegien tamainari begiratuta (17. taulako 7. eta 8. lerroak), aldiz, erraz ikus daiteke corpusa handitu ahala sarrerako hiztegiak ere handitu egiten direla. Sarrerako corpusaren hiztegiak gehiago handitzen dira irteerako corpusak baino: sarreran esaldi osoak daude; irteeran, aldiz, termino esanguratsuenak bakarrik. Gainera BPE gabeko hiztegia eta BPE erabilita sortzen diren sarrera hiztegien diferentzia handitu egiten da, corpusarekin batera. Ikusi 18. irudia:



Irudia 18: Entrenamenduaren corpusa handitu ahala, hiztegien tamaina ere handitu egiten da, BPErena gutxiago. Irteerako hiztegien marrak bata bestearen gainean daude, kopuru oso antzekoak dituztelako.

4.2.2 Emaitzak

*Itzulterm*en emaitzak *gold standard*arekin konparatuta ematen duen BLEU balioa 8. taulan ikusi da. Sortutako modelorik hoberenaren (TEM) BLEU balioa taula berean jarrita, erraz esan daiteke sare neuronalekin garatutako sistemak metodo estatistikoak aplikatutako sistemaren emaitzak hobetu dituela (ikusi 11. taula). BLEU metrikaren arabera 50 puntuko aldea atera dio TEMek *Itzulterm*i, TEB metrikaren arabera berriz 10 puntuko aldea. Bigarren metrikak gehiago erreparatzen dio terminologia-erauzketari itzulpenaren antzekotasunari baino. Ondorioz bi metriken emaitzak bata bestearen osagarriak izango dira.

Corpusa	**BLEU garbitu gabea**	**BLEU garbia**	**TEB**	$TEB * 100/6$
Itzulterm DEV	0,412	0,267	4,634	0,7698
Itzulterm TEST	0,392	**0,248**	4,619	**0,7699**
TEM DEV	0,863	0,778	5,272	0,879
TEM TEST	0,855	**0,766**	5,214	**0,869**

Taula 11: *Itzulterm*en eta TEM sistemaren emaitzak *Gold standard*arekiko, BLEU eta TEB metrikekin.

TEB sistemak **0,766**ko puntuazioa lortu du BLEU metrikan. *Itzulterm* tresnarekiko alderatuz, 50 puntuko aldea lortu du BLEU metrikan. *Modela* itzultzaile automatikoak (Etchegoyhen et al. (2018)), gaztelaniatik euskararako itzulpen automatiko neuronalak, estatistikoarekiko 4 puntuko aldea lortu zuen BLEU metrikan. Kualitatiboki aztertuta, aldiz, hobekuntza nabarmena ekarri zuen *Modelak es-eu* itzulpenetarako. Lau puntuko alde

horrek itzulpen automatikoarentzat ekarritako hobekuntza ikusita, TEB sistemak *Itzulterm*ekiko lortutako 50 puntuko diferentzia oso ona da. Kontuan izan behar da, hala ere, *Itzulterm* neurtzeko erabili den corpusa ez dela homogeneoa izan eta TEB sistemarena aldiz bai. *Itzulterm*en corpusa homogeneoa izango balitz 50 puntuko diferentzia hori txikiagotu egingo litzatekeela kontuan izanda ere; sare neuronalen teknikak terminologia-erauzketa ataza lantzeko baliogarriak direla frogatutzat eman daiteke.

Atal honetan TEMen emaitza kuantitatibo bat ikusi da; hurrengo bietan, adibide erreal batzuen bidez, emaitza kualitatiboki aztertuko da.

4.2.3 Adibide arrakastatsuak

Sortutako sistemarekin lortutako emaitzen adibideak ikusteko taula batzuk sortu dira, taula hauetako lerroak honela antolatu dira:

- **GS-sar**: *Gold standard*aren sarrera esaldia.
- **GS-irt**: *Gold standard*aren irteera, termino-bikote multzoa.
- **1-Best**: NBest sistemak emaitzarik onena bezela itzuli duen termino-bikote multzoa.
- **2-Best**: NBest sistemak bigarren emaitza onena bezala itzuli duen termino-bikote multzoa.
- **3-Best**: NBest sistemak hirugarren emaitza onena bezala itzuli duen termino-bikote multzoa.

	Sarrera/irteera testua
GS-sar	pago al asegurado de una cantidad de hasta 5 **millones** de pta según **baremo** de la invalidez zzzz aseguratuarentzat 5 **milioi** pezetaraino ezgaitasun **baremoaren** arabera
GS-irt	**millones** gggg **baremo** zzzz **milioi** gggg **baremoaren**
NBest	**Irteera testua**
1-Best	**millones** gggg **baremo** zzzz **milioi** gggg **baremoaren**
2-Best	**millones** gggg **baremo** zzzz **milioi** gggg **baremo**
3-Best	**pesetas** gggg **baremo** zzzz **milioi** gggg **baremoaren**

Taula 12: Sistemak itzulitako adibide arrakastatsu bat.

Lortu den emaitza arrakastatsu bat 12. taulan ikus daiteke. Ikusten denez, *Gold standard* (GS) corpusean dagoen emaitza ber-bera itzultzen du sistemak. Egia da, hala ere, GSaren sarrerako esaldian termino gehiago egon daitezkeela: (*asegurado-aseguratuarentzat, invalidez-ezgaitasun*); baina GSaren irteeran ere ez dira agertzen termino-bikote horiek. Adibide honek adierazten du sare neuronalek emandako adibideetatik ikasten dutela.

	Sarrera/irteera testua
GS-sar	los **grupos** establecidos en el presente **convenio** definen inicialmente los **puestos** de trabajo de la **empresa** zzzz **hitzarmen** honetan jarritako **taldeek enpresako lanpostuak** definitzen dituzte
GS-irt	**grupos** gggg **convenio** gggg **puestos** gggg **empresa** zzzz **taldeek** gggg **hitzarmen** gggg **lanpostuak** gggg **enpresako**
NBest	**Irteera testua**
1-Best	**grupos** gggg **convenio** gggg **puestos** gggg **empresa** zzzz **taldeek** gggg **hitzarmen** gggg **lanpostuak** gggg **enpresako**
2-Best	**grupos** gggg **convenio** gggg puesto gggg **empresa** zzzz **taldeek** gggg **hitzarmen** gggg lanpostuak gggg **enpresako**
3-Best	**grupos** gggg **convenio** gggg **puestos** gggg enpresako zzzz **taldeek** gggg **hitzarmen** gggg **lanpostuak** gggg enpresako

Taula 13: Sistemak itzulitako adibide arrakastatsu bat.

	Sarrera/irteera testua
GS-sar	**i** la disminución continuada y voluntaria en el **rendimiento** de **trabajo** normal o pactado zzzz **i** ohiko edo hitzartutako **lanean errendimendua** borondatez eta etengabe jaistea
GS-irt	**i** gggg **rendimiento** gggg **trabajo** zzzz **i** gggg **errendimendua** gggg **lanean**
NBest	**Irteera testua**
1-Best	**i** gggg **rendimiento** gggg **trabajo** zzzz **i** gggg **errendimendua** gggg **lanean**
2-Best	**i** gggg **rendimiento** gggg **trabajo** zzzz **i** gggg **errendimendua** gggg **lana**
3-Best	**i** gggg **rendimiento** gggg **lanean** zzzz **i** gggg **errendimendua** gggg **lanean**

Taula 14: Sistemak itzulitako adibide arrakastatsu bat.

GSaren irteeraren emaitza bera lortu da 13. eta 14. tauletako adibideetan ere. Eztabaidagarria izan daiteke 14. adibideko "**i**"tokena terminoa den edo ez; baina GSan horrelakoak termino bezala identifikatuta daudenez, hori da sistemak ikasi duena. GS garbiago batekin entrenatutako sare neuronal batek ez lituzke horrelakoak markatuko.

4.2.4 Errore analisiak

Erroreei dagokionez, aurkitu den kasuistikaren arabera antolatu dira adibideak:

HAP masterra

1. **Antzeko terminoak jatorri/helburu testuetan:**

Sarrerako esaldian irteerako hiztegian ez dagoen eta terminoa izateko hautagaia izan daitekeen proposamen bat aurkitzean, sistemak termino horretatik gertu dagoen beste bat proposatzen du. Portaera hori oso ohikoa da sare neuronalak erabiltzen dituzten sistemetan. Adibideak 15., 16. eta 17. tauletan ikus daitezke.

	Sarrera/irteera testua
GS-sar	la **fijación** de dicho tiempo se realizará por acuerdo de ambas partes zzzz denbora horren **finkapena** bi alderdien adostasunez egingo da
GS-irt	fijación gggg **tiempo** gggg **acuerdo** gggg **partes** zzzz **finkapena** gggg **denbora** gggg **adostasunez** gggg **alderdien**
NBest	**Irteera testua**
1-Best	consolidación gggg **tiempo** gggg **acuerdo** gggg **partes** zzzz **finkapena** gggg **denbora** gggg **adostasunez** gggg **alderdien**
2-Best	puesto gggg **tiempo** gggg **acuerdo** gggg **partes** zzzz **finkapena** gggg **denbora** gggg **adostasunez** gggg **alderdien**
3-Best	procedencia gggg **tiempo** gggg **acuerdo** gggg **partes** zzzz **finkapena** gggg **denbora** gggg **adostasunez** gggg **alderdien**

Taula 15: Errorea: antzeko terminoa jatorrizko testuan.

Goiko adibidean, 15. taulan, GSak itzultzen dituen termino-bikoteen artean "*fijación-finkapena*" termino-bikotea dago. Sistemak itzuli dituen hiru emaitzetan aldiz ez da "*fijación*" gaztelaniazko terminoa agertzen; hitz hori irteerako hiztegian ez duela aurkitu dirudi. Ondorioz, antzeko edo hitz horretatik gertu egon daitezkeen hitzak sortu ditu: "*consolidación*" lehenengo proposamenean, "*puesto*" bigarrenean eta "*procedencia*" hirugarrenean. Adibide honek erakusten du sistema gai dela hiztegian aurkitzen ez duen termino batetik gertu dagoen termino bat, sinonimo bat, itzultzeko. Ataza honetan, terminologia-erauzketan, testuetako terminologia berria erauztea da helburua, ez terminologiaren sorkuntza. Ondorioz, garrantzitsua da irteerako termino-bikoteetan sarrerako testuan dauden terminoak egotea; sinonimoek ez dute balio.

Entrenamendu garaian bi hiztegi erabili ordez bakarra erabili izan balitz ziurrenik erauzitako termino-bikoteetarako sinonimoen sorkuntza txikiagoa izango litzateke.

	Sarrera/irteera testua
GS-sar	**sondika** zzzz **sondika**
GS-irt	**sondika** zzzz **sondika**
NBest	**Irteera testua**
1-Best	**plentzia** zzzz **plentzia**
2-Best	**alonsotegi** zzzz **plentzia**
3-Best	**larrabetzu** zzzz **plentzia**

Taula 16: Errorea: antzeko terminoa jatorrizko eta helburuko testuetan.

Antzeko egoera ikus daiteke 16. taulan "*sondika-sondika*"termino-bikotearekin. Termino horiek ez ditu irteerako hiztegian topatu, ondorioz gertuko batzuk itzuli ditu, lehenengo proposamenean nahiko txukun "*plentzia-plentzia*", hurrengo bi proposamenetan aldiz traketsago: "*alonsotegi-plentzia*", "*larrabetzu-plentzia*". Hala ere, nolabait egin du herrien lotura: Sondika, Plentzia, Alonsotegi eta Larrabetzu. Itzuli dituen termino guztiak herri izenak dira. Lehen esan bezala, hemen ere termino logikoak sortzen dituen arren, ez da terminologia-erauzketa atazarentzat espero den portaera.

Sarrera hiztegian egongo da "*sondika*"terminoa, baina irteera hiztegian ez; eta sorkuntza irteera hiztegiarekiko egiten duenez, "sinonimoak"itzultzen ditu. Arazo hori hiztegi bakarra erabilita konponduko litzateke.

	Sarrera/irteera testua
GS-sar	**f** un **día** por **boda** de un **hijo** o **hermano** zzzz **f egun** bat **seme** alaba edo **neba** arrebaren baten **ezkontzagatik**
GS-irt	**f** gggg **día** gggg **boda** gggg **hijo** gggg **hermano** zzzz **f** gggg **egun** gggg **ezkontzagatik** gggg **seme** gggg **neba**
NBest	**Irteera testua**
1-Best	**f** gggg **día** gggg **matrimonio** gggg **hijo** gggg **padres** zzzz **f** gggg **egun** gggg **ezkontzagatik** gggg **seme** gggg **aitaren**
2-Best	**f** gggg **día** gggg **matrimonio** gggg **hijo** gggg **padres** zzzz **f** gggg **egun** gggg **ezkontzagatik** gggg **seme** gggg **aiton**
3-Best	**f** gggg **día** gggg **matrimonio** gggg **hijo** gggg **padres** zzzz **f** gggg **egun** gggg **ezkontzagatik** gggg **seme** gggg **anaia**

Taula 17: Errorea: antzeko terminoa jatorrizko eta helburuko testuetan.

Adibide honetan ere, 17. taula, proposatzen dituen terminoak gertutasun bat adierazten dute. "*Boda* terminoarentzat *matrimonio* sortu du; eta, "*hermano*"terminoarentzat "*padres*"; eta "*neba*"terminoarentzat "*aitaren*", "*aiton*"edo "*anaia*"terminoak itzuli ditu.

2. **Antzeko terminoak itzultzean okerreko adiera bat:**

Termino bat itzuli ordez antzekoa den beste termino bat itzultzean ez da beti asmatzen. Batzuetan jatorrizko terminoaren beste adiera bati dagokion terminoa itzultzen da.

	Sarrera/irteera testua
GS-sar	**servicio** militar zzzz **soldaduska**
GS-irt	zzzz
NBest	**Irteera testua**
1-Best	**servicio** zzzz **zerbitzua**
2-Best	**servicio** zzzz **zerbitzuan**
3-Best	**servicio** zzzz **zerbitzuen**

Taula 18: Errorea: antzeko terminoa baina okerreko adiera.

Adibidez, 18. taulan ikus daitekeen bezala, gaztelania ataleko termino batentzat ("*servicio*"), sarrerako esaldian dagoen terminoa ("*soldaduska*") ezagutzen ez duenez; gaztelania ataleko terminoarentzat ezagutzen duen ordain bat itzuli du "*zerbitzua*", baina kasu honetan ez da egokia.

3. **Termino-bikote gutxiago itzultzea:**

GSarekin konparatuta, sistemak, batzuetan, behar baino termino-bikote gutxiago itzultzen ditu. Ikusi adibidez, 19. taulako adibidea. GSeko irteeran dagoen "*fallecimiento-heriotzari*"termino-bikotea ez du itzuli.

	Sarrera/irteera testua
GS-sar	**c fallecimiento** del **cónyuge** ó **hijos** zzzz c **ezkontidearen** edo **seme** alaben **heriotzari**
GS-irt	**c** gggg **fallecimiento** gggg **cónyuge** gggg **hijos** zzzz **c** gggg **heriotzari** gggg **ezkontidearen** gggg **seme**
NBest	**Irteera testua**
1-Best	**c** gggg **cónyuge** gggg **hijos** zzzz **c ezkontidearen** gggg **seme**
2-Best	**c** gggg **cónyuge** gggg **hijo** zzzz **c ezkontidearen** gggg **seme**
3-Best	**c** gggg **cónyuge** gggg **hijas** zzzz **c ezkontidearen** gggg **seme**

Taula 19: Errorea: GSeko termino-bikote guztiak ez ditu itzuli.

4. **Termino ezezagunak, <unk>**:

Sarrerako esaldietako termino batzuk ezezagunak direnean, hau da, hiztegian topatzen ez dituenean, <unk>tokenaren bidez markatzen ditu. *Fairseq* sistemak erabiltzen duen metodoa da hori, hiztegia gehiegi handitzen denean maiztasun gutxien duten tokenak modu horretan katalogatzen ditu. Hiperparametro baten bidez desaktibatu daiteke aukera hori,

baina bere horretan utzi da (ikusi 20). taulako adibidea: sarreran ezezaguna duen terminoarentzat euskarazko hitz bat proposatu du. Gaztelaniazko "**embriaguez**"terminoarentzat "**mozkorkeria**"terminoa sortu du.

	Sarrera/irteera testua
GS-sar	**f** la **embriaguez** y **toxicomanía** no habitual durante la **jornada** de trabajo zzzz **f <unk>** eta **toxikomania lanaldian** ohikoa ez denean
GS-irt	**f** gggg **embriaguez** gggg **toxicomanía** gggg **jornada** zzzz **f** gggg **<<unk>>** gggg **toxikomania** gggg **lanaldian**
NBest	**Irteera testua**
1-Best	**f** gggg **embriaguez** gggg **jornada** zzzz **f** gggg **mozkorkeria** gggg **lanaldian**
2-Best	**f** gggg **embriaguez** gggg **jornada** zzzz **f** gggg **toxikomania** gggg **lanaldian**
3-Best	**f** gggg **toxicomanía** gggg **jornada** zzzz **f** gggg **mozkorkeria** gggg **lanaldian**

Taula 20: Errorea: termino ezezaguna.

BPE erabiltzen ez den modeloetan <unk>tokena gehiagotan agertzen da; hiztegia gehiago handitzen denez, maiztasun gutxiagoko terminoak sartzeko lekurik ez delako geratzen. BPE erabilita arazo hori konponduko litzateke.

5. **Helburuko proposamenean GSean baino termino-bikote gehiago erauztea:**

Beste batzuetan berriz, GSean ez dauden eta zuzenak izan daitezkeen termino-bikoteak itzultzen ditu sistemak. Ikusi adibidez 21. taulako adibidea.

	Sarrera/irteera testua
GS-sar	se mantiene éste **turno** mientras <unk>con el **turno** de **guardia** zzzz **txanda** honi eutsiko zaio **<unk> <unk>** tartekatzen den bitartean
GS-irt	**turno** gggg **turno** zzzz **txanda** gggg **txanda**
NBest	**Irteera testua**
1-Best	**turno** gggg **turno** gggg **guardia** zzzz **txanda** gggg **txanda** gggg **guardia**
2-Best	**turno** gggg **turno** gggg **guardia** zzzz **txanda** gggg **txanda** gggg **txanda**
3-Best	**turno** gggg **turno** gggg **guardia** zzzz **txanda** gggg **txanda** gggg **txandetan**

Taula 21: Errorea: GSean baino termino-bikote gehiago itzultzen dituen adibide bat.

Adibide horretan "*guardia-guardia*"termino-bikote berria itzuli du. Termino-bikote hori egokiak izan daiteke, baina egokitasun hori neurtzeko sistema automatikorik ez dago.

5 Ondorioak eta etorkizuneko lanak

Atal honetan, master amaierako lan honetako ondorioak (5.1.) eta sortutako sistema hobetzeko egin daitezkeen lanak (5.2.) azalduko dira.

5.1 Ondorioak

Master amaierako lan honetan, sare neuronalak erabiliz terminologia-erauzketa elebiduna landu da. Orain arte, euskara eta gaztelania hizkuntzen terminologia-erauzketa hurbilpena sistema estatistikoetan oinarrituta egin izan da. Lan honetan berriz, terminologia-erauzketa sekuentziatik sekuentziarako ataza modura planteatu da. Terminologia-erauzketaz gain lerrokatzea ere landu da, izan ere, bi hizkuntzetako terminologia-erauzketa egin denez erauzitako termino-bikoteak parekatuta itzultzea erronka bezala planteatu da. Azpimarratzekoa da bi hizkuntzetako erauzketa eta lerrokatze ataza aldi-berean ikasten dituela sare neuronalak. Orain arteko sistemetan, hizkuntza bakoitzeko terminologia aparte erauzten zen eta ondoren lerrokatzea egiten zen. Dena aldi-berean egiteak erroreen propagazioa ekiditen du. Hori egiteko, hizkuntzaren prozesamenduko hainbat atazarentzat baliogarria izan den sekuentziatik sekuentziarako *Transformer* arkitektura erabili da.

Sare neuronala entrenatzeko *Itzulterm* tresnak urteetan gorde dituen erauzketen datubasearekin osatutako corpusa erabili da. Sortutako corpusak 2 milioitik gora lerro zituen. Corpus horren lagin bat bakarrik erabili da: hasiera batean corpus handiagoa erabiltzea pentsatu bazen ere, azkenean 100.000 lerro besterik ez dira erabili, 75.000 lerrotik gora emaitzak okertu egiten zirela ikusi ondoren. *Itzulterm* tresnatik lortutako corpus gordina aberastu eta garbitu egin da lan honetarako.

Entrenamendurako corpusa aberasteaz gain, corpuseko lerroak nahastuta ere egin da saiakera. Saiakera honen motibazioa corpus anitzagoa lortzea zen, sistema alorrarekiko independenteagoa egiteagatik. Hau da, nahastu gabeko corpusa hartuta, gai batzuekiko oso emaitza onak eta beste batzuekin aldiz exkaxak lortuko zirela pentsatu zen. Horretarako, 2 milioiko corpuseko lerroak nahastu eta horietatik lehenengo 100.000 lerroak hartu ziren lagintzat. Saiakera honek emaitza txarrak eman zituen ordea. Lerroak itzulpen-memoria askotatik ekartzeak, corpusak barne hartzen dituen alorrak ere gehiago dira. Ondorioz, corpusaren hiztegia asko handitzeaz gain, hitzen agerpen kopurua askoz baxuagoa izango da. Hitz gutxitan esanda, aldakortasun eta dispertsio handiegia ekartzen zuen memoria txiki horretarako. Ondorioz, sistema ez zen ongi ikasteko gai.

Emaitza horiek ikusita, pausu hori baztertu eta *Itzulterm* tresnatik lortutako corpusetik lehenengo 100.000 lerroak erabili dira corpusak osatzeko. Horrela, corpuseko lerroak alor gutxiagokoak izanik, hiztegian sartutako terminoak ere gaiarekiko aldakortasun txikiagoa izango dute eta gai bakoitzean termino kopuru handiagoa izango du ikasteko.

Konparaziorako erabilitako sistema, *Itzulterm* tresna izan da berriro ere. *Itzulterm* tresnan landutako terminologiatik baldintza batzuk betetzen dituzten termino-bikoteak erabili dira entrenamendurako, terminologia bilduma osoa aldiz ebaluaziorako.

Ebaluaziorako probatutako lehenengo metrika BLEU izan da. Horren arabera *Itzulterm* tresnak 0.258 balioa du, lan honetan landutako sistemak (TEM), aldiz, 0,766. BLEU

metrikaren arabera 50 puntuko aldea atera dio TEM sistemak *Itzulterm*i. BLEU metrika itzulpen-automatikorako funtzionatzen duen sistema da, esaldien gertutasuna neurtzen duelako; BLEUk esaldien luzera, ordena, eta beste hainbat parametro kontuan hartzen ditu.

BLEU metrikan aurkitutako eragozpenak ikusita, terminologia-erauzketa ataza ebaluatzeko beste sistema bat eraiki da: TEB (*Termino Erauzle Balidazioa*). TEB algoritmoak ondoko parametroak hartzen ditu kontuan: sistemak itzultzen dituen termino-bikoteetatik zenbat dauden *gold standard*ean, zenbat ez; zenbat termino-bikote dauden sarrera esaldietan, zenbat ez; zenbat termino dauden sarrerako *gold standard*eko terminoetan, zenbat ez eta abar. Metrika honen arabera *Itzulterm* tresnak 0,767 balioa du, TEM sistemak aldiz 0.869. TEB metrika zehatza izan ez arren, BLEUrekin lortutako emaitzak osatzeko balio du.

Ondorioz, bi metrika probatu eta bietan aurreko sistemaren emaitzak hobetu ditu TEM sistemak. Ondorioz, aurretik eskura zeuden terminologia-erauzketa sistemak erabili ordez sare neuronalen teknologia erabiliz hobetu daitekeela frogatu da.

5.2 Etorkizunean egiteko geratu diren lanak

Sare neuronalak erabiliz terminologia-erauzketa ataza egikaritu daitekeela eta sistema estatistikoek baino emaitza hobeak lor daitezkeela frogatzen duen sistema bat aurkeztu ondoren, sistema hau hobetzeko puntu batzuk planteatuko dira:

1) Saio honetan modeloak corpus txikiekin entrenatu dira (10.000, 25.000, ... 100.000), corpus handiagoekin saiatuta emaitzak asko okertzen zirelako. Berez, sare neuronalen printzipioen arabera, gero eta datu gehiago izan, orduan eta emaitza hobeak lortu beharko lirateke. Kasu honetan ez da horrela gertatu, arrazoiak aztertu beharko lirateke. Ziurrenik corpusa sortzeko erabili den oinarriak eragina izan du honetan, alorraren arabera sailkatu gabeko itzulpen-memorietatik sortu baita corpusa. Alorraren araberako corpusak sortuta, entrenamendu garaian sortutako hiztegia ere alorraren araberakoa izango litzateke. Ondorioz, ondorengo termino erauzketa ere eraginkorragoa izango litzateke.

   Alorraz gain ondokoak ere aztertu beharko lirateke: erabilitako ingurunea egokia den; sortutako corpusa egokia den, gaikako corpusekin emaitza hobeak lortuko liratekeen, hiztegiaren tamaina gehiegi handitzen den eta horrek eragiten duen emaitzak kaskartzea eta abar.

2) Sistema honek hitz bakarreko terminoak erauzten ditu, hitz anitzeko terminoak erauzteko modua bilatu beharko litzateke.

3) Lan honetan garatutako sistema hizkuntzarekiko independentea da, entrenamenduko datuen araberakoa da. Lan honetan *es-eu* hizkuntza-bikotearekin bakarrik egin dira probak, hizkuntzekiko independentzia frogatzeko, beste hizkuntza batzuekin ere probatu beharko litzateke.

4) Sistema ebaluatzeko beste metrika batzuk probatu beharko lirateke, BLEUrekin egindako ebaluazioa lan honetarako egin den TEB algoritmoarekin osatu da. Baina TEB algoritmoak hobetzeko puntu batzuk baditu: adibidez, *gold standard*eko irteeran dauden termino-bikoteetatik sistemak itzultzen ez dituenak ez dira penalizatu.

5) NBest parametroen emaitzak itzuli dira. NBest-ekin lortutako 3 emaitzak aztertu dira eta hoberena aukeratzeko script bat egin ere bai (TEB metrikan oinarrituta), baina beti lehenengo aukera da onena. Ondorioz, ataza honetarako NBest hiperparametroa erabilgarria den aztertu beharko litzateke.

6) Sistema corpus tamaina desberdinekin probatzeaz gain, BPErekin eta BPE gabe probatu da. Ikusi da puntu batetik aurrera BPE aplikatuta emaitza hobetzen bada ere, 50.000 lerroetako corpusetik aurrera berriro emaitza okertzen hasten dela (ikusi 16. irudia). Okertze honen arrazoiak aztertu beharko lirateke: BPEk hiztegiaren tamaina mugatzen duenez, termino gehiago sartzeko aukera du; baina ez du espero zen emaitza eman. Corpus handiagoekin eta beste ingurune batean probatu beharko litzateke, teorian BPEren eragina positiboa izan beharko litzateke.

7) Itzulpen automatikoan entrenamendurako hiztegi bakarra sortzen da: sarrerako testutik lortutakoa eta irteerakotik lortutakoa biak hiztegi bakarrean eta BPE sistema bakarra erabiliz. Lan honetan bi hiztegi erabili dira, corpusaren sarrera testurako bat eta corpusaren irteera testurako beste bat. Horrek eragina izan duela ikusi da. Irteerako testuko hiztegia txikiagoa izanik, kasu batzuetan ez ditu espero zitezkeen terminoak erauzi, behar zituen terminoak hiztegian ez zeudelako. Etorkizuneko lan bat izan daiteke bi hiztegi erabili ordez hiztegi bakarra erabiltzea.

8) Sistema honen hobekuntza posible bat hizkuntzak banaka aztertzea litzateke. Hau da, hizkuntza bakoitzerako erauzle bat sortzea eta ondoren bi sistemen emaitzak parekatzea. Honela BPErekin ustez sortu den arazoa konponduko litzateke, erauzle bakoitzak bere hiztegia izango luke eta nahasmena ere txikitu egingo litzateke. Proposamen honetarako bi hizkuntzetako terminoak lerrokatzen dituen sistema bat bilatu beharko litzateke.

9) Sistemak, kasu batzuetan, terminologia erauzi beharrean sorkuntza egin duela ikusi da 4.2.4. atalean. Itzulpen automatikoa edo antzeko atazetarako terminoen sinonimoak itzultzea positiboa bada ere, terminologia-erauzketaren xedea testuetan dagoen terminologia identifikatzea da, ez sinonimoak topatzea. Sorkuntza lan hori ekiditeko teknikak aztertu beharko lirateke.

Erreferentziak

R. Agerri eta G. Rigau. Robust multilingual named entity recognition with shallow semi-supervised features. *Artificial Intelligence*, 238:63–82, 2016.

Rodrigo Agerri, Iñaki San Vicente, Jon Ander Campos, Ander Barrena, Xabier Saralegi, Aitor Soroa, eta Eneko Agirre. Give your text representation models some love: the case for basque. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May 2020. European Language Resources Association.

Jay Alammar. The illustrated transformer. July 2018. URL `http://jalammar.github.io/illustrated-transformer/`.

Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth Trippe, Juan Gutierrez, eta Krys Kochut. Text summarization techniques: A brief survey. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 8:397–405, 07 2017. doi: 10.14569/IJACSA.2017.081052.

Mouhamed Gaith AYADIa, Riadh BOUSLIMIa, eta Jalel AKAICHIs. A model for multilingual terminology extraction via a medical social network. *Natural Language Engineering*, 2017.

Vit Baisa, Barbora Ulipova, eta Michal Cukr. Sketch engine. *Natural Language Engineering*, 2015.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, eta Robert L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993. URL `https://www.aclweb.org/anthology/J93-2003`.

Yen-Lu Chow eta Richard Schwartz. The N-best algorithm: Efficient procedure for finding top N sentence hypotheses. In *Speech and Natural Language: Proceedings of a Workshop Held at Cape Cod, Massachusetts, October 15-18, 1989*, 1989. URL `https://www.aclweb.org/anthology/H89-2027`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, eta Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.747. URL `http://dx.doi.org/10.18653/v1/2020.acl-main.747`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, eta Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

HAP masterra

Chris Dyer, Victor Chahuneau, eta Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/N13-1073`.

A. Gurrutxaga Elhuyar Fundazioa, X. Saralegi, eta S. Ugartetxea. Itzulterm: gaztelania eta euskarazko corpus paraleloetatik terminologia erauzteko sistema. *Natural Language Engineering*, 2009.

Thierry Etchegoyhen, Eva Martínez García, Andoni Azpeitia, Gorka Labaka, Iñaki Alegría, Itziar Cortes, Amaia Jauregi, Igor Ellakuria, Maite Martin, eta Eusebi Calonge. Neural machine translation of basque. *Conference of the European Association for Machine Translation*, page 139–148, 2018.

Manaal Faruqui eta Shankar Kumar. Multilingual open relation extraction using cross-lingual projection. 2013.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, eta YannÑ. Dauphin. Convolutional sequence to sequence learning. In Doina Precup eta Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/gehring17a.html`.

A. Gurrutxaga, X. Saralegi, S. Ugartetxea, eta I. Alegria. Erauzterm: euskarazko terminoak erauzteko tresna erdiautomatikoa. *Natural Language Engineering*, 2004.

A. Gurrutxaga, X. Saralegi, S. Ugartetxea, eta I. Alegria. Elexbi, a basic tool for bilingual term extraction from spanish-basque parallel corpora. *Natural Language Engineering*, 2006.

Le Ha, Gabriela Fernandez, Ruslan Mitkov, eta Gloria Pastor Corpas. Mutual bilingual terminology extraction. *Natural Language Engineering*, 2008.

Franz Josef Och eta Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, eta Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv*, 1904.01038, 2019.

Lluís Padró, Miquel Collado, Samuel Reese, Marina Looberest, eta Irene Castellon. Freeling 2.1: Five years of open-source language processing tools. 2004.

HAP masterra

Kishore Papineni, Salim Roukos, Todd Ward, eta Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://www.aclweb.org/anthology/P02-1040`.

Eli Pociello, Antton Gurrutxaga, Eneko Agirre, Izaskun Aldezabal, eta German Rigau. Wnterm: Enriching the mcr with a terminological dictionary. 01 2008.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, eta Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. 2019.

Rico Sennrich, Barry Haddow, eta Alexandra Birch. Neural machine translation of rare words with subword units. pages 1715–1725, 01 2016. doi: 10.18653/v1/P16-1162.

Sonit Singh. Natural language processing for information extraction. *arXiv*, 1807.02383, 2018.

Ilya Sutskever, Oriol Vinyals, eta Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanÑ Gomez, Ł ukasz Kaiser, eta Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, eta R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf`.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, eta et al. Clinical information extraction applications: A literature review. 77(34–49), 2018.

Thomas Zenkel, Joern Wuebker, eta John DeNero. End-to-end neural word alignment outperforms giza++, 2020.