EUSKAL HERRIKO UNIBERTSITATEA

Lengoaia eta Sistema Informatikoak

Doktorego-tesia

Datuen Ustiapena Itzulpen Automatikorako

Andoni Azpeitia Zaldua

Donostia, 2021eko abendua

EUSKAL HERRIKO UNIBERTSITATEA

Lengoaia eta Sistema Informatikoak

Datuen Ustiapena Itzulpen Automatikorako

Andoni Azpeitia Zalduak Eneko Agirre Bengoaren eta Arantza del Pozo Echezarretaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Doktore titulua eskuratzeko aurkeztua.

Donostia, 2021eko abendua

Laburpena

Datuetan oinarritutako itzulpen automatikoa, lehendabizi itzulpen automatiko estatistikoarekin (SMT) eta gero itzulpen automatiko neuronalarekin (NMT), azken urteotan gailendutako paradigma da. Sistema hauek corpus paraleloak erabiliz (testu berbera bi hizkuntza ezberdinetan lantzen duten datu bildumak) elikatzen dira entrenamendu prozesu batean. Itzulpen automatikoaren abantaila nagusia itzulpen berriak egin ahal izateko jakintza automatikoki erauzten dela da, baina tamalez, jakintza orokortzeko ahalmena entrenamendurako corpuseko adibideek mugatzen dute.

Tesi honen helburu nagusia corpus paraleloen kalitatea hobetzea da hiru alderdi landuz: corpus tamaina handituz, corpusen datuak domeinura egokituz eta datu multzo zaratatsuak iragaziz. Kalitatezko corpusak sortzeko lau ikerketa lerrotan egindako lanak aurkezten dira: dokumentuen lerrokatzea, esaldien lerrokatzea, datuen aukeraketa eta esaldi paraleloen iragazpena. Ikerketa guztiak enpresek finantzatutako proiektuen testuinguruan egin dira, kalitateaz gain, eramangarritasun helburua ere oso kontutan eduki delarik tesian zehar.

Dokumentuak lerrokatzeko ikerketa lerroan, dokumentuen konparagarritasuna neurtzeko metrika bat proposatzen da. Metrika hau oso eraginkorra izan ezezik guztiz eramangarria da, inolako eredurik entrenatu behar izan gabe testuko terminoak konparatzen baititu hizkuntzarekiko independenteak diren metodoak erabiliz. Esaldien lerrokatzeari dagokionez, dokumentuak lerrokatzeko proposatutako antzekotasun metrika egokitu da, esaldietan dokumentuetan baino informazio gutxiago dagoela kontutan hartuz. Datuen aukeraketan, ugariagoak diren corpusetan testu multzo baliagarriak aukeratzeko testuko terminoen maiztasun erlatiboaren erabilera aztertu da. Aurreko ikerlerroetan proposaturiko metodoen antzera, emaitza lehiakorrak lortu dira eramangarritasuna alde batera utzi gabe. Azkenik, esaldi paraleloen iragazpena esaldien lerrokatzearen kasu berezi bat bezala landu da, bi esaldiren arteko antzekotasun maila iragazpena egiteko ustiatuz.

Egindako ikerketen baliagarritasuna aztertzeko esperimentu ugari egin dira artearen egoerako beste sistemekin konparaketak eginez eskuragarri dauden corpus libreak erabiliz, kasu askotan artearen egoera hobetzea lortu dela-

rik. Esaldi konparagarrien lerrokatzearen kasuan, nazioarteko ataza batean emaitza onenak lortu ziren bi urtez jarraian. Azkenik, garatutako esaldien lerrokatze metodoa erabiliz, albisteen domeinuan euskaraz eta gaztelaniaz idatzitako itzulpenak biltzen dituen ia 600.000 esaldi paraleloko corpus lerrokatu bat sortu eta komunitatearekin elkarbanatu da.

Abstract

Data-driven machine translation, first based on statistical machine translation (SMT) and later based on neural machine translation (NMT), has become the dominant approach in recent years. These type of systems are fed with parallel corpora (data collections with the same text written in two different languages) in a training process. The main advantage of machine translation is the ability to automatically extract knowledge from data, but in the same way, its capability to generalise knowledge is also conditioned by the examples observed in the training corpus.

The main goal of this thesis is to improve the quality of parallel corpora working on three different aspects: increasing corpora size, adapting corpora to the target domain and filtering noisy data. For this purpose, investigations carried out in the following four research fields are presented: document alignment, sentence alignment, data selection and parallel sentence filtering. Because all investigations have been performed in the context of real projects, the portability of the methods explored has been a pursued objective, in addition to quality improvement, throughout the thesis.

In the document alignment research line, a novel document similarity metric has been proposed. In addition to being effective, this metric does not require model training and it is language independent. Regarding sentence alignment, the similarity method developed for document alignment has been adapted taking into account that sentences contain less information than documents. For data selection, relative term frequencies have been explored to select valuable bitexts from more abundant corpora, also achieving high portability and competitive results. Finally, parallel sentence filtering has been treated as a particular case of sentence alignment, exploiting the similarity between sentences to filter out harmful data.

To test the usefulness of the proposed methods a wide variety of evaluations have been carried out against other state-of-the-art systems using corpora under free licence, improving the state-of-the-art in many cases. Regarding sentence alignment, best results were obtained in an international shared task for two consecutive years. Finally, a dataset composed by almost 600.000 parallel sentences with translations written in Basque and Spanish in the news

domain has been created using the developed sentence alignment methods and shared with the community.

Eskerrak

Tesiaren egile bezala pertsona bakarra azaltzen da, baina benetan, modu batean ala bestean, jende gehiagoren elkarlanaren emaitza da. Lehenik eta behin, eskerrak eman nahi nizkioke nire tutoreei, Eneko eta Arantzari, haien esperientziarik gabe zeharo galduta egongo nintzatekeelako. Inoiz ez didate ezetz esan, baina bai azaldu dizkidatela erabaki bakoitzaren alde onak eta ez hain onak, oraindik gogoan ditut izandako eztabaidak!

Beste pertsona garrantzitsu bat Eva da. Zure ideiak oso baliagarriak izan dira eta beti hitzegin duzu zintzo, baina alderdi zientifikotik at, oraindik eta garrantzitsuagoa izan da alderdi emozionala. Momentu onenetan oso ondo pasa dugu 😚 , baina batez ere momentu txarrenetan beti egon zara ni laguntzeko prest.

Ez dakit lagunak eskertzea komeni den..., tentazio ugariren iturri izan dira eta... Bromak aparte, eskerrik asko ni aguantatzeagatik, ospatuko dugu elkarrekin 🍺 !

Azkenik, gurasoak eskertu nahi nituzke. Zuei esker lana errazagoa izan da benetako altxorra eman didazuelako, denbora eta maitasuna. Besarkada handi handi bat ❤️ !

Eskerrik asko guztioi!

Aurkibidea

1. KAPITULUA

Sarrera

Corpus paraleloak, hau da, bi hizkuntza ezberdinetan testu berbera lantzen duten datu bildumak, datuetan oinarritutako itzulpen automatikorako funtsezkoak dira, corpusean baitago ikasketa automatikoaren bidez erauzi beharreko ezagutza. Corpusaren eragina aztertu ahal izateko, 90ko hamarkadan azaldutako itzulpen automatiko estatistikoaren inguruan (SMT) (Brown *et al.*, 1990) eta azken urteotan gailendutako itzulpen automatiko neuronalaren inguruan (NMT) (Bahdanau *et al.*, 2014) zenbait ikerketa egin dira (Imam *et al.*, 2011; Khayrallah eta Koehn, 2018), eta ondorioa berbera da: corpusaren kalitate eta tamainak itzulpenen kalitatea baldintzatzen du.

Hizkuntza naturalaren prozesamenduko beste zenbait alorretan ere corpus paraleloekiko menpekotasuna dago, esate baterako lexikoi eleanitzen sorkuntza automatikoan (Rapp, 1995). Hortaz, kalitatezko hizkuntza baliabide paraleloak sortzea berezko ikerketa lerroa da, eta hizkuntza ezberdinetan idatzitako informazio multzoak lerrokatu ahal izateko teknika ugari sortu dira (Tiedemann, 2011).

Zoritxarrez, nahiz eta urteetan zehar gero eta corpus gehiago sortu (Tiedemann, 2012), hizkuntza pare, estilo, edota domeinuaren arabera corpus paraleloak baliabide urriak dira. Hizkuntza pare gutxirako aurkitu daitezke tamaina eta kalitate egokiko corpus paraleloak, eta kasu askotan domeinu jakin batzuetarako bakarrik.

Arazo honi aurre egiteko aukera bat Interneten hazkundeari esker, corpus paraleloak baino ugariagoak diren hizkuntza baliabideak ustiatzea da, hau

da, corpus konparagarriak. Corpus konparagarriak nahiz eta itzulpen zuzenak ez izan, antzeko edo erlazionatutako informazioa partekatzen dituzten datu multzo eleanitzak dira (Sharoff *et al.*, 2013; Morin *et al.*, 2015).

Corpus konparagarrien bidez corpus paraleloak automatikoki sortu ahal izateko aukera asko daude, baina nagusienak honako bi hauek dira:

1. Corpusetako esaldiak hizkuntza bakoitzeko biltzea eta esaldiak lerrokatzea (produktu kartesiarraren bidez konbinazio guztiak aztertuz edo aurreprozesaketa baten bidez bilaketa espazioa murriztuz).

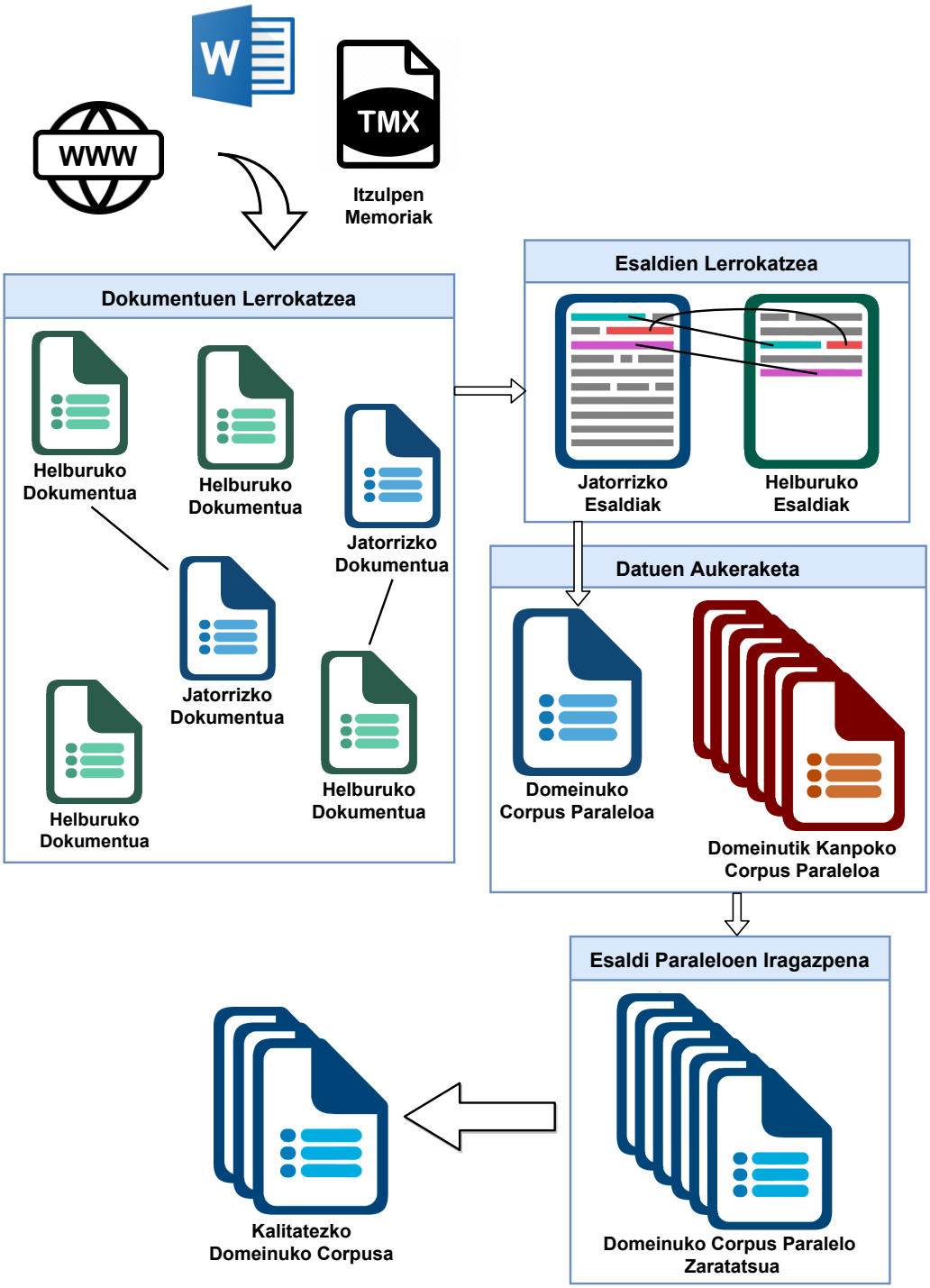
2. Normalean dokumentu bakoitzak alor jakin bat jorratzen du, hortaz, lerrokatze prozesua bi pausutan egitea: lehendabizi dokumentuka eta gero esaldika.

Bigarren aukera lehenengoa baino irtenbide osoagoa da, eta gainera, malgutasun handiagoa eskaintzen du. Horregatik, bigarren bideari helduz, tesi honetan kalitatezko corpus paraleloak sortzen laguntzeko zenbait ikerketa lerro aurkezten dira. Hauexek dira ikerketa lerroak:

- Dokumentuen lerrokatzea.
- Esaldien lerrokatzea.
- Datuen aukeraketa.
- Esaldi paraleloen iragazpena.

Tesi honetan ikertutako metodoei esker informazio iturri ezberdinak erabili daitezke kalitatezko corpus paraleloak sortzeko. Metodo hauek modu ezberdinetan konbinatzeko aukera dago eskuragarri dauden dokumentuen arabera, baina oro har, 1.1. irudian azaltzen den bezala erabili daitezke: lehendabizi datuak dokumentu mailan eta gero esaldi mailan lerrokatu, datuen aukeraketaren bidez lortutako datuak beste domeinuko corpusekin osatu, eta azkenik, datu zaratatsuenak iragazi.

1.1 irudia: Kalitatezko Domeinuko Corpusa.



Tesiko atalak honela antolatzen dira. Lehendabizi, sarrera honetan tesiaren gaia motibatu, tesiaren testuingurua azaldu, eta ikerketen oinarrizko helburu eta ekarpenak azaltzen dira. Jarraian, 2. kapituluan artearen egoera aztertzen da. Hurrengo 3, 4, 5 eta 6. kapituluetan tesiko ikerketen deskribapena egiten da. 7. kapituluan egindako ikerketak erabili direneko aplikazioak deskribatzen dira. Azkenik, 8. kapituluan azken ondorioak laburbiltzen dira.

1.1. Motibazioa

Itzulpen zerbitzu profesionalak oso garrantzitsuak dira merkatu global eleanitzaren garapen ekonomikoan, non oztopo linguistikoek merkatua zatitzen duten. Gainera, Euskal Autonomia Erkidego edo Europar Batasuna bezalako merkatuetan, eleaniztasuna berezko ezaugarria denez, arazoa areagotu egiten da[1].

Oztopo linguistikoak direla eta sortutako zailtasunei mundu gero eta globalizatuago batean itzuli beharreko informazioaren hazkunde esponentziala gehitu behar zaie, itzulpen eta egokitzapen linguistikorako prozesuak oso konplexu eta astunak baitira. Nahiz eta azken urteetan itzulpen automatikoaren kalitatea hobetu, adituen ezagutzaren beharra dago testuak hizkuntza eta kultura bakoitzaren berezitasunetara egokitzeko.

Itzulpen eleanitzerako teknologia osagarriak alor askotako behar dira: ikusentzunezko edukietan azpitituluak ipintzeko, informazio eta komunikazioaren industriarako edukiak sortzeko, edo dokumentu teknikoen itzulpena nazioarteko merkatu berriak atzemateko.

Esan bezala, azken bi hamarkadetan itzulpen automatikoaren kalitateak gora egin du eta itzulpen prozesuak asko hobetu dira. 2010. urtera arte, eredu estatistikoetan oinarritutako itzulpen automatikoa zen gehien erabilitako teknologia. Hala ere, SMT ereduek badituzte beren mugak, eta adituek orain ere postedizio lanak egin behar dituzte akatsak zuzentzeko. Sare neuronalen inguruan egindako ikerketetatik sare neuronaletan oinarritutako itzulpen

[1] Ikus 2008ko hizkuntzaren berdintasunaren txostena: `https://www.europarl.europa.eu/news/en/headlines/economy/20180621STO06335/multilingualism-online-language-barriers-are-a-major-challenge`

automatikoa gailendu da, eta itzulpenen kalitatea nabarmen hobetzea lortu da. Hasieran, NMT teknologia arlo zientifikoetan erabiltzen hasi zen, baina gaur egun arlo komertzialean ere guztiz finkatuta dago, eta besteak beste, Google[2], Microsoft[3] eta Systranek[4] NMT zerbitzuak eskaintzen dituzte.

Nahiz eta itzulpen automatikoaren teknologia hobetu, oraindik bide luzea dago itzultzaileen parte-hartzerik gabe itzulpen zuzenak modu egonkor batean lortu ahal izateko. Itzulpen automatikoan ereduek entrenamendu corpus paraleloen beharra dute uneko hizkuntza eta domeinuaren ezaugarri guztiak erauzi eta ikasi ahal izateko. Chomskyren esanetan, gizakiok perpausa kopuru infinitua sortu eta ulertzeko gaitasuna dugu (Escutia, 2013), baina tamalez, corpusak finituak dira.

Arazo honi aurre egiteko asmoz, tesi honen helburu nagusia automatikoki kalitatezko corpus paraleloak sortzen laguntzeko tresna berriak ikertu eta garatzea da. Lehen aipatu dugu itzuli beharreko informazioa gero eta handiago dela, baina era berean, gero eta informazio gehiago dago eskura. Domeinurako baliagarria den informazioa bilatuz eta modu egokian antolatuz, itzulpenerako egokienak diren esaldi pareak identifikatuz eta zaratatsuenak diren datuak iragaziz automatikoki kalitatezko corpus paraleloak sortu daitezke, eta corpusen kalitate eta tamaina handitzea lortzen bada, itzulpenen kalitateak ere gora egingo du.

1.2. Testuingurua

Tesi honetako ikerketak sare neuronaletan oinarritutako metodoen hazkundean kokatzen dira. Nahiz eta orokorrean sare neuronaletan oinarritako teknikak artearen egoeraren mugak bultzatu[5] (Young *et al.*, 2018), ikasketa auto-

[2] https://slator.com/technology/nearly-indistinguishable-from-human-translation-google-claims-breakthrough/

[3] https://www.microsoft.com/es-es/translator/business/machine-translation/

[4] https://www.systransoft.com/download/press-releases/systran-pr-purely-neural-mt-engine-a-revolution-for-the-machine-translation-market-2016-08-30.pdf

[5] https://paperswithcode.com/sota/named-entity-recognition-ner-on-conll-2003

matikorako ohiko teknikak baino menpekotasun handiagoa du etiketatutako datu kantitatean, eta zenbait testuingurutan, hizkuntza pare, domeinu eta corpusen tamainaren arabera, SMT ereduak egokiagoak izan daitezke (Dowling *et al.*, 2018; Mahata *et al.*, 2018; Veliz *et al.*, 2021).

Hala ere, tesiaren muina datuen ustiaketa da, eta bai SMT eta baita NMT ereduak, biak datuetan oinarritutako teknologiak izanik, entrenamendurako corpusetik erauzi eta ikasten dituzte uneko hizkuntza eta domeinuaren ezaugarriak. Beraz, tesi honetako ikerketak guztiz baliozkoak dira bi sistementzako.

Noski, artearen egoerak aurrera jarraitu du tesiko ikerketak bukatu ostean, eta aurkeztutako teknika batzuk gaindituak izan dira. Hala eta guztiz ere, ikerketak egindako unean sare neuronaletan oinarritutako metodoak azaldu ziren, baina hurrengo kapituluetako ebaluazioetan ikus daitekeen bezala, lortutako emaitzak lehiakorrak izan ezezik, zenbait kasuetan hobeak izan ziren.

1.3. Ikerketaren Helburuak

Helburu nagusia kalitatezko corpus paraleloak automatikoki sortu ahal izateko metodo berriak ikertu eta garatzea da. Helburu hau lortu ahal izateko bide bat dagoeneko eskura dauden corpus konparagarriak ustiatzea da, behar den informazio egokia bilatuz, antolatuz eta prozesatuz. Nahiz eta metodo batek emaitza onenak izan, ezingo da erabili metodoa astunegia bada, edota bere konfigurazioa konplexuegia bada, horregatik, tesi honetako ikerketek zenbait ezaugarri bete behar dituzte:

- Kalitatea. Helburu nagusia kalitatezko corpus paraleloak lortzea da, horregatik, ikerketa guztiak lehiakorrak izan behar dira artearen egoerarekiko. Helburu hau betetzen den egiaztatu ahal izateko, metodo guztiak artearen egoerarekin konparatzen dira. Are gehiago, garatutako zenbait metodorekin ataza ofizialetan parte hartu da lortutako emaitzak balioztatzeko asmoz.

- Konputazionalki eraginkorra. Aipatu bezala, metodo onenak ez du ezertarako balio tamaina egokiko corpusak zentzuzko denbora tarte ba-

tean prozesatzeko gai ez bada. Kalitatezko corpusak sortu ahal izateko datu multzo handiak prozesatu behar dira, beraz, helburu hau oso garrantzitsua da.

- Hizkuntza eta domeinuarekiko eramangarritasuna. Garatutako metodoak benetako egoeratan aplikatzea posible izan behar da, eta horretarako zenbait baldintza bete behar dira. Alde batetik, garatutako metodoak hizkuntzarekiko independenteak izan behar dira, bestela, metodoek eskaini ditzaketen onurak oso mugatuak egongo dira. Bestetik, metodoek mota ezberdinetako informazioa prozesatzeko gai izan behar dira, ez baita berdin corpus paralelo edo konparagarriak aztertzea, edota corpusak dokumentu edo esaldi mailan aztertzea. Azkenik, dokumentuak prozesatzeko testu edukia bakarrik hartu behar da kontutan, inolako metadatu edo egitura berezirik kontutan izan gabe. Adibidez, WIKIPEDIA-ko dokumentuek etiketa eta esteka ugari dituzte edukia aberasteko, baina metadatu berezi horiek prozesatzen dituen metodo bat garatuko balitz, WIKIPEDIA-tik kanpo metodoaren eraginkortasuna okerragoa izango litzateke. 7. kapituluan garatutako metodoak erabili direneko zenbait proiektu deskribatzen dira.

- Erabilterraza. Metodoak egoera bakoitzaren berezitasunetara egokitzeko konfigurazioa ahalik eta errazena izan behar da. Tesiko kapituluetan zehar lortutako emaitzak zein konfiguraziorekin lortu diren azaltzen da, eta metodo bakoitza hizkuntza, domeinu, eta tamaina ezberdineko corpusetan ebaluatu da.

1.4. Ekarpen Nagusiak

Atal honetan tesiaren ekarpen nagusiak laburbiltzen dira. Honako hauek dira:

- Dokumentuen lerrokatzea. Dokumentuak lerrokatzeko DOCAL sistema garatu da. Bi tekniketan oinarritzen da: itzulpen lexikoetan eta Jaccard koefizientean (Jaccard, 1901). 3. kapituluan azaltzen dira xehetasun guztiak.

Egindako argitalpenak honako hauek dira:

Etchegoyhen T. and Azpeitia A. A portable method for parallel and comparable document alignment. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 243-255, 2016.

Azpeitia A. and Etchegoyhen T. Docal-vicomtech's participation in the wmt16 shared task on bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 666-671, 2016.

Azpeitia A. and Etchegoyhen T. Efficient document alignment across scenarios. *Machine Translation*, 33(3):205-237, 2019.

- Esaldien lerrokatzea. DOCAL-en antzera, esaldiak lerrokatzeko STACC sistema garatu da. DOCAL sisteman zenbait egokitzapen egin dira lerrokatzeak esaldien berezitasunetara egokitzeko. 4. kapituluan deskribatzen da.

  Egindako argitalpenak honako hauek dira:

  Etchegoyhen T. and Azpeitia A. Set-theoretic alignment for comparable corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009-2018, 2016.

  Azpeitia A., Etchegoyhen T., and Martínez E. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41-45, 2017.

  Azpeitia A., Etchegoyhen T., and Martínez E. Extracting parallel sentences from comparable corpora with stacc variants. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*, pages 48-52, 2018.

- Datuen aukeraketa. Informazioa lerrokatuz lehen corpus paralelo bat sortu ostean, uneko domeinurako datu gehiago lortzeko aukera dago. Horretarako, domeinutik kanpoko corpus handiagoetan uneko domeinuko datuak biltzeko RFR metodoa garatu da. Terminoek bi domeinue-

tan (unekoa eta kanpokoa) duten maiztasun erlatiboa erabiltzen da. 5. kapituluan aurkezten da.

Egindako argitalpena honako hau da:

Etchegoyhen T., Azpeitia A., and Martínez E. Exploiting relative frequencies for data selection. In *Proceedings of MT Summit XVI, Volume 1: Research Track*, pages 170-184, 2017.

- Esaldi paraleloen iragazpena. SMT eta NMT ereduetarako kaltegarria da datu zaratatsuak egotea entrenamendurako corpusean (Khadivi eta Ney, 2005; Khayrallah eta Koehn, 2018). Arazo honi aurre egiteko STACC-en antzekotasun metrikan oinarritutako zenbait ikerketa aurkezten dira 6. kapituluan.

  Egindako argitalpena honako hau da:

  Azpeitia A., Etchegoyhen T., and Martínez E. Stacc, oov density and n-gram saturation: Vicomtech's participation in the wmt 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 860–866, 2018.

- EITB corpusa. Albisteen domeinuan esaldi mailan lerrokatutako gaztelaniaz eta euskaraz idatzitako albisteak biltzen dituen corpusa da. Corpusa sortzeko 4. kapituluan garatutako metodoak erabili dira. Ia 600.000 esaldi lerrokatu dira. 4.2. atalean azaltzen dira corpusaren xehetasunak.

  Honako argitalpen honetan deskribatzen da egindako lana:

  Etchegoyhen T., Azpeitia A., and Pérez N.. Exploiting a large strongly comparable corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

  Eta EITB corpusa honako estekaren bidez lortu daiteke: `http://metashare.elda.org/repository/search/?q=eitb+documents`

2. KAPITULUA

Artearen Egoera

Azken bi hamarkadetan datuetan oinarritutako itzulpen automatikoa izan da hurbilpen nagusia, lehendabizi itzulpen automatiko estatistikoa (SMT) (Brown *et al.*, 1990), eta gero itzulpen automatiko neuronala (NMT) (Bahdanau *et al.*, 2014). Teknika hauek esaldi paraleloen bilduma handiak behar dituzte itzulpenak behar bezala modelatu ahal izateko, baina askotan, hizkuntza parearen eta domeinu zehatzaren arabera, datu iturri horiek ez daude eskuragarri.

Gaur egun, Internet eta bestelako informazio iturriei esker datu multzo ugari daude eskuragai. Corpus elebakarrak aurki ditzakegu, eta baita corpus konparagarriak ere, hau da, itzulpen zuzenak ez diren baina antzeko edo erlazionatutako informazioa partekatzen duten datu multzo eleanitzak (Sharoff *et al.*, 2013; Morin *et al.*, 2015).

Corpus elebakar eta konparagarrietatik abiatuta corpus paraleloak automatikoki sortzeko zenbait teknika erabili daitezke hala nola dokumentu eta esaldien lerrokapena, datuen aukeraketa edo esaldi paraleloen iragazpena, besteak beste. Tesi honen helburu nagusia kalitatezko itzulpenak lortzeko corpus paraleloen kalitatea hobetzea da, eta aukeratutako bidea aipatutako teknikak hobetzea da. Hurrengo azpiataletan teknika hauen artearen egoera deskribatzen da.

2.1. Dokumentuen Lerrokapena

Urteetan zehar dokumentuak lerrokatzeko teknika ugari erabili izan dira dokumentuen paralelogarritasun eta konparagarritasunaren arabera.

Guztiz paraleloak diren dokumentuak lerrokatzeko, fitxategien izenak konparatzen dituzten teknikak izan daitezke eraginkortasun handiena dutenak, dokumentuen edukia prozesatu beharrik ez dagoelako. Teknika hauek erabili ahal izateko fitxategien izenek eredu jakin bat jarraitu behar dute hizkuntza ezberdinen artean, baina askotan baldintza hau ez da betetzen, ezta biltegi profesionaletan ere (Tiedemann, 2011). Horregatik, ohikoa da fitxategien izenak konparatzen dituzten teknikak dokumentuen edukia aztertzen duten beste teknika osagarriekin batera erabiltzea (Chen *et al.*, 2004). Bestalde, dokumentuen metadatuen erabilgarritasuna sakonki aztertu zuten Resni eta Smithek (Resnik eta Smith, 2003), URL propietateak eta etiketen egiturak ustiatu zituztelarik Internetetik lortutako HTML dokumentuak lerrokatzeko. Ikuspegi honek ere dokumentuen edukia ez aztertzearen abantaila du, baina dokumentuek egitura jakin bat jarraitu behar dute teknika mota hauek eraginkorrak izan daitezen. PTMINER sistemak ere metadatuen bidez lerrokatzen ditu dokumentuak, URL propietate, dokumentu tamaina eta hizkuntza identifikatzaileen erabilera konbinatuz (Chen eta Nie, 2000).

HAPAX-etan oinarritutako metodoak ere erabili dira dokumentu paraleloak lerrokatzeko (Enright eta Kondrak, 2007). Metodo hauen funtsa dokumentuetan elkarbanatutako termino bakarren kopurua erabiltzea da lerrokapenak egiteko. ADA-BOOST sailkatzaileen inguruan ikerketak egin dira dokumentuetatik ezaugarriak erauzteko eta lerrokatze ereduak entrenatzeko (Patry eta Langlais, 2005): dokumentuen tamaina, entitateak, puntuazio ikurrak, etab. Geroago, PARADOCS sistema aurkeztu zen (Patry eta Langlais, 2011a). PARADOCS honako modulu hauek osatzen dute: (1) HAPAX eta zenbakizko entitateetan oinarritutako informazio berreskurapen sistema bat; (2) dokumentuen hiru ezaugarrirekin entrenatutako sailkatzaile bat (izen-entitate kopurua, edizio distantzia eta edizio distantzian oinarritutako balio bitar bat); (3) dokumentu bat beste batekin bakarrik lerrokatuko dela bermatzen duen iragazpen modulu bat.

PARALLEL TEXT IDENTIFICATION sisteman fitxategien izenetan oinarritutako modulu bat eta terminoen maiztasunekin kalkulatutako bektoreetan oinarritutako beste modulu bat konbinatzen dira (Chen *et al.*, 2004). Bigarren moduluaren helburua dokumentuen antzekotasun semantikoa neurtzea da. Bestalde, BITS sistema (Ma eta Liberman, 1999) Internet arakatzeko gauza da dokumentu paraleloak lortzeko. Lerrokapenak egiteko, jatorrizko hizkuntzako dokumentuetako termino kopuruarekiko itzulitako terminoen proportzioa kalkulatzen du.

Dokumentu konparagarriak bereziki lerrokatzeko ere ikerketa ugari egin dira. Fung eta Cheungek, esate baterako, esaldi paraleloak erauzi zituzten hizkuntza ezberdinetako dokumentu elebakarretatik (Fung eta Cheung, 2004). Dokumentuak lerrokatzeko, lehendabizi TF-IDF (Salton eta McGill, 1986) teknikak erabiliz dokumentu bektoreak sortzen dira, eta gero, dokumentuen konparagarritasuna kosinu antzekotasunaren bidez neurtzen da. Erabilitako beste teknika bat jatorrizko dokumentuak itzuli eta da gero partekatutako n-gramak aztertuz lerrokapenak egitean datza (Uszkoreit *et al.*, 2010). EMACC sisteman (Ion *et al.*, 2011), *Expectation Maximization* (EM) (Moon, 1996) algoritmoa erabiliz lerrokatzen dira dokumentuak. Funtsean, termino mailan IBM ereduen bidez (Brown *et al.*, 1993) egiten den lerrokapenaren antzekoa da, automatikoki sortutako lexikoi elebidunak erabiltzen direlarik EM algoritmoa dokumentuetan aplikatzeko. DICTMETRIC sistema azkarragoa da (Ion, 2012), non itzulpen lexikoak, stemming teknikak, WORDNET (Miller, 1998) eta kosinu antzekotasuna konbinatzen diren.

Hiztegi elebidunak erabiliz itzulpena duten terminoen proportzioa kalkulatuz konparagarritasun maila kalkulatu daiteke (Li eta Gaussier, 2013). LINA sistema (Morin *et al.*, 2015) berriz HAPAX-etan oinarritzen da (Enright eta Kondrak, 2007), eta dokumentu bat lerrokapen batean baina gehiagotan agertzen bada, honako bi metodoekin aukeratzen da egokiena: *pigeonhole* arrazoiketaren bidez probabilitate handieneko lerrokapenak aukeratzen dira, eta WIKIPEDIA-ren hizkuntza arteko estekak erabiliz lerrokatutako bi dokumentuek hirugarren hizkuntzako dokumentu berberarekin lotzen badira, lerrokapen hori ontzat ematen da. AUT sisteman berriz (Zafarian *et al.*, 2015), lau teknika nagusi konbinatzen dira: dokumentuen bektoreak, TOPIC MO-

DELLING, izen-entitateen identifikazioa, eta terminoen lerrokapena itzulpen automatikoaren bidez.

WMT 2016-ko dokumentuen lerrokatze atazan[1] (Buck eta Koehn, 2016a), HTML dokumentuak lerrokatzeko zenbait teknika aurkeztu ziren. Dokumentuen lerrokatzean egindako ikerketekin ataza honetan parte hartu genuen (ikus 3. kapitulua). Horregatik, une horretako artearen egoera azaltzearren, atazako partehartzaile nagusienetakoen metodoak laburbiltzen dira.

Gomes eta Lopesek SMT sistema baten bidez prozesatutako segmentuen itzulpentauletan oinarritutako sistema bat aurkeztu zuten (Gomes eta Lopes, 2016). Partekatutako segmentuen proportzioa eta URL-ak erabiltzen dituzte lerrokapenak egiteko. YODA sisteman (Dara eta Lin, 2016), WMT 2016-ko itzulpenak erabiltzen dira partekatutako n-gramak bilatzeko, eta dokumentuen tamainan oinarritutako heuristiko bat ere erabiltzen da. Buck eta Koehnek ere WMT 2016-ko itzulpenak erabiltzen dituzte (Buck eta Koehn, 2016b). Lehendabizi, partekatutako n-grametatik TF-IDF bektoreak sortzen dira, eta gero, kosinu antzekotasunaren bidez lerrokapenak neurtzen dira. Germannek, terminoak errepresentatzen ditu espazio bektore batera proiektatuz, dokumentuak errepresentazio semantiko horren bidez indexatzen ditu, eta kosinu antzekotasunaren bidez eta URL-ak konparatuz dokumentuak lerrokatzen ditu Germann (2017).

WMT 2016-an aurkeztutako zenbait lanetan dokumentuen egitura erabiltzen da lerrokatzeak egiteko. Papavassiliou eta besteen lanean estekak, URL-ak eta HTML etiketak erabiltzen dira (Papavassiliou *et al.*, 2016). BITEXTOR (Esplà-Gomis eta Forcada, 2009) sistemaren bertsio berri bat ere aurkezten zen URL antzekotasuna, partekatutako estekak, eta BAG OF WORDS teknikak erabiltzen dituena (Esplà-Gomis *et al.*, 2016).

Zenbait metodo deskribatzen dira Le eta besteen ikerlanetan (Le *et al.*, 2016): termino berberen posizioan oinarritutako teknikak, bi termino batera agertzeko probabilitatea neurtzen duten teknikak, eta URL-en Levenshtein distantzia (Levenshtein, 1966). Analisi morfologikoa erabiltzen duen sistema bat ere aurkeztu zen (Medveď *et al.*, 2016). Bertan, analisia egin ostean,

[1] `http://www.statmt.org/wmt16/bilingual-task.html`

hitz gako eta itzulpen lexikoen gainean aplikatutako TF-IDF teknikak erabiltzen dira. Jakubina eta Langlaisek APACHE LUCENE-ren[2] bidez URL eta testu edukia indexatzen dute, eta galdeketa elebakar eta elebidunekin antzekoenak diren dokumentuak berreskuratzen dituzte (Jakubina eta Langlais, 2016). Azkenik, GALE-CHURCH lerrokatze algoritmoa eta hiztegi elebidunak erabiltzen dituen metodo bat aurkeztu zen (Mahata *et al.*, 2016).

Jaccard antzekotasuna, tesi honetan aurkeztutako lanaren oinarrizko osagaia, oso metrika erabilia da informazioaren berreskuratze eta testuaren laburbiltze metodoetan, eta baita antzekotasun semantikoa neurtzeko (Pilehvar *et al.*, 2013). Dokumentuen konparagarritasun maila neurtzeko ere erabili izan da (Paramita *et al.*, 2013). Lehendabizi, jatorrizko hizkuntzako esaldiak itzultzen dira, gero, zaratatsuak diren esaldiak erauzten dira. Azkenik, Jaccard metrikarekin antzekotasun handieneko dokumentua aukeratzen da atalase bat gainditzen badu.

Tesi honetan aurkeztutako metodoen ostean, dokumentuen lerrokapenaren inguruko ikerketek aurrera jarraitu dute, eta sare neuronalak erabiliz dokumentuen egitura errepresentatzeko metodoak aurkeztu dira. Jiang eta besteek, tamaina handiko dokumentuak lerrokatu ahal izateko, hainbat geruzako arreta mekanismoan oinarritutako sare neuronal errepikariekin (SMASH RNN), terminoen errepresentazioa ezezik, dokumentuen egitura semantikoa ere atzematen dute (Jiang *et al.*, 2019). Siamdar egitura batekin, dokumentu pareen errepresentazioak erauzi eta antzekotasun probabilitatea kalkulatzeko gai dira. Yang eta besteek, tamaina handiko dokumentuak hainbat geruzako transformer sare neuronal hierarkiko Siamdarren (SMITH) bidez errepresentatzen dituzte (Yang *et al.*, 2020). SMITH eredua esaldi mailan maskaratutako corpus batean aurreentrenatzen dute. Azkenik, Zhou eta besteek maila ezberdinetako egiturak konparatu ahal izateko metodo bat aurkeztu zuen (Zhou *et al.*, 2020). Horrela, dokumentuak beste dokumentuekin lerrokatu daitezke, eta baita dokumentuak esaldiekin. Lerrokapen horiek egiteko aurretik entrenatutako arreta kodetzaile hierarkikoez (HAN) hornitzen den eredu bat proposatzen dute. Dokumentuak lerrokatzeko orduan, aurreko bi metodoek emaitza hobeak lortzen dituzte, baina Zhouren ereduaren abantaila nagusia

[2] `https://lucene.apache.org/`

mota ezberdinetako testu unitateak lerrokatzeko malgutasuna da.

2.2. Esaldien Lerrokapena

Corpus konparagarrietatik esaldi paraleloak erauzteko teknika ugari erabili izan dira. Lehenengo metodoetako bat 2002. urtean aurkeztu zen (Zhao eta Vogel, 2002), bertan, esaldien tamaina eta hiztegi lexikoak konbinatzen dira egiantza handieneko estimatzailearekin. Munteanu eta Marcuk zuhaitz atzizkien erabilera ere ikertu zuten (Munteanu eta Marcu, 2002). Geroago, jatorrizko teknika hobetzea lortu zuten IBM eredu batetik lortutako itzulpen probabilitateak erabiliz (Brown *et al.*, 1993) sailkatzaile bitar bat entrenatuz (Munteanu eta Marcu, 2005). Fung eta Cheungek corpus elebakarretatik esaldi paraleloak erauzteko lehenengo metodoa proposatu zuten (Fung eta Cheung, 2004). Horretarako, dokumentuak lerrokatu ostean, kosinu antzekotasuna erabiltzen dute esaldi paraleloak aukeratzeko.

Hainbat metodok itzulpen automatikoa erabiltzen dute hiztegi lexikoen ordez. Esaldi paraleloak lortzeko aukera bat jatorrizko hizkuntzako esaldiak itzultzea da, eta gero, TER metrikaren bitartez (Snover *et al.*, 2006), itzulpena jatorrizko esaldiekin konparatzea (Rauf eta Schwenk, 2009). Beste aukera bat TER ordez BLEU metrika (Papineni *et al.*, 2002) erabiltzea da (Sarikaya *et al.*, 2009). Itzulpen automatikoa erabiltzeak hizkuntzaren konplexutasunak hobeto modelatzea ahalbidetzen du, esate baterako, testuinguruko terminoen eragina eta terminoen ugalkortasuna. Hala ere, hiztegi lexikoen bidez itzulpen lexiko gehiago lortu daitezke, eta hauen erabilera errazagoa da itzulpen ereduen entrenamendurik beharrik ez baitago.

Esaldien ezaugarrietan oinarritutako teknikak ere oso erabiliak izan dira. Teknika hauen abantaila nagusienetako bat hizkuntza pare eta domeinu ezberdinetara egokitzeko aukera da. Logistic regression eredu bat entrenatuz aurreko metodoekiko hobekuntzak lortu ziren (Ștefănescu *et al.*, 2012). Bost funtzio erabiliz esaldietatik bost ezaugarri erauzten dira, eta logistic regression eredu bat entrenatuz ezaugarri horien pisuak ezartzen dira. Beste aukera bat CRF ereduak entrenatzea da. Eredu hauekin sailkatzaile bitarren estandarrekiko hobekuntzak lortu ziren (Smith *et al.*, 2010).

BUCC 2017-ko (10th Workshop on Building and Using Comparable Corpora) esaldien lerrokatze atazan[3] (Zweigenbaum *et al.*, 2017) esaldiak lerrokatzeko zenbait metodo aurkeztu ziren. Ataza honen helburua inolako metadaturen laguntzarik gabe corpus elebakarretatik paraleloak diren esaldiak identifikatzea da.

Mahata eta besteek SMT eredu baten bidez itzultzen dituzte jatorrizko hizkuntzako esaldiak, esaldiak tamainaren arabera multzokatzen dituzte, eta azken sailkapena kosinu antzekotasunarekin egiten dute (Mahata *et al.*, 2017). SVM (Cortes eta Vapnik, 1995) sailkatzaileak erabiltzen dituen metodo bat ere aurkeztu zen (Zhang eta Zweigenbaum, 2017). Lehendabizi, jatorrizko hizkuntzako corpusa itzultzen dute MICROSOFT TRANSLATOR TEXT API[4] erabiliz; gero, APACHE SOLR[5] bidez jatorrizko esaldi bakoitzeko helburuko hizkuntzako kandidatuak lortzen dituzte; eta azkenik, SVM eredu baten bidez esaldi pareak sailkatzen dituzte. Grégoire eta Langlaisek ate errepikariko bi noranzkoko sare neuronal errepikari bat (BI-GRU) (Cho *et al.*, 2014) erabiltzen dute esaldien antzekotasunak neurtzeko (Grégoire eta Langlais, 2017). BI-GRU eredua jatorrizko eta helburuko hizkuntzekin entrenatzen dute, eta horrela, esaldi guztien bektore errepresentazioa kalkulatzeko eta esaldi pareen sailkapen bitar bat egiteko gai dira. Esaldi pareen bilaketa espazioa murrizteko kosinu antzekotasuna erabiltzen dute.

BUCC 2018-ko esaldien lerrokatze atazan[6] (Pierre Zweigenbaum eta Rapp, 2018) aurkeztutako metodoen kalitatea aurreko urtekoa baina hobea izan zen. Bouamor eta Sajjadek esaldi mailako bektoreak, NMT ereduen itzulpenak eta sailkatzaile bitar bat konbinatzen dituzte (Bouamor eta Sajjad, 2018). Esaldiak errepresentatzeko MULTIVEC eredu elebiduna erabiltzen dute (Bérard *et al.*, 2016). Jatorrizko hizkuntzako esaldi bakoitzeko N kandidatu aukeratzen dituzte bektore-antzekotasunaren arabera, eta bi irizpide erabiliz aukeratzen dute kandidatu onena: (1) NMT ereduaren itzulpena erreferentziatzat hartuta antzekotasun handiena duen kandidatua, edo (2) SVM sailkatzaile

[3] https://comparable.limsi.fr/bucc2017/bucc2017-task.html

[4] https://azure.microsoft.com/en-us/services/cognitive-services/translator-text-api/

[5] https://solr.apache.org/

[6] https://comparable.limsi.fr/bucc2018/bucc2018-task.html

bitar batek emandako probabilitate handiena duen kandidatua. Leong eta besteen metodoak bi osagarri nagusi ditu (Leong *et al.*, 2018). Lehendabizikoa autoencoder eredu berezi bat da (Ye *et al.*, 2016). Eredu honek lerrokatzerako kandidatuak lortzen ditu bektore errepresentazioak bektore espazio amankomun batera proiektatuz. Bigarrena entropia maximoaren sailkatzaile bat da (Nigam *et al.*, 1999) kandidatu guztietatik onena aukeratzen duena.

Tesi honetan aurkeztutako metodoak eta gero esaldien lerrokapenerako metodo gehiago aurkeztu dira, gehien bat sare neuronaletan oinarrituta. Schwenkek esaldien errepresentazio eleanitzetan oinarritu zen (Schwenk, 2018). Esaldien errepresentazioak lortu ahal izateko hiru geruzako BLSTM kodetzaileko NMT sistema entrenatzen dute, eta entrenatutako kodetzailearekin esaldiak bektoretan bihurtzen dituzte. Geroago, Artetxe eta Schwenkek metodo hori hobetu zuten marjinaren kontzeptuarekin (Artetxe eta Schwenk, 2018), hau da, esaldi bakoitzeko, helburuko hizkuntzako N kandidatu bilatzen dira, eta lerrokapen onenaren antzekotasuna kosinu antzekotasunaren eta gainerako kandidatuen kosinu antzekotasunaren arteko marjina izango da. Marjinaren erabilerak F1 metrikan 10 puntu baino gehiagoko hobekuntza dakar. Metodo honekin WIKIMATRIX sortu zen, WIKIPEDIA-ko artikuluetatik erauzitako 135 milioi esaldi pareko corpusa 1.620 hizkuntzetan (Schwenk *et al.*, 2019).

Esaldien lerrokatzea oraindik ere oso teknika erabilia da, esate baterako, PARACRAWL corpusa sortzerako orduan (Banón *et al.*, 2020), esaldiak lerrokatzeko hiru tresna konparatu ziren: HUNALIGN (Varga *et al.*, 2007), BLEU metrikan oinarritutako BLEUALIGN (Sennrich eta Volk, 2010), eta LASER esaldien errepresentazio eleanitzetan [7] (Artetxe eta Schwenk, 2019) oinarritutako VECALIGN (Thompson eta Koehn, 2019). Lerrokapen onenak lortu zituen metodoa azkena izan zen.

2.3. Datuen Aukeraketa

Corpus paraleloetatik azpimultzoak aukeratzea oso teknika erabilia da itzulpen automatikoa domeinura egokitzeko edota sistemen entrenamendua azkartzeko (Eetemadi *et al.*, 2015). Domeinuaren egokitzapenaren helburu na-

[7] `https://engineering.fb.com/2019/01/22/ai-research/laser-multilingual-sentence-embeddings/`

gusia corpusetako datuen erabilera optimizatzea da domeinu zehatzeko itzulpen automatikorako eraginkorrena den esaldi paraleloen azpimultzo minimoa aukeratuz. Horrela, hizkuntza baliabideak urriak diren kasuetan domeinuko datu gehiago lortzea lortu daiteke.

Datu elebidunak aukeratzeko hurbilpen ugari erabili dira. Esate baterako, TF-IDF pisuak erabili izan dira antzeko esaldiak pareak identifikatu edota entrenamendua baldintzatzeko (Lü *et al.*, 2007), eta baita n-grama maiztasunarekin konbinatuz entrenamendurako corpusaren tamaina murrizteko (Eck *et al.*, 2005). Foster eta besteek lehendabizi domeinutik kanpoko corpuseko esaldiak ordenatzen dituzte hizkuntza eredu batek emandako perplexitatearen arabera, eta gero, garapenerako corpus batean BLEU puntuazio onena lortzen duten lehenengo N esaldiak aukeratzen dituzte (Foster *et al.*, 2010). Geroago, esaldien aukeraketa prozesua hobetu zuten esaldien pisuen ikasketa metodoa hedatuz (Matsoukas *et al.*, 2009): esaldi pareetatik ezaugarriak erauzten dituzte konbinazio linealeko eredu bat entrenatzeko.

Perplexitatean oinarritutako metodoak dira datuen aukeraketarako erabilienak izan direnak. Esate baterako, Foster eta besteek jatorrizko domeinuko helburuko hizkuntzako corpusa erabiltzen dute hizkuntza eredu bat entrenatzeko, eta eredu horretatik lortutako perplexitatearen arabera ordenatzen dira domeinutik kanpoko esaldi pareak (Foster *et al.*, 2010). Beste aukera bat hizkuntza ereduen entropia itzulpen ereduenekin gurutzatzea da (Mansour *et al.*, 2011). Hiztegiaren asetasunaren metodoarekin ere (Lewis eta Eetemadi, 2013) ikerketak egin dira perplexitatearekin konbinatuz (Aydın eta Ozgür, 2014). Lehendabizi domeinutik kanpoko corpusa jatorrizko domeinuaren perplexitatearen arabera ordenatzen da, eta datuen azpimultzoak hiztegiaren asetasunaren metodoarekin aukeratzen dira. Metodo erabilienetako bat Axelrod eta besteena da (Axelrod *et al.*, 2011). Modified Moore-Lewis metodoa (Moore eta Lewis, 2010) (aurrerantzean MML) hedatu zuten domeinutik kanpoko corpuseko esaldi pareak ordenatzeko domeinuko eta domeinutik kanpoko hizkuntza ereduen entropia gurutzatu elebidunaren arabera. Metodo honen bidez, aldi berean domeinutik hurbil eta kanpoko domeinutik urruti dauden esaldi pareak aukeratzea lortzen da. Metodoaren zenbait aldakuntzekin ere ikerketak egin dira kategoria gramatikal eta termino klaseekin

hizkuntza ereduak entrenatuz (Axelrod *et al.*, 2015a,b).

Nahiz eta metodo gehienak domeinuen arteko antzekotasun metrika optimoen ikerkuntzan oinarritu, datuen aukeraketa itzulpen kalitatearen arabera bideratu daiteke. Estrategia posible bat domeinutik kanpoko corpusetik datu multzoak modu inkrementalean gehitzea da itzulpen sistema batera (Banerjee *et al.*, 2012). Datu multzoa aukeratzen da baldin eta soilik baldin garapenerako corpus batean itzulpen kalitatea hobetzen bada. Esaldien antzekotasunean oinarritzen ez den beste lan bat Daumé Iii eta Jagarlamudirena da, non lexikoaren estaldura hartzen den kontutan termino ezezagunak ustiatuz korrelazio analisi batean (Daumé Iii eta Jagarlamudi, 2011). Beste ikerlan batean berriz, jatorrizko domeinuko probabilitate-banaketa eta n-grama maiztasunak kontutan hartzen dituzte datu multzoak aukeratzeko orduan (Gascó *et al.*, 2012). Sare neuronal errepikariekin (*recurrent neural network*, RNN) ere saiakerak egin dira datu gehigarriak aukeratzeko. Adibidez, txinera-ingeleserako, perplexitatean oinarritutako sistemekin alderatuz hobekuntza nabariak lortu dira (Wong *et al.*, 2016).

Entropia gurutzatu elebidunean oinarritutako sistemak *de facto* estandartzat har daitezke eta oso erabiliak dira beste sistemekin konparaketak egiteko. Kirchhoff eta Bilmes (2014)-ek, funtzio azpimodularrak erabiliz hobekuntzak lortzen dituzte BLEU metrikan MML-rekin alderatuz. Peris *et al.* (2016)-ek lortutako hobekuntzak ez dira hain nabariak, baina sare neuraletan oinarritutako sailkatzaile bati esker erabili beharreko datu multzoak murriztea lortzen dute. Banerjee *et al.* (2013)-ek ere beraien metodoa MML-rekin konparatzen dute. Kalitatearen estimazioan oinarritzen dira, eta datu gutxiago erabiliz BLEU metrikan emaitzak apur bat hobetzea lortzen dute. Orokorrean, nahiz eta hobekuntzak estatistikoki esanguratsuak izan, MML gutxigatik gainditzen da, eta hortaz, oraindik baliozko metodoa da konparaketak egiteko.

Nahiz eta tesi honetan datuen aukeraketa elebidunaren inguruan jardun, datuen aukeraketa elebakarra ere oso erabilia da hizkuntza ereduak egokitzeko. Esate baterako, entropia gurutzatua hobetzea lortu daiteke domeinuz kanpoko datu multzo hobeak erauziz (Mediani *et al.*, 2014). Terminoen asoziazioa erabiliz aukeraketa prozesuari antzekotasun semantikoa gehitzea lortzen da. Mansour eta besteek iragazpen metodo bat deskribatzen dute, non entropia

gurutzatuaren zenbait konbinazio direla medio emaitzak hobetzen dituzten (Mansour *et al.*, 2011). Azkenik, sare neuronalen erabileraren bitartez termino ezezagunak hobeto kudeatzea posible dela egiaztatu da (Duh *et al.*, 2013).

Tesi honetan aurkeztutako lanen ostean datuen aukeraketaren ikerkuntzarekin jarraitu da (Ramponi eta Plank, 2020). Adibidez, Aharoni eta Goldberrek, BERT (Devlin *et al.*, 2018) DISTILBERT (Sanh *et al.*, 2019), ROBERTA (Liu *et al.*, 2019), GPT-2 (Radford *et al.*, 2018) eta XLNET (Yang *et al.*, 2019) ereduak erabiltzen dituzte oso datu multzo handietan aurreentrenatutako ereduek domeinuak berez klusterretan antolatzen dituztela erakusteko (Aharoni eta Goldberg, 2020). Guo eta besteek zenbait bektore distantzia metrikekin esperimentuak egin zituzten domeinu sailkatzaile bat entrenatuz (Guo *et al.*, 2020); distantzia euklidearra, kosinu distantzia, MMD (*Maximum Mean Discrepancy*) distantzia (Gretton *et al.*, 2012), FLD (*Fisher Linear Discriminant*) distantzia (Friedman *et al.*, 2001) eta CORAL (*Correlation Alignment*) distantziarekin (Sun eta Saenko, 2016) egin zituzten probak. Terminoen gainjartzea datu multzoen antzekotasunaren adierazle dela konprobatu ahal izan da zenbait ikerketan (Üstün *et al.*, 2019; Lin *et al.*, 2019). Azkenik, beste ikerlerro batean aurreentrenatutako ereduak bigarren aurreentrenamendu batean helburuko domeinura egokitzea onuragarria den aztertzen da. Gururangan eta besteek, aurreentrenatutako ROBERTA eredu bat (Liu *et al.*, 2019) lau domeinutara egokitzen dute datuen egokitzapena hiztegi gainjartzearen bidez eginez (Gururangan *et al.*, 2020).

2.4. Esaldi Paraleloen Iragazpena

Kalitatezko corpus paraleloak ez dira oso ugariak erabili beharreko hizkuntza parearen arabera. Corpus paraleloak lortzeko irtenbide bat Internetetik datu multzoak erauzi, eta zenbait prozesaketa egin ostean esaldi paraleloak sortzea da (Forcada *et al.*, 2016). Hala ere, Internet oso datu iturri zaratatsua da, eta automatikoki dokumentu eta esaldiak lerrokatuz erroreak gertatzen dira. Horregatik, litekeena da bukaerako corpusean akatsak egotea, itzulpen automatikorako kaltegarria izan daitekeena (Khadivi eta Ney, 2005; Khayrallah eta Koehn, 2018).

Corpus paraleloetatik datu zaratatsuak garbitzeko ataza hainbat ikerketetan jorratu da. Munteanu eta Marcuk, entropia maximoaren sailkatzaile bat entrenatzen dute entrenamendurako datu garbi eta zaratatsuak erabiliz (Munteanu eta Marcu, 2005). Esplà-Gomis eta Forcadak esaldien lerrokapen puntuazioa sartu zuten BITEXTOR sisteman, Internetetik automatikoki corpus paraleloak sortzeko tresna, zalantzazko esaldi pareak iragazteko asmoz (Esplà-Gomis eta Forcada, 2009). Khadivi eta Neyk bi hurbilpen ebaluatu zituzten, lehenengoa esaldien tamainan oinarritua, eta bigarrena, itzulpen lexikoaren egiantzean oinarritua (Khadivi eta Ney, 2005). Iragazitako corpusa erabiliz itzulpen kalitatea hobetzen dela egiaztatu ahal izan zuten. Esaldi zaratatsuak identifikatzeko metodo ez-gainbegiratuak ere erabili dira. Metodo hauei esker corpusaren iragazitako bertsioarekin hobekuntzak lortzen dira (Taghipour *et al.*, 2011). *Graph-based random walk* algoritmoaren bidez ere hobekuntzak lortu daitezke, esate baterako txinera-ingelesa corpusak iragazteko (Cui *et al.*, 2013). Trena interesgarri bat ZIPPORAH da, corpus zaratatsuetan datuen aukeraketa azkarra egiteko sistema (Xu eta Koehn, 2017). ZIPPORAH itzulpenen egokitasun eta jariakortasunean oinarritzen dira. Egokitasuna aztertzeko, esaldiak *bag-of-words* bidez errepresentatzen dituzte, eta hauen itzulpena helburuko hizkuntzako esaldien errepresentazioarekin konparatzen dute. Jariakortasuna neurtzeko berriz, hizkuntza ereduak erabiltzen dituzte.

WMT 2018-ko esaldi paraleloen iragazpenerako atazan[8] (Koehn *et al.*, 2018) SMT eta NMT ereduak entrenatzen dira iragazitako corpusarekin, eta domeinu ezberdineko corpusetan ebaluatzen dira, oso ataza aberatsa da beraz. Oso mota ezberdineko teknikak konbinatu ziren hasierako corpus zaratatsua garbitzeko.

Junczys-Dowmuntek zenbait teknika konbinatzen ditu esaldi paraleloak iragazteko (Junczys-Dowmunt, 2018): (1) hizkuntza identifikatzaile bat lehen iragazpen bat egiteko, (2) kalitatezko corpus paralelo batetik bi itzulpen eredu entrenatzen dira (bi eredu kontrako hizkuntza norabidearekin) eta eredu hauen entropia gurutzatuarekin beste iragazpen bat egiten da, (3) eta azken iragazpena MML metodoarekin egiten da (Moore eta Lewis, 2010).

[8] http://www.statmt.org/wmt18/parallel-corpus-filtering.html

Lok eta Littellek sistema gainbegiratu bat (Lo *et al.*, 2018) eta ez-gainbegiratu bat (Littell *et al.*, 2018) aurkeztu zuten. Sistema gainbegiratua YISI itzulpen automatiko semantikoaren ebaluazio metrikan (Lo, 2018) oinarritzen da. Sistema honetan ere lehendabiziko iragazpen bat egiten da hizkuntza detektatzaile eta zenbait heuristikorekin, eta gero, konbinazio linealeko eredu bat entrenatzen da YISI metrika, ezkutuko markov eredu (HMM) lerrokatze eredu (Vogel *et al.*, 1996), perplexitate eta esaldi bektoreekin ezaugarriak erauzi eta gero. Azken iragazpen bat egiten da bigramak berririk ez duten esaldi erredundanteak kenduz. Sistema ez-gainbegiratua berriz, hasierako iragazpena egin eta gero, MAHALANOBIS distantzian oinarritzen da (Mahalanobis, 1936) esaldi-bektoreen antzekotasuna kalkulatzeko. Azken iragazpena sistema gainbegiratuan egiten den berbera da.

Sánchez-Cartagena eta besteek hiru pausutan egiten dute corpusaren iragazpena (Sánchez-Cartagena *et al.*, 2018): (1) eskuz idatzitako erregelen bidezko (hizkuntza okerreko esaldien identifikazioa barne) lehen iragazpen bat egiten da, (2) ausazko basoaren sailkatzaile baten bidez (Breiman, 2001) oker lerrokatutako esaldi pareak detektatzen dira, eta (3) n-grama asetasunean oinarritutako metodoen bidez esaldi pareak berrordenatzen dira.

Lu eta besteek sistema hiru tekniketan oinarritzen da (Lu *et al.*, 2020): (1) GPT-2 eredu elebidunak (Radford *et al.*, 2019), (2) baldintzatutako entropia gurutzatu elebidunak (Junczys-Dowmunt, 2018), eta (3) FAST ALIGN tresnarekin entrenatutako terminoak lerrokatzeko IBM ereduak (Dyer *et al.*, 2013).

Rossenbach eta besteek, corpusaren aurreprozesaketa bat egin ostean, zenbait heuristiko erabiltzen dituzte esaldi paraleloen lehen aukeraketa bat egiteko (Rossenbach *et al.*, 2018): termino kopuru eta esaldi tamainan oinarritutakoak, Levenshtein distantzia (Levenshtein, 1966), eta datu erredundantzian oinarritutakoak. Aukeratutako esaldi pareen antzekotasuna neurtzeko bi geruzako LSTM hizkuntza eredu neuronal (Hochreiter eta Schmidhuber, 1997) eta TRANSFORMER itzulpen eredu neuronalak (Vaswani *et al.*, 2017) erabiltzen dituzte.

WMT 2018-ko emaitzetan ikuste denez, NMT ereduen kalitatea SMT ereduena baina altuagoa da (orokorrean +5 puntuko aldea BLEU metrikan 100M hitzeko corpusarekin), eta sistema gehienen ezberdintasunak ez dira oso handiak,

100M hitzeko corpusarekin partehartzaile gehienak BLEU metrikan 2 bi punturen barne baitaude.

Tesi honetako ikerketen ostean, esaldi paraleloen iragazpenaren ikerketa lerroak aurrera egin du. Horren adibide garbia WMT-ko esaldi paraleloen iragazpenerako atazaren arrakasta da, 2019 eta 2020. urteetan berriro ere plazaratu baitzen. Atazaren edizio hauek baliabide gutxiko hizkuntzetarako egokituta daude[9] (Koehn *et al.*, 2019, 2020).

WMT 2019-ko atazan metodo ugari aurkeztu ziren. Chaudhary *et al.* (2019)-ek, LASER esaldien errepresentazio eleanitzetan (Artetxe eta Schwenk, 2019) eta bektoreen antzekotasunen arteko marjinetan (Artetxe eta Schwenk, 2018) oinarritzen da. Erdmann eta Gwinnupek bestelako metodo bat erabili zuten (Erdmann eta Gwinnup, 2019): lehendabiziko aurreprozesamendu bat egin eta gero (esaldien luzera eta hizkuntza identifikatzaile batean oinarritutakoa), esaldien estaldura eta itzulpen kalitatea neurtzen dute. Itzulpenen kalitatea itzulpen eredu batek emandako METEOR metrikarekin (Denkowski eta Lavie, 2014) aztertzen dute. Sen eta besteek antzeko metodo bat aurkeztu zuten (Sen *et al.*, 2019). Jatorrizko hizkuntzako esaldia ingelesera itzuli ostean, antzekotasuna Levenshtein distantziarekin (Levenshtein, 1966) neurtzen dute. Azkenik, Bernier-Colborne eta Lok bi osagai nagusi konbinatzen dituzte (Bernier-Colborne eta Lo, 2019): terminoen errepresentazio elebidunak erabiltzen dituen YISI-2 antzekotasun metrika (Lo *et al.*, 2018), eta *tranfer learning* teknika aurreentrenatutako XLM eredu baten bidez (Lample eta Conneau, 2019). WMT 2019 atazako emaitzetan erreparatuz, deigarria da SMT ereduen emaitzak NMT ereduenak baina hobeak izatea. Emaitza hauek baliabideen urritasunarekin azaltzen dira, oraindik ere SMT ereduek bere esparrua dutela frogatuz.

WMT 2020-ko atazan antzeko metodoak aurkeztu ziren. Ia partaide guztiek heuristikoetan oinarritutako aurreprozesaketa bat egiten dute esaldi pare zaratatsuenak baztertzeko. Lu eta besteek hiru osagai konbinatzen dituzte (Lu *et al.*, 2020): (1) GPT-2 eredu elebidunak (Radford *et al.*, 2019), (2) NMT ereduek emandako entropia gurutzatua, eta (3) IBM ereduen terminoen itzulpen

[9] `http://www.statmt.org/wmt19/parallel-corpus-filtering.html` eta `http://www.statmt.org/wmt20/parallel-corpus-filtering.html`

probabilitateak (Dyer *et al.*, 2013). Acarcicek eta besteek, S-BERT (Reimers eta Gurevych, 2019), BERT (Devlin *et al.*, 2018) eta ROBERTA (Liu *et al.*, 2019) ereduen bertsio eleanitzak konparatzen dituzte, eta entrenamendu algoritmo bat proposatzen dute ereduei adibide "zailak" ezberdintzen erakusteko (Acarcicek *et al.*, 2020). Lo eta Joanisen metodoa YISI-2 antzekotasun metrikan (Lo *et al.*, 2018) oinarritzen da (Lo eta Joanis, 2020). YISI-2 metrikak esaldien errepresentazioak erabiltzen ditu antzekotasunak neurtzeko, eta errepresentazio horiek aurreentrenatutako XLM-ROBERTA (Conneau *et al.*, 2019) eredua atazarako berrentrenatuz lortzen dituzte. Bukatzeko, Esplà-Gomis eta besteak BICLEANER tresnan oinarritzen dira (Sánchez-Cartagena *et al.*, 2018). BICLEANER hedatzeko dute oso ausazko baso (ERT) (Geurts *et al.*, 2006) eta 7-gramako karaktere ereduak erabiltzen dituzte (Esplà-Gomis *et al.*, 2020).

3. KAPITULUA

Dokumentuen Lerrokatzea

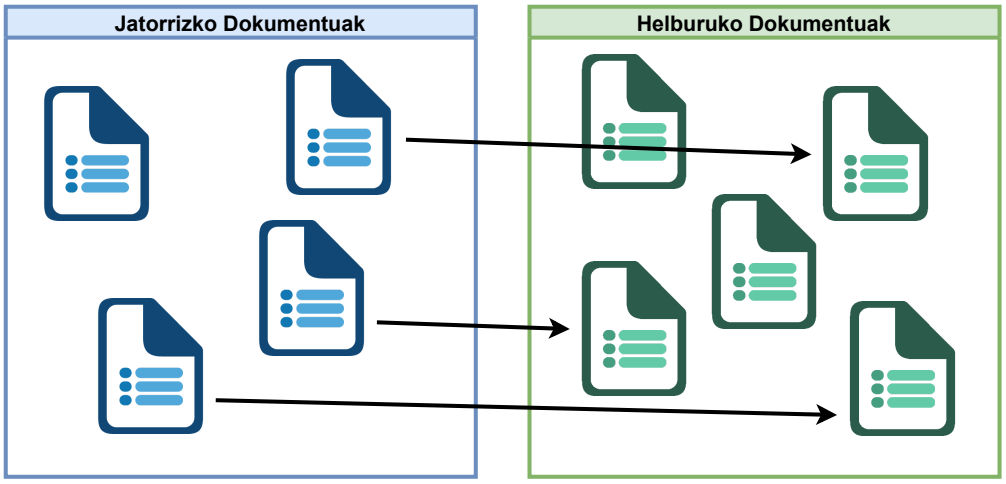
Azken urteotan, corpus paraleloen bilduma handiak sortu izan dira itzulpen automatikoaren ikerketa eta garapenerako (Tiedemann, 2012). Hala ere, hizkuntzaren berezko aldagarritasunarengatik eta itzuli beharreko domeinu ezberdinen ugaritasuna dela medio, corpus paralelo gehiago behar dira itzulpen automatikorako sistemak elikatzeko, bai domeinu zabaletarako eta baita domeinu zehatzetarako. Bi datu iturri nagusi erabili izen dira datu paraleloak sortzeko. Alde batetik, adituek itzulitako dokumentuak daude. Hauek kalitatezko corpus paraleloak sortzeko erabili daitezke, esate baterako, EUROPARL corpusa (Koehn, 2005). Bestetik, antzeko domeinu eta gaiak jorratzen dituzten dokumentuak ditugu, hau da, dokumentu konparagarriak. Azken hauen adibide garbia WIKIPEDIA dugu. Nahiz eta kalitate aldetik corpus paraleloak hobeak izan, corpus konparagarriak oso ugariak dira eta itzulpen automatikorako baliagarriak izan daitezke (Munteanu eta Marcu, 2005).

WIKIPEDIA corpusaren kasuan dokumentuak lerrokatuta daude hizkuntza esteken bidez, baina beste zenbaitetan lerrokapenik ez dago. Esate baterako, enpresen dokumentazio teknikoan dokumentu paraleloek antzeko fitxategi izena izan ohi dute, eta irtenbide bat espresio erregularrak erabiltzea izan daiteke. Tamalez, ezin da bermatu dokumentuen fitxategi izenek kasu guztietan egitura berbera jarraitzea, eta gainera, dokumentu multzo batetik bestera fitxategi izenen egitura ezberdina izan liteke.

Hortaz, datu iturriak ustiatu ahal izateko lehen pausua dokumentuen lerrokatzea izan ohi da. 3.1. irudian laburbiltzen den bezala, bi dokumentu bilduma

elebakar izanda, dokumentuen lerrokatzea jatorrizko hizkuntzako dokumentu bakoitzeko dagokion helburuko hizkuntzako dokumentua bilatzean datza. Jakina, bilaketa hori egin ahal izateko antzekotasun metrikaren bat behar da, eta dokumentu batekin konbinazio guztiak probatu ostean baliteke dokumentu horrek lerrokatzerik ez izatea metrikaren emaitzen arabera.

3.1 irudia: Dokumentuen lerrokatzea.



Kapitulu honetan DOCAL deritzon sistema da aurkezten da. DOCAL-ek dokumentuak lerrokatzen ditu antzekotasun metrika sinple baten bidez, multzo egituren arteko eragiketetan eta itzulpen lexikoetan oinarritzen dena. DOCAL-en eraginkortasuna neurtzeko, artearen egoerako beste sistemekin konparatu da, mota ezberdinetako dokumentuak dituzten esperimentu sakonak eginez. Esperimentu hauetan, DOCAL-en oinarrizko metodoa eta baita egindako zenbait optimizazio neurtu dira, horrela, etorkizuneko ikerketak zein norabidetik joan daitezkeen hausnartu daiteke.

Hurrengo atalak honela antolatzen dira. 3.1. atalean oinarrizko metodoa azaltzen da. Segidan, 3.2. atalean oinarrizko metodoa probatzeko egindako esperimentuak aurkezten dira. 3.3. atalean berriz, oinarrizko metodoaren gainean zenbait optimizazio egiten dira eta hauen ekarpena neurtzen da. Azkenik, 3.4. atalean, izandako emaitzak laburbiltzen dira eta azken ondorioak ateratzen dira.

3.1. Dokumentuen Antzekotasuna, DOCAL

DOCAL-ek (*DOCument ALignment*) dokumentuen antzekotasun metrika batean oinarritutako dokumentuak lerrokatzeko metodoa da. Helburua kalitatezko lerrokatzeak lortzea da eraginkortasun eta eramangarritasuna bermatuz. Funtsean, DOCAL bi tekniken pean oinarritzen da: dokumentu bakoitza errepresentatzeko termino multzo bat erabiltzen da, eta bi dokumentuen antzekotasuna neurtzeko beren termino multzoen gainean Jaccard koefizientea kalkulatzen da (Jaccard, 1901).

Lehenengo pausua dokumentu pare bakoitzeko termino multzoak erauztean datza. Testua tokenizatu eta gero, dokumentuko terminoak multzo egitura batean gordetzen dira. Ez da inolako filtrorik aplikatzen, beraz, hitz lexikoak, funtzionalak eta puntuazio-markak egongo dira token multzoetan.

Bigarren pausuaren helburua dokumentu bakoitzeko termino multzoak itzultzea da, eta horretarako aurrebaldintza itzulpen-taula bat entrenatzea da. Itzulpen-taulak corpus paraleloetatik erauzten dira IBM ereduek[1] emandako itzulpen probabilitateak kontutan hartuz. Termino bakoitzeko, probabilitate handiena duten k itzulpen onenak aukeratzen dira itzulpen termino multzoa sortzeko. Azpimarratu beharra dago ez direla itzulpen probabilitateak erabiltzen antzekotasunaren kalkuluan. Honetarako arrazoi nagusia probabilitateak domeinu ezberdinetako corpusetik erauzten direla da, eta oso ziurra da distribuzio lexikoa ezberdina izango dela bi domeinuetan (lerrokatu beharreko dokumentuen domeinua, eta itzulpen probabilitateak erauzteko erabili den corpusaren domeinua)[2].

Dokumentu baten itzulpen termino multzoa sortzeko, termino multzoko termino bakoitzeko probabilitate handieneko k itzulpen lexikoak bilatzen dira itzulpen-taulan eta itzulpen multzoan gordetzen dira. Gero, itzulpen termino multzoak honako bi eragiketa hauen bidez hedatzen dira:

[1] GIZA++ (Och eta Ney, 2003) erabiltzen dugu itzulpen lexikoen taulak sortzeko.

[2] Besterik adierazi ezean 5 itzulpen onenak hartzen dira: tamaina handiko baina fidagarritasun gutxiko multzoen eta tamaina txikiko baina fidagarritasun handiko itzulpenen artean oreka ona ematen baitu. Balio hau enpirikoki ezarri da aurretiko proba batzuk egin eta gero, baina garapenerako corpus bat erabiliz balio optimoa kalkulatzeko aukera ere badago

- Jatorrizko hizkuntzako termino multzoan dauden zenbakiak eta letra larriz hasten diren terminoak bere horretan ere gehitzen dira (zenbaki eta izen-entitateak kontuan hartzeko).

- Termino multzoak konparatzeko orduan, aldaketa morfologikoak kontutan hartu ahal izateko aurrizki komun luzeenak gehitzen dira (3.1.2. azpiatalean ikusiko dugu zehazki). Eragiketa hau hautazkoa da.

Azkenik, bi dokumentuen arteko antzekotasun ratioa neurtzen da, beraien termino multzoen ebakidura bildurarekin zatituz. Eragiketa hau bi norabideetan egiten da, hau da, jatorrizko hizkuntzatik helburuko hizkuntzara eta helburuko hizkuntzatik jatorrizko hizkuntzara. Azken emaitza bi Jaccard koefizienteen arteko batezbesteko aritmetikoa da.

Formalki, d_i eta d_j tokenizatutako bi dokumentu izanik l_1 eta l_2 hizkuntzetan hurrenez hurren, S_i d_i-ren termino multzoa, S_j d_j-ren termino multzoa, T_{ij} S_i multzoaren hedatutako termino multzoa l_2 hizkuntzara, eta T_{ji} S_j multzoaren hedatutako termino multzoa l_1 hizkuntzara. 3.1. ekuazioak d_i eta d_j dokumentuen arteko antzekotasuna nola kalkulatzen den deskribatzen du.

$$docal(d_i, d_j) = \frac{1}{2}\left(\frac{|T_{ij} \cap S_j|}{|T_{ij} \cup S_j|} + \frac{|T_{ji} \cap S_i|}{|T_{ji} \cup S_i|}\right) \tag{3.1}$$

Jaccard koefizienteak oso propietate interesgarriak ditu dokumentuak lerrokatzeko. Alde batetik, emaitza 0-tik 1-erako zenbaki erreal bat da. Bestetik, beste metrikekin konparatuta, DICE indizea (Dice, 1945) esate baterako, Jaccard koefizienteak gehiago zigortzen ditu elementu gutxi partekatzen dituzten multzoak. Propietate hau oso baliagarria da hitz funtzionalak bakarrik partekatzen dituzten multzoentzat. Kosinu antzekotasunarekin alderatuz, Jaccard koefizientearekin oso ezberdinak diren multzoek puntuazio okerragoa dute.

Hurrengo bi azpiataletan termino multzoen hedapenaren xehetasunak azaltzen dira.

3.1.1. Hiztegitik Kanpoko Terminoen Hedapena

Itzulpen-taulak corpuseko lexiko guztia estaltzea ezin da bermatu. Horregatik oso garrantzitsua da itzulpen multzoak hedatzea lerrokatzerako baliagarriak diren terminoak gehituz nahiz eta termino hauek itzulpen-tauletan ez egon. Testua letra xehetara pasatzen ez denez[3], letra larriz hasten diren terminoak erabili daitezke itzulpen multzoak hedatzeko betiere termino horiek itzulpen-tauletan agertzen ez badira[4]. Eragiketa sinple honekin izen-entitate posibleak gehitzen dira, eta izen-entitateak izanik testua lerrokatzeko ezaugarri adierazgarrienetakoak, lerrokatzeak hobetzea lortzen da.

Zehazki, honela egiten da hiztegitik kanpoko terminoen hedapena:

- d izanik l_1 hizkuntzako dokumentu bat, S d-ren termino multzoa eta T itzulpen multzoa l_1 hizkuntzatik l_2 hizkuntzara,
- S-ko w termino bakoitzeko:
  - w itzulpen-taulan badago, gehitu w-ren probabilitate handieneko k itzulpenak T multzoan.
  - Bestela, w letra larriz hasten bada edo zenbaki bat bada, gehitu w T multzoan.

Adibide batekin hedapen honen ekarpena ikusiko dugu. Suposa dezagun ingelesez idatzitako dokumentu batean *"John Doe"* dugula, eta *"John"* tokenak itzulpena duela baina *"Doe"* tokenak ez. Lehenengo terminoaren kasuan, hiztegiak itzulpena izango du eta T itzulpen multzoan *"John"* izango dugu. Bigarren terminoaren kasuan berriz, *"Doe"* ez dago hiztegian, baina teknika honi esker, termino hau ere T itzulpen multzoan izango dugu. Demagun *Doe*

[3]Tokenizazioa alde batera utzita, 3.3. azpiataleraino ez da inolako aurreprozesamendurik egiten ahalik eta metodologia arinena erabiltzearren.

[4]Itzulpen-taulan bilaketak eginez, termino bat hiztegitik kanpoko terminoa den ala itzulpena duen izen-entitatea den identifikatu daiteke. Bilaketa hau egitea berebizikoa da, bestela, termino okerrak gehitzeko arriskua dago. Adibidez, *"German"* hitza gazteleraz *"Alemania"* da, eta kasu honetan gaizki legoke *"German"* gehitzea gaztelaniazko itzulpen multzoan. Gainera, kontutan eduki behar da alemanez izen arruntak letra larriz hasten dira.

bakarrik dugula gaztelaniazko dokumentu batean, Hiztegitik kanpoko terminoen hedapena erabili izan ez bagenu, bi dokumentuen arteko lerrokatzea okerragoa izango litzateke.

Hiztegitik kanpoko terminoen hedapenaren ekarpena corpusaren araberakoa da, hau da, azaltzen diren izen-entitateen kantitatearen araberakoa. Horregatik 3.2.3. azpiatalean deskribatutako EITB testean emaitzak ez dira hobetzen, baina 3.2.2. azpiataleko BUCC 2015 atazan berriz 30 puntu baina gehiagoko aldea dago frantses-ingeles corpusean. Horregatik, lehenetsitako konfigurazioan teknika hau erabiltzen da kontrakoa esaten ez bada.

3.1.2. Aurrizki Komun Luzeenak

Datuetan oinarritutako metodoetan ohikoa da aldaketa morfologikoak kontutan eduki beharra, bestela, lema berbera duten hitz ezberdinak unitate independentetzat hartzen dira eta datuen sakabanaketa handituz. Arazo hau ekiditeko lematizazio edo stemming teknikak erabili ohi dira. Morfologikoki aberatsak diren hizkuntzetan ordea, lematizazioa egitea ez da batere erraza, batez ere lematizazioa ondo egin ahal izateko baliabide linguistikoak egokiak eskura ez badaude. Gainera, lematizazioa egiteak konputazio kostu handia ekar lezake.

Arazo hau ekiditeko, aukerazko hedapen teknika bat diseinatu da aurrizki komun luzeenetan oinarritua. Lehendabizi, erro komunak bilatzeko termino multzo minimoak sortzen dira. Hau da, demagun T_{ij} itzulpen multzoa eta S_j termino multzoa ditugula l_2 hizkuntzan, orduan, T'_{ij} eta S'_j multzoak honela kalkulatzen dira: $T'_{ij} = T_{ij} - S_j$ eta $S'_j = S_j - T_{ij}$. Gero, T'_{ij}-ko elementu bakoitzeko eta S'_j-ko elementu bakoitzeko, n karaktere baina gehiagoko aurrizki komun guztiak T_{ij} eta S_j multzoetan gehitzen dira[5].

Adibide bezala, jo dezagun multzo bat dugula *"sample"* terminoarekin, eta beste multzo batean *"sampling"* dugula. Bi multzoen arteko diferentzia eginez, *"sample"* terminoa duen multzoa lortuko dugu, eta aurrizkien bilaketa egiterakoan *"sampl"* gehituko dugu hasierako bi multzoetan. Beraz, bi multzoen antzekotasuna kalkulatu eta gero, Jaccard koefizienteak balio handia-

[5]Esperimentuetan $n = 3$ erabili dugu.

goa izango du gehitutako *"sampl"* aurrizkiari esker.

Izen-entitateen kasuan bezala, aurrizki komun luzeen emaitza domeinuaren araberakoa da. Hala ere, hiztegitik kanpoko terminoen hedapenarekin alderatuz, konputazionalki aurrizki komun luzeenen kalkulua dezente astunagoa da, eta gainera, 3.2. kapituluan egindako esperimentuen arabera, emaitzak ez dira gehiegi hobetzen. Arrazoi guzti hauengatik, kontrakoa adierazten ez bada azaldutako emaitzetan ez da aurrizki komun luzeenik erabiltzen.

3.1.3. Dokumentuen Indexazioa

Aplikazio domeinu batzuetan, baliteke produktu kartesiarraren bidez dokumentu konbinazio guztiak aztertzea irtenbide optimoa izatea; corpusean dokumentu gutxi badaude bilaketa espazioa txikia izango da eta dokumentu konbinazio guztiak aztertu ahal izango dira. Hala ere, produktu kartesiarrak $O(n^2)$ konplexutasuna du, eta dokumentu kantitatea handitzen den heinean konputazio kostua ere gero eta gehiago handitzen da. Egindako esperimentuen arabera, 260 milioi konbinaziotik gora 64 GB RAM-eko zerbitzari bat produktu kartesiarrarekin ito egiten da.

Dokumentu kopurua oso handia den corpusetarako, hizkuntza arteko informazio-eskuratzea metodologia bat diseinatu da. Lehendabizi, helburuko hizkuntzako dokumentuak indexatzen dira dokumentu bakoitzeko terminoak gordez APACHE LUCENE[6] bilaketa motorra erabiliz. Gero, jatorrizko hizkuntzako dokumentu bakoitzeko bilaketa bat egiten da itzulpen multzoko terminoak erabiliz helburuko hizkuntzako dokumentu antzekoenak eskuratzeko. DOCAL-eko antzekotasun metrika eskuratutako dokumentu horien gainean kalkulatzen da.

EUROPARL corpusaren zazpigarren bertsioko corpuseko gaztelania-ingelesa hizkuntza parean[7], 9.433 eta 9.672 dokumentu izanik hurrenez hurren, dokumentuen lerrokatzeak 223 minutu eta 13 segundo irauten du produktu kartesiarrarekin (91.235.976 konbinazio); indexazioarekin berriz, dokumentu bakoitzeko 100 hautagai bilatuz, lerrokatzeak 33 minutu eta 45 segundo

[6] `https://lucene.apache.org`
[7] `http://www.statmt.org/europarl/`

irauten du (9.433.000 konbinazio).

Bi metodologiak erabiltzen dira 3.2 eta 3.3. azpiataletako esperimentuetan: indexazioa aplikatzen da BUCC 2015 eta WMT 2016 corpusetan, eta produktu kartesiarra gainerako guztietan. Bata edo bestea erabiltzeko erabakia dokumentu konbinazio posibleen araberakoa da, ahal den kasuetan produktu kartesiarra hobetsiz.

3.1.4. Lerrokatze Onenaren Optimizazioa

Dokumentuen lerrokatzea jatorrizko hizkuntzatik helburuko hizkuntzara egiten da, beraz, litekeena da helburu hizkuntzako dokumentu bat jatorrizko hizkuntzako dokumentu bat baina gehiagorekin lerrokatzea. Fenomeno honek lerrokatze zuzenak ezkutatu ditzake, baina kasu askotan, bi lerrokatzeen (lerrokatze zuzena eta metrikak adierazitako lerrokatze onena) puntuazioen arteko diferentzia oso txikia da metrikak adierazitako lerrokatzearen alde. Honetarako irtenbide sinple bat dago: dokumentu berbera behin baino gehiagotan lerrokatzen baldin bada, puntuazio handieneko lerrokatzea mantentzea eta gainerakoak baztertzea.

Zehatzago esanda, d_i jatorrizko hizkuntzako dokumentua, d_j helburuko hizkuntzako dokumentua, eta sim_{ij} d_i eta d_j-ren arteko antzekotasun emaitza izanik, lerrokatze onenak (d_i, d_j, sim_{ij}) lerrokatzea ezabatuko du beste (d_k, d_j, sim_{kj}) lerrokatze bat existitzen bada non $sim_{kj} > sim_{ij}$ den.

Optimizazio honek askotan hobekuntza altua dakar. Izan ere, 3.2. azpitituluko esperimentuetan lerrokatze onenaren optimizazioak 10 puntu baino gehiagoko aldea baitu kasu batzuetan. Horregatik, lerrokatze onenaren optimizazioa esperimentu guztietan erabiltzen da kontrakoa adierazten ez bada.

3.2. Oinarrizko Metodoaren Esperimentuak

Gure lerrokatze teknika bestelakoekin konparatzeko, domeinu, hizkuntza eta dokumentuen arteko antzekotasun maila ezberdineko hiru esperimentu egin dira.

Lehendabizi, EUROPARL corpusarekin esperimentuak egin dira lehen emai-

tza batzuk izateko dokumentu paraleloak lerrokatzen. Gero, WIKIPEDIA-ko dokumentuak lerrokatu dira BUCC 2015 atazan (Sharoff *et al.*, 2015) non beste sistema ugarik parte hartzen duten. Horretaz gain, Euskaraz idatzitako dokumentuekin probak egin dira albisteen domeinuan EITB corpusarekin. Euskararen kasua oso interesgarria da DOCAL-en emaitzak aztertzeko morfologikoki oso aberatsak diren hizkuntzetan. Azkenik, DOCAL-en partehartzea aztertzen da First Conference on Machine Translation (WMT 2016) Buck eta Koehn (2016a) konferentzian, non Internetetik lortutako dokumentuak lerrokatu behar diren.

Esperimentu guztietan konfigurazio berbera erabili da. DOCAL-ek ez du inolako entrenamendurik behar, beraz, konfigurazio bakarra $k = 5$ finkatzea da itzulpen lexikoen taulan 5 itzulpen onenak eskuratzeko. Kontrakoa adierazten ez bada itzulpen lexikoen taulak sortzeko GIZA++ aplikatu da JRC-ACQUIS COMMUNAUTAIRE corpusean[8].

3.2.1. EUROPARL

EUROPARL corpusa (Koehn, 2005) eskura dauden corpus paralelo nagusienetakoa da. Europako parlamentuko aktak biltzen ditu eta Europako hizkuntza ofizialetara itzulita dago itzultzaile profesionalen lanari esker. Beraz, kalitatezko corpusa izanik, dokumentu paraleloen lerrokatzea probatzeko baliabide aproposa da. Literaturan corpus honen bertsio ezberdinak erabili dira neurketak egiteko, horregatik EUROPARL corpusaren bi bertsio erabili dira proba hauetarako, 2. eta 5. bertsioak hain zuzen ere[9] (aurrerantzean EU2 eta EU5 deituko diegu).

EU5 corpusean, lehenagoko bertsioari tamaina oso txikiko dokumentu ugari gehitu zitzaizkion. Dokumentu hauen eragina ikertzeko asmoz EU5 corpusaren beste bertsio bat sortu da esaldi bakarra duten dokumentuak filtratuz, EU5.2 corpusa. EUROPARL corpusaren tamaina 3.1 taulan azaltzen da.

Esperimentu hau EUROPARL corpuseko dokumentuak lerrokatzean datza lor-

[8] 2015-eko Azaroko bertsioa erabili da. OPUS biltegian (Tiedemann, 2012) dago eskura: `http://opus.nlpl.eu/JRC-Acquis.php`

[9] 2016-ko otsailean eskura zegoen bertsioa erabili da: `http://www.statmt.org/europarl/`

3.1 taula: EUROPARL corpusaren dokumentu kopura.

CORPUSA	ES-EN		FR-EN		NL-EN	
	ES	EN	FR	EN	NL	EN
EU2	488	488	488	488	488	488
EU5	6.199	6.199	6.203	6.203	6.206	6.206
EU5.2	4.611	4.611	4.638	4.638	4.652	4.652

tutako emaitzak jatorrizko lerrokatzearekin konparatuz. Hizkuntza pareei dagokionez, honako hauek erabiltzen dira esperimentu honetan: gaztelania-ingelesa, frantsesa-ingelesa eta nederlandera-ingelesa. Hizkuntza pare horiek aukeratu dira bestelako sistemen emaitzak corpus horien gainean argitaratu direlako.

Konparaketa egiteko, Enright eta Kondrak-ek 2007-an aurkeztutako metodoa inplementatu da (Enright eta Kondrak, 2007). Bertan, dokumentuak bag-of-words teknikarekin prozesatzen dira 1-eko frekuentziako eta gutxienez 4 karaktereko terminoak erabiliz dokumentuak errepresentatzeko. Termino horien kopuru gehien partekatzen dituzten dokumentuak izango dira lerrokatzen direnak. Metodo honi HAPAX deituko diogu. Egindako inplementazioa konprobatzeko, gaztelania-ingelesa hizkuntza pareko dokumentuak lerrokatu dira eta lortutako emaitzak argitaratutakoekin alderatuta berberak dira.

3.2, 3.3 eta 3.4. tauletan laburbiltzen dira hiru hizkuntza pareetan lortutako emaitzak. Konparaketa egiteko aukeratutako metrikak doitasuna (*precision*), estaldura (*recall*) eta F1 dira[10].

EU2 corpusean bi metodoek oso emaitza onak izan dituzte. Izan ere, gaztelania-ingelesa hizkuntza parean bi sistemek hutsik dagoen dokumentu bakarrean egiten dute oker, eta ezberdintasun bakarra hizkuntzak nahastuta dituen dokumentu bakarrean dago. DOCAL-ek bakarri asmatzen du dokumentu horretan.

Beste corpus handiagoei dagokionez, HAPAX-en emaitzak nabarmen jaisten diren bitartean DOCAL-enak oso onak izaten jarraitzen dute . HAPAX-en

[10]Tesi guztian zehar emaitza onenak letra lodiz agertzen dira.

3.2 taula: Emaitzak EUROPARL gaztelania-ingelesa corpusean.

SISTEMA	CORPUSA	DOITASUNA	ESTALDURA	F1
HAPAX	EU2	99,6	99,6	99,6
DOCAL	EU2	**100,0**	**99,8**	**99,9**
HAPAX	EU5	54,2	54,2	54,2
DOCAL	EU5	**95,6**	**74,4**	**83,7**
HAPAX	EU5.2	72,9	72,9	72,9
DOCAL	EU5.2	**99,3**	**92,5**	**95,8**

3.3 taula: Emaitzak EUROPARL frantsesa-ingelesa corpusean.

SISTEMA	CORPUSA	DOITASUNA	ESTALDURA	F1
HAPAX	EU2	99,6	99,6	99,6
DOCAL	EU2	**100,0**	**99,8**	**99,9**
HAPAX	EU5	54,5	54,5	54,5
DOCAL	EU5	**95,7**	**72,6**	**82,6**
HAPAX	EU5.2	72,5	72,5	72,5
DOCAL	EU5.2	**99,2**	**91,0**	**94,9**

3.4 taula: Emaitzak EUROPARL nederlandera-ingelesa corpusean.

SISTEMA	CORPUSA	DOITASUNA	ESTALDURA	F1
HAPAX	EU2	99,6	99,6	99,6
DOCAL	EU2	**100,0**	**99,8**	**99,9**
HAPAX	EU5	50,3	50,3	50,3
DOCAL	EU5	**96,2**	**74,0**	**83,7**
HAPAX	EU5.2	67,2	67,2	67,2
DOCAL	EU5.2	**99,2**	**92,4**	**95,7**

arazoak bi arrazoirengatik izan daitezke: dokumentu kopurua hamar aldiz handiagoa delako edota dokumentu askoren tamaina oso txikia delako. EU5 corpusean lortutako emaitzek, Enright eta Kondrak-ek ateratako ondorioarekin kontrajarriz (EU2 corpusa erabili zuten beraien esperimentuetan), EUROPARL corpusa baliagarria izan daitekeela adierazten dute dokumentu paraleloen lerrokatzea ebaluatzeko.

EU5.2 corpusean berriz, non esaldi bakarra duten dokumentuak filtratu diren, bi metodoen emaitzek gora egiten dute. Hala ere, DOCAL-ek lerrokatze hobeak lortzen ditu: batezbestekoz 24,1 puntuko aldea F1 metrikan. Corpus hau sortzeko arrazoia errealitatean aurkitu daitezkeen dokumentuekin antzekotasun handiagoa izatea da, eta aldi berean, bilaketa eremua handitzeak dokumentuak lerrokatzeko metodoetan zein eragina duen aztertzea. Lortutako emaitzek erakusten dutenez, DOCAL-ek dokumentu kopurua handiekin ere lerrokatze onak lortzen ditu. HAPAX baino metodo egonkorragoa da beraz.

Patry eta Langlais-ek ere beraien PARADOX metodoa erabili zuten EU5 corpusean (Patry eta Langlais, 2011b). Hala ere, ez da erraza emaitzak DOCAL-ekin konparatzea. PARADOX-en esperimentuetan, Patry eta Langlais-ek okerrak dituzten dokumentuak kendu zituzten (encoding arazoak, gaizki amaitutako dokumentuak, etab.), eta corpusaren bertsio ezberdinekin egin zuten proba dokumentuak iragaziz esaldi kopuruaren arabera, betiere 10 esaldi baina gehiagoko dokumentuak hartuz. Frantses-ingelesa eta nederlandera-ingelesa hizkuntza pareetan 95 eta 93 puntuko emaitzak lortu zituzten hurrenez hurren F1 metrikan gehienez 100 esaldiko dokumentuak hartuz. DOCAL-ekin, esaldi bat baina gehiagoko dokumentu guztiak hartuz, 94,9 eta 95,7 puntuko emaitzak lortu dira hizkuntza pare berberetan.

3.2.2. BUCC 2015

2015-eko Building and Using Comparable Corpora ataza (BUCC 2015) lehendabizikoa izan zen dokumentuen antzekotasuna neurtzeko corpus amankomun bat sortzen. BUCC 2015 corpusa hizkuntza anitzetako dokumentuek osatzen dute eta guztira milioi erdi dokumentu baina gehiago daude (Sharoff *et al.*, 2015).

3.5 taula: BUCC 2015 corpusaren dokumentu kopurua.

CORPUSA	EN	FR	DE	ZH	DE-EN	FR-EN	ZH-EN
TEST	314.283	114.428	147.221	21.473	-	-	-
GOLD	-	-	-	-	147.515	114.802	21.473

Dokumentuak WIKIPEDIA-tik erauzi ziren hizkuntzen arteko estekak erabiliz betiere antzeko tamainako dokumentuak lerrokatzen zirela ziurtatuz. Dokumentu konparagarriak dira beraz. Atazaren helburua jatorrizko dokumentu bakoitzeko antzekotasun handien duten helburuko hizkuntzako bost dokumentu ematea da, antzekotasunaren arabera ordenatuta.

Nahiz eta atazan bost hizkuntza pare egon, hiru hizkuntza pare bakarrik erabili zituzten partehartzaileek. Hizkuntza pare horiek frantsesa-ingelesa, txinera-ingelesa eta alemana-ingelesa dira. Hizkuntza bakoitzeko dokumentu kopurua 3.5. taulan azaltzen da bai testerako eta baita gold estandarrerako.

Hiru metrika erabiltzen dira lerrokatzeak neurtzeko TREC[11] konferentziako estandarra jarraituz:

- SUCCESS@1. Jatorrizko hizkuntzako dokumentuetan dagokien helburuko dokumentua zein proportzioan aurkitzen den adierazten du.
- SUCCESS@5. SUCCESS@1 metrikaren antzekoa da, baina dagokien helburuko dokumentua bost antzekoenen artean dagoen ala ez neurtzen da. Berdin dio lehenengo ala azkeneko posizioan dagoen.
- MRR. *Mean Reciprocal Range*. Jatorrizko dokumentuen $1/N$-ren batezbestekoa, non N helburuko dokumentuaren posizioa den. Hau da, dokumentu baterako dagokion lerrokatzea 3. posizioan badago, orduan dagokion emaitza 1/3 izango da.

Hiru sistemek hartu zuten parte. LINA (Morin *et al.*, 2015) sistema HAPAX-etan oinarritua dago (Enright eta Kondrak, 2007). LINA-ren ekarpen

[11]Text Retrieval Conference: `https://trec.nist.gov/`

nagusia lerrokatze amankomunak ebazteko bi algoritmoak dira: *pigeonhole* arrazoiketa (LINA.P aurrerantzean) eta hirugarren hizkuntza batekiko esteken erabilera (LINA.CL aurrerantzean). Bigarren partehartzailea, CCNUNLP, Li eta Gaussier-en ikerlanetan oinarritua dago (Li eta Gaussier, 2013). Azkenik, AUT (Zafarian *et al.*, 2015) sisteman, dokumentuen bektoreak, dokumentuen izena, *Topic Modelling* eta itzulpen automatikoa erabiltzen dira dokumentuen antzekotasuna neurtzeko. AUT sistemak erroreak izan zituen dokumentuen prozesamenduan eta lortutako emaitzak ez ziren izan haiek esperotakoak. Hortaz, konparaketa honetan ez dugu AUT erabili.

Arestian aipatu bezala, lerrokatze onenaren optimizazioarekin emaitzak nabarmen hobetzen dira (ikus 3.1.4. azpiatala), baina aurrizki komun luzeenekin normalean ez dago alde gehiegirik (ikus 3.1.2. azpiatala). Horregatik, DOCAL-en lehenetsitako bertsioak lerrokatze onenaren optimizazioa egiten du, baina ez ditu aurrizki komun luzeenak erabiltzen. Konfigurazio horien pisua neurtzeko, DOCAL.EZLO bertsioa ebaluatzen da lerrokatze onenaren optimizazioa kenduta, eta DOCAL.EZLO.AKL bertsioa aurrizki komun luzeenekin.

3.6. taulan frantsesa-ingeleseko emaitzak aurkezten dira. Hizkuntza pare honetan DOCAL da emaitza onenak dituen sistema ia 20 puntuko aldearekin SUCCESS@1 metrikan. DOCAL-en bertsio guztiekin ere gainerako metrika guztietan emaitza hobeak lortzen dira. DOCAL.EZLO eta DOCAL.EZLO.AKLren arteko aldea txikiak dira, corpus honetarako aurrizki komun luzeenak erabiltzeak ia eraginik ez du beraz. Honek ez du zertan esan nahi aurrizki komun luzeenak ez duela balio, baliteke tamaina txikiagoko dokumentuekin emaitzak ezberdinak izatea.

3.7. taulan ikusten denez, aleman-ingelesa hizkuntza parean antzeko ondorioak atera ditzakegu. DOCAL-en bertsio guztiak dira onenak dituzten sistemak, eta begi-bistakoa da lerrokatze onenaren optimizazioak duen onura 20 puntu baino gehiagoko aldearekin.

Azkenik, 3.8. taulan txinera-ingelesa corpuseko emaitza aurkezten dira. Esperimentu honetarako MULTIUN corpusa erabili da itzulpen-taulak entrenatzeko (itzulitako Nazio Batuen dokumentuak (Eisele eta Chen, 2010)). MULTIUN corpusak 10 milioi esaldi pare baino gehiago ditu, horregatik, entre-

3.6 taula: Emaitzak BUCC 2015 frantsesa-ingelesa corpusean.

SISTEMA	SUCCESS@1	SUCCESS@5	MRR
LINA.P	30,0	37,4	32,9
LINA.CL	57,7	60,6	59,0
CCNUNLP	60,7	76,4	66,9
DOCAL.EZLO	63,3	69,3	65,7
DOCAL.EZLO.AKL	63,6	69,3	65,9
DOCAL	**79,5**	**79,5**	**79,5**

3.7 taula: Emaitzak BUCC 2015 aleman-ingelesa corpusean.

SISTEMA	SUCCESS@1	SUCCESS@5	MRR
LINA.P	24,9	35,5	29,0
LINA.CL	60,7	63,9	62,2
DOCAL.EZLO	62,1	68,8	64,9
DOCAL	**81,9**	**81,9**	**81,9**

namendua bizkortzearren, lehenengo 2 milioi esaldi pareak bakarrik erabili dira. Testu txinatarrak segmentatzeko berriz STANFORD SEGMENTER erabili da (Tseng *et al.*, 2005). Hau da hizkuntza pare bakarra non DOCAL-ek ez dituen emaitza onenak lortzen, 17 puntuko aldea baitago SUCCESS@5 metrikan, 7 puntuko aldea MRR metrikan, eta 2 puntu baina gutxiagoko aldea SUCCESS@1 metrikan. DOCAL bezala, CCNUNLP sistema informazio lexikoan oinarritzen denez interesgarria izango litzateke erabilitako itzulpen-taulek duten doitasuna eta estaldura aztertzea. Tamalez, ikerketa hori egin ahal izateko CCNUNLP sistemaren taulei buruzko informazioa falta da.

Beharrezkoa da aipatzea corpus txinatarraren segmentazioa egitea ez dela batere erraza gainerako hizkuntzetan egin den tokenizazioarekin alderatuta. Egindako segmentazioak nolabaiteko pisua du corpus txinatarrean izandako emaitzetan, eta hortaz, segmentazio mota ezberdinekin ikerketa egitea interesgarria izango litzateke. 3.3.1. azpiatalean zenbait optimizaziorekin lortutako emaitzak aurkezten dira nahiz eta segmentazio metodo berbera

3.8 taula: Emaitzak BUCC 2015 txinera-ingelesa corpusean.

SISTEMA	SUCCESS@1	SUCCESS@5	MRR
CCNUNLP	**71,0**	**86,1**	**76,9**
DOCAL.EZLO	56,2	60,2	57,8
DOCAL	69,6	69,6	69,6

erabili.

3.2.3. EiTB

Kasu honetan, morfologikoki oso aberatsak diren hizkuntzekin zer nolako emaitzak lortzen diren neurtu nahi da. EITB[12] corpusa konparagarritasun maila altua duten dokumentuen bilduma bat da. Berrien domeinuan albiste berberak jasotzen dituzten dokumentuak sortzen dira euskaraz eta gaztelaniaz, baina dokumentuak ez dira itzulpenak, lantalde ezberdinek sortzen baitituzte dokumentuak hizkuntzaren arabera[13] (xehetasun guztiak 4.2. azpiatalean).

Konparaketa egiteko EMACC (Ion *et al.*, 2011) eta DICTMETRIC (Pinnis *et al.*, 2012) sistemak erabili dira. Bi metodo hauek dokumentu paraleloak lerrokatzeko diseinatu ziren ACCURAT proiektuaren testuinguruan[14]. EMACC sistema *Expectation Maximization* algoritmoan oinarritzen da terminoak lerrokatzeko. Doitasun altua izateko diseinatu zen nahiz eta konputazionalki kostu handiko algoritmoa izan. DICTMETRIC berriz algoritmo arinagoa da eta itzulpen lexikoak, stemming, WORDNET-eko informazioa (Miller, 1995) eta kosinu antzekotasunean oinarritzen da.

Testerako corpus zatian eskuz lerrokatutako 299 dokumentu daude. Hiru sistemetan produktu kartesiarra erabili da dokumentuen konbinazio guztiak aztertzeko, eta konfigurazio guztia lehenetsitako balioekin utzi da. Itzulpentaulak GIZA++-ekin sortu dira 645.223 esaldi paraleloko IVAP (Instituto Vas-

[12]Euskal Irrati Telebista: `https://www.eitb.eus`

[13]Corpusa lortzeko jo esteka honetara: `http://metashare.elda.org/repository/search/?q=eitb+documents`

[14]`http://www.accurat-project.eu/`

3.9 taula: Emaitzak EITB-ko gaztelania-euskara corpusean.

SISTEMA	DOITASUNA	ESTALDURA	F1	DENBORA
DICTMETRIC	58,5	58,5	58,5	**0m07,431s**
EMACC	90,7	84,6	87,5	7m09,693s
DOCAL.EZLO	90,0	**90,0**	90,0	0m44,975s
DOCAL	**91,1**	89,3	**90,2**	0m44,110s

co de Administración Pública) corpusa erabiliz. IVAP corpusa administrazio publikoko dokumentuen bilduma bat da eta itzulpenak adituen bidez eskuz egin dira (ikus 4.3.1. azpiatala).

3.9. taulan erakusten dira emaitzak, doitasun, estaldura eta F1 metrikekin. Sistema guztiak makina berberean exekutatu dira (8/16 nukleo/hariko prozesatzailea eta 48 GB RAM) dokumentuak lerrokatzeko denbora neurtzeko asmoz.

Sistema azkarrena DICTMETRIC da. Hala ere, 58,5-eko emaitzak bakarrik lortzen ditu eta gainerako sistemek dokumentuak askoz hobeto lerrokatzen dituzte. EMACC eta DOCAL-ek emaitza onak lortzen dituzte, baina DOCAL hiru metriken arabera hobeto aritzen da 91,1-eko doitasun, 89,3-ko estaldura eta 90,2 puntuko emaitzarekin F1 metrikan. Exekuzio denborak konparatzeko orduan, DOCAL ere EMACC baino azkarragoa da: DOCAL-ek minutu bat baina gutxiago irauten du EMACC-ek 7 minutu baino gehiago behar dituen bitartean. Corpus honetako emaitzen arabera, DOCAL da sistema egokiena, bai exekuzio denboraren aldetik, eta baita lerrokatzeen kalitatearen aldetik.

Beste ondorio interesgarri bat lerrokatze onenaren optimizazioaren erabilera da. Orain arte egindako esperimentuetan lerrokatze onenaren optimizazioa erabiliz emaitzak asko hobetzen dira, baina ez da EITB corpusaren kasua. Corpusaren tamaina txikiagoa izanik lerrokatzeen bilaketa espazioa nahiko txikia da, eta beraz, lerrokatze onenaren optimizaziorik gabe DOCAL lerrokatze gutxiagotan nahasten da. Dena den, lerrokatze onenaren optimizazioaren erabilerak ez ditu emaitzak ia okertzen tamaina txikiko corpusetan.

Bukatzeko, DOCAL sistemak euskara bezala morfologikoki aberatsak diren

hizkuntzetan ere emaitza onak lortzen ditu. Esperimentu honetarako konfigurazioa prestatzeko ez da aparteko lanik egin, lehenetsitako parametroak erabiltzea nahikoa izan da.

3.2.4. WMT 2016

WMT 2016ko dokumentuen lerrokatze atazan[15] (Buck eta Koehn, 2016a), ingelesez eta frantsesez idatzitako dokumentu sorta bat ematen da eta helburua itzulpenak diren dokumentu pareak identifikatzea da. Dokumentuak Internetetik erauzi dira eta web-domeinuka antolatuta daude. Dokumentuen fitxategi izena beraien URL-a da, beraz, emaitza URL pareak dira.

Ebaluatzeko orduan, estaldura bakarrik neurtzen da, hortaz, okerrak diren lerrokatzeek ez dituzte azken emaitza okertzen. Hala ere, lerrokatzeekin kontuz ibili behar da, ebaluazioa egin baino lehen 1-1 erregela aplikatzen baita dokumentu pareetan, hau da, jatorrizko hizkuntzako dokumentu batek helburuko hizkuntzako dokumentu bakarrarekin lerrokatu daiteke. Adibidez, (d_1, d_2) eta (d_1, d_3) dokumentu pareetan 1-1 erregela aplikatu eta gero, (d_1, d_2) edo (d_1, d_3) dokumentu parea izango dugu, baina ez biak. 1-1 erregelarako hash taulak erabiltzen direnez, dokumentu pareen aukeraketa arbitrariotzat har genezake. Gold estandarrak 2.402 dokumentu pareko tamaina du.

Lau web-domeinutarako bakarrik erabili da dokumentuen indexazioa lerrokatze kandidatuak lortzeko, eta gainerakoetan konbinazio guztiak aztertu dira produktu kartesiarrarekin. 260 milioi konbinazio arte produktu kartesiarra egitea posible da erabilitako zerbitzarian[16]. Dokumentuen indexazioa erabiliz erroreak aztertzea zailagoa da, akatsak antzekotasun metrikarengatik edo indexazioarengatik gertatzen diren aztertu behar baita.

Atazako 21 sistemen emaitza guztiak Buck eta Koehn-en argitalpenean aurkitu daitezke (Buck eta Koehn, 2016a). 3.10. taulan DOCAL-en emaitza ofizialak ikus daitezke sailkapenean hobeto geratu diren sistemekin batera.

DOCAL 21 sistemetatik 5.na geratu da emaitza ofizialean; 2.128 lerrokatze zuzen aurkitzen ditu. Interesgarria da nola partehartzaile gehienek, bereziki

[15] `http://www.statmt.org/wmt16/bilingual-task.html`

[16] 8/16 nukleo/ariko prozesatzailea eta 64 GB RAM-eko zerbitzaria.

3.10 taula: WMT 2016ko emaitza ofizialak. Lerrokatze kopurua 1-1 erregela aplikatu ostean geratzen dena da.

SISTEMA	LERR. KOPURUA	LERR. ZUZENAK	ESTALDURA
NOVALINCS-URL-COV	235.812	2.281	95,0
YODA	318.568	2.256	93,9
UEDIN1_COSINE	368.260	2.140	89,1
NOVALINCS-COV	235.763	2.129	88,6
DOCAL	191.993	2.128	88,6

hobeto geratu direnak, dezente lerrokatze gehiago aurkitzen dituzten. Honek DOCAL-en doitasuna hobea dela adierazi lezake. Hala ere, doitasuna konparatzea ezinezkoa da, gold estandarretik kanpoko lerrokatzeak zuzenak diren banan-banan aztertu beharko lirateke eta.

Erroreen analisi bat egin eta gero, gutxienez 100 kasu aurkitu dira non gold estandarrean lerrokatze okerrak dauden (errore hauek 3.11. taulan azaltzen dira). Identifikatutako erroreetan DOCAL-ek dokumentu zuzena bilatzen du. Errore kopurua gold estandarraren %4,16-koa da, eta azken emaitzan eragin nabaria du: gold estandarra konpontzen bada DOCAL-ek 2.228 lerrokatze zuzen ditu %92.8-ko estaldurarekin (ikus 3.12. taula). Noski, baliteke gainerako sistemek ere arazo berbera edukitzea eta haien emaitzak ere hobeak izatea. Dena den, helburuko hizkuntzako dokumentuen antzekotasuna dela medio, atazako antolatzaileek beste proba bat egin zuten gold estandarrean adierazitako eta sistemak adierazitako lerrokatzeen artean %5 baino gutxiagoko diferentzia badago (edizio distantziaren bidez neurtuz) lerrokatze horiek zuzentzat hartzeko. Proba honek nolabait gold estandarrean aurkitutako erroreak konpentsatzen ditu, eta beraz, konparaketa hobeagotzat har genezake. Emaitza hauek 3.13. taulan erakusten dira.

Azken proba honetan DOCAL-en sailkapena hobetzen da, 4. postura igotzen da 93,1-eko estaldurarekin. Gold estandarreko erroreak konponduta 92,8-eko emaitza lortzen denez, proba honetan izandako hobekuntzak erroreen konponketagatik lortzen direla hausnartu daiteke.

Emaitza ofizialei helduz, bost sistema onenen konparaketa interesgarria da.

3.11 taula: WMT 2016ko gold estandarreko erroreak.

JATORRIZKOA	http://artfactories.net/Espace-Linga-Tere.html
GOLD ESTAND.	http://artfactories.net/-Republique-centrafricaine-.html
ZUZENA	http://artfactories.net/Espace-Linga-Tere-Bangui.html
JATORRIZKOA	http://www.ipu.org/hr-e/169/Co121.htm
GOLD ESTAND.	http://www.ipu.org/hr-f/168/Co121.htm
ZUZENA	Correct:http://www.ipu.org/hr-f/169/Co121.htm
JATORRIZKOA	http://www.lifegrid.fr/en/projets/projects/biomedicale-search.html
GOLD ESTAND.	http://www.lifegrid.fr/fr/projets/appel-a-projets-e-nnovergne-lifegrid-2006/recherche-biomedicale.html
ZUZENA	http://www.lifegrid.fr/fr/projets/31-recherche-biomedicale.html
JATORRIZKOA	http://www.nserc-crsng.gc.ca/Prizes-Prix/Excellence-Excellence/Profiles-Profilseng.asp?ID=1008
GOLD ESTAND.	http://www.nserc-crsng.gc.ca/Prizes-Prix/Herzberg-Herzberg/Profiles-Profilsfra.asp?ID=1003
ZUZENA	http://www.nserc-crsng.gc.ca/Prizes-Prix/Excellence-Excellence/Profiles-Profilsfra.asp?ID=1008
JATORRIZKOA	http://www.rfimusique.com/musiqueen/articles/060/article6465.asp
GOLD ESTAND.	http://www.rfimusique.com/musiquefr/articles/060/article14625.asp
ZUZENA	http://www.rfimusique.com/musiquefr/articles/060/article13250.asp
JATORRIZKOA	http://www.rfimusique.com/musiqueen/articles/129/article8397.asp
GOLD ESTAND.	http://www.rfimusique.com/musiqueen/articles/128/article18057.asp
ZUZENA	http://www.rfimusique.com/musiqueen/articles/129/article18094.asp
JATORRIZKOA	http://www.lalettrediplomatique.fr/contribution.php?choixlang=2&id=10&idrub=X
GOLD ESTAND.	http://www.lalettrediplomatique.fr/contribution.php?id=10&idrub=X
ZUZENA	http://www.lalettrediplomatique.fr/contribution.php?choixlang=1&id=10&idrub=X
X =	5, 7, 11-12, 15-17, 23, 28-31, 35, 37-39, 43, 45-46, 50-52, 56-58, 61-65, 69, 83-84, 86, 89, 91-94, 96-100, 103-107, 109-111, 114-115, 119-120, 123-125, 127-130, 133-135, 137-141, 144, 146, 149-152, 155-156, 158, 160-163, 165-167, 169, 173, 175, 177, 194, 197

3.12 taula: DOCAL-en WMT 2016ko emaitzak gold estandarreko 100 erroreak zuzenduta.

GOLD ESTANDARRA	LERR. KOPURUA	LERR. ZUZENAK	ESTALDURA
OFIZIALA	191.993	2.128	88,6
ZUZENDUTA	191.993	2.228	92,8

Haietatik bik atazako entrenamendurako dokumentuak erabiltzen dituzte: YODA (Dara eta Lin, 2016) eta UEDIN1 (Buck eta Koehn, 2016b). Entrenamendurako dokumentuak itzulpen automatiko bidez sortu dira, normalean egoera errealetan eskura ez daudenak, eta ez da batere praktikoa itzulpen automatiko bidez dokumentuak sortzea, eta are gehiago DOCAL bezalako metodo batekin antzeko emaitzak lortzen badira. Gainerako bi sistemak NOVALINCS-en (Gomes eta Lopes, 2016) bertsioak dira; NOVALINCS-COV, non segmentuen itzulpen-tauletan (ingelesez *phrase table*) oinarritzen den segmentuen arteko estaldura ratioak kalkulatzeko, eta NOVALINCS-URL-COV, URL-ak analizatzen diren bertsioa. Lehendabizikoak ia emaitzak berberak lortzen ditu DOCAL-ekin konparatuta, baina okerrago dabil antzeko lerrokatzeak onar-

3.13 taula: WMT 2016ko emaitzak antzeko lerrokatzeak onartzen badira.

SISTEMA	LERROK. ZUZENAK	ESTALDURA
YODA	2.307	96,0
NOVALINCS-URL-COV	2.303	95,9
UEDIN2_LSI-V2	2.281	95,0
DOCAL	2.235	93,1
NOVALINCS-COV	2.192	91,3

tzen badira. Bestea sistema onena da emaitza ofizialean URL-en prozesamenduak emandako bultzadari esker. Antzera, Germannek ataza honetan aurkeztutako sistema guztien artean, URL-ak erabiltzen dituenUEDIN2_LSI-V2 bertsioa da emaitza onenak dituena (Germann, 2017).

URL-etan oinarritutako heuristikoek dokumentuen edukiak analizatzen dituzten sistemak osatu ditzakete, eta ataza honetan emaitzak hobetzeko faktore garrantzitsua da. Hala ere, askotan ez da posible URL-ak erabiltzea, Internetetik erauzitako dokumentuekin bakarrik egin baitateke. Arrazoi honegatik interesgarria da dokumentuen edukietara mugatzen diren sistemak bakarrik aztertzea. 3.14. taulak, antzeko lerrokatzeak onartuz, dokumentuen edukietara mugatzen diren sistemen emaitzak aurkezten ditu (entrenamendurako dokumentuak erabiltzen dituzten sistemak ere agertzen dira taulan).

Ikus daitekeen bezala egoera horien pean DOCAL da sistema onena. Gainera, beste sistema batzuek baina metodologia sinpleagoarekin eta entrenamendurako dokumentuak erabili beharrik gabe. Segmentuen itzulpen-taularik ere ez da behar, sinpleagoak diren itzulpen-taulekin nahikoa baita. Hortaz, lortutako emaitzak gogobetekoak dira. Hala ere, 3.3. azpiatalean oraindik eta emaitza hobeagoak aurkezten dira optimizazio batzuen ostean.

3.3. Metodoaren Hobekuntzak

Aurreko 3.2. azpiatalean DOCAL-en eraginkortasuna frogatu da egoera ezberdineko zenbait esperimentutan. Arestian aipatu bezala, gure sistemaren lehenetsitako konfigurazioa erabili da; JRC corpusaren bitartez sortutako

3.14 taula: WMT 2016ko emaitzak antzeko lerrokatzeak onartzen badira. Dokumentuen edukietara mugatzen diren sistemak bakarrik hartzen dira kontutan.

SISTEMA	LERROK. ZUZENAK	ESTALDURA
DOCAL	2.235	93,1
NOVALINCS-COV	2.192	91,3
UEDIN1_COSINE	2.140	89,1
YSDA	2.102	87,5
BADLUC	2.062	85,9
UFAL1	2.060	85,8
MEDVED	1.986	82,7
UFAL2	1.954	81,4
ADAPT	726	30,2
ADAPT-V2	733	30,5
JIS	48	2,0

itzulpen-taulak erabiltzen ditu eta tokenizazioa kenduta testuaren gainean inolako aurreprozesamendurik ez da egiten. Atal honetan ordea, jatorrizko metodoari zenbait optimizazio egiten zaizkio eta esperimentu berdinetan probatzen dira lehenetsitako konfigurazioarekin konparatuz. Azpiatal honen helburua optimizazio bakoitzaren berezitasunak eta ekarpenak aztertzea da.

3.3.1. Pisu Lexikoak

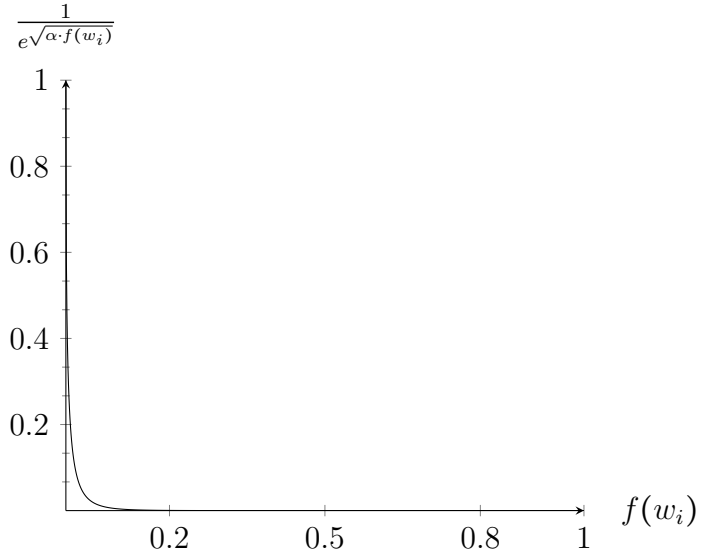
Jatorrizko antzekotasun metrikarekin, termino multzo guztietan (jatorrizkoak eta hedatutakoak) termino guztiek pisu berbera dute. Nahiz eta intuitiboki zentzua izan termino bakoitzak bere klasearen arabera (adibidez, hitz lexikoak eta funtzionalak, izen-arruntak eta determinatzaileak, zenbakiak, etab.) pisu ezberdina izatea, benetan eraginkorra den pisuak esleitzeko metodo bat diseinatzea ez da erraza. Estandarrak diren TF-IDF metodoan oinarritutako teknika batzuek stop-word listak behar dituzte, eta hizkuntzaren arabera zaila izan liteke lista hori lortzea. Beste aukera bat stop-word listak eskuz sortzea da, baina termino bat lista horretan sartu behar den edo ez erabakitzea konplexua izan daiteke. Beste irtenbide bat dokumentuko termi-

noak esanguratsuak diren ala ez determinatzea da metodo ez-gainbegiratuen bidez (Gelbukh *et al.*, 2010), zeintzuk terminoen maiztasunak konparatzen dituzten aplikazio domeinuan eta domeinu generikoen artean. Metodo hauekin ordea, pisuen arteko oreka egokia finkatzea ez da erraza eta termino konkretu batzuen agerpenak gehiegi baldintzatu lezake antzekotasun metrika.

Ikerlan honetan bestelako irtenbide bat bilatu da. Mikolov-en ikerlanean inspiratuta (Mikolov *et al.*, 2013). Bertan, ohiko termino eta termino arraroen arteko oreka bi aldagaiko funtzio baten bidez kontrolatzen da: enpirikoki ezarritako atalase bat eta terminoaren maiztasuna. Gure metodorako 3.2. ekuazioan erakusten den funtzioa erabiltzen da, non $f(w_i)$ w_i terminoaren maiztasun erlatiboa den, eta α kurbaren leuntasuna kontrolatzeko parametroa. 3.2. irudian funtzioaren itxura erakusten da $\alpha = 300$ parametroarekin[17].

$$W(w_i) = \frac{1}{e^{\sqrt{\alpha \cdot f(w_i)}}} \tag{3.2}$$

3.2 irudia: Pisu lexikoen funtzioa $\alpha = 300$ parametroarekin.



[17]Hau lehenetsitako balioa egindako esperimentuetan.

Funtzio honek gure beharretarako egokiak diren propietateak ditu. Lehendabizi, funtzioko balioak 0-tik 1-era doaz. Hau jatorrizko metodoarekin bateragarria da, terminoak termino multzoaren partaide izatea baita antzekotasun metrikaren muina. Bigarrenik, termino lexikoei 1-etik hurbil dauden pisuak esleitzen zaizkie hitz funtzionalei oso pisu baxuak ematen zaizkien bitartean. Pisuen kalkulua hizkuntza bakoitzeko egiten da, hau da, jatorrizko hizkuntzako pisuak kalkulatzeko jatorrizko dokumentu guztiak elkartzen dira eta termino guztien maiztasun erlatiboa kalkulatzen da, eta helburuko hizkuntzako pisuak kalkulatzeko gauza berdina egiten da helburuko dokumentuekin.

Dokumentuen antzekotasuna neurtzeko pisu lexikoak integratu behar dira jatorrizko Jaccard antzekotasun formulan. Izan bedi S termino multzoa eta T itzulitako termino multzoa, 3.3. ekuazioan pisu lexikoak nola integratzen diren erakusten da. Azkenik, $docal_w$ antzekotasun metrika berria 3.4. ekuazioan azaltzen den bezala kalkulatzen da.

$$WJ(T,S) = \frac{\sum\limits_{w_m \in \{T \cap S\}} W(w_m)}{\sum\limits_{w_n \in \{T \cup S\}} W(w_n)} \tag{3.3}$$

$$docal_{lex}(d_i,d_j) = \frac{1}{2}\left(WJ(T_{ij},S_j) + WJ(T_{ji},S_i) \right) \tag{3.4}$$

Nahiz eta pisu lexikoen kalkulua bestelako corpusetan kalkulatu ahal izan, tesi honetako esperimentu guztietan lerrokatu behar diren dokumentuen gainean kalkulatzen dira pisuak. Kalkulua horrela eginda pisu lexikoak esperimentuko domeinura egokituta egongo dira.

3.15. taulan jatorrizko metodoarekin konparaketa egiten da BUCC 2015 atazan[18]. Hiru hizkuntza pareetan oinarrizko metodoa hobetzen da nahiz eta hobekuntzak nahiko txikiak izan. Normalean, dokumentuen hitz kopurua nahiko handia da, informazio lexiko nahikoarekin, eta tamaina honek nolabait berdintzen du hitz funtzionalek ekar lezakeen desoreka pisu lexikoak

[18]Lerrokatze onenaren optimizazioarekin dokumentu bat beste batekin bakarrik lerrokatu daitekeenez, SUCCESS@1, SUCCESS@5 eta MRR metriketan lortzen diren emaitzak berberak dira.

erabili behar izan gabe.

3.15 taula: Pisu lexikoen eragina BUCC 2015 atazan.

SISTEMA	HIZKUNTZA PAREA	SUCCESS@1	Δ
DOCAL	FR-EN	79,52	-
DOCAL.LEX	FR-EN	**80,63**	+1,11
DOCAL	DE-EN	81,92	-
DOCAL.LEX	DE-EN	**82,78**	+0,86
DOCAL	ZH-EN	69,55	-
DOCAL.LEX	ZH-EN	**70,24**	+0,69

3.16, 3.17 eta 3.18. tauletan EUROPARL corpusean lortutako emaitzak ikus daitezke. BUCC 2015 corpusean gertatzen den bezala, emaitza guztiak hobetzen dira alde txikiarekin. DOCAL-en emaitzak oso onak ziren (+80 puntu hizkuntza pare guztietan corpusaren bertsio guztietan) eta ez dago tarterik hobekuntza altuak izateko.

Azkenik, pisu lexikoak EITB corpusean probatu dira 3.19. taulan erakusten diren emaitzak lortuz. Corpus honetan pisu lexikoek emaitzak hobetu ez ezik, zertxobait okerragoak dira. Corpus honen kasua nahiko berezia da dokumentu kopuru txikiarengatik. Pisu lexikoak kalkulatzeko informazio gutxi dagoenez, pisuen balioa ez da hain zehatza eta lerrokatzeak ez dira hain onak.

Emaitza guztiak ikusita, pisu lexikoek berez oso onak diren lerrokatzeak are gehiago hobetu ditzakete. Salbuespen bakarra EITB corpusa da, non dokumentu kopuru txikiarengatik pisuak ez diren hain zehatzak. Hobekuntza potentzial handiena beraz, dokumentu txiki askoko aplikazio domeinuetan dago. Bestalde, pisu lexikoen kalkulua hizkuntzarekiko independentea da, ez dira hizkuntza baliabide gehiago behar, eta prozesamendu denborak ez du gainkarga handirik. Propietate hauek eramangarritasun helburuarekin guztiz bat datoz, eta lerrokatze arazoaren testuinguruaren arabera, pisu lexikoak baliagarriak izan daitezke.

3.16 taula: Pisu lexikoen eragina EUROPARL gaztelania-ingelesa corpusean.

SISTEMA	CORPUSA	DOITASUNA	ESTALDURA	F1
DOCAL	EU2	**100,0**	**99,8**	**99,9**
DOCAL.LEX	EU2	**100,0**	**99,8**	**99,9**
DOCAL	EU5	**95,6**	74,4	83,7
DOCAL.LEX	EU5	**95,6**	**75,2**	**84,2**
DOCAL	EU5.2	**99,3**	92,5	95,8
DOCAL.LEX	EU5.2	**99,3**	**93,3**	**96,2**

3.17 taula: Pisu lexikoen eragina EUROPARL frantsesa-ingelesa corpusean.

SISTEMA	CORPUSA	DOITASUNA	ESTALDURA	F1
DOCAL	EU2	**100,0**	**99,8**	**99,9**
DOCAL.LEX	EU2	**100,0**	**99,8**	**99,9**
DOCAL	EU5	95,7	72,6	82,6
DOCAL.LEX	EU5	**96,0**	**74,2**	**83,7**
DOCAL	EU5.2	99,2	91,0	94,9
DOCAL.LEX	EU5.2	**99,3**	**92,7**	**95,9**

3.18 taula: Pisu lexikoen eragina EUROPARL nederlandera-ingelesa corpusean.

SISTEMA	CORPUSA	DOITASUNA	ESTALDURA	F1
DOCAL	EU2	**100,0**	**99,8**	**99,9**
DOCAL.LEX	EU2	**100,0**	**99,8**	**99,9**
DOCAL	EU5	96,2	74,0	83,7
DOCAL.LEX	EU5	**96,5**	**75,1**	**84,5**
DOCAL	EU5.2	**99,2**	92,4	95,7
DOCAL.LEX	EU5.2	**99,2**	**93,6**	**96,3**

3.19 taula: Pisu lexikoen eragina EITB gaztelania-euskara corpusean.

SISTEMA	DOITASUNA	ESTALDURA	F1
DOCAL	**91,1**	**89,3**	**90,2**
DOCAL.LEX	90,4	88,6	89,5

3.3.2. Itzulpen-Taulak

DOCAL-ek itzulpen-taulak behar ditu dokumentuen termino multzoak itzultzeko, hau izanik behar den hizkuntza baliabide bakarra. Oinarrizko esperimentuetarako JRC corpusa erabili da itzulpen-taulak sortzeko. Azpiatal honetan itzulpen-taularen garrantzia neurtzen da itzulpen-taula generiko handiagoekin konparatuz. Helburua itzulpen multzo hobeak sortzea da itzulpen-taulen estaldura lexikoa handituz. OPUS biltegian (Tiedemann, 2012) eskura dauden corpusak erabili dira itzulpen-taula generikoak sortzeko.

Erabili daitezkeen corpusen tamaina oso ezberdina izan daiteke, eta honek domeinu batzuek besteek baino pisu handiagoa izatea dakar. Itzulpen-taula generikoak sortzeko, domeinu bakoitzeko esanguratsuenak diren esaldiak erabili dira esaldi kopuru jakin bat lortu arte. Behean azaltzen da nola sortzen den corpus generikoa.

Corpus bakoitzeko, lehendabizi corpuseko esaldi paraleloak sailkatzen dira beren perplexitatearen arabera, txikienetik handienera. Perplexitateak lortzeko 5-gramako bi hizkuntza eredu bat entrenatzen dira corpusaren zati elebakarrak erabiliz[19]. Esaldi paraleloaren perplexitatea helburuko hizkuntzako eta helburuko hizkuntzako perplexitatearen batezbestekoa eginez kalkulatzen da.

Azken corpus generikoa sortzeko, aurreko pausuan sailkatutako corpus bakoitzetik lehenengo n esaldi pareak hartzen dira. n balioaren ezarpena kontu handiz egin behar da; oso altua bada, corpus txikienen lexikoak ez du eraginik izango, eta oso txikia bada, corpus generikoaren tamaina txikiegia izango da. Gure esperimentuetarako n = 500.000 da, corpus bakoitzaren garrantziaren eta corpus generikoaren tamainaren arteko oreka egokia ematen baitu. 3.20. taulak corpus generikoa sortzeko erabilitako corpusak eta aukeratutako esaldi pare kopurua azaltzen du. Itzulpen-taula generikoak sortzeko 1.692.551 eta 1.969.494 esaldi pareko corpus generikoak sortu dira aleman-ingeleserako eta frantses-ingeleserako hurrenez hurren.

DOCAL-en oinarrizko emaitzekin konparatzeko egin den lan bakarra itzulpen-

[19] KENLM bidez entrenatzen dira hizkuntza ereduak (Heafield, 2011)

3.20 taula: Corpus generikoaren esaldi pare kopurua.

CORPUSA	DE-EN		FR-EN	
	GUZTIRA	HAUKERATUAK	GUZTIRA	HAUKERATUAK
OPENSUBS	11.473.328	500.000	28.024.360	500.000
MULTIUN	103.490	103.490	9.142.161	500.000
EUROPARL	1.776.292	500.000	1.826.770	500.000
JRC	449.818	449.818	708.896	316.327
TED	138.243	139.243	153.167	153.167
GENERIKOA	13.941.171	1.692.551	39.855.354	1.969.494

3.21 taula: Itzulpen-taula generikoen eragina BUCC 2015 atazan.

SISTEMA	HIZKUNTZA PAREA	SUCCESS@1	Δ
DOCAL	FR-EN	79,52	-
DOCAL.LEX	FR-EN	80,63	+1,11
DOCAL.LEX.GEN	FR-EN	**82,64**	+2,01
DOCAL	DE-EN	81,92	-
DOCAL.LEX	DE-EN	82,78	+0,86
DOCAL.LEX.GEN	DE-EN	**84,45**	+1,67

taulen aldaketa da[20]. 3.21. taulak erakusten duenez, BUCC 2015 atazan, bai frantses-ingelesa eta bai aleman-ingelesa corpusetan nahiz eta alde handiegia ez egon (ia puntu bateko aldea), emaitzak beti hobeak dira. Dokumentuen hitz kopurua altua izan ohi da, eta pisu lexikoekin gertatzen den antzera (ikus 3.3.1. azpiatala), termino bat itzulpen-taulan ez agertu harren, gainerako terminoen agerpenek informazio galera hori berdindu dezakete. Hala ere, pisu lexikoen eta itzulpen-taula generikoen hobekuntzak pilatzen dira, beraz, bi teknika hauen konbinazioa onuragarria da.

Itzulpen-taula generikoak WMT 2016 atazan ere erabili dira; 3.22 eta 3.23. tauletan daude emaitzak. Kasu honetan itzulpen-taula generikoen erabilera

[20]Taula generikoaren erabilera .GEN atzizkiarekin adierazten da, eta emaitzen diferentzia DOCAL sistemarekiko konparatzen da.

3.22 taula: Itzulpen-taula generikoen eragina WMT 2016 atazan.

SISTEMA	LERROK. ZUZENAK	ESTALDURA	Δ
DOCAL	2.128	88,6	-
DOCAL.GEN	2.175	**90,5**	+1,9
DOCAL.LEX.GEN	2.163	90,0	+1,4

3.23 taula: Itzulpen-taula generikoen eragina WMT 2016 atazan, gold estandarrean erroreak dituzten 100 lerrokatzeak zuzenduta.

SISTEMA	LERROK. ZUZENAK	ESTALDURA	Δ
DOCAL	2.228	92,8	-
DOCAL.GEN	2.275	**94,7**	+1,9
DOCAL.LEX.GEN	2.263	94,2	+1,4

konparatu da pisu lexikoekin eta gabe. BUCC 2015-eko esperimentuan bezala, itzulpen-taulekin emaitza guztiak hobetzen dira (+1.9 puntu), baina kasu honetan emaitza onenak pisu lexikoak erabili gabe lortzen dira. Emaitza hauekin, DOCAL.LEX.GEN sistema hirugarren posiziora igoko litzateke emaitza ofizialetan WMT 2016erako espreski egindako bi sistema onenekin zuzenean lehiatuz.

Esperimentu guztiak egin eta gero, itzulpen-taula generikoen erabilerarekin emaitza hobeak lortzen direla ondorioztatu dezakegu, gainera, inolako konfigurazio berezirik ez da egin. Domeinurako bereziki diseinatutako itzulpen-taulak sortuz litekeena da emaitzak hobetzen jarraitzea, baina lan astuna izan liteke eta ez dator bat DOCAL-en eramangarritasun helburuarekin.

3.3.3. Dokumentuen Indexazioa

Arestian aipatu bezala, dokumentu multzo handiak lerrokatu behar diren egoeratan bilaketa espazioa handiegia izan daiteke produktu kartesiarraren bidez dokumentu konbinazio guztiak aztertzeko, eta horregatik kasu hauetan indexazioa erabiltzea komeni da. Dokumentuen indexazioan jatorrizko hizkuntzako dokumentu bakoitzeko helburuko hizkuntzako 100 dokumentu ego-

kienak berreskuratzen dira. Dokumentuen indexazioa eta galdeketa APACHE LUCENE-ren bidez egiten da.

Noski, LUCENE-ren bidez egindako galdeketetan dokumentu egokiak aurkitzen ez badira azken emaitzak ez dira onak izango. Azpiatal honetan dokumentuak indexatzeko eta berreskuratzeko beste metodo batzuk aztertzen dira ea lerrokatze hobeak lortzen diren konprobatzeko asmoz.

Lehenetsitako konfigurazioan, LUCENE-k TF-IDF metodoan oinarritzen den algoritmo bat erabiltzen du dokumentuak berreskuratzeko. Beste ohiko aukera bat OKAPI BM25 metodoa da (Jones *et al.*, 2000), bertan dokumentuen tamaina eta terminoen maiztasuna bezalako propietateak hartzen dira kontutan antzekotasun probabilitateak kalkulatzeko.

Berreskuratutako dokumentuen antzekotasuna hobetzeko beste aukera bat dokumentuak indexatzeko orduan tamainaren araberako dokumentuen kategorizazio zehatzago bat egitea da. DOCAL-en lehenetsitako konfigurazioan dokumentuak bi atributu erabiliz kategorizatzen dira: *small* eta *large*. Izan bedi l dokumentuaren tamaina, $\bar{x}$ batezbesteko aritmetikoa eta σ desbiderapen estandarra, bi atributuak honela kalkulatzen dira: $small : l \leq (\bar{x} + \sigma)$ eta $large : l \geq (\bar{x} - \sigma)$.

Dokumentuen kategorizazio zehatzago bat egiteko SSLL sailkapena diseinatu da. Bertan, bi atributu erabili ordez lau atributu erabiltzen dira. 3.24. taulan azaltzen dira. Optimizazio sinple honek ez du gainkarga berezirik oinarrizko metodoan, eta aldi berean, dokumentu antzekoagoak berreskuratzen lagundu lezake.

BM25 eta SSLL tekniken ekarpena neurtzeko BUCC 2015 atazako frantses-ingeles eta txinera-ingeles corpusak erabili dira. 3.25. taulan azaltzen dira emaitzak[21]. Ohiko SUCCESS@1 metrikaz aparte, LUCENE metrika azaltzen da indizetik lerrokatze egokia zein ehunekotan berreskuratzen den jakiteko.

Bi hizkuntza pareetan, BM25 bidez dokumentu egokia nahiko ehuneko handiagoan aurkitzen da, %10-etik eta %20-erainoko hobekuntzarekin. Frantses-ingeles hizkuntza parean SSLL ere probatu da, baina ez du lortzen BM25 metodoak bezain besteko onura dokumentu egokia berreskuratzeari dagokionez.

[21]Emaitzen diferentzia DOCAL sistemarekin konparatzen da.

3.24 taula: Dokumentuak indexatzeko eta berreskuratzeko SSLL metodoa. Dokumentuak tamainaren araberako lau atributurekin kategorizatzen dira.

ATRIBUTUA	BALDINTZA
small	$l \leq (\bar{x} - \sigma)$
smallish	$l \leq (\bar{x} + \sigma)$
largish	$l \geq (\bar{x} - \sigma)$
large	$l \geq (\bar{x} + \sigma)$

3.25 taula: Emaitzak BUCC 2015 atazan indexazio metodoa aldatuz.

SISTEMA	Hizk. Parea	SUCCESS@1	LUCENE (%)	Δ
DOCAL.LEX.GEN	FR-EN	82,64	77,53	-
DOCAL.LEX.GEN.BM25	FR-EN	80,97	**88,31**	-1,67
DOCAL.LEX.GEN.SSLL	FR-EN	**83,13**	79,38	+0,49
DOCAL.LEX	ZH-EN	70,24	61,43	-
DOCAL.LEX.BM25	ZH-EN	**71,27**	**80,99**	+1,03

Kasu guztietan, BM25 metodoaren bidez dokumentu zuzen gehiago berreskuratzen dira, baina hobekuntza honek ez du eraginik azken lerrokatzeetan. Txinera-ingeles corpusean bakarrik hobetzen dira emaitzak, baina puntu bateko aldearekin bakarrik. SSLL-rekin berriz, indexazioa ez da gehiegi hobetzen, baina lerrokatze emaitzak gora egiten du puntu erdiko onurarekin. Emaitza irregular hauen arrazoia berreskuratutako dokumentuen antzekotasunean egon liteke, DOCAL-ek ezberdintasun gutxiago aurkitzen baititu dokumentuen artean, eta lerrokatze okerrak diskriminatzea zailagoa da. Emaitzen aldakortasuna ikusita, lehenetsitako indexazio metodoa erabiltzen jarraitzen da gainerako esperimentuetan. Etorkizunari begira, esperimentu gehiago egin daitezke indexazio algoritmoaren gainean egindako optimizazioei buruz. Behar bada, dokumentuen antzekotasun metrika hobetuz BM25 metodoak dituen onurak dokumentuen lerrokatzeetan eragin handiagoa izan dezake.

3.3.4. Testuaren Aurreprozesamendua

3.1.1. azpiatalean, oinarrizko metodoa ahalik eta arinen mantentzearren, tokenizazioa alde batera utzita inongo aurreprozesamendurik egiten ez dela aipatzen da. Azpiatal honetan ordea, truecasing-aren eragina aztertzen da hasierako erabakia zuzena den ala ez eztabaidatzeko asmoz.

Truecasing-a esaldiko lehenengo terminoaren forma probabilitate handieneko formarekin aldatzean datza; terminoa letra xehez hasteak probabilitate handiagoa badu, lehenengo terminoa ere letra xehez hasiko da. Honek dokumentuen lexikoaren tamaina eta datuen sakabanaketa murrizten du.

Lehenetsitako konfigurazioan truecasing ez egiteko erabakia bi arrazoirengatik motibatuta dago. Alde batetik, honek gainkarga bat dakar dokumentuen aurreprozesamenduan; lehendabizi terminoen formen probabilitateak kalkulatu behar dira, eta gero truecasing-a egin behar da dokumentuko esaldi guztietan. Gainera, truecasing metodoaren arabera, truecasing-a egin aurretik dokumentuko paragrafoak esalditan zatitu behar dira. Dokumentu kopuruaren arabera lan astuna izan liteke. Noski, aurreprozesamendua lerrokatze prozesutik kanpo egin liteke, baina ala ere metodoaren eramangarritasuna nolabait oztopatzen da. Beste alde batetik, demagun esaldi batek batezbes-

tekoz 20 termino dituela, orduan, dokumentuko terminoen %5-a aldatuko da gehienez, beraz, truecasing-aren eragina mugatua da.

Truecasing-a egiteko honako bi pausu hauek jarraitu dira:

1. Dokumentuetako paragrafoak esalditan banatzen dira MOSES-ko (Koehn *et al.*, 2007) *split-sentences.perl* scriptaren bidez[22].

2. Dokumentuetako esaldietan truecasing-a egiten da MOSES-ko *truecase.perl* scripta erabiliz[23]. Script hori erabili aurretik eredu bat entrenatu behar da terminoen formen probabilitateak biltzen dituena. Eredu hori entrenatzeko dokumentu elebakarrak erabiltzea posiblea da, baina atazaren arabera baliteke dokumentu gutxi izatea eta entrenatutako probabilitateak kaxkarrak izatea. Esperimentu hauetarako OPUS biltegiko corpusekin[24] lehendik entrenatutako ereduak erabili dira.

Probatarako BUCC 2015 eta WMT 2016 corpusak erabiliko dira, non nahiz eta DOCAL-en oinarrizko bertsioak emaitza onak izan, badago tartea hobekuntzarako.

BUCC 2015 corpuseko emaitzak 3.26. taulan azaltzen dira, bai corpus ofizialean eta baita 100 lerrokatze okerrak zuzenduta dituen corpusean. Truecasing aurreprozesamendua .TC atzizkiak adierazten du. Bi hizkuntza pareetan emaitzak pixka bat okerragoak dira hasieran egindako hausnarketa balioztatuz. Hala ere, truecasing-arekin ere emaitzak onak izaten jarraitzen dute beherapen bat izan harren.

WMT 2016 corpusean izandako emaitzak 3.27. taulan azaltzen dira. Konparaketa errazteko, DOCAL-en lehenetsitako bertsioa eta konfigurazio onena duen bertsioa ere azaltzen dira. Emaitzen diferentziak jatorrizko metodoarekiko kalkulatzen dira.

[22] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/ems/support/split-sentences.perl`

[23] `https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl`

[24] `http://opus.nlpl.eu/`

3.26 taula: Emaitzak BUCC 2015 atazan truecasing eginez.

SISTEMA	HIZK. PAREA	SUCCESS@1	Δ
DOCAL.LEX.GEN	FR-EN	**82,64**	-
DOCAL.LEX.GEN.TC	FR-EN	82,40	-0,24
DOCAL.LEX.GEN	DE-EN	**84,45**	-
DOCAL.LEX.GEN.TC	DE-EN	84,30	-0,15

3.27 taula: Emaitzak WMT 2016 atazan truecasing eginez.

SISTEMA	GOLD ESTANDAR	LER. ZUZENAK	ESTALDURA	Δ
DOCAL	OFIZIALA	2.128	88,6	-
DOCAL.GEN	OFIZIALA	2.175	90,5	+1,9
DOCAL.GEN.TC	OFIZIALA	2.253	**93,8**	+5,2
DOCAL	ZUZENDUA	2.228	92,8	-
DOCAL.GEN	ZUZENDUA	2.275	94,7	+1,9
DOCAL.GEN.TC	ZUZENDUA	2.313	**96,3**	+3,5

BUCC 2015 corpusean gertatzen den ez bezala, truecasing-ak eragin handia du corpus ofizialean eta zuzendutako corpusean (+3,3 eta +1,6 puntu hurrenez hurren). Corpusa aztertzen bada, dokumentuek tamaina oso txikiko esaldi asko dituztela ikus daiteke (termino pare batekoak), eta gainera esaldi horiek beraien artean antzekotasun handia dute. Emaitzak ikusita, zenbait kasutan esaldietako lehen terminoa funtsezkoa izan da antzeko dokumentuak ezberdintzeko. Arrazoi honengatik, truecasing-aren eragina tipologia honetako dokumentuetara mugatzen dela ondorioztatu genezake, salbuespeneko kasu bezala har daitekeena, zeren eta normalean dokumentuetako esaldiek informazio gehiago dute.

Itzulpen-taula generikoak eta truecasing-a erabilita, DOCAL hirugarren sistema da emaitza ofizialetan bigarrenarekiko 0,1 puntuko aldearekin. DOCAL-en oinarrizko bertsioak izandako emaitzak ikusita, antzeko lerrokatzeak onartuz lortzen diren emaitzak eta guk zuzendutako gold estandarrean lortzen diren emaitzak oso antzekoak dira (ikus 3.2.2. azpiatala). Zuzendutako gold es-

tandarrean DOCAL.GEN.TC sistemak 96,3 puntu lortu ditu eta YODA-k 96,0 antzeko lerrokatzeak onartzen badira. Bi emaitza hauek ezin dira zuzenean konparatu, baina bai erakusten dutela DOCAL.GEN.TC oso konpetitiboa dela WMT 2016 corpusean, eta are gehiago DOCAL-ek dokumentuen URL-ak eta metadatuak erabiltzen ez dituela kontutan hartzen badugu.

3.4. Ondorioak

3.1. atalean dokumentuak lerrokatzeko metodo eraginkor bat deskribatzen da, guztiz erabilgarria dena dokumentu gutxi edo ugari dauden egoeretan, eta baita dokumentu paralelo edo konparagarriak dituzten corpusetan.

Aurkeztutako metodoa, DOCAL, dokumentuen edukian bakarrik oinarritzen da. Honako teknika eta baliabideak erabiltzen ditu:

- Jaccard koefizientea.
- Itzulpen-taulen bidez lortutako itzulpen lexikoak.
- Termino multzoen hedapena aurrizki komunak eta izen-entitateak bilatuz.
- APACHE LUCENE-ren erabilpena dokumentuen bilaketa espazioa mugatzeko.
- Lerrokatze onenaren optimizazioak dokumentu bat beste batekin bakarrik lerrokatuko dela bermatzen du.

Teknika guzti horien konbinazioarekin aldi berean oso eraginkorra eta sinplea den metodoa lortzen da. Izan ere, konplexuagoak diren metodoen emaitzak berdindu edo hobetzea lortzen da. Proposatutako metodoa edozein dokumentu kopuruko atazetara moldatu daiteke produktu kartesiarraren bidez dokumentu konbinazio guztiak aztertuz, edo lehen iragazpen bat eginez helburuko hizkuntzako dokumentuak indexatuz eta antzekoenak direnak berreskuratuz.

DOCAL dokumentuen lerrokatzerako ataza ugaritan probatu da. Propietate oso ezberdinetako sei hizkuntzatako dokumentuak erabili dira: alemana, ingelesa, hizkuntza erromanikoak, txinera eta euskara, morfologia aberatseko hizkuntza dena. Ataza hauek benetako egoeratan aurkitu litezkeen dokumentuak dituzte: EUROPARL eta WMT 2016 corpusetako dokumentu paraleloak, eta EITB eta BUCC 2015 corpusetako dokumentu konparagarriak. Orokorrean, esperimentu ugari egin dira bestelako sistemekiko bidezko konparaketak eginez.

Esperimentu konfigurazio honen pean, lehendabizi EUROPAL corpuseko dokumentu paraleloekin egin da proba. Corpus honetan DOCAL beste metodoak baino eraginkorragoa dela ikusi dugu. Hurrengo esperimentuan BUCC 2015 atazako WIKIPEDIA-ko dokumentu konparagarriak lerrokatu dira. Frantses-ingeles eta aleman-ingeles hizkuntza pareetan DOCAL-ek lortu ditu emaitza onenak nahiko alde handiarekin, baina txinera-ingeles corpusean gutxigatik sistema onenaren atzetik geratu da. Euskara-gaztelaniari dagokionez, EITB corpusean DOCAL-ek emaitza onenak lortu ezezik konputazionalki ere oso metodo eraginkorra dela erakutsi du. DICTMETRIC azkarragoa da dokumentuak lerrokatzen, baina bere emaitzak nahiko kaxkarrak dira. Azkenik, DOCAL sistema onenetarikoa izan da WMT 2015-eko Internetetik erauzitako dokumentuetan, non beste zenbait sistemek atazarako bereziki prestatutako algoritmoak erabiltzen dituzten. Esate baterako, dokumentuen URL helbideak konparatzeko teknikak.

Oinarrizko metodoari zenbait optimizazio egin zaizkio metodoaren hobekuntzak zein norabidetik joan daitezkeen jakiteko. Pisu lexikoak probatu dira, non termino bakoitzak bere maiztasun erlatiboaren arabera pisu handiagoa edo txikiagoa izango duen antzekotasun metrikan. Pisu lexikoek emaitzak hobetu dituzte, nahiz eta hobekuntzak txikiak izan. Bigarren optimizazio bat itzulpen-taula generikoen erabilera da. Itzulpen-taula generikoak sortzeko domeinu ezberdinetako corpusak konbinatu dira modu balantzatu batean, eta helburua dokumentuen terminoen itzulpenen bilaketetan zehaztasun handiagoa izatea da. Itzulpen-taulekin pisu lexikoekin baino arrakasta handiagoa izan dugu. Dokumentuen indexazioarekin ere esperimentuak egin dira indizetik berreskuratutako dokumentuek jatorrizko hizkuntzako doku-

mentuarekiko antzekotasun handiagoa izan dezaten. Berreskuratutako dokumentuen antzekotasuna hobetzea lortu da, baina hobekuntza hori ez da ematen azken emaitzetan, izan ere, zeinbait kasutan emaitzak apur bat okertzen dira. Azken optimizazio batekin ere egin da proba dokumentuen aurreprozesamenduaren eragina neurtzeko, truecasing-aren eragina hain zuzen ere. Emaitzak nola-halakoak izan dira BUCC 2015 corpusean, baina corpusaren arabera emaitzak asko hobetzen direla erakutsi du WMT 2016 atazan egindako esperimentuak. Beraz, orokorrean, egindako optimizazioek hobekuntza esanguratsuak erakutsi dituzte, batez ere WMT 2016 atazan.

Proposatutako metodoak ez du inolako entrenamendurik behar eta zuzenean erabili daiteke edozein motako dokumentuen gainean. Badira konfigurazio batzuk dokumentu mota batzuetarako hobeak direnak beste batzuetarako baino, baina lehenetsitako konfigurazioaren emaitzak oso egonkorrak eta gogobetekoak dira. Beraz, hasierako eraginkortasun eta erabilgarritasun helburuak guztiz bete dira.

4. KAPITULUA

Esaldien Lerrokatzea

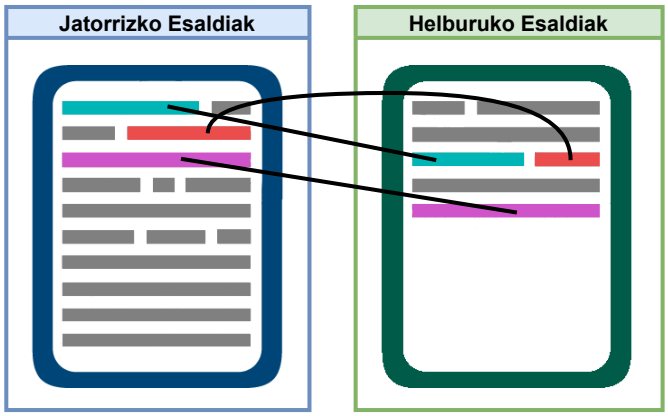
Datuetan oinarritutako itzulpen automatikoaren hazkundearekin tamaina gero eta handiagoko corpus paraleloen beharra ere hazi da. Nahiz eta azken urteotan kalitatezko corpus paraleloak sortu izan diren (Tiedemann, 2012), corpus paraleloen sorkuntza oso lan astuna da, askotan adituen laguntza behar baita testuak eskuz lerrokatzeko edota itzulpenak zuzentzeko. Zailtasun hauek gainditu ahal izateko aukera bat corpus konparagarriak ustiatzea da corpus paraleloak automatikoki sortzeko.

Lerrokatu beharreko corpusaren arabera modu ezberdinetan bideratu daiteke esaldien lerrokapena. Corpusaren tamaina txikia bada (50K-100K esaldi hizkuntza bakoitzerako) esaldien lerrokapena zuzenean egin liteke esaldi konbinazio guztiak aztertuz. Aldiz, corpusa dokumentutan antolatuta badago, eta corpusaren tamaina handia bada edota dokumentuek gai ezberdinak jorratzen badituzte, esaldien lerrokapena egin aurretik dokumentuen lerrokatzea egitea gomendagarria izan daiteke (ikus aurreko 3. kapitulua).

Dokumentu moten arabera ere badira ezberdintasunak esaldien lerrokatzean. Dokumentu paraleloetan esaldi bakoitza dagokionarekin lerrokatu behar da. Dokumentuak paraleloak izan harren badira kasuak non dokumentuetan akatsak dauden (hizkuntza okerreko esaldiak, hitzik gabeko esaldiak...) eta esaldi batzuek lerrokapenik ez izatea. Azkenik, corpus konparagarrietan (paraleloak izan ez harren informazio amankomuna partekatzen duten corpusak) paralelogarritasun handieneko esaldiak lerrokatu eta gainerakoak baztertzen dira. Funtsean, esaldien lerrokatzea 4.1. irudian laburbiltzen da. Bukaerako

corpus paraleloa sortzeko esaldien konbinazio guztiak (ala konbinazioen azpimultzo bat) konparatu beharra dago antzekotasun metrika batekin. Antzekotasun metrika horrek determinatuko du lerrokatze baten "paralelotasun" maila.

4.1 irudia: Esaldien lerrokatzea.



Dokumentuak eta esaldiak lerrokatzeko erabilpen kasuak nahiko antzekoak dira, horregatik kapitulu honetako helburuak aurrekoan azaldutakoen berdintsuak dira: kalitatezko lerrokatzeak behar dira corpus paralelo egokiak sortzeko, konputazionalki metodo eraginkorra izan behar du, eta eramangarritasuna bermatu behar da edozein hizkuntza edo esaldirako.

Kapitulu honetan STACC deritzon sistema aurkezten dugu. STACC eta aurreko kapituluan azaldutako DOCAL sistema metodo berdinean oinarritzen dira, eta beraz, lerrokatzeak egiteko metodoa oso antzekoa da. Metodoaren funtsa esaldiak token multzoen bidez errepresentatzea da token multzoen terminoak konparatuz antzekoenak diren esaldi pareak identifikatzeko. Eraginkortasun eta eramangarritasun helburuak betetzen direla egiaztatzeko, STACC artearen egoerako beste sistemekin konparatzen da. Zenbait optimizaziorekin ere ikertzen da, eta optimizazio hauen eragina aztertzeko BUCC 2017 eta BUCC 2018 atazetan parte-hartzen da.

STACC-en baliagarritasuna aztertzeko benetako erabilpen kasu bat aurkezten da. Erabilpen kasu horren helburua berrien domeinuko testuak euskaratik gaztelaniara (edo kontrako norantzan) itzultzeko SMT ereduak entrenatzea

da. Berrien domeinuan ez daude behar adina corpus paralelo kalitatezko itzulpenak lortu ahal izateko, hortaz, proposatutako irtenbidea EITB corpus konparagarria STACC erabiliz esaldi mailan lerrokatzea da.

Hurrengo atalak honela antolatzen dira. 4.1. atalean oinarrizko metodoa azaltzen da. Hurrengo, 4.2. atalean esaldi mailan lerrokatutako EITB corpusa aurkezten da. 4.3. atalean berriz, lehenengo esperimentuak erakusten dira. Gero, 4.4. atalean oinarrizko metodoaren gainean egindako zenbait optimizazio deskribatzen dira eta hauen eragina neurtzen da. Azkenik, 4.5. azpiatalean, izandako emaitzak laburbiltzen dira eta azken ondorioak ateratzen dira.

4.1. Esaldien Antzekotasuna, STACC

Corpus konparagarrietatik abiatuta corpus paraleloak sortzeko prozesuan, dokumentuen lerrokatzea egin daiteke gero dokumentuetako esaldiak lerrokatzeko. Prozesamendua horrela eginez, esaldien bilaketa espazioa murriztu ezezik, esaldien lerrokatzea uneko dokumentuan egiten denez erroreak minimizatzea ere lortzen da. Nahiz eta prozesamendu kateko osagaien ordenagatik zentzua izan dokumentuen lerrokatzea esaldien lerrokatzea baino lehen azaltzea, kronologikoki STACC DOCAL baino lehenago garatu zen, eta nahiz eta 3. kapituluan oinarrizko osagaiak azaldu, benetan STACC garatzeko unean sortu ziren.

EITB corpusa lerrokatzeko orduan (4.2. atalean azaltzen da), esaldiak lerrokatzeko sistemen bilaketa egin zen eta LEXACC sistema (Ștefănescu *et al.*, 2012) aurkitu zen. Orokorrean, LEXACC oso metodo eraginkorra da eta 8 hizkuntza paretako corpusak lerrokatzeko balio du, nahiko metodo eramangarria da beraz (gainerako hizkuntza pareentzako lehenetsitako konfigurazio bat erabili daiteke). Hala ere, EITB corpusarekin egindako esperimentutan lerrokatze okerrak gertatzen zirela ikusi zen. Errore horiek konpontzeko bi ikerketa lerro diseinatu ziren: LEXACC-en antzekotasun funtzioak euskarara egokitzea, eta metodo berri bat sortzea LEXACC-en funtzioek inspiratuta. Lehenengo bidetik LEXACC.EU sortu zen, eta bigarrenetik berriz STACC.

Kapitulu honetako kontribuzio nagusia STACC denez eta bere oinarria LE-

XACC-en dagoenez, lehendabizi LEXACC azaltzen da.

4.1.1. LEXACC

LEXACC (Ștefănescu *et al.*, 2012) hizkuntza arteko informazio-eskuratze tekniketan oinarritzen da esaldiak lerrokatzeko. APACHE LUCENE[1] indexazio motorra erabiltzen du bi pausutan: helburuko esaldiak indexatzeko, eta indexatutako dokumentuak berreskuratzeko jatorrizko esaldien termino esanguratsuen itzulpenekin. Terminoen itzulpenak lortzeko corpus paraleloen bidez elikatutako IBM ereduak erabiltzen dira (Brown *et al.*, 1993). ACCURAT proiektuaren barne[2] besteak beste bi tresna nagusi garatu ziren lerrokapenak egiteko: dokumentuak lerrokatzeko EMACC (Ion *et al.*, 2011), eta esaldiak lerrokatzeko LEXACC. LEXACC zuzenean erabili daiteke esaldiak lerrokatzeko, baina bereziki prestatuta dago EMACC-en ostean erabiltzeko.

Indizetik antzekoenak diren helburuko esaldiak berreskuratu eta gero, antzekotasun metrika bat aplikatzen da lerrokatze egokiena zein den kalkulatzeko. Antzekotasun metrika hori esaldien ezaugarriak neurtzen dituzten bost funtzioen batura orekatua da. Izan bitez s eta t jatorrizko eta helburuko esaldiak hurrenez hurren, $f_1, ..., f_5$ esaldien ezaugarriak erauzteko funtzioak, eta $\theta_1, ..., \theta_5$ funtzio bakoitzaren pisuak, $sim(s,t)$ antzekotasun metrika 4.1. ekuazioan azaltzen den bezala kalkulatzen da.

$$sim(s,t) = \sum_{i=1}^{5} \theta_i \cdot f_i(s,t) \tag{4.1}$$

Bost funtzioak honako hauek dira:

- f_1. Esaldietako termino lexikoen arteko antzekotasun probabilitatea kalkulatzen du. Bi esaldietako termino lexiko guztiak berberak badira, funtzio honen emaitza 1-ekoa izango da, aldiz, termino lexikoek ez badute zerikusirik, emaitza 0 izango da. Termino lexikoak puntuazio iku-

[1] `https://lucene.apache.org/`
[2] `http://www.accurat-project.eu/`

rrak ez diren eta stop-word listetan[3] agertzen ez direnak dira. Termino lexikoen arteko antzekotasun probabilitatea lau mailatan kalkulatzen da:

1. Bi termino lexiko berberak badira, emaitza 1-eko izango da.
2. Bestela, karaktereen antzekotasuna neurtzen da normalizatutako Levenshtein distantziarekin (Levenshtein, 1966). Antzekotasuna atalase bat baino altuagoa bada, hori izango da funtzioaren irteera.
3. Bestela, bi terminoak itzulpen-taula lexikoetan[4] bilatzen dira. Itzulpen probabilitatea atalase bat baino altuagoa bada, hori izango da funtzioaren irteera.
4. Bestela, emaitza 0 izango da.

- f_2. Ikuspegi sintaktiko batetik, lerrokatutako bi termino lexikoren inguruan lerrokatutako termino funtzionalak (esanahi lexikorik gabeko terminoak) ere egongo dira. Funtzio honek, f_1 funtzioaren lerrokatze bakoitzaren inguruan (±3 terminoko distantzia batean) lerrokatutako termino funtzional bat egoteko probabilitate kalkulatzen du.

- f_3. f_1 funtzioko lerrokatzeen zeihartasuna neurtzen du (Tufiş *et al.*, 2006). Horretarako, Pearson korrelazio koefizientean oinarritutako metrika bat definitzen da. Pearson koefizientea kalkulatu ahal izateko, jatorrizko eta helburuko esaldiak bektoretan bihurtzen dira f_1 funtzioko lerrokatzeak erabiliz.

- f_4. Esaldiak termino lexiko berberekin hasi eta bukatzen diren adierazten duen funtzio bitarra da. Lau termino lexiko aztertzen dira: lehenengo biak eta bukaerako biak.

- f_5. Esaldiak puntuazio ikur berberarekin bukatzen diren adierazten duen funtzio bitarra.

[3] LEXACC-ekin batera zenbait stop-word lista ematen dira, baina aukera dago norberarenak erabiltzeko.

[4] LEXACC-ekin batera zenbait itzulpen-taula lexiko ematen dira, baina norberarenak erabiltzeko edo entrenatzeko aukera dago (GIZA++ erabiliz adibidez (Och eta Ney, 2003)).

Funtzio guztien pisuak optimizatzeko logistic regression eredu bat entrenatzen da. Labur esanda, logistic regression ereduak adibide positiboak eta negatiboak egokien banatzen dituen hiperplanoaren pisuak ikasten ditu. LEXACC-en kasuan, adibide bakoitzaren hiperplanoko posizioa $f_1, .., f_5$ funtzioek zehazten dute.

Pisu optimoak entrenatu ostean, LEXACC-en bidez jatorrizko hizkuntzako esaldi bakoitza antzekotasun handieneko helburuko hizkuntzako esaldi batekin lerrokatutako da. Hala ere, batez ere corpus zaratatsu edo konparagarrien kasuan, bi esaldi lerrokatuta egoteak ez du zertan esan nahi paraleloak direnik. Esaldi paraleloak lortzeko atalase batetik gorako antzekotasuna duten lerrokapenak hartzen dira. Ștefănescu *et al.* (2012)-ek atalase guztiak probatu zituzten 0-tik 1-era 0,01-eko gehikuntzarekin.

Hizkuntza bakoitzaren ezaugarri sintaktikoek eragin zuzena dute bost funtzioen pisuetan. Adibidez, hizkuntza bakoitzaren hitz ordena propioak θ_1 eta θ_3 pisuetan eragiten du. Horregatik pisu guztiak hizkuntza pare bakoitzerako optimizatu behar dira. Gainera, itzulpen-taulen probabilitateak erabiltzen direnez, LEXACC-en antzekotasun metrika ez da simetrikoa, eta beraz, pisuen optimizazioa hizkuntza parearen norabidearen arabekoa da.

4.1.2. LEXACC.EU

STACC-en aurrekari bezala, LEXACC euskararen ezaugarri morfosintaktikoetara egokitzeko bost funtzioetan zenbait aldaketa egin dira. *Lerrokatzeen zeihartasuna* (f_3) eta *esaldien hasiera eta bukaerako termino lexikoak* (f_4) neurtzen dituzten funtzioak ez dira oso egokiak euskararen hitzen ordena librearengatik. Litekeena da corpus batean hitzen ordenak koherentzia bat izatea, eta funtzioen pisuak corpus horretara egokitzen badira emaitzak onak izan daitezke, baina orduan pisu horiek ez lirateke optimoak izango beste corpus batean bestelako hitzen ordena erabiltzen bada.

Termino funtzionalen lerrokatze funtzioak (f_2) ere arazoak ematen ditu. Euskara hizkuntza aglutinatiboa izanik terminoak morfologikoki oso aberatsak dira. Horregatik, termino funtzionalen lerrokatze zuzen bat egin ahal izateko analisi morfologiko bat egin beharko litzateke euskarazko esaldietan. Arazo

hau argiago ikustearren demagun *"voy a donosti"* esaldia dugula gaztelaniaz eta *"donostira noa"* esaldia euskaraz, f_1 funtzioak *"donosti-donostira"* eta *"voy-noa"* lerrokatzeak aurkituko ditu, baina f_2 funtzioak oker egingo du *"a"* termino funtzionala ezingo duelako lerrokatu.

Funtzioen zenbait konbinazio eta konfiguraziorekin esperimentuak egin dira. Honako hauek dira aukeratutako funtzioak:

- $f_1^{'}$. LEXACC-en *termino lexikoen lerrokatze* funtzioa (f_1).
- $f_2^{'}$. $f_1^{'}$ funtzioaren berdina, baina kontrako norabidean, hau da, helburuko esalditik jatorrizko esaldira.
- $f_3^{'}$. Amankomunean dauden izen-entitateen kantitatea neurtzen du. Izen-entitateak letra xehez hasten diren eta itzulpen-taulan ez dauden terminoak dira. Izan bitez E_s eta E_t s eta t esaldietako izen-entitateen multzoa hurrenez hurren, $f_3^{'}$ funtzioa 4.2 ekuazioan definitzen da.

$$f_3^{'}(s,t) = \frac{|E_s \cap E_t|}{|E_s \cup E_t|} \tag{4.2}$$

- $f_4^{'}$. LEXACC-en f_5 funtzioa, esaldiak puntuazio ikur berberarekin bukatzen diren aztertzen duena.

Horretaz gainera, beste aldaketa bat ere egin da: $f_1^{'}$, $f_2^{'}$ eta $f_3^{'}$ funtzioetan: itzulpen-taulen probabilitateak erabili ordez 1-eko probabilitatea ematea itzulpen-taulako agerpenei. Aldaketa honentzako arrazoi nagusia IVAP eta EITB corpusen domeinuen arteko ezberdintasuna da: domeinu batean lortutako probabilitateek ez dute zertan bat egin behar beste domeinu bateko banaketa lexikoarekin. 4.3.1. azpiatalean EITB corpusean izandako emaitzek LEXACC.EU sistemarako egindako aldaketek hobekuntza nabariak dakartzatela erakusten dute.

4.1.3. STACC

STACC (*Set-Theoretic Alignment of Comparable Corpora*) esaldiak lerrokatzeko garatutako metodoa da eta DOCAL-en azaldutako antzekotasun metrikan

oinarritzen da (ikus 3.1. atala). Metodoarentzako inspirazioa LEXACC-en *termino lexikoen lerrokatze* funtziotik dator (f_1). Hasierako esperimentuetan funtzio horrekin bakarrik lerrokatze egoki gehienak lortzen zirela ikusi zen, eta funtzio horren inguruko ikerketetatik sortu zen STACC.

Esaldietan dokumentuetan baino informazio gutxiago dago, horregatik STACC eta DOCAL-en artean zenbait ezberdintasun daude. Esaldien tamaina txikiagoarengatik lehenengo terminoaren formak garrantzi handiagoa du. Lehenengo terminoaren forma normalizatzeko esaldiak tokenizatu eta gero truecasinga egiten da.

Izan bitez s_i eta s_j tokenizatutako eta hauen gainean truecasing egindako bi esaldi l_1 eta l_2 hizkuntzetan hurrenez hurren, S_i s_i-ren token multzoa, S_j s_j-ren token multzoa, T_{ij} S_i multzoaren hedatutako (ikus 3.1.1 eta 3.1.2. azpiatalak) token multzoa l_1 hizkuntzatik l_2 hizkuntzara, eta T_{ji} S_j multzoaren hedatutako token multzoa l_2 hizkuntzatik l_1 hizkuntzara. 4.3. ekuazioak s_i eta s_j esaldien arteko antzekotasuna nola kalkulatzen den deskribatzen du.

$$stacc(s_i, s_j) = \frac{1}{2}\left(\frac{|T_{ij} \cap S_j|}{|T_{ij} \cup S_j|} + \frac{|T_{ji} \cap S_i|}{|T_{ji} \cup S_i|}\right) \tag{4.3}$$

Hurrengo azpiatalean STACC-en berezitasunak azaltzen dira.

4.1.3.1. Aurrizki Komun Luzeenak

Dokumentuen lerrokatzean, 3.1.2. azpiatalean, aurrizki komun luzeenen kalkulua azaldu da. Token multzoak konparatzeko orduan terminoen agerpenak kontatzen dira inolako lematizazio edo stemming teknikarik erabili gabe. Aurrizki komun luzeenen kalkuluak terminoen morfologia kontutan hartzeko aukera ematen du terminoen aurrizkiak token multzoetan gehituz.

DOCAL-en kasuan aurrizki komun luzeenen kalkuluak gainkarga nabari bat suposatzen du token multzoen tamainarengatik. Gainera, 3.2. ataleko esperimentuetan ez da hobekuntza handirik ikusten. STACC-en kasuan berriz, aurrizki komun luzeenekin emaitzak hobeak lortzen direla konprobatu ahal izan da: EITB corpusean +2,9 puntuko aldea dago F1 metrikan eta WIKIPEDIA-ko bulgariera-ingelesa corpusean +3,7, +2,6 eta +5,5 puntuko hobe-

kuntzak lortzen dira. Emaitza hauek ikusita, aurrerantzean aurrizki komun luzeenak erabiliko dira kontrakoa esaten ez bada.

4.1.3.2. Lerrokatze Onenaren Optimizazioa

DOCAL-en oinarrizko konfigurazioan lerrokapen onenaren optimizazioa erabiltzen da (ikus 3.1.4. azpiatala). Optimizazio honi esker, helburuko hizkuntzako esaldi bat jatorrizko hizkuntzako esaldi batekin bakarrik lerrokatzea lortzen da antzekotasun handiena duten lerrokatzeak mantenduz eta gainerakoak baztertuz.

Dokumentuak lerrokatzerako orduan lerrokatze onenaren optimizazioarekin emaitzak nabarmen hobetzen dira (+10 puntu baino gehiagoko hobekuntzak esperimentuaren arabera). Esaldiak lerrokatzeko ordea, hobekuntzak lortzen diren harren ez dira hain onak (+4,9 puntuko hobekuntzak gehienez).

4.1.3.3. Atalasea

STACC-en esaldi pareen artean nolako antzekotasuna dagoen adierazten da. Bi esaldi oso antzekoak diren ala ez azaltzen du, baina ez da sailkatzaile bitar bat, hots, ez ditu bereizten paraleloak diren eta ez diren lerrokatzeak. Orduan, lerrokatzeak paraleloak diren ala ez determinatzeko atalase bat erabiltzen da.

Atalaserako hiru aukera daude:

- Lehenetsitako balio bat erabiltzea.
- Atalaserik gabe lerrokatze guztiak kalkulatzea, antzekotasun metrikaren arabera lerrokatzeak ordenatzea, eta orduan ezartzea atalasea lerrokatzeen kalitatea okertzen den puntuan (doitasuna hobetsi nahi bada atalasea handituz, eta estaldura hobetsi nahi bada atalasea txikituz).
- Garapenerako corpus batekin atalasea optimizatzea.

4.3 eta 4.4. ataletan azaltzen da erabilitako atalasea.

4.1.3.4. Itzulpen Automatikoa

Itzulpen token multzoak sortzeko jatorrizko token multzoko termino bakoitzarekin bilaketak egiten dira itzulpen-tauletan[5], eta probabilitate handieneko itzulpenak itzulpen token multzoan sartzen dira. Itzulpen token multzoak sortzeko beste aukera bat itzulpen automatikoa erabiltzea da: jatorrizko esaldia itzuli eta itzulpeneko terminoekin itzulpen token multzoak sortzea.

Metodo honek itzulpen ereduak entrenatzea eskatzen du eta erabilgarritasunaren aldetik atzera pausu bat dakar. Hala ere, itzulpen-taulak entrenatzeko corpus paraleloak behar dira, hortaz, itzulpen ereduen entrenamendurako ez da hizkuntza baliabide osagarririk behar. Gainera, kontutan eduki behar da itzulpen-taulen entrenamendua SMT ereduen entrenamenduaren parte dela. Bestalde, itzulpen automatikoaren erabilerak zenbait abantaila izan ditzake. Alde batetik, itzulpenak inguruko hitzengatik baldintzatuta daudenez itzulpen zehatzagoak sortzeko aukera dago. Beste alde batetik, hitz baten itzulpenetik termino bat baina gehiago lortzeko aukera dago.

Oinarrizko metodoaren aldaera honen ezaugarriak ebaluatzeko zenbait esperimentu egiten dira 4.3.1. azpiatalean. Aurrerantzean metodo honi STACC.IA deituko zaio.

4.2. EiTB Corpusa

Esaldien lerrokatzearen ekarpen garrantzitsu bat EITB corpusa da. Orokorrean, itzulpen automatikorako tamaina handiko kalitatezko corpus paraleloak behar dira. Tamalez, domeinuaren arabera, euskaraz ez daude eskura baliabide linguistiko nahikoak. Testuinguru honetan kokatzen da EITB corpusa, berrien domeinuan esaldi mailan lerrokatutako gaztelaniaz eta euskaraz idatzitako albisteak biltzen dituen corpusa[6].

Esan bezala, nahiz eta tesi honetan lehendabizi dokumentuen lerrokatzea aurkeztu, kronologikoki esaldien lerrokatzearen inguruan egin ziren lehenengo

[5]Itzulpen-taulak corpus paraleloetatik erauzten dira IBM ereduek emandako terminoen itzulpen probabilitateak kontutan hartuz.

[6]EITB corpusa lortzeko jo esteka honetara: `http://metashare.elda.org/repository/search/?q=eitb+documents`

ikerketak, eta horregatik, EITB corpuseko dokumentuak lerrokatzeko unean bestelako irtenbide bat hautatu zen. Atal honetan ematen dira xehetasun guztiak.

EITB corpusa euskaraz eta gaztelaniaz idatzitako albisteen bilduma bat da. Bi hizkuntzetan bildutako albisteak berberak dira baina lantalde ezberdinek idatzitakoak dira, beraz, berriak ez dira itzulpen zuzenak. Terminologia estandarra jarraituz (Skadiņa *et al.*, 2012), EITB corpusak konparagarritasun maila altua duela esan daiteke.

Jatorrizko corpusak 59 XML dokumentu ditu euskaraz eta beste 57 gaztelaniaz, eta 2009. urtetik 2013. urtera bitartean gertatutako albisteak biltzen ditu. XML dokumentu bakoitzak hilabete bateko albisteak jasotzen ditu, eta albiste bakoitzak honako egitura du:

- <id>. Albistearen identifikazio zenbakia. Garrantzitsua da jakitea identifikatzaile hau bakarra dela eta ez duelako inolako loturarik hizkuntzen artean.
- <title>. Albistearen titulua.
- <link>. Jatorrizko HTML esteka.
- <pubDate>. Argitalpen data.
- <description>. Albistearen edukia.
- <category>. Albistearen kategoria (kultura, kirolak, ekonomia, etab.).

EITB corpusaren tamainaren xehetasunak 4.1. taulan aurkezten dira. Hizkuntza bakoitzeko milioi bat esalditik gorako corpusa izanik, EITB corpusa oso baliagarria izan daiteke gaztelania-euskara hizkuntza parerako itzulpen sistemak sortzeko. Gainera, politika, kirolak eta bestelako gaiak jorratzen dituenez nahiko terminologia aberatsa du.

Dokumentuen lerrokatzea. Albiste bakoitza independentea da eta ez dago modurik automatikoki errorerik gabe jatorrizko hizkuntzako albisteak helburuko hizkuntzako albisteekin lotzeko. Esaldiak lerrokatzeko aukera bat

4.1 taula: Jatorrizko EITB corpusa.

URTEA	ALBISTEAK		ESALDIAK		TOKENAK	
	ES	EU	ES	EU	ES	EU
2009	18.759	18.552	223.323	236.753	4.672.018	3.068.989
2010	17.979	17.462	204.004	216.043	4.325.927	2.778.677
2011	19.037	18.856	216.240	216.240	4.948.890	3.083.384
2012	18.972	19.344	213.730	229.270	4.932.887	3.043.726
2013	13.601	13.484	160.908	164.363	3.557.014	2.160.011
GUZTIRA	88.348	87.698	1.018.205	1.077.331	22.436.736	14.134.787

produktu kartesiarraren bidez konbinazio guztiak aztertzea da (Ion, 2012), baina corpusaren tamaina kontutan izanda ez da irtenbide egokia (bilioi bat konbinazio baino gehiago daude). Bilaketa espazioa murrizteko irtenbide bat dokumentuen lerrokatzea egitea da (Fung eta Cheung, 2004; Ion *et al.*, 2011), edota indexazio teknikak erabiliz lerrokatze kandidatuak lortzea indize baten gainean galdeketak eginez (Rauf eta Schwenk, 2011; Munteanu eta Marcu, 2005; Ștefănescu *et al.*, 2012). EITB corpusa lerrokatzeko lehenengo irtenbidea jarraitzea erabaki da.

Albisteen lerrokatzea egiteko EMACC erabili da (Ion *et al.*, 2011) (DOCALekiko konparaketa bat aurkezten da 3.2.3. azpiatalean). EMACC Expectation Maximization algoritmoan oinarritzen da eta doitasun altua lortzen du, baina konputazionalki kostu handiko algoritmoa da. Bilaketa espazioa gehiago mugatzeko (7.000 milioi lerrokatze posible baino gehiago daude), lerrokatzeak XML dokumentu bakoitzeko egin dira, hau da, hilabete bakoitzeko, horrela, bilaketa espazio mugatu ezezik lerrokatzeen kalitatea hobetzea ere lortzen da.

EMACC-en bidez dokumentuak lerrokatzeko itzulpen-taulak behar dira. Itzulpen-taulen kalitateak eragin zuzena du lortutako lerrokatzeetan, eta horregatik garrantzitsua da itzulpen-taulak entrenatzeko corpus egokiak erabiltzea. Itzulpen-taulak sortzeko IVAP corpusa erabili da. IVAP corpusak adituek eskuz lerrokatutako itzulpenak biltzen ditu, hortaz, corpusaren kalitatea aproposa da ataza honetarako (4.3.1. azpiatalean azaltzen dira IVAP corpusaren

4.2 taula: EITB-ko ebaluaziorako corpusak.

CORPUSA	ESALDIAK ES	ESALDIAK EU
500:500	500	500
1000:1000	1.000	1.000
1000:1500	1.000	1.500

xehetasunak).

Lerrokatzeak lehenetsitako parametroekin egin dira, jatorrizko dokumentu bakoitza helburuko hiru dokumenturekin lerrokatuz. Lortutako emaitzak eskuz aztertu eta gero, konfigurazio honekin lerrokatze kopuruaren eta lerrokatze kalitatearen artean oreka egokia lortzen dela hausnartu da. Guztira, gaztelaniaz idatzitako dokumentuen %94-a lerrokatu da euskarazko dokumentuen %92-arekin.

Esaldien lerrokatzea. Esaldien lerrokapenaren inguruan sistemen eraginkortasuna neurtzeko ebaluaziorako corpusak prestatu dira. Doitasuna neurtzeko EITB corpusetik 2009ko urtarrileko euskarazko 500 esaldi hartu dira eta eskuz lerrokatu dira. Estaldura neurtzeko berriz beste bi zati prestatu dira lehengo 500 esaldi pareei lerrokatzerik gabeko esaldiak gehituz: lehendabiziko zatian lerrokatzerik ez dituzten 500 esaldi gehitu dira gaztelaniaz eta beste 500 euskaraz, eta bigarren zatirako lerrokatzerik gabeko 500 esaldi gehitu dira gaztelaniaz eta beste 1.000 euskaraz. 4.2. taulan azaltzen dira ebaluaziorako corpusak. EITB corpuseko esaldiak lerrokatzen lortutako emaitzak berriz 4.3.1. azpiatalean aurkezten dira.

4.3. Oinarrizko Metodoaren Esperimentuak

Azpiatal honetako esperimentuetan lau corpus erabiltzen dira. Lerrokatzeak EITB, ACCURAT eta WIKIPEDIA corpusetan ebaluatzen dira, eta IVAP corpusa beste hiru atazetarako erabili da: (1) EITB corpusa lerrokatu ahal izateko itzulpen-taulak sortzeko, (2) LEXACC eta LEXACC.EU sistemen pisuak gaztelania-euskara hizkuntza parerako egokitzeko, eta (3) oinarrizko SMT ere-

duak entrenatzeko.

Aukeratutako corpusak eskuz lerrokatuta daude eta proportzio ezberdinarekin zarata gehitu zaie lerrokatzerik gabeko esaldiekin. Horrela, sistemen portaera neurtu daiteke zarata maila ezberdineko egoeratan.

Tesi honen helburuetako bat eramangarritasuna bermatzea denez, STACC sistemaren portaera ebaluatu da 10 hizkuntza ezberdinetan. Morfologikoki aberastasun maila ezberdina duten hizkuntzak aukeratu dira (ingelesa eta euskara), eta eskura dauden hizkuntza baliabideen kantitatearen ikuspegitik egoera ezberdinean dauden hizkuntzak ere aukeratu dira (gaztelania eta letoniera).

Esaldien lerrokatzeak doitasun, estaldura eta F1 metrikekin neurtzen dira. Oro har, corpus konparagarri batetik esaldi paraleloak erauzteko orduan, helburua kalitatezko ahalik eta datu gehien bilatzea da, horregatik, sistema onena zein den erabakitzeko F1 metrika erabiltzen da.

Sistemen Konfigurazioa. Bidezko konparaketa bat egitearren, itzulpen-taula lexikoak berberak dira hauek erabiltzen dituzten sistema guztietarako. Gaztelania-euskararako IVAP bidez entrenatutako itzulpen-taulak erabili dira, eta gainerako hizkuntza pareetan JRC-ACQUIS COMMUNAUTAIRE corpusa erabili da itzulpen-taulak entrenatzeko[7]. LUCENE-ren konfigurazioari dagokionez, jatorrizko dokumentu bakoitzeko 100 dokumentu berreskuratzen dira.

4.1.3. azpiatalean azaltzen denez, itzulpen token multzoak sortzeko itzulpen-tauletatik k itzulpen onenak bilatzen dira. Azpiatal honetako esperimentuetan $k = 5$ itzulpen onenak erabiltzen dira. Balio honen aukeraketa guztiz arbitrarioa da; aurretiko proba batzuen ostean balio honek balantza egokia erakutsi du terminoen itzulpenen kalitatearen eta itzulpen multzoen tamainaren artean, baina ez da optimizazio berezirik egin. Izan ere, konfigurazio hau ez da optimoa kasu guztietarako, esate baterako, $k = 2$ erabiliz ACCURAT ingeles-greziera corpusean 2,9 puntuko hobekuntza dago.

Lerrokatze onenaren optimizazioa ere ez da erabiltzen kontrakoa esaten ez bada. Lerrokatze onenaren optimizazioa dokumentu guztien gainean antze-

[7] DOCAL-en erabiltzen diren itzulpen-taulak berrerabili dira ahal izan den kasuetan.

kotasun metrika aplikatu ostean egiten da, beraz, prozesu independentetzat har dezakegu. Hain zuzen ere, gainerako sistemetan ere optimizazio hau egitea posible da. Horregatik, azpiatal honetan neurtu nahi dena STACC-en antzekotasun metrika da, inolako optimizazio berezirik egin gabe.

LEXACC sistemaren konfigurazioari dagokionez, LUCENE indizearen konfiguraziorako lehenetsitako balioak erabili dira bi salbuespenekin. Lehendabizi, lehenetsitako konfigurazioan jatorrizko eta helburuko esaldiak berdinak badira ez dira inoiz lerrokatzen. Konfigurazio hau ez da egokia EITB corpuserako, zenbait kirol albistetan erroreak ematen baitira (adibidez, *"Touring 6, Oiartzun 0"* esaldia). Bigarrenik, aurreprozesamendu bat egiten da non jatorrizko eta helburuko esaldien arteko luzera ratioa 1,5 baino handiagoa bada lerrokatze hori ez den onartzen. Proba batzuen ostean EITB corpuseko errore batzuk ratio horrengatik gertatzen direla konprobatu da, horregatik ratioa 7,5-era igo da.

Atalasea. Atalase optimoa determinatzeko testean 0-tik 1-era balio ezberdinekin probatu da aldiro atalasea 0,01-eko balioarekin handituz. Noski, errealitatean ez dago aukera atalasea optimizatzeko, eta lehenetsitako atalase bat erabili beharko litzateke. Hala ere, lehenetsitako atalaseak erabiltzeak ez luke sistemen bidezko konparaketa bat egitea ahalbidetuko. Beste aukera bat corpus bakoitzeko garapenerako zati bat erabiltzea izango litzateke atalase optimoa kalkulatzeko, baina esperimentu hauetako corpusetan ez dago garapenerako zatirik. Dena den, EITB corpusaren ebaluazioan doitasun eta estalduraren bilakaera erakusten da atalasearen arabera.

4.3.1. EITB

EITB corpusarekin bi esperimentu egin dira. Lehendabizi, EITB-ko ebaluaziorako hiru corpusetako esaldiak lerrokatzen izandako emaitzak aurkezten dira, eta gero, EITB corpuseko esaldi guztiak lerrokatuz corpus berri bat sortu da SMT ereduak entrenatzeko.

IVAP corpusa. IVAP corpusa (Instituto Vasco de Adiministración Pública) gaztelania-euskara hizkuntza parerako itzulpen memorien bilduma bat da.

4.3 taula: IVAP corpusa.

CORPUSA	LERROK. ESALDIAK	TOKENAK ES	TOKENAK EU
ENTRENAMENDUA	645.223	9.717.604	7.556.964
GARAPENA	2.000	48.492	37.908
EBALUAZIOA	2.000	49.056	38.081

Itzulpen memoriak administrazio publikoko adituek egindakoak dira. IVAP corpusa bi atazetarako erabili da: gaztelania-euskara hizkuntza parerako itzulpen-taulak entrenatzeko eta oinarrizko SMT ereduak entrenatzeko.

Corpusaren garbiketa egin eta gero (karaktere okerrak eta gaizki etiketatutako itzulpen unitateak zeuden), corpus paraleloa denez esaldiak HUNALIGN bidez (Varga *et al.*, 2007) lerrokatu dira. HUNALIGN-ek esaldi pareak banaka-banaka ordenan prozesatzen ditu. HUNALIGN-ek itzulpen probabilitateak kalkulatzen ditu, eta probabilitatearen arabera esaldiak baztertu edo elkartzen ditu (baliteke esaldi baten itzulpena helburuko hizkuntzako n esalditan egotea). 4.3. taulan lortutako corpusaren tamaina aurkezten da.

Entrenamenduak egiteko esaldi bakoitza tokenizatu da eta truecasing-a egin da. Truecasing ereduak entrenatzeko IVAP corpuseko zati elebakarrak erabili dira. Corpusa entrenamendu, garapen eta ebaluaziorako zatitan banatu da phrase-based SMT ereduak entrenatzeko (Koehn *et al.*, 2003) MOSES erabiliz (Koehn *et al.*, 2007). SMT ereduen entrenamenduaren lerrokatze pausuan itzulpen-taulak sortzen dira, beraz, entrenamendu bakarrarekin azpiatal honetako esperimentuetarako eredu guztiak entrenatzen dira.

Konparatutako sistemak. EITB corpusa lerrokatzeko lau dira konparatzen diren sistemak: LEXACC, euskarako egokitutako LEXACC.EU, STACC, eta STACC.IA, itzulpen token multzoak sortzeko itzulpen automatikoa erabiltzen duen STACC-en bertsio berezia (ikus aurreko 4.1.3.4. azpiatala).

4.1.1. azpiatalean esan bezala, LEXACC sistemak bost funtzio erabiltzen ditu esaldietatik ezaugarriak erauzteko. Ezaugarri horiek uneko hizkuntza parearen araberakoak dira, horregatik, funtzio bakoitzaren eragina orekatzeko bost

pisu entrenatu behar dira. Gaztelania-euskarako pisuak IVAP corpusaren bidez optimizatu dira. IVAP corpusetik 9.500 adibide positibo eta beste 9.500 negatibo erabili dira; pisuen ebaluazioa 500 adibide positibo eta negatiboren gainean egin da 0,95-eko zehaztasuna lortuz.

Emaitzak. 4.4. taulan esaldien lerrokapenean lortutako emaitzak erakusten dira eta baita metrika onenak zein atalaserekin lortzen diren. Hiru corpusetan STACC da emaitza onenak lortzen dituen sistema, batez ere itzulpen automatikoa erabiltzen ez duen bertsioa. STACC.IA sistemarekin lortutako itzulpen token multzoak IVAP corpuseko domeinura egokituta daude, eta lortutako emaitzek frogatzen dutenez, egokiagoa da itzulpen token multzo *malguagoak* sortzea itzulpen-tauletatik k itzulpen onenak bilatuz. Gainera, itzulpen automatikoaren erabilera eramangarritasun helburuaren aurka doa, beraz, emaitza hauek oso gogobetekoak dira.

Euskarara egokitutako LEXACC.EU bertsioak ere emaitza onak lortzen ditu hasieran egindako hipotesiak balioztatuz. Arrazoi posible bat euskararako kaltegarriak diren funtzioak kendu izana da. Baliteke ere *termino lexikoen lerrokatze* funtzioetan (f_1' eta f_2') egindako aldaketek zerikusia izatea, hots, itzulpen-tauletan azaltzen diren probabilitateak erabili ordez 1-eko puntuazioa erabiltzea.

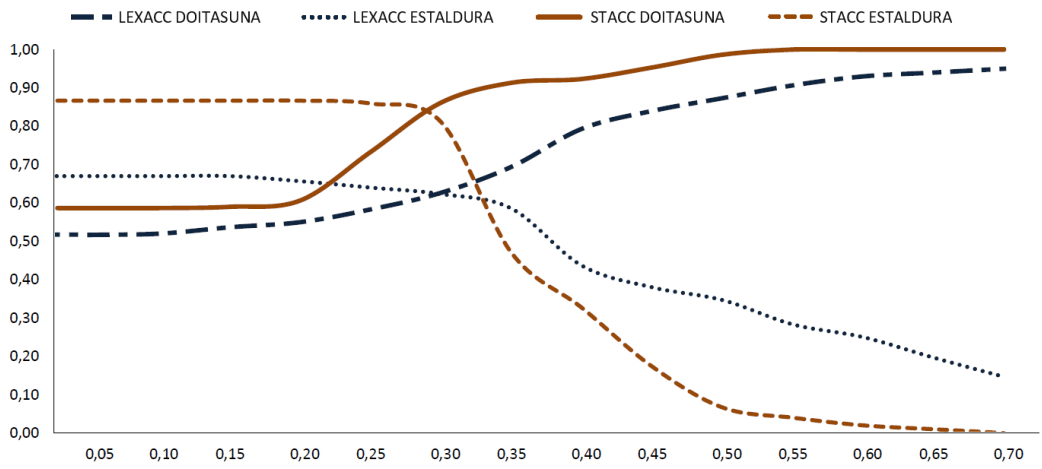
4.2. irudian azaltzen den doitasun eta estalduraren bilakaera ere interesgarria da. Bertan azaltzen diren datuak ebaluaziorako hiru corpusetan izandako emaitzen batezbestekoa eginez kalkulatu dira. LEXACC eta STACC sistemek antzeko bilakaera erakusten dute, non ezberdintasun nagusia LEXACC-en estalduraren beherapena lehunagoa dela den. Kurben berdintasun honek STACC-en antzekotasun metrika LEXACC-ena baino eraginkorragoa dela erakusten du, emaitza onak maximo lokal batengatik lortzen ez direla frogatuz.

Oro har, STACC da balantza egokiena erakusten duen sistema lortzen diren lerrokatzeen kantitatearen eta kalitatearen artean. 4.5. taulan erakusten denez, EITB corpus guztian zehar lerrokatze gehiago lortzen dira STACC-ekin LEXACC-ekin baino, hurrenez hurren 596.492 eta 368.184 lerrokatze lortuz. Esaldi pare horiek erauzteko, sistema bakoitzarentzat corpus zaratatsuenean (1000:1500 corpusean) lortutako atalase optimoa erabili da.

4.4 taula: Lerrokapen emaitzak EITB corpusean. LEXACC.EU euskararako egokitutako LEXACC-en bertsioa da. STACC.IA itzulpen automatikoa erabiltzen duen STACC-en bertsioa da.

SISTEMA	CORPUSA	ATALASEA	DOITASUNA	ESTALDURA	F1
LEXACC	500:500	0,12	80,5	74,2	77,2
LEXACC.EU	500:500	0,05	85,4	85,4	85,4
STACC.IA	500:500	0,15	88,6	87,0	87,8
STACC	500:500	0,17	**91,0**	**90,8**	**90,9**
LEXACC	1000:1000	0,31	63,3	55,6	59,2
LEXACC.EU	1000:1000	0,36	76,3	69,0	72,5
STACC.IA	1000:1000	0,24	80,2	70,6	75,1
STACC	1000:1000	0,30	**84,4**	**81,2**	**82,8**
LEXACC	1000:1500	0,33	59,6	50,2	54,5
LEXACC.EU	1000:1500	0,36	72,2	66,6	69,3
STACC.IA	1000:1500	0,24	79,0	67,6	72,8
STACC	1000:1500	0,30	**81,1**	**78,0**	**79,5**

4.2 irudia: LEXACC eta STACC sistemen doitasunaren eta estalduraren bilakaera atalasearen arabera EITB corpusean. Azaltzen diren datuak hiru ebaluaziorako corpusetan izandako emaitzen batezbestekoak dira.



4.5 taula: Lerrokatutako esaldi kopurua EITB corpusean.

SISTEMA	ATALASEA	ESALDIAK	TOKENAK ES	TOKENAK EU
LEXACC	0,33	368.184	7.605.144	5.756.838
STACC	0,30	596.492	10.875.355	10.875.355

Corpus paraleloak sortzeko arrazoi nagusia itzulpen automatikoaren bidez lortzen diren itzulpenen kalitatea hobetzea da. Horregatik, EITB corpusaren lerrokatzeak itzulpen automatikoan duen eragina aztertu da. Horretarako, hiru SMT eredu pare entrenatu dira bi norabideetan:

- IVAP corpusaren bidez entrenatutako ereduak. Ereduen parametroak optimizatzeko corpus bereko garapenerako zatia erabili da.
- LEXACC bidez EITB corpusetik erauzitako esaldi pareekin entrenatutako ereduak. Parametroak optimizatzeko corpus beretik ausaz 2.000 esaldi pare aukeratu dira.
- Aurreko ereduen antzera STACC erabiliz EITB corpuseko esaldiak lerrokatzeko.

Eredu guztien ebaluazioa EITB corpusarekin egin da eskuz egiaztatutako 1.678 esaldi erabiliz. Ebaluazioa egiteko aukeratutako metrika BLEU da (Papineni *et al.*, 2002).

Guztiak phrase-based SMT ereduak dira eta entrenamendurako MOSES erabili da. Ereduen parametroak honako hauek dira: 5-eko segmentu tamaina maximoa eta 6 terminoko esaldien birrordenaketa muga maximoa. Uneko corpuseko helburuko hizkuntzako zati elebakarrarekin hizkuntza ereduak entrenatu dira KENLM erabiliz (Heafield, 2011). Zehazki, 5-gramako ereduak entrenatu dira aldatutako KNESER-NEY leuntzearekin (Heafield *et al.*, 2013). Itzulpen ereduen parametroak MERT bidez optimizatu dira (Och eta Ney, 2003). Ebaluaziorako berriz MULTEVAL erabili da (Clark *et al.*, 2011).

4.6. taulan ikus daitezke entrenatutako SMT ereduen emaitzak. Argi geratzen da EITB corpusetik lerrokatutako esaldi pareak erabiliz eredu hobeak

4.6 taula: EITB corpusarekin izandako hobekuntzak SMT ereduetan. Emaitzak BLEU metrikarekin azaltzen dira.

EREDUA	ES → EU	EU → ES
IVAP	9,1	13,7
LEXACC	17,8	23,7
STACC	**19,1**	**25,5**

lortzen direla. STACC-en lerrokatzeen kalitateak SMT ereduetan eragin zuzena du: LEXACC-ekin alderatuta 1,3 puntuko hobekuntza dago euskararanzko norabidean, eta 1,8 puntuko aldea gaztelaniaranzko norabidean.

Laburbilduz, STACC izan da EITB corpusarekin egindako esperimentu guztietan emaitza onenak lortu dituen sistema. LEXACC-ekin gertatzen ez den bezala, STACC-ekin ez da aparteko entrenamendurik egin behar izan. Beraz, STACC sistema eraginkorragoa izan ezezik eramangarriagoa ere bada.

4.3.2. ACCURAT

ACCURAT zazpi hizkuntza paretarako esaldiak biltzen dituen corpus konparagarria da. Horregatik, corpus aproposa da STACC-en eramangarritasunaren inguruan esperimentuak egiteko.

Corpusa. ACCURAT corpusa izen bereko proiektuaren barne sortutako corpus librea da[8]. Bertan, berrien domeinuko testuak biltzen dira eta eskuz lerrokatutako ebaluaziorako corpusak sortu dira. Corpusak zazpi hizkuntza pare ditu, guztiak ingelesaren inguruan.

Zarataren aurrean sistemen portaera aztertzeko asmoz, lerrokapenik gabeko esaldiak gehitu dira ACCURAT corpuseko ebaluaziorako corpusean. Guztira, hiru corpus erabiltzen dira:

- 1:1. ACCURAT proiektuko ebaluaziorako corpusa.

[8]ACCURAT proiektua: `http://www.accurat-project.eu/`. Corpusa esteka honen bidez dago eskura: `http://metashare.elda.org/repository/search/?q=accurat`

- 2:1. 1:1 corpuseko esaldi bakoitzeko beste lerrokatzerik gabeko esaldi pare bat gehitu da ACCURAT proiektuko jatorrizko corpus konparagarritik.
- 100:1. 2:1-en antzekoa, baina kasu honetan 1:1 corpusari lerrokatzerik gabeko beste 99 esaldi pare gehitu zaizkio.

Lerrokatzerik gabeko esaldiak aukeratzeko orduan, jatorrizko hizkuntzako esaldiak ACCURAT-eko jatorrizko corpus konparagarriko lehenengo esaldiak hartu dira, eta helburuko hizkuntzarako berriz, ACCURAT corpuseko azkeneko esaldiak. 100:1 corpusean zenbait hizkuntza paretarako ez daude behar adina esaldi pare proportzio hori bete ahal izateko, kasu hauetan EUROPARL corpusa erabili da falta diren esaldi pareak lortzeko. 4.7. taulan laburbiltzen da sortutako corpusaren informazioa.

Konparatutako sistemak. ACCURAT corpusa lerrokatzeko STACC eta LEXACC sistemak erabili dira. Bi sistema hauek eramangarriak izateko diseinatu dira, beraz, esperimentu hauetarako ez da konfigurazio berezirik egin behar izan. LEXACC-en bost funtzioen pisuetarako egileek optimizatutako pisuak erabili dira (Ștefănescu *et al.*, 2012).

Emaitzak. 4.8. taulan STACC eta LEXACC sistemek ACCURAT corpusean izandako emaitzak aurkezten dira. 21 ebaluaziorako corpusetatik bitan bi sistemen emaitzak berberak dira, beste bitan LEXACC-ek ditu emaitza onenak, eta gainerako guztietan STACC da sistema eraginkorrena. Zarata gehien duten corpusak bakarrik hartzen baditugu kontutan (100:100 corpusak), STACC da emaitza onenak dituen sistema hizkuntza pare guztietan. Izan ere, corpusei zarata gehiago gehitzen zaien heinean, STACC eta LEXACC-en arteko diferentzia areagotu egiten da. Hobekuntza handiena ingelesa-aleman 100:100 corpusean ematen da +17,1 puntuko aldearekin STACC-en alde.

Arestian aipatu dugu sistemen arteko bidezko konparaketa bat egiteko lerrokapen onenaren optimizaziorik ez dela egiten. Optimizazio hori esaldi guztien artean antzekotasun metrika aplikatu ostean egiten da, eta beraz, sistema guztietan aplikatu liteke. Dena den, ACCURAT corpusean hizkuntza pare ugari daudenez, corpus honen gainean lerrokapen onenaren eragina

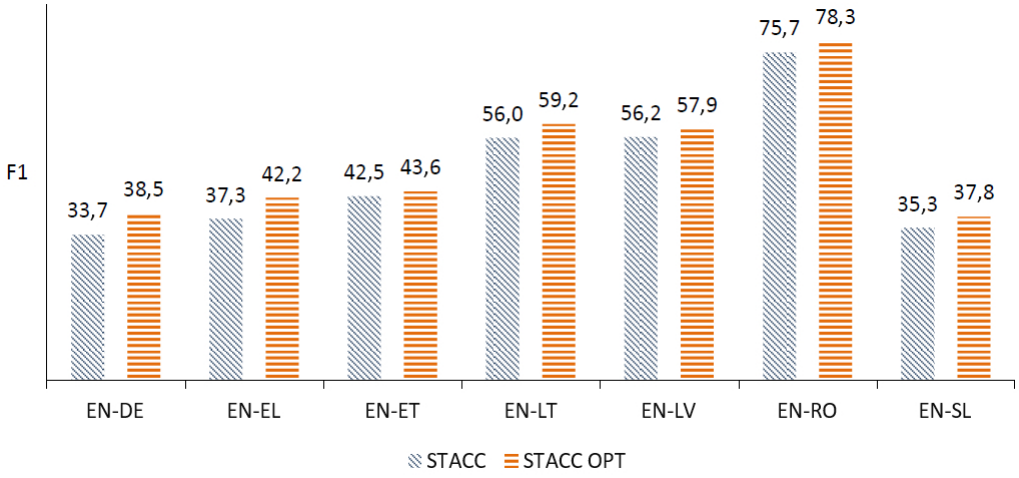
4.7 taula: ACCURAT corpusaren esaldi pare tamaina. Corpusak sortzeko hiru datu iturri erabili dira: (1) ACC. TEST, ACCURAT proiektuko ebaluaziorako corpusa, (2) ACC. JATOR., ACCURAT proiektuko jatorrizko corpus kongaragarria, eta (3) EUROP., EUROPARL corpusa.

HIZK.	CORPUSA	ITURRIA			GUZTIRA
		ACC. TEST	ACC. JATOR.	EUROP.	
EN-DE	1:1	512	-	-	512
	2:1	512	512	-	1.024
	100:1	512	6.891	43.797	51.200
EN-EL	1:1	512	-	-	512
	2:1	512	512	-	1.024
	100:1	512	24.276	26.412	51.200
EN-ET	1:1	512	-	-	512
	2:1	512	512	-	1.024
	100:1	512	50.688	-	51.200
EN-LT	1:1	512	-	-	512
	2:1	512	512	-	1.024
	100:1	512	50.688	-	51.200
EN-LV	1:1	512	-	-	512
	2:1	512	512	-	1.024
	100:1	512	50.688	-	51.200
EN-RO	1:1	512	-	-	512
	2:1	512	512	-	1.024
	100:1	512	50.688	-	51.200
EN-SL	1:1	512	-	-	512
	2:1	512	512	-	1.024
	100:1	512	15.857	34.831	51.200

4.8 taula: Lerrokapen emaitzak ACCURAT corpusean F1 metrikarekin neurtuta.

HIZK. PAREA	CORPUSA	LEXACC	STACC	Δ
EN-DE	1:1	96,0	**96,7**	+0,7
	2:2	83,4	**89,2**	+5,8
	100:100	16,6	**33,7**	+17,1
EN-EL	1:1	**89,5**	88,8	-0,7
	2:2	**83,2**	**83,2**	0,0
	100:100	22,7	**37,3**	+14,6
EN-ET	1:1	88,9	**92,0**	+3,1
	2:2	73,9	**79,9**	+6.0
	100:100	34,2	**42,5**	+6.3
EN-LT	1:1	93,1	**96,1**	+3,0
	2:2	81,2	**86,9**	+5.7
	100:100	45,1	**56,0**	+10,9
EN-LV	1:1	95,0	**96,6**	+1,6
	2:2	83,8	**88,2**	+4,4
	100:100	45,1	**56,2**	+11,1
EN-RO	1:1	**99,4**	98,8	-0,6
	2:2	**95,3**	**95,3**	0,0
	100:100	70,4	**75,7**	+5,3
EN-SL	1:1	88,5	**89,5**	+1,0
	2:2	81,6	**82,3**	+0,7
	100:100	24,9	**35,3**	+10,4

4.3 irudia: STACC eta lerrokapen onenarekin optimizatutako konfigurazioaren arteko konparaketa ACCURAT corpusean.



neurtu da. 4.3. irudian ikus daitezke lortutako emaitzak 100:100 corpusetan. Hizkuntza pare guztietan optimizatutako bertsioarekin hobekuntzak lortzen dira, non diferentzia handiena grezierarekin ematen den +4,9 puntuko aldearekin.

Azpiatal honetan STACC sistemaren emaitzak egonkorrak direla ikusi dugu LEXACC-ekin konparatuz zazpi hizkuntza paretan. Oro har, STACC da emaitza onenak lortzen dituena hizkuntza eta zarata maila ezberdinetako corpus guztietan salbuespen gutxi batzuk alde batera utzita.

4.3.3. WIKIPEDIA

Azpiatal honetan WIKIPEDIA-tik erauzitako esaldiak lerrokatzen izandako emaitzak aurkezten dira. Corpus honek STACC beste hiru hizkuntza paretan esperimentuak egiteko eta beste sistema berri batekin konparaketak egiteko aukera ematen du.

Corpusa. Esperimentu hauetarako corpusa WIKIPEDIA-tik erauzi da. WIKIPEDIA-ko artikuluak esaldi mailan eskuz lerrokatu dira eta hiru hizkuntza-

4.9 taula: WIKIPEDIA corpusaren esaldi pare tamaina. Corpusak sortzeko hiru datu iturri erabili dira: (1) WIKI, WIKIPEDIA corpusa, (2) EUROP., EUROPARL corpusa, eta (3) NEWS C., NEWS COMMENTARY corpusa.

HIZK.	CORPUSA	ITURRIA			GUZTIRA
		WIKI	EUROP.	NEWS C.	
BG-EN	1:1	516	-	-	512
	100:1	516	51.084	-	51.200
DE-EN	1:1	314	-	-	512
	100:1	314	-	31.086	51.200
ES-EN	1:1	500	-	-	512
	100:1	500	-	49.500	51.200

paretako esaldiak daude: ingelesa-alemana, ingelesa-gaztelania eta ingelesa-bulgariera (Smith *et al.*, 2010)[9].

ACCURAT corpusaren kasuaren antzera (ikus aurreko 4.3.2. azpiatala), WIKIPEDIA corpusari zarata gehitu zaio 1:1 eta 100:1 proportzioekin. Ingelesa-bulgarierarako EUROPARL corpusa erabili da, eta ingelera-aleman eta ingelera-gaztelania hizkuntza pareetarako berriz NEWS COMMENTARY corpusa[10]. Corpusaren tamaina 4.9. taulan azaltzen da.

Konparatutako sistemak. Hiru dira konparatzen diren sistemak: STACC, LEXACC eta lehen mailako Conditional Random Field kate linealetan (Lafferty *et al.*, 2001) oinarritutako beste sistema bat (Smith *et al.*, 2010) (aurrerantzean CRF).

CRF sisteman jatorrizko esaldi bakoitzeko ezkutuko aldagai batek dagokion helburuko esaldia identifikatzen du (*null* lerrokatzerik ez badago). Sistema honen egileek sailkatzaile bitar batekin (Munteanu eta Marcu, 2005) konpa-

[9]Esperimentuak egiteko unean corpusa esteka honen bidez zegoen eskura: `http://research.microsoft.com/en-us/people/chrisq/wikidownload.aspx`. Tamalez, esteka ez dabil.

[10]`http://www.statmt.org/wmt13/translation-task.html`

raketak egin zituzten eta WIKIPEDIA corpusean CRF izan zen emaitza onenak lortu zituen sistema.

LEXACC-en konfigurazioari dagokionez, aleman-ingeles hizkuntza parerako optimizatutako pisuak erabili dira (Ștefănescu *et al.*, 2012), baina bulgariera-ingelesa eta gaztelania-ingelesa hizkuntza pareetarako optimizatutako pisuak ez daude eskura, horregatik kasu hauetarako lehenetsitako pisuak erabili dira.

Emaitzak. WIKIPEDIA corpusaren emaitzak 4.10. taulan aurkezten dira. Ingeles-gaztelania eta ingeles-aleman corpusetan STACC eta LEXACC-en emaitzak antzekoak dira 1:1 corpusean, baina 100:100 corpus zaratatsuan STACC da sistema eraginkorrena +7,1 eta +4,8 puntuko aldearekin hurrenez hurren.

Ingeles-bulgariera hizkuntza pareko emaitzak deigarriak dira, hau baita kasu bakarra non LEXACC den sistema onena 1:1 eta 100:100 corpusetan. Baliteke zarata gehitzeko erabilitako corpusak emaitzetan eragina izatea. Datu zaratatsu guztiak EUROPARL corpusetik erauzi dira, eta EUROPARL-eko datuen domeinua gertuago dago itzulpen-taulak sortzeko erabilitako JRC corpusetik WIKIPEDIA corpusetik baino. STACC gehiago oinarritzen da itzulpen-tauletan LEXACC baino, non bost funtzioetatik bitan bakarrik erabiltzen diren itzulpen-taulak. Gainera, bi funtzio horietan itzulpen-taulak Levenshtein distantzia aplikatu ostean bakarrik erabiltzen da. Beraz, LEXACC-en funtzioek STACC-ek baino gaitasun handiagoa dute itzulpen-taulen ahuleziak konpentsatzeko.

CRF sistemarekin konparaketak egiteko egileek argitaratutako emaitzak erabili dira. Bertan, lerrokapenak ebaluatzeko bestelako metrika bat erabili zuten: atalasea 80 eta 90 puntuko doitasuna lortzeko finkatzen dute, eta atalase horrekin nolako estaldura lortzen den neurtzen da. 4.11. taulan azaltzen dira 1:1 corpusean lortutako emaitzak[11]. Zenbait kasutan, STACC eta LEXACC sistemekin atalase baxuenarekin ere 80 eta 90 puntutik gorako doitasuna lortzen da. Kasu hauek ↑ ikurrarekin adierazten dira. Ikus daitekeenez, STACC eta LEXACC sistemek CRF-k baino emaitza hobeak lortzen dituzte. Ebaluazio honetan ere F1 metrikarekin neurtutako kasuaren antzeko egoera dugu, hau

[11] 100:100 corpusa azpiatal honetarako bereziki prestatu da, beraz, corpus horretarako emaitzak ez daude eskura CRF sistemarako.

4.10 taula: Lerrokapen emaitzak WIKIPEDIA corpusean F1 metrikarekin neurtuta.

HIZK. PAREA	CORPUSA	LEXACC	STACC	Δ
EN-BG	1:1	**87,1**	84,9	-2,2
	100:100	**27,6**	16,6	-11,0
EN-DE	1:1	**82,7**	82,0	-0,7
	100:100	31,0	**35,8**	+4,8
EN-ES	1:1	98,2	**99,7**	+1,5
	100:100	66,2	**73,3**	+7,1

4.11 taula: Lerrokapen emaitzak WIKIPEDIA corpusean. 80 eta 90 puntuko doitasunarekin nolako estaldura lortzen den adierazten da. Atalase baxuenarekin 80 eta 90 puntuko doitasuna hobetzen diren kasuak ↑ ikurrarekin adierazten dira.

HIZK. PAREA	CRF		LEXACC		STACC	
	E@90	E@80	E@90	E@80	E@90	E@80
EN-BG	72,0	81,8	↑ **80,4**	↑ 80,4	80,2	↑ **81,6**
EN-DE	58,7	68,8	**75,2**	78,7	68,8	↑ **81,8**
EN-ES	90,4	93,7	↑ 97,0	↑ 97,0	↑ **99,6**	↑ **99,6**

da, LEXACC-ek 90 puntuko doitasunarekin STACC-ek baino estaldura hobea lortzen du bulgariera eta alemanarekin. Gainerako kasu guztietan STACC da sistema eraginkorrena.

Esan bezala, hau da STACC eta LEXACC gertuen dauden esperimentua. Batez ere interesgarria da ingelera-bulgariera corpusaren kasua, hau baita aurkitutako lehen kasua non LEXACC sistema eraginkorrena den. Hala ere, lortutako emaitzak positiboak dira STACC sistemarako, aipatutako kasua alde batera utzita STACC baita emaitza onenak lortu dituena, eta gainera, beste sistemek baino teknika sinpleagoak erabiltzen ditu inolako entrenamendurik behar izanik gabe.

4.4. Metodoaren Hobekuntzak

Aurreko 4.3. azpiatalean STACC sistemarekin zenbait esperimentu egin dira eta izandako emaitzek STACC-en eraginkortasuna eta eramangarritasuna frogatu dute. Atal honen helburua oinarrizko metodoaren gainean zenbait optimizazio egitea da, eta optimizazio horien hobekuntzak neurtzea oinarrizko metodoarekin konparaketak eginez. Bi dira egindako optimizazioak: pisu lexikoak eta izen-entitateen zigorra.

4.4.1. Pisu Lexikoak

Pisu lexikoak 3.3.1. azpiatalean azaltzen dira dokumentuen lerrokatzerako. Labur esanda, pisu lexikoen ideia token multzoko terminoak orekatzea da termino bakoitzak corpusean duen garrantziaren arabera. Adibidez, demagun $T = \{$ *"katua"*, *"eta"*, *"zakurra"*$\}$ token multzoa lerrokatu nahi dugula eta bi kandidatu ditugula, $S_1 = \{$ *"hori"*, *"eta"*, *"hura"*$\}$ eta $S_2 = \{$ *"katua"*, *"ta"*, *"txakurra"*$\}$ token multzoak. Bi lerrokatzeek termino bakarra dute amankomunean, $T \cap S_1 = \{$ *"eta"*$\}$ eta $T \cap S_2 = \{$ *"katua"*$\}$, baina intuitiboki behintzat *"katua"* terminoak *"eta"* terminoak baino pisu handiagoa izan beharko luke.

Dokumentuak lerrokatzen egindako esperimentuetan gehienez +1.11 puntuko hobekuntza lortzen da F1 metrikan (oro har pisu lexikorik erabili gabe 80 puntu baino gehiagoko emaitzak lortzen dira), baina esaldietan dokumentuetan baino informazio gutxiago dagoenez bestelako emaitzak lortu litezke. Izan ere, informazio lexiko nahikoa izan ohi da dokumentu bakoitzean, eta tamaina honek nolabait berdintzen du hitz funtzionalek ekar lezaketen desoreka pisu lexikoak erabili behar izan gabe.

Pisu lexikoen eragina aztertzeko BUCC 2017 atazan parte hartu genuen. Hurrengo 4.4.1.1. azpiatalean azaltzen dira parte-hartzearen xehetasunak.

4.4.1.1. BUCC 2017

BUCC 2017ko (10th Workshop on Building and Using Comparable Corpora) esaldien lerrokatze ataza[12] (Zweigenbaum *et al.*, 2017), esalditan banatutako

[12] `https://comparable.limsi.fr/bucc2017/bucc2017-task.html`

bi corpus elebakar emanda, esaldi paraleloak identifikatzean datza. Lerrokapenik gabeko esaldiak WIKIPEDIA-tik erauzi dira[13], eta esaldi paraleloak NEWS COMMENTARY corpusetik[14]. Atazan lau hizkuntza paretako corpusak ematen dira: aleman-ingelesa, frantses-ingelesa, errusiera-ingelesa eta txinatarra-ingelesa. Gure kasuan aleman-ingeles eta frantses-ingeles hizkuntza pareetan bakarrik parte hartzea erabaki genuen[15].

4.12. taulan azaltzen dira corpusaren xehetasunak. Kontuan izan behar da jatorrizko corpusean zenbait esaldi bikoiztuta daudela, horregatik taulako estatistikak apur bat aldatzen dira datu ofizialekin alderatuta. Ikus daitekeenez, esaldi guztien artetik %2-%4 bakarrik dira paraleloak, beraz, zarata maila handiko corpusak dira.

4.12 taula: BUCC 2017 corpusa.

CORPUSA	HIZK.	CORPUS ELEBAKARRA			CORPUS PARALELOA		
		SAMPLE	TRAIN	TEST	SAMPLE	TRAIN	TEST
DE-EN	DE	32.593	413.869	413.884	1.037	9,573	9.550
	EN	40.354	399.337	396.534	1.037	9,573	9.550
FR-EN	FR	21.497	271.874	276.833	929	9.080	9.043
	EN	38.069	369.810	373.459	929	9.080	9.043

STACC-ek itzulpen-taulak behar ditu jatorrizko token multzoak itzultzeko. Orain arte JRC-ACQUIS COMMUNAUTAIRE corpusa erabili da itzulpen-taula horiek entrenatzeko, baina BUCC 2017 atazarako, itzulpen-taulen estaldura lexikoa ahalik eta gehien handitzearren, itzulpen-taula generikoak erabili dira (DOCAL-entzako erabiltzen diren berberak, ikus 3.3.2. azpiatala).

Konfigurazioari dagokionez, parametroak TRAIN eta SAMPLE corpusetan probak eginez optimizatu dira. Itzulpen-tauletatik $k = 4$ itzulpen onenak bila-

[13]2016-ko abenduaren 1-eko bertsioa: `http://ftp.acc.umu.se/mirror/wikimedia.org/dumps/`

[14]11. bertsioa: `http://www.casmacat.eu/corpus/news-commentary.html`

[15]BUCC 2018 atazan ere parte hartu genuen, baina urte hartan hizkuntza pare guztiak aukeratu genituen. 4.4.2.1. azpiatalean aurkezten dira parte-hartzearen nondik-norakoak.

tzen dira. LUCENE indizetik gehienez 100 esaldi berreskuratzen dira. Lerrotze onenaren optimizazioa egiten da. α-ren balio ezberdinekin ere probak egin dira pisu lexikoen funtzioaren kurbaren leuntasuna kontrolatzeko. Azkenik, lerrokatze atalase ezberdinak erabili dira konfiantza baxuko lerrokatzeak baztertzeko.

Parte-hartzaile bakoitzak hiru sistema/konfigurazioren emaitzak bidali zitzakeen. Gure kasuan STACC-en hiru konfigurazio bidaltzea erabaki zen bakoitza lerrokatze atalase ezberdinarekin. Atalase bakoitzak metrika bat optimizatzeko helburua du:

- STACC.LEX.F. TRAIN corpusean emaitza onenak lortzen dituen atalasea F1 metrikan. .LEX atzizkiak pisu lexikoen erabilera adierazten du.
- STACC.LEX.D. Aurrekoaren antzera TRAIN corpusean doitasun onena lortzen duen atalasea.
- STACC.LEX.E. TRAIN corpusean estaldura onena lortzen duen atalasea.

Hiru konfigurazio hauekin doitasun, estadura eta F1 metrika optimizatzeak STACC sisteman duen eragina neurtzea da.

4.13 eta 4.14. tauletan ikus daitezke emaitza ofizialak. Aleman-ingeles corpusean STACC izan zen parte-hartzaile bakarra, baina frantses-ingeles corpusean hiru parte-hartzaileetatik STACC-en konfigurazio guztiek lortu dituzte emaitza onenak nahiko alde handiarekin. STACC-en bertsio bakoitza da optimizatutako metrikan emaitza onenak lortzen dituena.

Ezjakina da corpus konparagarrian dauden esaldi guztietatik zehazki zeintzuk diren paraleloak. Horregatik, baliteke STACC-ek izandako gezurrezko positiboen artean benetako esaldi paraleloak egotea. Izan ere, BUCC 2017 atazaren antolatzaileek honen inguruan zenbait ebaluazio egin zituzten. Frantses-ingeles corpusean STACC-ek 978 gezurrezko positibo ditu. Hauetatik 60 lerrokatze eskuz analizatu ziren, eta horien artean 3-5 itzulpen bikain eta beste 8-13 itzulpen ia bikain aurkitu ziren (Zweigenbaum *et al.*, 2017).

4.15 eta 4.16. tauletan corpus guztietan izandako emaitzen xehetasunak aurkezten dira. Lehenik eta behin, corpus guztietan pisu lexikoek bultzada

4.13 taula: Emaitzak BUCC 2017-ko aleman-ingeles corpusean.

SISTEMA	LERROK. KOP.	DOITASUNA	ESTALDURA	F1
STACC.LEX.F	8.640	88	80	**84**
STACC.LEX.E	9.949	82	**85**	**84**
STACC.LEX.D	7.586	**92**	73	82

4.14 taula: Emaitzak BUCC 2017-ko frantses-ingeles corpusean.

SISTEMA	LERROK. KOP.	DOITASUNA	ESTALDURA	F1
STACC.LEX.F	8.831	80	79	**79**
STACC.LEX.D	7.569	**87**	73	**79**
STACC.LEX.E	10.768	70	**83**	76
RALI2	47.576	12	63	20
RALI1	57.761	10	66	18
RALI3	66.201	9	63	15
JUNLP1	38.736	3	11	4
MIN	7.569	9	63	15
MEDIAN	29.172	41	70	48
MEAN	33.118	45	71	48
MAX	66.201	87	83	79
STDDEV	24.062	34	7	30

4.15 taula: Emaitzak BUCC 2017-ko aleman-ingeles corpusean. LUC indizetik esaldi egokia berreskuratzeko ehunekoa da, ATAL atalase optimoa da, DOI doitasuna eta EST estaldura.

CORPUSA	SISTEMA	α	ATAL	LUC	DOI	EST	F1
SAMPLE	STACC.LEX.F	100	0,16	99,04	95,46	91,32	**93,35**
	STACC.LEX.D	100	0,17	99,04	**97,95**	87,75	92,57
	STACC.LEX.E	100	0,15	99,04	88,27	**93,64**	90,88
	STACC	-	0,22	99,04	91,84	80,33	85,70
TRAIN	STACC.LEX.F	250	0,17	98,50	86,99	78,96	**83,33**
	STACC.LEX.D	250	0,18	98,50	**90,89**	73,41	81,23
	STACC.LEX.E	250	0,16	98,50	80,21	**85,55**	82,79
	STACC	-	0,23	98,50	79,26	69,16	73,87
TEST	STACC.LEX.F	250	0,17	98,63	88,15	79,75	**83,74**
	STACC.LEX.D	250	0,18	98,63	**92,10**	73,16	81,55
	STACC.LEX.E	250	0,16	98,63	81,93	**85,35**	83,60

handia ematen dutela ikus daiteke, gutxi gora behera +10 puntuko hobekuntzarekin. Beste ondorio garrantzitsu bat lortutako emaitzen egonkortasuna da; corpus guztietan 80 puntutik gorako emaitzak lortzen dira optimizatutako metrikan.

Erakutsitako tauletan ez da azaltzen, baina α parametroaren eragina nahiko ahula da: 100 eta 500 arteko balioekin lortutako emaitzak ia berdinak dira; tarte horretatik oso urrun dauden balioekin bakarrik okertzen dira emaitzak.

DOCAL-ekin gertatzen ez den bezala, atal honetako esperimentuetan izandako emaitzek pisu lexikoekin lerrokatzeak nabarmen hobetzen direla adierazten dute. Gainera, pisu lexikoak automatikoki kalkulatzen dira uneko corpuseko terminoen maiztasunarekin eta ez da aparteko konfiguraziorik egin behar. Beraz, pisu lexikoen erabilera guztiz bat dator eramangarritasun eta eraginkortasun helburuekin.

4.16 taula: Emaitzak BUCC 2017-ko frantses-ingeles corpusean. LUC indizetik esaldi egokia berreskuratzeko ehunekoa da, ATAL atalase optimoa da, DOI doitasuna eta EST estaldura.

CORPUSA	SISTEMA	α	ATAL	LUC	DOI	EST	F1
SAMPLE	STACC.LEX.F	500	0,14	99,46	90,51	91,39	**90,95**
	STACC.LEX.D	500	0,15	99,46	**93,74**	86,98	90,23
	STACC.LEX.E	500	0,13	99,46	83,13	**93,33**	87,93
	STACC	-	0,22	99,46	89,36	75,03	81,57
TRAIN	STACC.LEX.F	250	0,16	96,84	78,43	79,23	**78,83**
	STACC.LEX.D	250	0,17	96,84	**84,36**	73,40	78,50
	STACC.LEX.E	250	0,15	96,84	68,51	**83,83**	75,40
	STACC	-	0,23	96,84	72,69	63,12	67,57
TEST	STACC.LEX.F	250	0,16	96,81	80,41	78,52	**79,46**
	STACC.LEX.D	250	0,17	96,81	**87,08**	72,89	79,35
	STACC.LEX.E	250	0,15	96,81	69,82	**83,14**	75,90

4.4.2. Izen-Entitateen Penalizazioa

STACC-en oinarrizko metodoak izen-entitateak kontutan hartzen ditu itzulpen token multzoak sortzeko orduan. Termino bat letra larriz hasten bada eta itzulpen-taulan agertzen ez bada, orduan termino hori itzulpen token multzoan sartzen da. Intuitiboki behintzat, izen-entitateek pisu handia izan behar lukete esaldien arteko antzekotasunean. Izen-entitateen pisua handitzeko asmoz, jatorrizko antzekotasun metrikari esaldien artean partekatzen ez diren izen-entitateen agerpena zigortzeko funtzio bat gehitzen zaio.

Esaldi bakoitzaren izen-entitateak gordetzeko, esaldi bakoitzeko N izen-entitateen multzoa sortzen da letra larriz hasten diren terminoekin. Izan bitez s_i eta s_j jatorrizko eta helburuko esaldiak hurrenez hurren, S_i eta S_j s_i eta s_j esaldien token multzoak, N_i eta N_j s_i eta s_j esaldien izen-entitateen multzoak. s_i eta s_j esaldien artean izen-entitate ezberdinen agerpena zigortzeko $zig(s_i, s_j)$ funtzioa 4.4. ekuazioan azaltzen den bezala kalkulatzen da.

$$zig(s_i, s_j) = \frac{|\,(N_i - N_j) \cup (N_j - N_i)\,|}{|\,S_i \cup S_j\,|} \tag{4.4}$$

Esaldietan izen-entitaterik ez badago funtzioaren balioa 0 izango da, hortaz, jatorrizko antzekotasun metrikan ez luke inongo eraginik izango. Esaldien artean zenbat eta izen-entitate ezberdin gehiago egon, funtzioaren balioa 1-etik orduan eta hurbilago egongo da.

Jatorrizko antzekotasun metrikan zigortze funtzioa 4.5. ekuazioan azaltzen den bezala integratzen da.

$$stacc_{zig}(s_i, s_j) = stacc(s_i, s_j) - zig(s_i, s_j) \tag{4.5}$$

Era berean, 4.6. ekuazioak zigortze funtzioa pisu lexikoekin batera nola erabiltzen den adierazten du.

$$stacc_{lex\_zig}(s_i, s_j) = stacc_{lex}(s_i, s_j) - zig(s_i, s_j) \tag{4.6}$$

Izen-entitateen zigortze funtzioaren eragina BUCC 2018 atazan aztertzen da.

4.4.2.1. BUCC 2018

BUCC 2018ko (11th Workshop on Building and Using Comparable Corpora) esaldien lerrokatze atazan[16] (Pierre Zweigenbaum eta Rapp, 2018), esalditan banatutako bi corpus elebakar emanda, esaldi horietatik paraleloak zeintzuk diren identifikatu behar da. 2017 eta 2018. urteko corpusak berberak dira, beraz, BUCC 2018 atazan ere WIKIPEDIA eta NEWS COMMENTARY corpusetik erauzitako esaldiak lerrokatu behar dira lau hizkuntza paretarako: alemaningelesa, frantses-ingelesa, errusiera-ingelesa eta txinatarra-ingelesa. BUCC 2018 atazan STACC-en optimizazioen eraginkortasuna neurtzen da hizkuntza pare guztietan.

4.17. taulan azaltzen dira corpusaren zehaztasunak. Oro har, esaldi pare guztietan zarata maila handia dago, corpus elebakarreko esaldien artean %2-%4 bakarrik lerrokatzen baitira.

4.17 taula: BUCC 2018 corpusa.

CORPUSA	HIZK.	CORPUS ELEBAKARRA			CORPUS PARALELOA		
		SAMPLE	TRAIN	TEST	SAMPLE	TRAIN	TEST
DE-EN	DE	32.593	413.869	413.884	1.038	9,580	9.550
	EN	40.354	399.337	396.534	1.038	9,580	9.550
FR-EN	FR	21.497	271.874	276.833	929	9.086	9.043
	EN	38.069	369.810	373.459	929	9.086	9.043
RU-EN	RU	45,459	460.853	457,327	2,374	14,435	14,330
	EN	72.766	558.401	566.356	2.374	14.435	14.330
ZH-EN	ZH	8.624	94.637	91.824	257	1.899	1.896
	EN	13.589	88.860	90.037	257	1.899	1.896

Terminoen itzulpenak bilatzeko itzulpen-taulak erabiltzen dira. Alemaningeles eta frantses-ingeles hizkuntza pareentzat DOCAL-en itzulpen-taula generikoak erabili dira (ikus 3.3.2. azpiatala), eta txinera-ingeleserako BUCC

[16] https://comparable.limsi.fr/bucc2018/bucc2018-task.html

2015 atazan MULTIUN (Eisele eta Chen, 2010) bidez entrenatutako itzulpen-taulak (ikus 3.2.2. azpiatala). Errusiera-ingeleserako itzulpen-taula berriak sortu behar izan dira MULTIUN bidez.

STACC-en konfiguraziorako 2017an erabilitako parametro berberak erabili dira: itzulpen-tauletatik $k = 4$ itzulpen onenak bilatzen dira, LUCENE indizetik gehienez 100 esaldi berreskuratzen dira eta lerrokatze onenaren optimizazioa egiten da. 2017an pisu lexikoen erabilerak jatorrizko metodoa nabarmen hobetzen duela ondorioztatu zen, beraz, BUCC 2018an ere pisu lexikoak erabiltzen dira, baina kasu honetan $\alpha = 250$ balioarekin. Corpus txinatarra berezia da, tokenak segmentutan banatu behar baitira. Horretarako STANFORD SEGMENTER erabili da (Tseng *et al.*, 2005).

2017. urtean bezala, BUCC 2018 atazan parte-hartzaile bakoitzak hiru sistema/konfigurazioren emaitzak bidali zitzakeen. Bidalitako hiru konfigurazioak honako hauek dira:

- STACC.LEX.ZIG.F. Pisu lexiko eta izen-entitateen zigortze funtzioa erabiltzen dituen bertsioa. Erabilitako atalaseak F1 metrika optimizatzen du TRAIN corpusean.

- STACC.LEX.ZIG.D. Aurrekoaren antzera TRAIN corpusean doitasun onena lortzen duen atalasearekin.

- STACC.LEX.F. TRAIN corpusean F1 metrikan emaitza onena lortzen duen atalasearekin, baina zigortze funtzioa erabili gabe.

4.18, 4.19, 4.20 eta 4.21. tauletan ikus daitezke emaitza ofizialak. Hizkuntza pare guztietan STACC da emaitza onenak lortzen dituen sistema, eta gainera, lehengo urteko emaitzak hobetzen dira kasu guztietan. F1 metrikan %80-tik gorako emaitzak lortzen dira aleman-ingeles, frantses-ingeles eta errusiera-ingeles corpusetan, eta %90-etik gorako doitasuna lortzen da corpus berberetan. Txinera-ingeles corpuseko emaitzak zertxobait okerragoak diren harren, F1 metrikan ehunekoa 77-koa da, eta doitasunean 89-koa.

4.18 taula: Emaitzak BUCC 2018-ko aleman-ingeles corpusean.

SISTEMA	LERROK. KOP.	DOITASUNA	ESTALDURA	F1
STACC.LEX.ZIG.F	9.271	87	**84**	**86**
STACC.LEX.ZIG.D	8.265	**91**	79	85
STACC.LEX.F	8.769	88	81	84
STACC.LEX.F 2017	8.640	88	80	84

4.19 taula: Emaitzak BUCC 2018-ko frantses-ingeles corpusean.

SISTEMA	LERROK. KOP.	DOITASUNA	ESTALDURA	F1
STACC.LEX.ZIG.F	8.136	86	77	**81**
STACC.LEX.ZIG.D	7.173	**91**	72	80
STACC.LEX.F	8.887	80	**79**	80
H2@BUCC18_1	7.947	82	72	76
H2@BUCC18_2	9.607	71	75	73
H2@BUCC18_3	8.300	70	64	67
STACC.LEX.F 2017	8.831	80	79	79

4.20 taula: Emaitzak BUCC 2018-ko errusiera-ingeles corpusean.

SISTEMA	LERROK. KOP.	DOITASUNA	ESTALDURA	F1
STACC.LEX.ZIG.F	11.010	86	77	**81**
STACC.LEX.F	11.370	79	**79**	79
STACC.LEX.ZIG.D	10.127	**90**	71	79

4.22, 4.23, 4.24 eta 4.25. tauletan SAMPLE, TRAIN eta TEST corpusetan izandako emaitzak azaltzen dira. Izen-entitateen zigortze funtzioak berez onak diren emaitzak hobetzea lortzen du. Hala ere, hobekuntza horiek ez

4.21 taula: Emaitzak BUCC 2018-ko txinera-ingeles corpusean.

SISTEMA	LERROK. KOP.	DOITASUNA	ESTALDURA	F1
STACC.LEX.F	1.763	80	**75**	**77**
STACC.LEX.ZIG.F	1.680	80	71	75
STACC.LEX.ZIG.D	1.373	**89**	64	74
NLP2CT1	1.169	73	45	55
NLP2CT2	1.209	72	46	56
Z_NLP1 2017	1.985	42	44	43

dira hain nabarmenak pisu lexikoekin lortzen direnekin alderatuta, gehienez F1 metrika +2 puntutan hobetzea lotzen baita. Txinera-ingelesa corpusa da salbuespen kasua. Corpus horretan emaitzak -3 puntutan okertzen dira, hala ere, emaitza hau ez da ustekabekoa, alfabeto txinatarrean izen-entitateak ezin baitira antzeman terminoen lehendabiziko karakterea aztertuz.

Laburbilduz, izen-entitateen zigortze funtzioak lerrokatzeen kalitatea handitzen du. Nahiz eta emaitzak ez diren gehiegi hobetzen, konputazionalki oso metodo arina da eta hobekuntzak nahiko egonkorrak dira. Hala ere, txinera bezalako hizkuntzetan ez da bere erabilpena gomendatzen, izen-entitateak detektatzeko bestelako teknikak erabili beharko lirateke eta.

4.22 taula: Emaitzak BUCC 2018-ko aleman-ingeles corpusean. LUC indizetik esaldi egokia berreskuratzeko ehunekoa da, ATAL atalase optimoa da, DOI doitasuna eta EST estaldura.

CORPUSA	SISTEMA	α	ATAL	LUC	DOI	EST	F1
SAMPLE	STACC.LEX.ZIG.F	250	0,15	99,04	97,36	89,01	93,00
	STACC.LEX.ZIG.D	250	0,16	99,04	**99,21**	85,54	91,87
	STACC.LEX.F	250	0,15	99,04	95,09	**91,51**	**93,27**
TRAIN	STACC.LEX.ZIG.F	250	0,16	98,50	84,81	**83,74**	**84,27**
	STACC.LEX.ZIG.D	250	0,17	98,50	**89,86**	78,28	83,67
	STACC.LEX.F	250	0,17	98,50	87,00	79,96	83,33
TEST	STACC.LEX.ZIG.F	250	0,16	98,65	86,81	**84,27**	**85,52**
	STACC.LEX.ZIG.D	250	0,17	98,65	**91,47**	79,16	84,87
	STACC.LEX.F	250	0,17	98,65	88,06	80,86	84,31

4.23 taula: Emaitzak BUCC 2018-ko frantses-ingeles corpusean. LUC indizetik esaldi egokia berreskuratzeko ehunekoa da, ATAL atalase optimoa da, DOI doitasuna eta EST estaldura.

CORPUSA	SISTEMA	α	ATAL	LUC	DOI	EST	F1
SAMPLE	STACC.LEX.ZIG.F	250	0,14	99,46	92,26	**91,07**	**91,66**
	STACC.LEX.ZIG.D	250	0,15	99,46	**95,33**	87,84	91,43
	STACC.LEX.F	250	0,15	99,46	92,44	89,45	90,92
TRAIN	STACC.LEX.ZIG.F	250	0,16	96,84	83,93	77,58	**80,63**
	STACC.LEX.ZIG.D	250	0,17	96,84	**87,81**	71,69	78,93
	STACC.LEX.F	250	0,16	96,84	78,43	**79,23**	78,83
TEST	STACC.LEX.ZIG.F	250	0,16	96,87	86,01	77,39	**81,47**
	STACC.LEX.ZIG.D	250	0,17	96,87	**90,62**	71,88	80,17
	STACC.LEX.F	250	0,16	96,87	80,27	**78,89**	79,58

4.24 taula: Emaitzak BUCC 2018-ko errusiera-ingeles corpusean. LUC indizetik esaldi egokia berreskuratzeko ehunekoa da, ATAL atalase optimoa da, DOI doitasuna eta EST estaldura.

CORPUSA	SISTEMA	α	ATAL	LUC	DOI	EST	F1
SAMPLE	STACC.LEX.ZIG.F	250	0,14	97,81	96,46	**88,37**	**92,24**
	STACC.LEX.ZIG.D	250	0,15	97,81	**97,94**	84,16	90,53
	STACC.LEX.F	250	0,15	97,81	95,42	86,98	91,01
TRAIN	STACC.LEX.ZIG.F	250	0,16	96,64	84,87	77,26	**80,89**
	STACC.LEX.ZIG.D	250	0,17	96,64	**88,05**	71,02	78,63
	STACC.LEX.F	250	0,16	96,64	77,69	**79,77**	78,72
TEST	STACC.LEX.ZIG.F	250	0,16	96,81	86,31	76,83	**81,30**
	STACC.LEX.ZIG.D	250	0,17	96,81	**89,91**	70,67	79,14
	STACC.LEX.F	250	0,16	96,81	79,44	**79,34**	79,39

4.25 taula: Emaitzak BUCC 2018-ko txinera-ingeles corpusean. LUC indizetik esaldi egokia berreskuratzeko ehunekoa da, ATAL atalase optimoa da, DOI doitasuna eta EST estaldura.

CORPUSA	SISTEMA	α	ATAL	LUC	DOI	EST	F1
SAMPLE	STACC.LEX.ZIG.F	250	0,12	100,00	95,79	**70,82**	**81,43**
	STACC.LEX.ZIG.D	250	0,13	100,00	**98,82**	65,37	78,69
	STACC.LEX.F	250	0,12	100,00	91,27	89,49	90,37
TRAIN	STACC.LEX.ZIG.F	250	0,13	97,05	79,26	70,62	**74,69**
	STACC.LEX.ZIG.D	250	0,14	97,05	**86,23**	64,61	73,87
	STACC.LEX.F	250	0,14	97,05	78,27	**74,72**	76,45
TEST	STACC.LEX.ZIG.F	250	0,13	97,15	79,82	70,73	**75,00**
	STACC.LEX.ZIG.D	250	0,14	97,15	**88,64**	64,19	74,46
	STACC.LEX.F	250	0,14	97,15	80,37	**74,74**	77,45

4.5. Ondorioak

4.1. atalean esaldiak lerrokatzeko metodo bat proposatzen da, STACC deritzona. Funtsean, STACC eta DOCAL metodo berdinean oinarritzen dira lerrokatzeak egiteko. Esaldiak lerrokatzeko zenbait egokitzapen egin dira; dokumentuetan esaldietan baino termino gehiago daude, eta horregatik, STACC-ek terminoen gainean prozesaketa gehiago egin behar ditu esaldiak ezberdintzeko.

STACC eta DOCAL-en arteko berezitasun nagusiak honako hauek dira:

- Aurreprozesaketa. Esaldietan termino gutxiago daudenez, esaldietako lehenengo terminoaren formak garrantzi handiagoa du. Horregatik truecasing-a egiten da.
- Aurrizki komun luzeenak. Token multzoek termino gutxiago dituztenez aurrizkien kalkuluak konputazionalki kostu txikiagoa du. Gainera, aurrizki komunen erabilpenak emaitzak nabarmen hobetzen ditu.
- Lerrokatze onenaren optimizazioa. Esaldien lerrokatzean ere emaitzak hobetzen ditu, hala ere, hobekuntzak ez dira hain handiak.
- Pisu lexikoak. Token multzoetan informazio gutxiago dagoenez esaldiek termino gutxiago partekatzen dituzte. Horregatik garrantzitsua da esanahi lexiko handiagoa duten terminoek pisu handiagoa izatea. Lortutako emaitzek pisu lexikoek lerrokatzeen kalitatea nabarmen hobetzen dutela erakusten dute.

STACC-en eraginkortasuna 12 hizkuntza pare biltzen dituzten 4 corpusetan probatu da. Orokorrean, lortutako emaitzak oso onak izan dira. EITB eta ACCURAT corpusetan LEXACC sistemarekin baino emaitza hobeak lortzen dira. WIKIPEDIA corpuseko emaitzak ere positiboak dira, bulgariera-ingeles corpusaren salbuespena alde batera utzita STACC-ekin lortzen baitira emaitza onenak. BUCC 2017 eta BUCC 2018 atazetan ere parte hartu da, eta hizkuntza pare guztietan STACC izan da emaitza onenak lortu dituen sistema nahiko alde handiarekin.

Beste bi esperimentu egin dira: itzulpen-taula lexikoak beharrean itzulpen automatikoa erabiltzea esaldietako terminoak itzultzeko, eta partekatzen ez diren izen-entitateen agerpenak zigortzeko funtzio baten erabilpena. Lehenengo esperimentuak itzulpen ereduak entrenatzea dakar, beraz, ez da hain metodo eramangarria. Gainera, egindako proben arabera emaitzak okerragoak dira. Izen-entitateen zigortze funtzioari dagokionez, BUCC 2018 atazan lortutako emaitzen arabera lerrokatzeak kasu guztietan hobetzen dira. Salbuespen bakarra txinera-ingelesa corpusa da, non txinatar alfabetoaren ezaugarriengatik izen-entitateak ez diren ondo identifikatzen.

STACC-ek ez du inolako entrenamendurik behar eta emaitzak oso egonkorrak dira. Badira konfigurazio parametro batzuk, baina hauek optimizatu beharrik ez dago emaitza onak lortzeko. Behar diren hizkuntza baliabide bakarrak itzulpen-taulak dira. Beraz, eraginkortasun eta erabilgarritasun helburuak betetzen dira.

Beste kontribuzio garrantzitsu bat STACC-en bidez lerrokatutako EITB corpusa da[17]. Gaztelania eta euskaraz idatzitako albisteak biltzen dituen corpusa da, eta guztira ia 600K esaldi pare daude. Eskuz lerrokatutako ebaluaziorako corpus bat ere sortu da esaldiak lerrokatzeko metodoen eraginkortasuna neurtzeko.

[17] EITB corpusa lortzeko jo esteka honetara: `http://metashare.elda.org/repository/search/?q=eitb+documents`

5. KAPITULUA

Datuen Aukeraketa

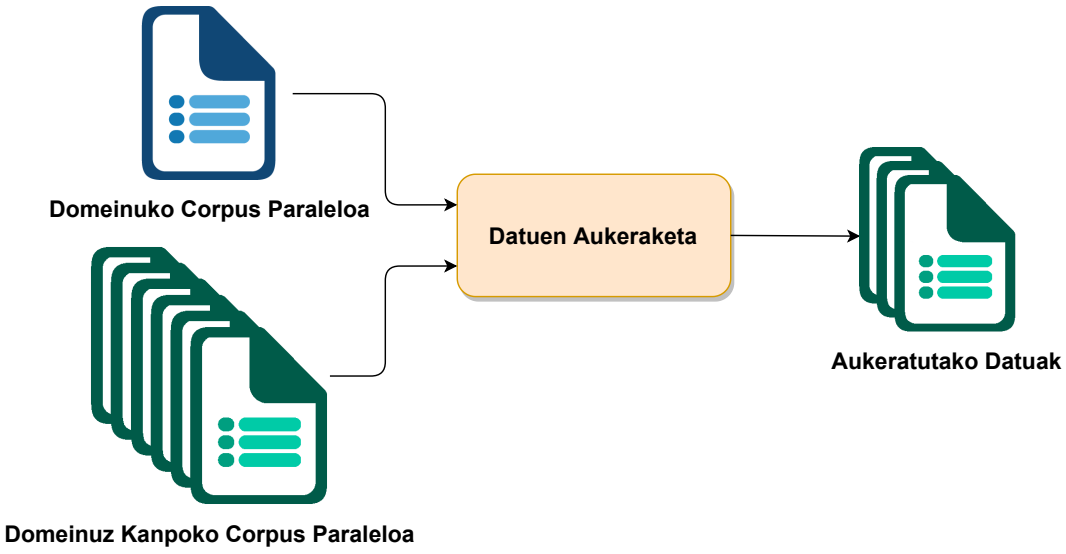
Datuetan oinarritutako itzulpen automatikoan kalitatezko corpus paralelo ugari behar dira ganorazko entrenamenduak egin ahal izateko. Askotan, entrenamendua domeinu jakin bateko corpusak erabiliz bakarrik egiten bada, estaldura eta zehaztasun arazoak izaten dira datuen urritasuna dela medio. Domeinu orokorreko corpus paraleloak ugariagoak izan ohi dira, horregatik, corpus hauetatik baliagarriak diren datu multzoak aukeratzea domeinua egokitzeko oso teknika erabilia da.

5.1. irudian ikus daiteke datuen aukeraketaren eskema orokorra. Domeinuko corpus paralelo bat eta domeinuz kanpoko beste corpus handiago bat izanda, bien arteko berdintasun eta ezberdintasunak neurtuz, domeinuz kanpoko corpusean baliagarrienak diren azpimultzoak aukeratzean datza.

Nahiz eta azpimultzo optimoak aukeratzeko metodo zehatzik oraindik ez egon, azpimultzo horiek bi ezaugarri nagusi bete behar dituztela esan daiteke. Lehendabizi, aukeratutako datu multzoek domeinuko hutsune lexiko eta sintaktikoak betetzen lagundu behar dute. Ezaugarri hau neurtzeko aukera bat lortutako hitz ezezagun berrien kantitatea ebaluatzea da. Bigarrenik, datu berriek ez lukete zaratarik gehitu behar. Domeinu batekiko datuen egokitasun hori neurtzeko, hizkuntza ereduek emandako perplexitatea erabili daiteke.

Aipatutako bi ezaugarri horiek aldi berean betetzen dituzten datu multzoak aukeratzea ez da erraza. Datu optimoek jatorrizko domeinutik hurbil egon behar dute, baina informazio berri nahikoa ez bada gehitzen, itzulpenetan ez

5.1 irudia: Datuen aukeraketa.



da eragin nabaririk egongo.

Kapitulu honetan, terminoen maiztasun erlatiboan oinarritutako datuen aukeraketa metodo bat aurkezten da, hau da, RFR metodoa (*Relative Frequency Ratios*). Esaldi bateko terminoek domeinuko eta domeinutik kanpoko corpusetan nolako maiztasun erlatiboa duten aztertuz, esaldi bakoitzaren baliagarritasuna neurtzen du. Metodo honek ez du inolako entrenamendu ezta hizkuntza baliabiderik behar, beraz, oso metodo arina eta eramangarria da. Proposatutako metodoa artearen egoeran oso erabilia den beste metodo batekin konparatzen da, termino ezezagunen kopurua aztertuz, aukeratutako esaldien perplexitatea neurtuz eta itzulpen metrika automatikoekin ebaluazio sakon bat eginez.

Hurrengo atalak honela antolatzen dira. 5.1. atalean terminoen maiztasun erlatiboan oinarritutako metodoa aurkezten da. 5.2. atalean berriz, egindako esperimentuetan erabilitako corpusa, ereduak eta emaitzak erakusten dira. Azkenik, 5.3. atalean ondorioak azaltzen dira.

5.1. Maiztasun Erlatiboak Ustiatzen, RFR

Perplexitatean oinarritutako datuen aukeraketa tekniketan, domeinutik kanpoko esaldi guztien artean jatorrizko domeinuarekiko antzeko n-grama banaketa duten tamaina txikiko esaldiak aukeratzeko joera dago (Sethy *et al.*, 2009). Nahiz eta ezaugarri hau zenbait corpusetarako onuragarria izan daitekeen (maiztasun handiko tamaina txikiko esaldiak dituztenak: filmen azpitituluak, eskuliburu teknikoak, etab.), orohar domeinuarekin erlazioa duten n-grama banaketa ezberdineko esaldien ekarpena alde batera uzten da. Kasurik okerrenean, baliteke perplexitatean oinarritutako datu aukeraketa sistemek dagoeneko domeinuko corpusean dauden esaldi berberak aukeratzea, eta esaldi hauek ez lukete corpusaren estalduran inolako eraginik izango. Oinarrizko hipotesia, itzulpen automatikoari begira ikuspegi lexiko eta sintaktikotik aberasgarrienak diren datuak aukeratu beharko liratekeela da. Hipotesi hau probatzeko, domeinutik kanpoko corpuseko esaldiak domeinuarekiko baliagarritasun lexiko eta sintaktikoaren arabera antolatzen dituen metodo bat diseinatu da (RFR, *Relative Frequency Ratios*), betiere esaldien antolakuntzan domeinuko n-grama banaketa kontutan eduki gabe.

5.1.1. Maiztasun Erlatiboen Ratioa

Proposatutako datuen aukeraketa teknikak esaldien arteko antzekotasuna estimatzen du domeinuko eta domeinutik kanpoko esaldien terminoen maiztasun erlatiboak kalkulatuz. Lehenengo pausua corpus bakoitzeko terminoen maiztasun erlatiboa kalkulatzea da[1]. Zehazki esanda, lehendabizi c corpuseko w hitz bakoitzaren maiztasun erlatiboa 5.1. ekuazioan azaltzen den bezala kalkulatzen da, non $C(w)$ funtzioa w terminoaren agerpen kopurua den.

$$\phi_c(w) = \frac{C(w)}{\sum_{i=1}^{|c|} C(w_i)} \tag{5.1}$$

Domeinuz kanpoko (s, t) esaldi pare bakoitzeko, non s jatorrizko hizkuntzako terminoen multzoa den eta t helburuko hizkuntzako terminoen multzoa,

[1]Esaldietan ez da inolako iragazpenik egiten, hots, puntuazio ikurrekin ere egiten da kalkulua.

domeinuarekiko antzekotasun metrika termino bakoitzaren maiztasun erlatiboak batuz kalkulatzen da, 5.2. ekuazioan azaltzen den antzekotasun metrika erabiliz.

$$rfr(s,t) = \frac{1}{2}\left(\sum_{i=1}^{|s|} \frac{\phi_d(w_i)}{\phi_o(w_i)} + \sum_{i=1}^{|t|} \frac{\phi_d(w_i)}{\phi_o(w_i)} \right) \tag{5.2}$$

5.2. ekuazioan, ϕ_d eta ϕ_o jatorrizko domeinuko eta domeinutik kanpoko maiztasun erlatiboak dira. Esaldien tamainaren arteko ezberdintasunak mugatzearren, antzekotasun kalkulua terminoen multzoen gainean kalkulatzen da, hau da, termino berbera esaldian bi aldiz agertzen bada, termino horrentzako behin bakarrik kalkulatuko da maiztasun erlatiboaren ratioa. Domeinutik kanpoko termino batek agerpenik ez badu domeinuan, maiztasun erlatiboaren ratioa 0-koa izango da.

Metrika honekin jatorrizko domeinuan domeinutik kanpo baino agerpen gehiago dituzten terminoz osatutako esaldiak hobesten dira. Gainera, maiztasun handiko terminoak ere kontutan hartzen dira (hala nola, termino funtzionalak), estilo, erregistro eta lexikoari dagokionez, esaldien antzekotasuna neurtzeko termino guztiak garrantzitsuak baitira. Azkenik, domeinutik kanpo bakarrik agertzen diren terminoek ez dute inolako eraginik RFR metrikan. Horrela, ezagunak diren terminoekin batera agertzen diren termino ezezagunak dituzten esaldiak aukeratzen dira, corpusaren estaldura handitzea lortuz.

5.1.2. Termino Ezezagunen Aukeraketa

Arestian esan bezala, domeinuan agertzen ez diren terminoek ez dute inolako eraginik RFR antzekotasun metrikan. Beraz, hiztegitik kanpoko terminoak inguruko termino ezagunen arabera bakarrik aukeratuko dira. Baliteke ezaugarri hau egokia ez izatea ondorengo bi ikuspuntu hauen arabera.

Lehendabizi, domeinutik kanpoko corpusak zaratatsuak izan ohi dira. Tamaina handiko corpusak sortzeko askotan Internet bezalako datu iturriak erabiltzen dira, eta esaldien lerrokatzea automatikoki egiten da. Adibidez, corpusean hirugarren hizkuntza batean idatzitako esaldiak egongo balira, antzekotasun metrikak termino ezagunak bakarrik hartuko lituzke kontutan,

normalean puntuazio ikurrak.

Bigarrenik, oinarrizko helburua ereduen estaldura lexiko eta sintaktikoa handitzea da domeinuan zentzua duen lexiko berriarekin. Beraz, estaldura lexiko eta sintaktiko handiagoa duten esaldiak aukeratu beharko lirateke, termino ezagunak bakarrik dituzten esaldiak baino.

Aurreko bi ikuspuntuak aintzat hartuz, azpiatal honetan aurkezten den tekniken helburua termino ezezagunak dituzten esaldiak hobestea da, betiere modu kontrolatu batean. Helburu hau lortzeko, jatorrizko RFR metrikari funtzio bat gehitzen zaio termino ezezagunen ehunekoa orekatzeko. Izan bedi u domeinutik kanpoko esaldiko termino ezezagunen ehunekoa, termino ezezagunak orekatzeko funtzioa 5.3. ekuazioan azaltzen da.

$$W(u) = sin(\alpha \cdot u^k) \tag{5.3}$$

Funtzio sinusoidal bat erabiliz, termino ezezagunen ehunekoa balio jakin bat baino handiagoa bada, funtzioak balio negatiboa izango du; era berean, esaldian termino ezezagunik ez badago, funtzioren balioa hutsa izango da.

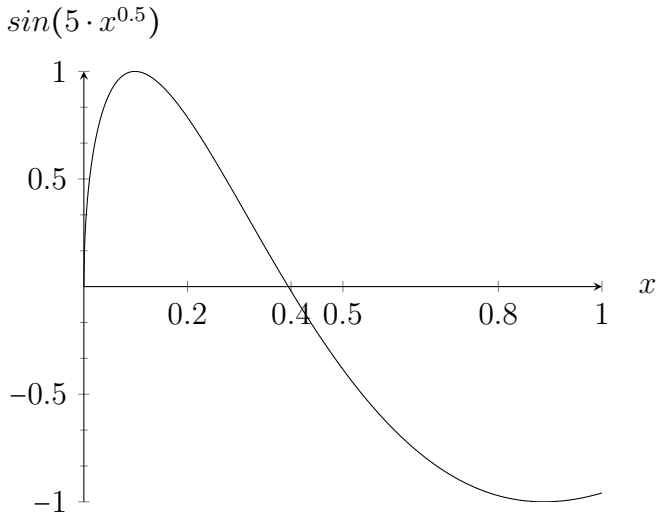
α eta k parametroak enpirikoki ezarri behar dira. Bi parametro horiekin zenbat termino ezezagun lortu nahi diren kontrolatzen da. 5.2. irudian $\alpha = 5$ eta $k = 0.5$ parametroekin kurbak duen itxura azaltzen da, balio hauek baitira 5.2. azpiataleko esperimentuetan erabili direnak.

Goiko parametroekin, %40 baino termino ezezagun baino gutxiago dituzten esaldiak aukeratuko lirateke. Oreka funtzioa 5.4. ekuazioan erakusten den bezala txertatzen da jatorrizko metrikan.

$$wrfr(s,t) = \frac{1}{2}\left(\exp\big(W(u_s)\big) \cdot \sum_{i=1}^{|s|} \frac{\phi_d(w_i)}{\phi_o(w_i)} + \exp\big(W(u_t)\big) \cdot \sum_{i=1}^{|t|} \frac{\phi_d(w_i)}{\phi_o(w_i)} \right) \tag{5.4}$$

Jatorrizko RFR metrika (5.2. ekuazioa) maiztasun erlatiboen baturan oinarritzen da, baina $W(u)$ funtzioaren balioak positibo nahiz negatiboak izan daitezke, horregatik erabiltzen da esponentziala, $W(u)$ funtzioaren emaitza beti

5.2 irudia: Termino ezezagunak orekatzeko funtzioaren itxura $\alpha = 5$ eta $k = 0.5$ parametroekin.



positiboa izan dadin. $exp(W(u))$ funtzioaren eragina ezberdina da termino ezezagunen ehunekoaren arabera (goiko $\alpha = 5$ eta $k = 0.5$ parametroetarako):

- $u = 0$. Funtzioak 1-eko balioa du. Kasu honetan WRFR eta RFR baliokideak dira.
- $u \in (0, 0.4)$. 1-etik gorako balia du, tarte honetan dauden esaldiak hobesten dira.
- $u \in [0.4, 1]$. 1-etik beherako balioa du, esaldi hauen emaitzek beherakada bat jasaten dute.

Hurrengo 5.2. azpiatalean funtzio honen eragina aztertuko dugu.

5.2. Esperimentuak

Azpiatal honetan deskribatutako esperimentuak benetako egoeratan datuen aukeraketarako metodoak konparatzeko diseinatu dira, non domeinutik kanpoko corpusetik datu guztien zati bat soilik hartzen den. Domeinutik kanpo-

ko corpusa metodo bakoitzaren metrikaren arabera ordenatu da, eta corpusaren zenbait lagin hartuz nolako emaitzak lortzen diren ikusiko dugu. Laginak corpusaren %1-etik %50-erainokoak izango dira. Esperimentu hauetako laginak Axelrod-ek bere ikerketetan erabilitakoen oso antzekoak dira (Axelrod *et al.*, 2011, 2012).

Esperimentuetan RFR eta termino ezezagunak orekatzeko funtzioa duen WRFR bertsioa *Modified Moore-Lewis* (aurrerantzean MML) metodoarekin konparatzen dira (Axelrod *et al.*, 2011), hau baita artearen egoeran metodorik erabilienetako bat bere eramangarritasun eta emaitzen sendotasunarengatik[2].

5.2.1. Corpusa

Jatorrizko domeinuko corpus bezala, ingeles-gaztelania hizkuntza parerako WMT albisteen itzulpen atazako NEWS COMMENTARY corpusa erabili da, NEWS TEST 2012 izanik garapenerako corpusa eta NEWS TEST 2013 ebaluaziorako corpusa. Ingeles-frantses hizkuntza parerako berriz, WMT medikuntza domeinuko itzulpen atazako corpusa erabiliko da, hau da, EMEA corpusa entrenamendurako eta KHRESMOI-SUMMARY corpusa garapen eta ebaluaziorako[3] (ikus 5.1. taula).

Domeinutik kanpoko corpus moduan (ikus), WMT itzulpen atazetako hiru corpus bildu dira: COMMON CRAWL, EUROPARL eta MULTIUN (ikus 5.2. taula). Hiru corpusak gehituz corpus handi berri bat sortu da, esperimentuetako metodoak erabiliz ordenatu beharrekoa. Corpusen estatistikak, bikoiztutako eta 60 termino baino gehiagoko esaldiak filtratu eta gero, 5.1 eta 5.2. tauletan azaltzen dira.

Azaldutako corpusak aukeratzeko arrazoiak bi dira. Lehendabizi, datuen aukeraketa lizentzia librepean eskura dauden corpusen gainean egin da, bakoitza bere azpi domeinu propio eta zarata maila ezberdinarekin. Corpus hauek aurkeztutako metodoen sendotasuna ebaluatzeko aukera ematen dute.

Bigarrenik, aukeratutako domeinuko corpusak oso ezberdinak dira, bata al-

[2]MML metodoaren inplementaziorako XENC tresna erabili ohi da (Rousseau, 2013): `https://github.com/antho-rousseau/XenC`

[3]Domeinuko corpusa `http://www.statmt.org/wmt13/` eta `http://www.statmt.org/wmt14/` web helbideen bitartez lortu da.

5.1 taula: Domeinuko corpusa.

HIZKUNTZA	CORPUSA	ZATIA		
		ENTREN.	GARAPEN	EBALUAZIOA
EN-ES	NEWS COMMENTARY	207.137	3.003	3.000
EN-FR	EMEA	354.288	500	1.000

5.2 taula: Domeinutik kanpoko corpusa.

HIZKUNTZA	CORPUSA			
	COMMON CRAWL	EUROPARL	MULTIUN	GUZTIRA
EN-ES	1.814.883	1.842.496	8.079.790	11.661.326
EN-FR	3.065.194	1.826.770	9.142.161	13.864.506

bisteen domeinukoa eta bestea medikuntzaren domeinukoa. Domeinutik kanpoko corpusa aldiz berbera da bi kasuetan. Beraz, domeinutik kanpoko corpusa ez da aukeratu domeinuarekiko hurbiltasunaren arabera, eta esperimentuetan corpus berbera erabiltzeak dakarren eragina ikusi ahal izango dugu.

5.2.2. Aukeratutako Datuen Analisia

5.3. taulan erakusten den bezala, metodo bakoitzak datu ezberdinak aukeratzen ditu. Espero bezala, ezberdintasun nagusia MML eta WRFR metodoen artean gertatzen da. RFR eta WRFR metodoek aukeratutako esaldien %80 baino gehiago berbera da ingeles-frantseserako, baina ingeles-gazteleraren kasuan berriz ezberdintasun gehiago ikusten dira laginaren arabera. Honek albisteen domeinuan termino ezezagun gehiago daudela erakusten du. Atera daitekeen beste ondorio bat laginaren tamaina handitzen den heinean metodoen arteko ezberdintasunak gutxitzen direla da.

Albiste eta medikuntza domeinuen arteko egoera ezberdina da. EUROPARL eta MULTIUN corpusen eta NEWS COMMENTARY corpusaren artean antzeko datu asko daude, baina EMEA corpusaren kasuan, domeinutik kanpoko corpusean antzekoak diren esaldiak sakabanatuagoak daude, eta horregatik datu

5.3 taula: Metodo ezberdinen bidez aukeratutako datuen artean amankomunean dauden esaldien ehunekoa.

LAGINA	MML-RFR		MML-WRFR		RFR-WRFR	
	EN-ES	EN-FR	EN-ES	EN-FR	EN-ES	EN-FR
%1	11,72	24,55	5,16	21,09	44,81	84,20
%2	15,88	24,30	7,19	21,30	43,72	83,66
%5	23,86	24,72	12,17	22,08	45,42	82,59
%10	32,59	27,31	19,12	24,88	49,65	82,59
%20	44,12	34,45	30,85	32,13	57,51	83,70
%30	52,88	42,08	41,39	39,94	64,40	85,04
%40	60,59	49,90	51,30	48,04	70,81	86,39
%50	67,91	57,95	60,79	56,49	77,13	87,81

amankomun gehiago daude ingeles-frantses corpusean.

5.4 taula: Metodo ezberdinen bidez aukeratutako esaldien tamainen batezbestekoa.

LAGINA	MML		RFR		WRFR	
	EN-ES	EN-FR	EN-ES	EN-FR	EN-ES	EN-FR
%1	17,50	14,20	40,11	28,44	39,41	29,77
%2	19,30	15,70	40,03	29,72	40,04	31,04
%5	21,60	17,90	39,18	31,86	39,85	33,12
%10	23,20	19,70	37,90	33,14	38,86	34,19
%20	24,60	21,80	36,03	33,56	37,07	34,29
%30	25,40	23,00	34,63	33,10	35,52	33,60
%40	25,80	23,70	33,44	32,31	34,08	32,63
%50	26,10	24,20	32,30	31,35	32,70	31,53

Arestian aipatu dugunez, perplexitatean oinarritutako metodoek esaldi motzen aukeraketa hobesten dute, eta MML metodoak domeinuen arteko perplexitatea neurtzen duenez, aukeratutako datuek portaera hori erakusten dute. 5.4. taulan metodo ezberdinek aukeratutako esaldien tamainaren batezbestekoa azaltzen da. RFR eta WRFR metodoek tamaina handiagoko esaldiak

aukeratzen dituzte. Noski, laginak handitzen diren heinean ezberdintasunak txikiagoak dira.

5.5. taulan azaltzen dira %1-eko laginean aukeratutako adibide batzuk. Datuei begiratu azkar bat emanez, badirudi hiru metodoek domeinuarekin erlazionatutako esaldiak aukeratzen dituztela. Hurrengo azpiataletan zehatzago aztertzen da aukeratutako datuen antzekotasuna, lortutako termino ezezagunak eta perplexitatea aztertuz, eta itzulpen automatikorako ereduak entrenatuz.

5.5 taula: Aukeratutako esaldien adibideak %1-eko laginean albisteen domeinuan.

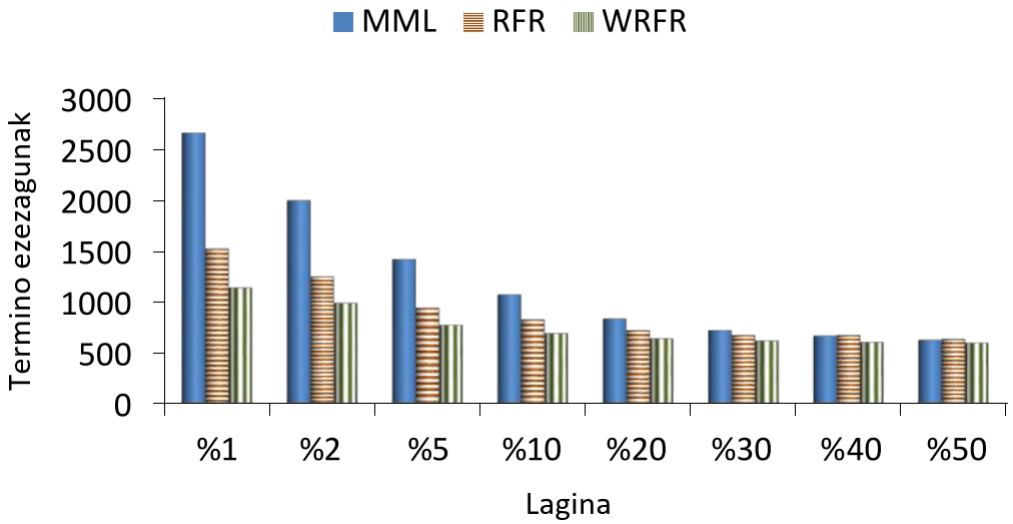
HIZK.	METODOA	ESALDIAK
EN-ES	MML	where are we heading ? - trillions of dollars more are waiting in the wings . - the implications are dire .
	RFR	the assumption that only an enlightened minority is in a position to respect human rights and freedoms . - greenhouse gas emissions can be cut through the use of nuclear energy , clean coal and low carbon-emitting renewable energies . - coupled with extensive deregulation of financial markets and excess liquidity , these imbalances encouraged investors to engage in leveraged risk-taking in search of profits .
	WRFR	during that period , their debt actually increased from $ 618 billion in 1980 to $ 3.25 trillion in 2006 . - Mr. Snowden (United States of America) said that the Commission for Sustainable Development had galvanized action and helped shape the agendas of a wide range of organizations around the world . - there has been a temptation for the West – Europe and the United States – to stress continuity and so-called stability .
EN-FR	MML	avoid contact with skin , eyes or clothing . - the unused portion should be discarded . - peel open the package with dry hands and place the tablet on your tongue .
	RFR	in terms of public health , the environmental impact of the new medicinal prod- ucts should be assessed . - antiretroviral treatment can be effective only if it is administered and monitored by health professionals working in a well-functioning national health system . - finally , it recognises the need for studies on vaccines and anti-viral medications that are independent of the pharmaceutical industry , including with regard to the monitoring of vaccination coverage .
	WRFR	during the final process , an operator peers through a microscope at the die surfaces , polishing them carefully with a diamond abrasive tool head that is vibrated by supersonic waves . - concentrations of petroleum contaminants in fish and crab tissue , as well as contamination of shellfish could have potentially significant adverse effects on health . - the first three , namely glycerine , brake fluid and anti-freeze , are considered to present the most extreme incompatibility with calcium hypochlorite .

5.2.3. Termino Ezezagunak

Perplexitatean oinarritutako metodoekin alderatuta, proposatutako metodoaren motibazio nagusienetako bat domeinuarekin erlazionatutako esaldi luzeagoak aukeratzea da. Horrela, termino ezezagun berri gehiago eskuratzeko aukera dago. Gainera, WRFR metodoaren kasuan, esaldietako termino ezezagunen kopurua kontrolatzeko mekanismoak erabili ahal dira.

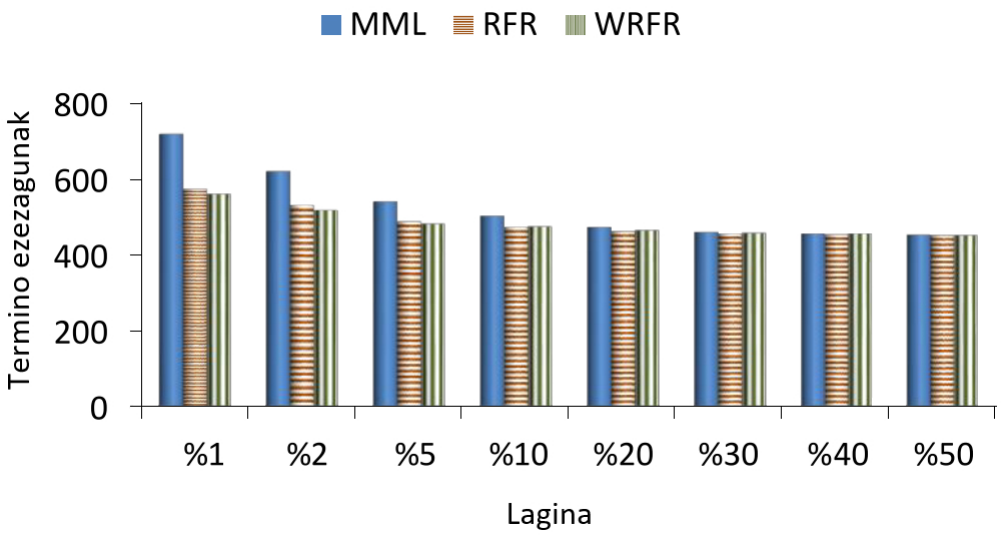
Hiru metodoekin lortzen diren termino ezezagunen onura neurtzeko honako ebaluazioa egin da: lagin bakoitzeko, hiru metodoekin aukeratutako corpusa jatorrizko hizkuntzako ebaluaziorako corpusarekin konparatu da, eta ebaluaziorako corpusarekiko ezezaguna den termino kopurua aztertu da. Emaitzak 5.3 eta 5.4. tauletan aurkezten dira.

5.3 irudia: Termino ezezagunak ingeles-gaztelania ebaluazio corpusean.



Taulak ikusita, argi dago RFR eta WRFR-ren bidez terminoen estaldura hobetzea lortzen dela. Alde handiena ingeles-gaztelania corpusaren kasuan lortzen da non %1-eko laginarekin 1.000 termino baino gehiagoko aldea dagoen. Lagina handitzen doan heinean, hiru metodoen arteko ezberdintasunak murrizten doaz bestelako esaldiak aukeratzeko probabilitatea txikiagoa delako. Espero bezala, WRFR metodorako egindako aldaketek termino ezezagun

5.4 irudia: Termino ezezagunak ingeles-frantses ebaluazio corpusean.



gehiago aukeratzen dituzte, %1-eko laginean RFR-rekin alderatuta ia 400 terminoko aldea baitago. Ingeles-frantses corpusak joera berbera aurkezten du, baina kasu honetan hiru metodoen artean ezberdintasun gutxiago daude. Izan ere, EMEA corpusaren lexikoa ez da NEWS COMMENTARY corpusarena bezain aberatsa, eta horregatik termino ezezagunak egoteko aukera gutxiago daude.

5.2.4. Perplexitatea

Aurreko azpiatalean ikusi dugu hiru metodoek ezberdintasun nabariak dituztela terminoak aukeratzeko orduan. Beste metrika garrantzitsu bat perplexitatea da, hau da, aukeratutako esaldiek hizkuntzaren modelatze estatistikoan duten eragina ebaluatzea. 5.6. taulan lagin bakoitzeko helburuko hizkuntzan hiru metodoek ebaluazio lortzen dituzten perplexitateak azaltzen dira, hizkuntza ereduak aukeratutako datuekin entrenatuz eta ebaluazio corpusean neurtuz. Entropiaren kalkuluan termino ezezagunak kontutan hartu dira.

MML metodoa jatorrizko domeinuan perplexitate baxuko eta domeinutik kanpo perplexitate altuko esaldiak aukeratzeko diseinatu zen, horregatik ez da harritzekoa orokorrean emaitza onenak metodo honekin lortzea. Hala ere,

5.6 taula: Lagin bakoitzeko helburuko hizkuntzako esaldien perplexitatea termino ezezagunak kontutan edukita.

LAGINA	GAZTELANIA			FRANTSESA		
	MML	RFR	WRFR	MML	RFR	WRFR
%1	335,55	281,53	**257,67**	**151,90**	153,63	157,64
%2	295,46	252,06	**232,76**	**147,55**	154,66	158,75
%5	249,95	224,52	**211,72**	**151,63**	163,54	166,89
%10	217,95	210,23	**202,32**	**161,38**	173,54	177,64
%20	**196,59**	201,26	197,93	**175,67**	187,74	191,15
%30	**188,95**	198,49	197,15	**187,42**	197,88	200,44
%40	**186,60**	197,38	196,85	**196,67**	205,13	207,20
%50	**186,60**	197,17	196,79	**203,95**	210,47	212,14

gaztelaniaren kasuan %1-etik %10-arte, RFR eta WRFR metodoekin emaitza hobeak lortzen dira. Frantseserako berriz, MML da lagin guztietan emaitza onenak lortzen dituen metodoa. Hala ere, orokorrean ezberdintasunak ez dira oso handiak.

RFR eta WRFR metodoek abantailak dituzte termino ezezagunak lortzeko orduan, eta baliteke abantaila hauek izatea lortutako emaitza konpetitiboen arrazoia. Termino ezezagunen eragina alde batera uzteko 5.7. taulan lortutako perplexitateak aurkezten dira entropiaren kalkuluan termino ezezagunak kontutan izan gabe. Ikus daitekeenez, termino ezezagunak kontutan hartu gabe antzeko portaera dute hiru metodoek, ezberdintasunak apur bat murrizten dira, baina lagin eta hizkuntza guztietan MML da emaitza onenak lortzen dituena.

Orokorrean, RFR eta WRFR metodoek perplexitate onak lortzen dituzte nahiz eta MML-ren emaitzetara ez iritsi. Kontutan hartzen badugu MML perplexitatea optimizatzeko diseinatuta dagoela, lortutako emaitzak oso positiboak direla esan dezakegu.

5.7 taula: Lagin bakoitzeko helburuko hizkuntzako esaldien perplexitatea termino ezezagunak kontutan eduki gabe.

LAGINA	GAZTELANIA			FRANTSESA		
	MML	RFR	WRFR	MML	RFR	WRFR
%1	221,38	217,07	**211,53**	**116,81**	123,81	127,86
%2	211,69	202,41	**196,18**	**116,01**	125,72	128,82
%5	193,56	186,75	**182,75**	**121,36**	132,94	135,70
%10	177,30	178,65	**176,67**	**129,67**	140,82	144,10
%20	**165,77**	173,77	173,97	**140,90**	151,35	154,29
%30	**162,42**	172,45	173,61	**149,75**	159,15	161,20
%40	**161,72**	172,28	173,51	**157,47**	164,47	166,20
%50	**161,72**	172,57	173,38	**163,06**	168,46	169,95

5.2.5. Itzulpen Automatikoa

Bukatzeko, azpiatal honetan metodo bakoitzarekin aukeratutako esaldiek itzulpen automatikoan duten eragina aztertzen da SMT ereduak entrenatuz. Lehendabizi, erreferentzia bezala bi eredu entrenatu dira: domeinuko corpuseko entrenamendurako zatia erabiliz entrenatutako eredua, eta domeinuko eta domeinuz kanpoko corpusen entrenamendurako zatiak osorik hartuta entrenatutako eredua. Ondoren, metodo bakoitzarekin aukeratutako esaldien lagin bakoitzeko eredu bat entrenatu da.

Eredu guztiak phrase-based SMT ereduak dira (Koehn *et al.*, 2003). Entrenamendurako MOSES erabili da (Koehn *et al.*, 2007) lehenetsitako parametroekin eta gehienez 5 terminoko segmentuekin. Segmentuen itzulpen-taulak adierazgarritasun estatistikoaren arabera kimatu dira (Johnson *et al.*, 2007) eta parametroak MERT bidez optimizatu dira (Och eta Ney, 2003). Hizkuntza ereduak entrenatzeko KENLM erabili da (Heafield, 2011); zehazki 5-gramako ereduak entrenatu dira aldatutako KNESER-NEY leuntzearekin (Heafield *et al.*, 2013).

Jatorrizko domeinuko eredua metodo eta lagin bakoitzarekin entrenatutako ereduarekin konbinatu da. Konbinazio hau egiteko domeinuko segmentuen itzulpen-taula domeinutik kanpokoarekin osatu da, domeinua adieraz-

5.8 taula: Lagin bakoitzeko BLEU emaitzak ingeles-gaztelania corpusean.

LAGINA	CORPUSA			METODOA			
	NEWS	COMM	GUZTIA	AUSAZ	MML	RFR	WRFR
%100		23,29	27,75	-	-	-	-
%1		-	-	24,07	23,64	†‡ 24,55	†‡⋆ **24,82**
%2		-	-	24,25	24,12	†‡ 25,10	†‡⋆ **25,46**
%5		-	-	25,19	25,00	†‡ 26,00	†‡ **26,14**
%10		-	-	26,27	26,10	†‡ 26,56	†‡⋆ **26,91**
%20		-	-	27,01	26,88	‡ 27,17	†‡ **27,26**

5.9 taula: Lagin bakoitzeko BLEU emaitzak ingeles-frantses corpusean.

LAGINA	CORPUSA		METODOA			
	EMEA	GUZTIA	AUSAZ	MML	RFR	WRFR
%100	27.10	37.96	-	-	-	-
%1	-	-	31.56	† 33.70	†‡ 34.79	†‡⋆ **35.12**
%2	-	-	31.75	† 34.81	†‡ **35.48**	†‡ 35.33
%5	-	-	33.23	† 35.91	†‡ 35.98	† **36.30**
%10	-	-	34.52	† 36.54	†‡⋆ **37.44**	†‡ 36.99
%20	-	-	**37.43**	**37.43**	37.28	37.27

ten duen aldagai bitar bat erabiliz Bisazza *et al.* (2011).

Azpiatal honetan esaldiak ausaz aukeratzen dituen beste erreferentziazko metodo bat erabiltzen da (aurrerantzean AUS). Horrela, beste hiru metodoek benetako hobekuntzak izan ditzaketen egiaztatu daiteke. Ereduen ebaluazioa BLEU metrikarekin egin da (Papineni *et al.*, 2002). Emaitzak 5.8 eta 5.9 tauletan azaltzen dira[4].

Orokorrean, RFR eta WRFR metodoak MML baino eraginkorragoak dira. Salbuespen bakarra ingelesa-frantses corpuseko %20-ko lagina da. Beste kasu

[4]Adierazgarritasun estatistikoa bootstrap resampling testarekin kalkulatu da (Koehn, 2004). † ikurrak adierazgarritasun estatistikoa adierazten du AUS-ekiko; ‡ ikurrak RFR edo WRFR eta MML artean; eta ⋆ ikurrak RFR eta WRFR artean; guztiak $p < 0.05$ balioarekin.

guztietan emaitza onenak ezezik adierazgarritasun estatistikoa ere lortzen da. %1-eko laginean WRFR metodoak 1,2 eta 1,4 puntuko aldea du MML-rekiko ingeles-gaztelania eta ingeles-frantses corpusetan hurrenez hurren.

Adierazgarriak dira ingeles-gaztelania corpusean MML metodoak izandako emaitzak, ausaz esaldiak aukeratuz emaitza hobeak lortzen baitira. MML metodoak aldez aurretik domeinuan dauden datuak aukeratzeko joera du, eta badirudi domeinutik kanpoko corpusaren zati batean NEWS COMMENTARY corpusaren antzeko esaldiak daudela. Ingeles-frantses corpusaren kasuan medikuntza domeinua lantzen denez, datu gutxiago partekatzen dira domeinutik kanpoko corpusarekin.

Deigarria da nola corpus guztia hartuta emaitza onenak lortzen diren. Zenbait ikerketetan atal honetan baino lagin handiagoak erabilita corpus guztia hartuta baino emaitza hobeak lortzen dira (Banerjee *et al.*, 2012; Wong *et al.*, 2016); beste ikerketa batzuetan berriz, kontrakoa gertatzen da (Peris *et al.*, 2017). Corpus guztia hartuz lortzen diren emaitzak hobetzea beraz ez da batere erraza, eta emaitzak uneko corpusen ezaugarrien araberakoak dira; gure kasuan, domeinutik kanpoko corpusaren tamaina oso handia da (aipatutako ikerketan erabilitakoekin konparatuz), eta beraz, atal honetako esperimentuetan beste ikerketetan baino egoera zailagoak aurkezten dira. Dena dela, entrenamendurako askoz datu gutxiago erabiliz antzeko emaitzak lortzen dira, datuen aukeraketaren erabilpena justifikatuz.

5.3. Ondorioak

Kapitulu honetan itzulpen automatikoan datuen aukeraketarako teknika berri bat deskribatzen da (RFR). Metodo honetan, terminoen maiztasun erlatiboak konparatzen dira jatorrizko domeinuko eta domeinutik kanpoko corpusen artean. Proposatutako metodoa artearen egoerako *Modified Moore-Lewis* (MML) metodoarekin konparatzen da, eta BLEU metrikaren aldetik, perplexitatearen aldetik eta baita lortzen den domeinuko lexiko berriaren aldetik, gure RFR metodoak emaitza eraginkorragoak lortzen ditu.

Beste ekarpen garrantzitsu bat kanpoko domeinuan termino ezezagunak meatzeko (WRFR) metodoa da. Izan ere, jatorrizko RFR metodoarekin konbina-

tuz gero lortzen dira emaitzarik onenak kasu gehienetan. WRFR antzekotasun metrikan termino ezezagunek duten pisua enpirikoki erazarritako bi parametrorekin kontrolatzen da.

Proposatutako metodoa sinplea da, hizkuntza-baliabide gehigarri edota konfigurazio konplexuen beharrik ez baitauka, eta hortaz, edozein domeinu egokitzapen egoeratan erabili liteke. Erabilgarritasun helburua betetzen den konprobatzeko NEWS COMMENTARY eta EMEA corpusekin probak egin dira, eta bietan lortutako emaitzak gogobetekoak izan dira, EMEA corpuseko %20-ko laginean izan ezik beste guztietan MML metodoaren emaitzak hobetuz.

6. KAPITULUA

Esaldi Paraleloen Iragazketa

Itzulpen automatikoan, kalitatezko itzulpenak lortu ahal izateko ereduak tamaina handiko corpus paraleloekin elikatu behar dira. Hala ere, eskuz egindako itzulpenak biltzen dituzten kalitatezko corpusak ez dira ugariak, eta informazio iturri ezberdinetatik lortutako corpusak garbitzeko behar handia dago.

Web domeinuak informazio iturri eleaniztun aberatsak dira, eta bertatik baliagarriak izan daitezkeen corpus paraleloak sortzeko aukera dago (Forcada *et al.*, 2016). Hala ere, dokumentu eta esaldiak automatikoki lerrokatuz erroreak sortzen dira, eta are gehiago informazio iturria zaratatsua bada. Zehaztasun gutxiko datu multzoak erabiltzea kaltegarria da kalitatezko entrenamenduak egin ahal izateko (Khadivi eta Ney, 2005; Khayrallah eta Koehn, 2018).

6.1. irudian esaldi paraleloen iragazketaren eskema bat erakusten da. Teknikaren funtsa, esaldi paraleloen corpus bat izanda kalitate txarreko esaldi pareak kenduz corpus berri bat sortzea da. Esaldi paraleloak iragazteko irizpideak ugari izan daitezke (adibidez, karaktere okerrak dituztenak, hizkuntza nahasiak, lerrokapen txarrak, etab.), baina azken helburua ereduen entrenamendurako kaltegarriak diren esaldi pareak iragaztea da.

Kapitulu honetan proposatzen den metodoa STACC (ikus 4. kapitulua) antzekotasun metrikan oinarritzen da. Erabaki honen helburua STACC-en eramangarritasuna, errendimendua eta lortutako emaitza onak esaldi paraleloen iragazpenean ere ematen diren egiaztatzea da.

6.1 irudia: Esaldi Paraleloen Iragazketa.

WMT 2018-ko esaldi paraleloen iragazpenerako atazak [1] (Koehn *et al.*, 2018) metodo ezberdinak egoera berberean konparatzeko aukera ematen du, SMT eta NMT sistemak entrenatuz eta domeinu eta mota ezberdineko corpusetan beren kalitatea neurtuz. Kapitulu honetan aurkeztutako metodoen baliagarritasuna neurtzeko asmoz WMT 2018 atazan parte hartu genuen.

Hurrengo atalak honela antolatzen dira. 6.1. atalean proposatutako metodoa azaltzen da. Gero, 6.2. atalean, WMT 2018 atazaren eta metodoaren konfigurazioaren xehetasunak deskribatzen dira. 6.3. atalean berriz lortutako emaitzak aurkezten dira. Azkenik, 6.4. atalean azken hausnarketak egiten dira.

6.1. Proposatutako Metodoa

Tesiaren 4. kapituluan esaldien lerrokatzearen inguruan zenbait ikerketa aurkeztu dira, eta STACC esaldi pareen antzekotasun metrika oso eraginkorra eta eramangarria dela egiaztatu ahal izan da. Funtsean, proposatutako esaldi paraleloen iragazpen metodoa esaldi pareak antzekotasun metrika baten arabera sailkatzean datza, hortaz, STACC metrika esaldien iragazpenerako ere erabili da ataza honetara egokitzeko aldaketa batzuk egin ostean.

Hurrengo azpiataletan STACC metodoari egindako egokitzapenak aurkezten dira.

[1] `http://www.statmt.org/wmt18/parallel-corpus-filtering.html`

6.1.1. Termino Ezezagunen Dentsitatea

WMT 2018 atazako esaldi paraleloen iragazpenerako corpusa oso zaratatsua da. WMT 2018 corpusa zenbait web domeinuetatik deskargatutako testuen bilduma da, eta lerrokapenen zehaztasuna baino, kantitatea izan zen hobetsitakoa. Huda Khayrallah eta Philipp Koehn-ek PARACRAWL proiektuan[2] sortutako corpusaren lagin bat eskuz aztertu zuten (Khayrallah eta Koehn, 2018), WMT 2018 corpusaren antzekoa dena, eta gaizki lerrokatutako esaldiak, hizkuntza okerreko esaldiak, itzulpenik gabeko esaldiak eta karaktere oker edo HTML etiketak dituzten esaldiak identifikatu zituzten. Kalitatezko beste corpus paralelo batetik erauzitako terminoen hiztegi bat izanda, esaldietako termino ezezagunen proportzioak lau errore mota horiek identifikatzen lagundu dezake. Adibidez, kalitatezko hiztegi lexiko batekin, *"Nire aitak katua ikusi zuen"* esaldiko terminoak itzuli daitezke *"mi mis padre papa vio dibisó el los gato felino"* terminoak lortuz (bi itzulpen onenak hartuz), eta termino horiek ez lukete lerrokapenik edukiko *"My father saw the cat"* hizkuntza okerreko esaldiarekin edota *"M43is 687&& —- "* esaldi zaratatsuarekin.

STACC metodoa zarata maila gutxiago duten corpus konparagarriak lerrokatzeko pentsatuta dago eta, kasu horietan, zenbakizko eta letra larriz hasitako hiztegitik kanpoko terminoen lerrokapenak antzekotasunaren adierazleak dira (ikus 3.1.1. atala). Baina baliteke datu multzo zaratatsuetan termino berezi hauek kaltegarriak izatea. Adibidez, gerta liteke ausazko zenbakien sekuentziek ustekabeko lerrokapenak izatea.

Ahal den neurrian, ataza zehatzetarako bereziki prestatutako garbiketa heuristikoak saihestu nahi dira. Esate baterako, errendimendu galera nabarmen bat eragingo lukeen hizkuntza-identifikatzaile baten erabilpena edo termino berezien iragazketa. Horren ordez, esaldiko termino ezezagunen agerpena zigortzen duen funtzio bat diseinatu da. Termino bat ezezaguna dela adierazteko, kalitatezko beste corpus paralelo batetik erauzitako hiztegi bat erabiliko da. $|oov|$ izanik s esaldiaren termino ezezagunen kopurua, zigortze funtzioaren kalkulua 6.1. ekuazioan azaltzen da.

[2] `https://paracrawl.eu/`

$$p(s) = 1 - \frac{|oov|}{|s|} \tag{6.1}$$

Corpuseko (s_i, s_j) esaldi pare bakoitzarentzat STACC.OOV antzekotasun metrika 6.2. ekuazioan erakusten da.

$$stacc.oov(s_i, s_j) = stacc(s_i, s_j) \cdot \frac{p(s_i) + p(s_j)}{2} \tag{6.2}$$

Beraz, termino ezezagun gutxiko esaldiak, estaldura zabaltzen lagundu lezaketenak, jatorrizko STACC metrikatik hurbil egongo dira, eta termino ezezagun askoko esaldiek 0 inguruko antzekotasuna izango dute.

6.1.2. N-gramen Asetasuna

WMT 2018 atazan antzekotasun kalkulurako esaldiak kontutan hartzea ahalbidetzen da. Horrela, erredundanteak diren esaldiak detektatu daitezke antzekotasunaren kalkuluan.

Datu erredundantziarekin esperimentuak egiteko n-grama estalduran oinarritutako metodo bat inplementatu da. Funtsean, metodo hau Eck eta Lewisen ikerketen antzekoa da (Eck *et al.*, 2005; Lewis eta Eetemadi, 2013) (ikus artearen egoera, 2.3. atalean). Ezaugarrien narriadura teknikekin ere erlazionatuta dago (Biçici eta Yuret, 2011), jatorrian SMT ereduetan erabilia, eta geroago NMT sistemetan ebaluatua (Poncelas *et al.*, 2018).

N-grama asetasuna kalkulatzeko, lehendabizi corpusa STACC.OOV metrikaren arabera ordenatzen da, antzekotasun handienetik txikienera. Jarraian, ordenatutako corpusa prozesatzen da jatorrizko hizkuntzako esaldi bakoitzeko n-gramak eraucteko (k orden jakin bat arte), lortutako n-gramak T PATRICIA TRIE (Morrison, 1968) egitura batean gordetzeko, eta esaldi bakoitzaren n-grama berriak kalkulatzeko. s jatorrizko hizkuntzako esaldi bakoitzarekin emandako pausuak honako hauek dira:

1. s esaldiaren n-grama guztiak lortu.
2. T egituran ez dauden n-gramak identifikatu (lehenengo esaldirako T

hutsik dago, beraz, n-grama guztiak izango dira berriak).

3. s esaldiko n-grama berrien proportzioa kalkulatu, ng izanik s esaldiko n-gramen multzoa eta k n-gramen ordena, 6.3. ekuazioan azaltzen den bezala.

$$ngsat(s) = \frac{\sum_{n=1}^{k} ng_k \notin T}{\sum_{n=1}^{k} ng_k \in T} \tag{6.3}$$

4. N-grama berriak gehitu T egituran.

Azkenik, (s_i, s_j) esaldi pare bakoitzaren STACC.OOV.NGSAT antzekotasuna 6.4. ekuazioan azaltzen den bezala kalkulatzen da.

$$stacc.oov.ngsat(s_i, s_j) = stacc.oov(s_i, s_j) \cdot ngsat(s_i) \tag{6.4}$$

N-grama berriak ez dituzten esaldiak 0-ko emaitza izango dute, eta era berean, esaldi baten n-grama guztiak berriak badira $ngsat(s)$ funtzioak ez du eraginik izango. Adibidez, i. esaldia *"Nire aitak txakurra ikusi du"* bada, $i+1$. esaldia *"Nire amak katua ikusi du"*, eta $i+2$. esaldia *"Nire aitak katua ikusi du"*, $i+2$. esaldiak ez du n-grama berririk eta hortaz bere antzekotasun balio berria hutsekoa izango da. Aldiz, demagun $i+3$. esaldia *"Ostegunean tesia aurkeztuko dut"* dela, bertako n-grama guztiak berriak dira eta n-grama asetasunak ez du eraginik edukiko.

Metodo hau Eck-en lanarekin bi modutan ezberdintzen da: n-grama frekuentziak ez dira erabiltzen, eta normalizazio faktorea esaldiaren tamaina izan ordez esaldiko n-grama kopurua erabiltzen da. Gainera, STACC.OOV.NGSAT metrikaren konplexutasuna lineala da karratua izan beharrean, esaldi bakoitzaren kostua ez baita birkalkulatu behar esaldi bakoitza prozesatu ostean. Lewis-en ikerketekin ere badira ezberdintasunak, ez baitago n-grama kopuruaren atalaserik. Azkenik, FEATURE DECAY metodoak ez bezala, ez da inongo eredurik entrenatzen antzekotasunaren kalkulurako.

N-grama asetasunaren helburua jatorrizko STACC metodoan konplexutasun nabarmenik gehitu gabe datu erredundanteak kontutan hartzea da. Meto-

doaren errendimendua hobetzeko n-gramen kalkulua jatorrizko hizkuntzara mugatu da, antzeko esaldiek antzeko erredundantzia baitute bi hizkuntzetan, eta $ngsat(s)$ funtzioaren zeregina erredundantzia hori neurtzea da, antzekotasunaren kalkulua STACC.OOV metrikaren esku utziz.

6.2. Esperimentuak

Egindako ikerketak probatzeko WMT 2018 atazan parte hartu genuen. Hurrengo azpiataletan atazaren xehetasunak eta bidalitako bertsioen konfigurazioak azaltzen dira.

6.2.1. WMT 2018 Ataza

WMT 2018-ean corpus paraleloen iragazketaren arazoari aurre egiteko ataza berri bat aurkeztu zen. Corpus paralelo zaratatsu bat izanda (zenbait web domeinuetatik lortutakoa), kalitate oneneko esaldi pareak identifikatu behar dira.

Iragazi beharreko corpusa PARACRAWL proiektuaren barne, aleman-ingeleserako prestatutako RELEASE 1 bertsioa da. Corpusa sortzeko orduan estaldura handia lortzea izan zen helburua, eta beraz, oso zaratatsua da. Izan ere, esaldi pareen guztien % 23-a bakarrik eman daiteke ontzat (Khayrallah eta Koehn, 2018). Ebaluaziorako corpusak sei dira, guztiak WMT 2018 albisteen itzulpen atazakoak. Helburua ebaluazio sakon bat egitea da mota eta domeinu ezberdineko corpusak erabiliz. Corpusen xehetasunak 6.1. taulan azaltzen dira.

Partaideen metodoak ebaluatzeko PARACRAWL corpusa ordenatu eta zenbait eredu entrenatzen dira. Honako hauek dira ebaluazioa egiteko eman beharreko pausuak:

1. Partaideek PARACRAWL corpuseko esaldi pareak ordenatu behar dituzte kalitate handienetik txikienera.

2. Corpusetik bi lagin ateratzen dira: kalitate oneneko esaldi pareak hartzen dira 10 milioi hitz lortu arte ingelesez, eta gauza berdina 100 milioi lortu arte.

6.1 taula: WMT 2018 atazako corpusa.

CORPUSA	ESALDI PAREAK	HITZAK INGELESEZ
PARACRAWL R1	104M	1.000M
NEWSTEST 2018	2.998	58.628
IWSLT 2017	1.138	18.162
ACQUIS	2.862	98.624
EMEA	3.000	93.071
GLOBALVOICES	3.000	54.930
KDE	3.000	109.716

3. Bi laginekin lau eredu entrenatzen dira: bi SMT eredu (SMT 10M eta SMT 100M), eta beste bi NMT eredu (NMT 10M eta NMT 100M).

4. Lau ereduen emaitzak sei corpusetan ebaluatzen dira C-BLEU metrikarekin.

6.2.2. Konfigurazioa

Esaldi paraleloak iragazteko parametro gutxi batzuk ezarri behar dira. STACC bi parametro hauekin konfiguratu da: aurrizki komun luzeena 3 karaktere tamainarekin erabiltzen da eta 5 lerrokatze onenak hartzen dira kontutan. STACC.OOV.NGSAT metodoan 1-etik 3-rako n-gramak erauzten dira.

STACC-en hiztegi lexikoak FAST ALIGN (Dyer *et al.*, 2013) tresnaren bidez entrenatu dira WMT 2018 albisteen itzulpen atazako corpusa erabiliz. Beraz, entrenamenduak EUROPARL V7, COMMON CRAWL, NEWS COMMENTARY eta RAPID corpusak biltzen ditu. PARACRAWL corpusa nahiko zaratatsua denez entrenamendutik kanpo uztea erabaki zen. Bikoiztutako esaldiak kendu eta gero, entrenamendurako corpusak 5.626.721 esaldi pare ditu.

WMT 2018 atazako esaldi paraleloak iragazteko 64 hariko zerbitzari bat erabili da. 104 milioi esaldi pareen iragazketak 57 minutu iraun du STACC.OOV metodoarekin eta 11,3 GB RAM behar izan ditu. STACC.OOV.NGSAT bertsioarekin prozesatze denbora 5 aldiz mantsoagoa izan da eta 100 GB RAM behar izan ditu, gehien bat TRIE egitura gorde eta prozesatu ahal izateko.

Aurretiko esperimentu guztiak WMT 2018 atazako garapenerako corpusean egin dira. Jatorrizko STACC metodoaren zenbait aldaerekin ere probak egin dira, bereziki pisu lexikoak erabiltzen dituen bertsioarekin. Nahiz eta emaitzak oso ezberdinak ez izan, jatorrizko metodoarekin lortu dira emaitza onenak. Emaitza hauek ez dira harritzekoak, WMT 2018 corpusa oso zaratatsua baita eta maiztasun gutxiko termino okerrek pisu lexikoa altua baitute.

6.3. Emaitzak

WMT 2018 atazako emaitza ofizialak 6.2. taulan azaltzen dira (Bojar *et al.*, 2018). Orokorrean, STACC.OOV metodoaren emaitzak onak dira, non 48 partaideetatik SMT 10M eta NMT 10M ereduetan 16 eta 13. lekuan geratu den hurrenez hurren, eta beste SMT 100M eta NMT 100M ereduetan 24 eta 27. lekuan. Gure metodoa nahiko sinplea eta eraginkorra izanik, eta sistema onenekiko desberdintasun txikiak ikusita, emaitza hauek gogobetekoak dira.

Ebaluaziorako corpus guztietan zehar emaitzak berdintsuak dira KDE corpusaren salbuespen bakarrarekin. Bertan, STACC.OOV sistema SMT 10M, SMT 100M eta NMT 10M ereduetan 10 onenen artean aurkitzen da, eta beste NMT 100M ereduan 20 onenen artean dago. Emaitza hauek azaldu daitezke gure sistema esaldien antzekotasuna neurtzeko sortu delako, eta ez hain beste esaldi pareek uneko domeinurako informazio gehigarri nahikoa duten ala ez aztertzeko. Horregatik, informazio teknikoa duten itzulpen zuzenek puntuazio altua izango dute nahiz eta hizkera hain teknikoa ez duten domeinuetarako garrantzitsuak ez izan.

N-grama asetasuna neurtzen duen STACC.OOV.NGSAT metodoaren emaitzak nahiko antzekoak dira SMT ereduetan, baina NMT ereduetan berriz emaitzek behera egiten dute. Emaitza hauek ikusita eta n-gramen kalkuluak errendimendua okertzen duela kontutan hartuz, jatorrizko STACC.OOV bertsioa da optimoa kasu guztietarako. NMT ereduetan izandako beherakada hainbat arrazoirekin azaldu daiteke. Alde batetik, SMT sistemak segmentuen aldaerekiko egonkorragoak dira esaldiak itzultzeko orduan itzulitako segmentuak independentetzat hartzen direlako. Bestetik, NMT sistemen muina terminoen errepresentazioa da *word embeddings*-en bidez, eta errepresentazioak kalita-

6.2 taula: WMT 2018 atazako emaitza ofizialak C-BLEU metrikarekin neurtuta. Sistemen sailkapena emaitzen batezbestekoarekin kalkulatu da.

EREDUA	CORPUSA	ONENA	STACC.OOV	STACC.OOV.NGSAT
SMT 10M	BATEZBESTEKOA	24,58	23,25	23,29
	SAILKAPENA	1/48	16/48	13/48
	NEWS	29,59	27,48	27,52
	IWSLT	22,16	20,42	19,80
	ACQUIS	21,45	19,33	19,33
	EMEA	28,28	26,51	26,84
	GLOBAL	22,67	21,20	21,12
	KDE	25,51	24,55	25,14
SMT 100M	BATEZBESTEKOA	26,50	25,91	25,80
	SAILKAPENA	1/48	24/48	29/48
	NEWS	31,35	30,47	30,17
	IWSLT	23,17	22,47	22,39
	ACQUIS	22,51	22,16	22,12
	EMEA	31,45	30,30	30,03
	GLOBAL	24,00	23,43	23,36
	KDE	26,93	26,63	26,70
NMT 10M	BATEZBESTEKOA	28,62	26,35	25,64
	SAILKAPENA	1/48	13/48	17/48
	NEWS	36,04	32,33	31,25
	IWSLT	25,23	22,57	21,81
	ACQUIS	25,30	22,55	20,67
	EMEA	32,72	28,96	29,09
	GLOBAL	26,72	24,28	23,48
	KDE	28,25	27,39	27,56
NMT 100M	BATEZBESTEKOA	32,06	30,40	24,91
	SAILKAPENA	1/48	27/48	40/48
	NEWS	39,85	37,08	27,23
	IWSLT	27,43	26,35	22,44
	ACQUIS	28,36	26,81	23,15
	EMEA	36,70	34,54	26,92
	GLOBAL	29,26	27,74	22,94
	KDE	30,79	29,89	26,76

6.3 taula: Gainerako partaideen emaitzekiko ezberdintasunak.

EREDUA	METRIKA	STACC.OOV	STACC.OOV.NGSAT
SMT 10M	Δ BATEZBESTEKOA	+1,83	+1,87
	Δ MEDIANA	+0,74	+0,79
	Δ ONENA	-1,33	-1,29
SMT 100M	Δ BATEZBESTEKOA	+1,03	+0,92
	Δ MEDIANA	+0,03	-0,08
	Δ ONENA	-0,59	-0,71
NMT 10M	Δ BATEZBESTEKOA	+4,51	+3,80
	Δ MEDIANA	+1,79	+1,09
	Δ ONENA	-2,27	-2,98
NMT 100M	Δ BATEZBESTEKOA	+2,47	-3,03
	Δ MEDIANA	-0,27	-5,77
	Δ ONENA	-1,65	-7,15
BATEZBESTEKOA	Δ BATEZBESTEKOA	+2,46	+0,89
	Δ MEDIANA	+0,57	-0,99
	Δ ONENA	-1,46	-3,03

tezkoak izateko garrantzitsua da antzeko segmentuak izatea.

Gure metodoa gainerako partaideekin hobeto konparatzeko batezbesteko emaitzen diferentziak kalkulatu dira. 6.3. taulan azaltzen dira partaideen emaitzen batezbesteko (Δ BATEZBESTEKO), mediana (Δ MEDIANA) eta partaide onenarekiko (Δ ONENA) ezberdintasunak. Eredu guztiekin izandako emaitzen batezbestekoarekiko ezberdintasunak ere aurkezten dira.

STACC.OOV metodoaren emaitzak batezbestekotik gora daude, batez ere NMT ereduetan, non emaitzetan +4,51 eta +2,47 puntuko hobekuntzak ikusten diren. STACC.OOV.NGSAT metodoaren emaitzak ere onak dira, baina NMT 100M ereduan -3.03 puntuko beherakada dago. Medianaren kasuan ezberdintasunak txikiagoak dira, baina orokorrean jatorrizko metodoaren emaitzak hobeak dira NMT 100M ereduaren salbuespen bakarrarekin (-0,27).

Sistema onenarekin konparatuz, NMT ereduetan ezberdintasunak handiagoak

dira SMT ereduetan baino. Batezbestekoz, STACC.OOV metodoa sistema onena baino -1,46 puntu okerrago dabil, eta partaide guztien batezbestekoak baino emaitza hobeak ditu +2,46 puntuko aldearekin. Kontutan eduki behar da STACC.OOV metodoaren eraginkortasuna, 104 milioi esaldi paralelo ordu batean prozesatzeko gai da inolako entrenamendu edo aparteko hizkuntza baliabideen beharrik gabe. Beraz, gure metodoa corpus paralelo zaratatsuak iragazteko aukera erabilgarri eta fidagarria dela esan dezakegu.

6.4. Ondorioak

Kapitulu honetan esaldi paraleloen iragazketaren inguruan egindako ikerketak azaltzen dira. Aurkeztutako metodoa 4. kapituluko STACC metodoan oinarritzen da, zein funtsean itzulpen lexikoen taulak bakarrik erabiltzen dituen esaldien antzekotasuna neurtzeko. Oinarrizko metodoa esaldien iragazpenera egokitzeko termino ezezagunen dentsitatea aztertzen duen zigortze funtzio bat gehitu zaio antzekotasun metrikari. Ideia nagusia esaldi zaratatsuak baztertzea da nahiz eta itzulpen zuzenak izan. Gainera, datuen erredundantzia kontutan hartzeko n-gramen asetasunaren inguruan ere zenbait ikerketa egin dira.

Egindako ikerketak ebaluatzeko WMT 2018-ko esaldi paraleloen iragazpen atazan parte hartu da bi sistema aurkeztuz: termino ezezagunen dentsitatea neurtzen duen STACC.OOV sistema, eta n-gramen asetasuna aztertzen duen STACC.OOV.NGSAT sistema. Metodoen eraginkortasun konputazionala eta sinpletasuna kontutan hartuz, izandako emaitzak positiboak izan dira. Izan ere, itzulpen lexikoen taulak bakarrik behar dira, eta hauek oso modu eraginkorrean sortu daitezke GIZA++ (Och eta Ney, 2003) edo FASTALIGN (Dyer *et al.*, 2013) bezalako tresnak erabiliz. Orokorrean, emaitza lehiakorrak lortu dira, partaide guztien lehen erdian sailkatuz eta bakarrik sistema onena baino -1,5 puntu gutxiago lortuz. N-grama asetasuna aztertzen duen bertsioaren emaitzak ez dira hain onak; konputazionalki metodo pisutsuagoa da eta orokorrean emaitzak okerragoak dira, batez ere NMT ereduetan.

Bukatzeko, esan beharra dago WMT 2018 corpuseko 104 milioi esaldi pareak prozesatzeko ordu bat nahikoa dela proposatutako STACC.OOV metodoarekin, konputazionalki oso metodo arina da beraz. Lortutako emaitzak

eta STACC.OOV metodoaren eraginkortasuna kontutan hartzen baditugu, benetako egoeratan aurki daitezkeen corpus zaratatsuak iragazteko baliozko metodoa dela esan daiteke.

7. KAPITULUA

Aplikazioak

Tesi honetako ikerketak benetako egoeratan erabili dira eta zenbait enpresen beharrei erantzuna ematen lagundu dute. Askotan, enpresek haien web-orriak, dokumentazio teknikoa edota bestelako dokumentuak itzuli behar dituzte, eta gainera, itzulpenak enpresen domeinuetara egokitu behar dira. Tamalez, normalean enpresek ez dituzte tamaina egokiko corpus paraleloak itzulpen automatikoa elikatzeko, edo dokumentu paraleloak ez daude formatu egokian edo behar bezala antolatuta. Egindako ikerketei esker corpus paraleloen tamaina eta kalitatea hobetzea lortu da.

Hona hemen azaldutako ikerketak erabili direneko proiektu nagusiak:

- **ADAPTA**. ADAPTA proiektua (*Traducción Automática Totalmente Personalizada basada en la Explotación de Datos Heterogeneos* - RTC-2015-3627-7) *Retos Colaboración* programaren bidez finantzatutako proiektua da (2015 - 2018). Proiektuaren helburu nagusia itzulpen automatiko estatistikoa erabiltzailearen beharretara egokitzea da. Egokitzapenak domeinu edota lexiko mailan egin daitezke. Proiektu hau izan zen DOCAL, STACC eta RFR sistemen euskarrietako bat.

- **TRADIN**. TRADIN proiektua (*Traducción Automática Personalizada para el Sector Industrial Basada en la Explotación Intensiva de Datos* - IG-2015/0000347) HAZITEK programaren bidez finantzatutako proiektua da (2015 - 2017). Proiektu hau ADAPTA proiektuaren antzekoa da, ezberdintasun nagusia foku industrialean dago; ohikoa da enpresek do-

kumentazio teknikoa izatea, eta nahiz eta dokumentu hauek batzuetan itzulita ez egon, enpresen domeinua definitzeko erabili daitezke. Proiektu hau izan zen DOCAL, STACC eta RFR sistemen beste euskarrietako bat.

- **MODELA**. MODELA proiektua[1] (*Modelaketa estatistikoa eta Deep Learninga kalitate handiko itzulpen automatikorako*) ELKARTEK programaren bidez finantzatutako proiektua da (2016 - 2017). MODELA proiektuaren helburua sare neuronaletan oinarritutako azken teknikak erabiliz euskara-gaztelaniarako NMT ereduak entrenatzea eta SMT ereduekin konparatzea da. MODELA proiekturako corpus paraleloak besteak beste DOCAL, STACC eta RFR erabiliz sortu ziren.

- **ELRI**. ELRI proiektua[2] (Etchegoyhen *et al.*, 2018) (*European Language Resource Infrastructure* - INEA/CEF/ICT/A2016/1330962), CEF[3] programaren bidez finantzatutako proiektua da (2017 - 2019). Proiektuaren helburua, administrazio publikoek azpiegitura deszentralizatu baten bidez hizkuntza baliabideak partekatzea da. ELRI-ri esker, erabiltzaileek elkarbanatutako corpusak automatikoki prozesatzen dira corpus paraleloak sortzeko eta ELRC-SHARE[4] gordailuaren bidez atzigarriak dira. ELRI proiektuaren barne prozesaketa kate bat diseinatu zen, eta bertan, DOCAL erabiltzen da dokumentuak lerrokatzeko.

[1] http://investigacion.modela.eus/
[2] http://www.elri-project.eu/
[3] https://ec.europa.eu/inea/en/connecting-europe-facility
[4] https://elrc-share.eu/

8. KAPITULUA

Ondorioak

Tesian zehar itzulpen automatikorako baliabideen sorkuntzaren arazoari aurre egiteko egindako ikerketak aurkeztu dira. Tesiaren muina itzulpen automatikoan oinarritutako sistemak elikatzen dituen corpus paraleloen tamaina eta kalitatea hobetzea da, bide honetatik itzulpen egokiagoak lortu ahal izateko. Gainera, corpus paraleloak domeinu jakin batera egokitzeko tresnak garatu dira, itzulpenak ere domeinu horren estilo eta lexikora moldatzeko.

Entrenamendurako corpusak hobetzeak onura ugari ditu. Datuetan oinarritutako itzulpen automatikoa da, lehendabizi itzulpen automatiko estatistikoarekin eta gero itzulpen automatiko neuronalarekin, gailendu den paradigma, eta sistema hauek datuetatik erauzten dute jakintza guztia, beraz, corpusean dago benetako altxorra. Hortaz, entrenamendurako corpusak aberastuz, SMT sistemak ezezik NMT sistemak ere hobetuko dira.

Tesiaren ikerlerro nagusiak honako hauek dira: dokumentuan lerrokatzea, esaldien lerrokatzea, datuen aukeraketa eta esaldi paraleloen iragazpena.

Dokumentuen lerrokatzea. Dokumentuak lerrokatzeko DOCAL sistema aurkeztu da, dokumentuen edukian bakarrik oinarritzen den sistema. Metodoaren oinarriak Jaccard koefizientea eta itzulpen-taulen bidez lortutako itzulpen lexikoak dira. Gero, metodoa osatzeko, dokumentuetako terminoen aurrizki komunak bilatzen dira, izen-entitateak aztertzen dira, eta bilaketa espazioa murrizteko APACHE LUCENE indizea erabiltzen da. Dokumentu bat beste bat baino gehiagorekin lerrokatzen den kasuetan, antzekotasun metrika onena duen lerrokatzea da aukeratzen dena.

DOCAL sistemaren ezaugarri garrantzitsu bat bere malgutasuna da. DOCAL-en bitartez dokumentu konparagarri zein paraleloak lerrokatu daitezke, eta ez du inolako konfigurazio zehatzik behar uneko hizkuntza edo domeinura egokitzeko.

Eraginkortasuna aztertu ahal izateko ebaluazio sakonak egin dira artearen egoerako beste sistemekin konparatuz. Gainera, WMT 2016-ko dokumentuen lerrokatze atazako parte hartzea ere azaltzen da. Guztira, 4 corpus eta 6 hizkuntza paretan lortutako emaitzak aurkezten dira.

Orokorrean, DOCAL sistemaren emaitzak oso onak dira. Kasu gehienetan emaitza onenak lortu dituen sistema izan da, eta gainerako besteetan, sistema onenetik oso hurbil geratu da. DOCAL-en beste abantaila bat bere errendimendua da, 45 segundotan ia 90.000 lerrokapen aztertzeko gauza baita. Gainera, kontutan eduki behar da konparaketak egiteko mota ezberdinetako metodoak aukeratu direla. Metodo batzuek entrenatu beharreko ereduak erabiltzen dituzte, beraz, metodo hauen eramangarritasuna DOCAL-ena baino okerragoa da. Beste batzuek dokumentuen metadatuak behar dituzte, eta egindako proben arabera, corpusaren arabera abantaila garbia da, baina tamalez, benetako egoeratan ez da ohizkoa metadatu horiek eskura egotea edo metadatuen egitura aurreikusitakoa izatea.

Oinarrizko metodoari zenbait optimizazio egin zaizkio. Hasteko, esaldiko termino esanguratsuenak hobesteko pisu lexikoak kalkulatzeko metodo bat sortu da. Bigarren optimizazio bat edozein domeinura egokitzeko itzulpen-taula generikoen erabilera da. Dokumentuen indexazioarekin ere zenbait esperimentu egin dira indizetik dokumentu egokiagoak berreskuratzeko asmoz. Azkenik, dokumentuen aurreprozesaketaren inguruan zenbait esperimentu egin dira. Optimizazioen eragina neurtzeko ebaluazio guztiak errepikatu dira, eta emaitza arrakastatsuenak itzulpen-taula generikoekin lortu dira. Orokorrean, dokumentuen aurreprozesaketarekin lortutako emaitzak antzekoak dira, baina termino gutxiko esaldiak dituzten corpusetan emaitzak asko hobetzen direla egiaztatu ahal izan da.

Esaldien lerrokatzea. Dokumentuen lerrokatzearen ostean esaldiak lerrokatu behar dira. Dokumentuak lerrokatu gabe esaldiak zuzenean lerrokatzea posible da, baina dokumentuen lerrokatzearekin bilaketa espazioa murriztea

lortzen da azken lerrokatzeen kalitatea hobetuz.

Esaldiak lerrokatzeko STACC metodoa garatu da. Funtsean, STACC eta DOCAL teknika berberetan oinarritzen dira. Dokumentuek eta esaldiek, batez ere termino kopurua dela eta, ezaugarri ezberdinak dituzte, eta horregatik STACC eta DOCAL-ek zenbait ezberdintasun dituzte.

Esaldietan dokumentuetan baino termino gutxiago daudenez, terminoen normalizazioak eragin handiagoa du, eta beraz, esaldien aurreprozesaketak dokumentuetan baino garrantzia handiagoa du. Termino kopuruarengatik, termino komunen kalkuluak dokumentuetan baino kostu konputazional gutxiago du, eta gainera, emaitzetan eragin handiagoa du. Era berean, pisu lexikoek esaldien errepresentazioak elkarren artean gehiago bereiztea ahalbidetzen du. Azkenik, esaldietako izen-entitateen garrantzia areagotzeko, esaldien artean partekatzen ez diren izen-entitateen agerpenak zigortzeko funtzio bat garatu da.

STACC artearen egoerarekiko nola kokatzen den aztertzeko esperimentu ugari egin dira. Guztira, 7 corpus eta 12 hizkuntza pareekin egin da ebaluazioa. Orokorrean, lortutako emaitzak oso onak dira, salbuespen bat alde batera utzita baldintza guztietan artearen egoera hobetzea lortu baita.

Esaldien errepresentazioak itzultzeko itzulpen-taulak erabili beharrean, jatorrizko esaldiak itzulpen automatikoaren bidez zuzenean itzuliz nolako lerrokatzeak lortzen diren aztertu da. Kasu honetan emaitza okerragoak lortu dira, itzulpen automatikoarekin itzulpen-taulekin baino errepresentazio zurrunagoak lortzen baitira.

Gainera, ikerlerro honetatik sortutako kontribuzio garrantzitsu bat EITB corpusa da. STACC-en bidez EITB-ko albisteetatik sortutako corpus konparagarria lerrokatu da, eta ia 600.000 esaldi paraleloko corpusa lortu da. Gainera, EITB corpusa lizentzia librepean eskuratu daiteke[1].

Datuen aukeraketa. Datuen aukeraketari esker domeinu jakin baterako datu gehiago eskuratu daitezke datu multzo egokiak domeinuz kanpoko beste corpus ugariagoetan bilatuz. Horretarako, terminoen maiztasun erlatiboetan

[1]Corpusa lortzeko jo esteka honetara: `http://metashare.elda.org/repository/search/?q=eitb+documents`

oinarritutako RFR (*Relative Frequency Ratios*) metodoa garatu da.

RFR metodoa nahiko sinplea da: esaldi pare bakoitzeko terminoek domeinuko eta domeinutik kanpoko corpusean duten maiztasun erlatiboa konparatzen da domeinuarekiko antzekoenak diren esaldiak aukeratzeko. RFR metodoaren hobekuntza bat ere aurkezten da, WRFR (*Weighted Relative Frequency Ratios*) alegia. WRFR metodoaren helburua domeinua lexiko berriarekin aberastea da. Horretarako, oinarrizko metodoari funtzio bat gehitzen dio esaldiko termino ezezagunen ehunekoa aztertzen duena.

(W)RFR artearen egoerako MML (*Modified Moore-Lewis*) metodoarekin konparatu da (esaldien perplexitatean oinarritzen dena). Esperimentuak bi jatorrizko domeinutan egiten dira kanpoko domeinuko corpus berbera erabiliz, eta kanpoko domeinuko lagin ezberdinak aukeratuz lortutako emaitzak aurkeztu dira. Bai BLEU metrikaren aldetik eta baita datuen sakabanaketaren aldetik, emaitzak hobetzea lortu da. Kasu gehienetan WRFR metodoarekin lortu dira emaitza onenak.

Esaldi paraleloen iragazketa. Esaldi paraleloen iragazpena itzulpen automatikorako ereduen entrenamendurako kaltegarriak diren datuak iragaztean datza, horrela, itzulpenen kalitatea hobetzea lortzen da. Zaratatsuak diren corpusek zenbait arazo izan ditzakete, adibidez, hizkuntza okerreko esaldiak egon daitezke eta baliteke esaldiek karaktere okerrak izatea.

Esaldi paraleloen iragazpena esaldien antzekotasunaren ikuspegitik hartu daiteke, hau da, jatorrizko eta helburuko hizkuntzako esaldiek ez badute zerikusirik, orduan esaldi parea zaratatsua da. Esaldien lerrokatzean izandako emaitza onak ikusita, STACC metodoa esaldien iragazpenerako egokitzeko zenbait ikerketa egin dira. Zehazki, bi aldaerekin egin dira esperimentuak: termino ezezagunen dentsitatea aztertzen duen zigortze funtzio batekin (STACC.OOV), eta n-grama asetasuna neurtzen duen algoritmo batekin (STACC.OOV.NGSAT).

Egindako ikerketak balioztatzeko WMT 2018-ko esaldi paraleloen iragazpen atazan[2] (Koehn *et al.*, 2018) parte hartu zen. Tesiaren helburuetako bat eramangarritasuna da, eta beraz, atazarako egokitzapen berezirik ez egiteko

[2] `http://www.statmt.org/wmt18/parallel-corpus-filtering.html`

erabakia hartu zen, hau da, ez da inolako aurreprozesaketarik egin esaldi pare zaratatsuenak iragazteko. Guzti hau kontutan hartuta, emaitza lehiakorrak lortu ziren, STACC.OOV metodoa WMT 2018-ko sistema onenetik gertu geratu baitzen -1,5 puntuko aldearekin. STACC.OOV.NGSAT metodoaren emaitzak okerragoak izan ziren.

Azken hitzak. Egindako ikerketa guztiek bi ikuspegitik izan dute eragina: alderdi zientifikoa eta alderdi praktikoa. Alderdi zientifikoari dagokionez, mota ezberdinetako argitalpenak egin dira. Atariko ikerketekin pisu ezberdineko kongresutan egin dira, eta lortutako emaitzak balioztatzeko eta hauen gainean hobekuntzak egiteko zenbait partekatutako atazetan parte hartu da artearen egoerarekin lehiatuz. Ikerketak biltzeko kongresu eta aldizkarietan azken argitalpenak egin dira. Ikuspegi praktikotik, 7. kapituluan azaltzen den bezala, garatutako tresnak benetako egoeratan erabili dira itzulpen automatikorako ereduak entrenatzeko zenbait proiekturen testuingurutan. Proiektu hauetan barne, itzulpen automatikoa enpresa ezberdinen domeinuetara egokitu ahal izan da.

Egindako ikerketak laburbilduz, hasieran planteatutako helburuak bete dira, artearen egoerarekiko emaitza lehiakorrak lortu ezezik, oso metodo eramangarriak lortu baitira. Ikerketa hauei esker, benetako egoeratan SMT eta NMT eredu hobeak entrenatu ahal izan dira entrenamendurako corpusen kalitatea hobetuz (ikus 7. kapitulua).

Glosarioa

Arreta mekanismo Attention mechanism. 15

Aurrizki komun luzeena Longest common prefix. 30

Datuen aukeraketa Data selection. 107

Datuen sakabanaketa Data sparseness. 59

DOCAL DOCument ALignment. 28

Dokumentuen lerrokatzea Document Alignment. 27

EM Expectation Maximization. 13

Entropia maximoa Maximum entropy. 18

Esaldi paraleloen iragazketa Parallel sentence filtering. 125

Esaldien lerrokatzea Sentence Alignment. 67

Ezaugarrien narriadura Feature decay. 128

Hapax Corpusean behin bakarrik azaltzen diren terminoak. 36

Hizkuntza arteko informazio-eskuratzea Cross-language information retrieval (CLIR). 33

Hiztegitik kanpoko terminoen hedapena Out-of-vocabulary expansion. 31

Itzulpen-taula Lexical translation table. 29

IVAP Instituto Vasco de Adiministración Pública. 79

Lerrokatze onenaren optimizazioa Best alignment optimisation. 34

Levenshtein distantzia Levenshtein distance. 14

Logistic regression . 70

MML Modified Moore-Lewis. 113

NMT Neural Machine Translation. 1

Perplexitatea Perplexity. 54

Pisu lexikoak . 49

RFR Relative Frequency Ratios. 108

Sare neuronal errepikari Recurrent neural network. 20

SMT Statistical Machine Translation. 1

SSLL Small Smallish Largish Large. 57

STACC Set-Theoretic Alignment of Comparable Corpora. 66

Stop-word Stop-word. 69

SVM Support Vector Machines. 17

Termino funtzionalak Esanahi lexikorik gabeko terminoak. 69

TF-IDF Term frequency – Inverse document frequency. 13

Truecasing Esaldiko lehenengo terminoari probabilitate handieneko forma ematea (letra larriz ala xehez).. 59

Zuhaitz atzizki Suffix tree. 16

Bibliografia

Acarcicek H., Çolakoğlu T., Hatipoğlu P.E.A., Huang C.H., eta Peng W. Filtering noisy parallel corpus using transformers with proxy task learning. *Proceedings of the Fifth Conference on Machine Translation*, 940–946, 2020.

Aharoni R. eta Goldberg Y. Unsupervised domain clusters in pretrained language models. *arXiv preprint arXiv:2004.02105*, 2020.

Artetxe M. eta Schwenk H. Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*, 2018.

Artetxe M. eta Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019.

Axelrod A., He X., eta Gao J. Domain adaptation via pseudo in-domain data selection. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 355–362, 2011.

Axelrod A., Li Q., eta Lewis W.D. Applications of data selection via cross-entropy difference for real-world statistical machine translation. *International Workshop on Spoken Language Translation (IWSLT) 2012*, 2012.

Axelrod A., Resnik P., He X., eta Ostendorf M. Data selection with fewer words. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 58–65, 2015a.

Axelrod A., Vyas Y., Martindale M., Carpuat M., eta Hopkins J. Class-based n-gram language difference models for data selection. *IWSLT (International Workshop on Spoken Language Translation)*, 180–187, 2015b.

Aydın B. eta Ozgür A. Expanding machine translation training data with an out-of-domain corpus using language modeling based vocabulary saturation. *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA), Vancouver, BC, Canada*, 2014.

Bahdanau D., Cho K., eta Bengio Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Banerjee P., Naskar S.K., Roturier J., Way A., eta van Genabith J. Translation quality-based supplementary data selection by incremental update of translation models. *Proceedings of COLING 2012*, 149–166, 2012.

Banerjee P., Rubino R., Roturier J., eta van Genabith J. Quality estimation-guided data selection for domain adaptation of smt. *MT Summit XIV: proceedings of the fourteenth Machine Translation Summit*, 101–108, 2013.

Banón M., Chen P., Haddow B., Heafield K., Hoang H., Espla-Gomis M., Forcada M.L., Kamran A., Kirefu F., Koehn P., *et al.*. Paracrawl: Web-scale acquisition of parallel corpora. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4555–4567, 2020.

Bérard A., Servan C., Pietquin O., eta Besacier L. Multivec: a multilingual and multilevel representation learning toolkit for nlp. *The 10th edition of the Language Resources and Evaluation Conference (LREC)*, 2016.

Bernier-Colborne G. eta Lo C.k. Nrc parallel corpus filtering system for wmt 2019. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 252–260, 2019.

Biçici E. eta Yuret D. Instance selection for machine translation using feature decay algorithms. *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 272–283, 2011.

Bisazza A., Ruiz N., eta Federico M. Fill-up versus interpolation methods for phrase-based smt adaptation. *International Workshop on Spoken Language Translation (IWSLT) 2011*, 2011.

Bojar O., Chatterjee R., Federmann C., Fishel M., Graham Y., Haddow B., Huck M., Yepes A.J., Koehn P., Monz C., *et al.*. Proceedings of the third conference on machine translation. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018.

Bouamor H. eta Sajjad H. H2@ bucc18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. *Proc. Workshop on Building and Using Comparable Corpora*, 2018.

Breiman L. Random forests. *Machine learning*, 45(1):5–32, 2001.

Brown P.F., Cocke J., Della Pietra S.A., Della Pietra V.J., Jelinek F., Lafferty J., Mercer R.L., eta Roossin P.S. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

Brown P.F., Della Pietra S.A., Della Pietra V.J., eta Mercer R.L. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

Buck C. eta Koehn P. Findings of the wmt 2016 bilingual document alignment shared task. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 554–563, 2016a.

Buck C. eta Koehn P. Quick and reliable document alignment via tf/idf-weighted cosine distance. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 672–678, 2016b.

Chaudhary V., Tang Y., Guzmán F., Schwenk H., eta Koehn P. Low-resource corpus filtering using multilingual sentence embeddings. *arXiv preprint arXiv:1906.08885*, 2019.

Chen J. eta Nie J.Y. Parallel Web Text Mining for Cross-language IR. *Content-Based Multimedia Information Access - Volume 1*, 62–77, Paris,

France, 2000. Centre des hautes études internationales d'informatique documentaire.

Chen J., Chau R., eta Yeh C.H. Discovering Parallel Text from the World Wide Web. *Proceedings of the Second Workshop on Australasian Information Security, Data Mining and Web Intelligence, and Software Internationalisation*, 157–161, Dunedin, New Zealand, 2004.

Cho K., Van Merriënboer B., Gulcehre C., Bahdanau D., Bougares F., Schwenk H., eta Bengio Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Clark J.H., Dyer C., Lavie A., eta Smith N.A. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 176–181, 2011.

Conneau A., Khandelwal K., Goyal N., Chaudhary V., Wenzek G., Guzmán F., Grave E., Ott M., Zettlemoyer L., eta Stoyanov V. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Cortes C. eta Vapnik V. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Cui L., Zhang D., Liu S., Li M., eta Zhou M. Bilingual data cleaning for smt using graph-based random walk. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 340–345, 2013.

Dara A.A. eta Lin Y.C. Yoda system for wmt16 shared task: Bilingual document alignment. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 679–684, 2016.

Daumé Iii H. eta Jagarlamudi J. Domain adaptation for machine translation by mining unseen words. *Proceedings of the 49th Annual Meeting of the As-*

sociation for Computational Linguistics: Human Language Technologies, 407–412, 2011.

Denkowski M. eta Lavie A. Meteor universal: Language specific translation evaluation for any target language. *Proceedings of the ninth workshop on statistical machine translation*, 376–380, 2014.

Devlin J., Chang M.W., Lee K., eta Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dice L.R. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

Dowling M., Lynn T., Poncelas A., eta Way A. Smt versus nmt: Preliminary comparisons for irish. 2018.

Duh K., Neubig G., Sudoh K., eta Tsukada H. Adaptation data selection using neural language models: Experiments in machine translation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 678–683, 2013.

Dyer C., Chahuneau V., eta Smith N.A. A simple, fast, and effective reparameterization of ibm model 2. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648, 2013.

Eck M., Vogel S., eta Waibel A. Low cost portability for statistical machine translation based on n-gram frequency and tf-idf. *International Workshop on Spoken Language Translation (IWSLT) 2005*, 2005.

Eetemadi S., Lewis W., Toutanova K., eta Radha H. Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29(3-4): 189–223, 2015.

Eisele A. eta Chen Y. Multiun: A multilingual corpus from united nation documents. *LREC*, 2010.

Enright J. eta Kondrak G. A Fast Method for Parallel Document Identification. *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 29–32, Rochester, New York, USA, 2007.

Erdmann G. eta Gwinnup J. Quality and coverage: The afrl submission to the wmt19 parallel corpus filtering for low-resource conditions task. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 267–270, 2019.

Escutia M. Chomsky, la naturaleza humana, el lenguaje y las limitaciones de la ciencia y una propuesta complementaria inspirada en c. s. lewis. 2013.

Esplà-Gomis M. eta Forcada M.L. Bitextor, a free/open-source software to harvest translation memories from multilingual websites. *Proceedings of MT Summit XII*, 1–8, Ottawa, Canada, 2009.

Esplà-Gomis M., Forcada M.L., Ortiz-Rojas S., eta Ferràndez-Tordera J. Bitextor's participation in WMT'16: shared task on document alignment. *Proceedings of the First Conference on Machine Translation*, 685–691, Berlin, Germany, 2016.

Esplà-Gomis M., Sánchez-Cartagena V.M., Zaragoza-Bernabeu J., eta Sánchez-Martínez F. Bicleaner at wmt 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. *Proceedings of the Fifth Conference on Machine Translation*, 952–958, 2020.

Etchegoyhen T., Anza Porras B., Azpeitia A., Martínez Garcia E., Vale P., Fonseca J.L., Lynn T., Dunne J., Gaspari F., Way A., *et al.*. Elri. european language resource infrastructure. 2018.

Forcada M.L., Espla-Gomis M., Perez-Ortiz J.A., *et al.*. Stand-off annotation of web content as a legally safer alternative to crawling for distribution. 2016.

Foster G., Goutte C., eta Kuhn R. Discriminative instance weighting for domain adaptation in statistical machine translation. *Proceedings of the*

2010 conference on empirical methods in natural language processing, 451–459, 2010.

Friedman J., Hastie T., Tibshirani R., *et al.*. *The elements of statistical learning*, 1 lib. Springer series in statistics New York, 2001.

Fung P. eta Cheung P. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and e. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 57–63, 2004.

Gascó G., Rocha M.A., Sanchis-Trilles G., Andrés-Ferrer J., eta Casacuberta F. Does more data always yield better translations? *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 152–161, 2012.

Gelbukh A., Sidorov G., Lavin-Villa E., eta Chanona-Hernandez L. Automatic term extraction using log-likelihood based comparison with general reference corpus. *International conference on application of natural language to information systems*, 248–255. Springer, 2010.

Germann U. Bilingual document alignment with latent semantic indexing. *arXiv preprint arXiv:1707.09443*, 2017.

Geurts P., Ernst D., eta Wehenkel L. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

Gomes L. eta Lopes G. First steps towards coverage-based document alignment. *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 697–702, 2016.

Grégoire F. eta Langlais P. Bucc 2017 shared task: a first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 46–50, 2017.

Gretton A., Borgwardt K.M., Rasch M.J., Schölkopf B., eta Smola A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1): 723–773, 2012.

Guo H., Pasunuru R., eta Bansal M. Multi-source domain adaptation for text classification via distancenet-bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 lib., 7830–7838, 2020.

Gururangan S., Marasović A., Swayamdipta S., Lo K., Beltagy I., Downey D., eta Smith N.A. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Heafield K. Kenlm: Faster and smaller language model queries. *Proceedings of the sixth workshop on statistical machine translation*, 187–197, 2011.

Heafield K., Pouzyrevsky I., Clark J.H., eta Koehn P. Scalable modified kneser-ney language model estimation. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 690–696, 2013.

Hochreiter S. eta Schmidhuber J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Imam A.H., Arman M.R.M., Chowdhury S.H., eta Mahmood K. Impact of corpus size and quality on english-bangla statistical machine translation system. *14th International Conference on Computer and Information Technology (ICCIT 2011)*, 566–571. IEEE, 2011.

Ion R. Pexacc: A parallel sentence mining algorithm from comparable corpora. *LREC*, 2181–2188. Citeseer, 2012.

Ion R., Ceauşu A., eta Irimia E. An expectation maximization algorithm for textual unit alignment. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 128–135, 2011.

Jaccard P. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272, 1901.

Jakubina L. eta Langlais P. BAD LUC@WMT 2016: a Bilingual Document Alignment Platform Based on Lucene. *Proceedings of the First Conference on Machine Translation*, 703–709, Berlin, Germany, 2016.

Jiang J.Y., Zhang M., Li C., Bendersky M., Golbandi N., eta Najork M. Semantic text matching for long-form documents. *The World Wide Web Conference*, 795–806, 2019.

Johnson H., Martin J., Foster G., eta Kuhn R. Improving translation quality by discarding most of the phrasetable. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 967–975, 2007.

Jones K.S., Walker S., eta Robertson S.E. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840, 2000.

Junczys-Dowmunt M. Dual conditional cross-entropy filtering of noisy parallel corpora. *arXiv preprint arXiv:1809.00197*, 2018.

Khadivi S. eta Ney H. Automatic filtering of bilingual corpora for statistical machine translation. *International Conference on Application of Natural Language to Information Systems*, 263–274. Springer, 2005.

Khayrallah H. eta Koehn P. On the impact of various types of noise on neural machine translation. *arXiv preprint arXiv:1805.12282*, 2018.

Kirchhoff K. eta Bilmes J. Submodularity for data selection in machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 131–141, 2014.

Koehn P. Statistical significance tests for machine translation evaluation. *Proceedings of the 2004 conference on empirical methods in natural language processing*, 388–395, 2004.

Koehn P. Europarl: A parallel corpus for statistical machine translation. *MT summit*, 5 lib., 79–86. Citeseer, 2005.

Koehn P., Chaudhary V., El-Kishky A., Goyal N., Chen P.J., eta Guzmán F. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. *Proceedings of the Fifth Conference on Machine Translation*, 726–742, 2020.

Koehn P., Guzmán F., Chaudhary V., eta Pino J. Findings of the wmt 2019 shared task on parallel corpus filtering for low-resource conditions. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 54–72, 2019.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., *et al.*. Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics, 2007.

Koehn P., Khayrallah H., Heafield K., eta Forcada M.L. Findings of the wmt 2018 shared task on parallel corpus filtering. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 726–739, 2018.

Koehn P., Och F.J., eta Marcu D. Statistical phrase-based translation. Barne-txostena, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 2003.

Lafferty J., McCallum A., eta Pereira F.C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Lample G. eta Conneau A. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Le T., Vu H.T., Oberländer J., eta Bojar O. Using Term position Similarity and Language Modeling for Bilingual Document Alignment. *Proceedings of the First Conference on Machine Translation*, 710–716, Berlin, Germany, 2016.

Leong C., Wong D.F., eta Chao L.S. Um-paligner: Neural network-based parallel sentence identification model. *11th Workshop on Building and Using Comparable Corpora*, page 53, 2018.

Levenshtein V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10 lib., 707–710, 1966.

Lewis W. eta Eetemadi S. Dramatically reducing training data size through vocabulary saturation. *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 281–291, 2013.

Li B. eta Gaussier E. Exploiting comparable corpora for lexicon extraction: measuring and improving corpus quality. *Building and Using Comparable Corpora*, 131–149. Springer, 2013.

Lin Y.H., Chen C.Y., Lee J., Li Z., Zhang Y., Xia M., Rijhwani S., He J., Zhang Z., Ma X., *et al.*. Choosing transfer languages for cross-lingual learning. *arXiv preprint arXiv:1905.12688*, 2019.

Littell P., Larkin S., Stewart D., Simard M., Goutte C., eta Lo C.k. Measuring sentence parallelism using mahalanobis distances: The nrc unsupervised submissions to the wmt18 parallel corpus filtering shared task. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 900–907, 2018.

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., eta Stoyanov V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lo C.k. The nrc metric submission to the wmt18 metric and parallel corpus filtering shared task, 2018.

Lo C.k. eta Joanis E. Improving parallel data identification using iteratively refined sentence alignments and bilingual mappings of pre-trained language models. *Proceedings of the Fifth Conference on Machine Translation*, 972–978, 2020.

Lo C.k., Simard M., Stewart D., Larkin S., Goutte C., eta Littell P. Accurate semantic textual similarity for cleaning noisy parallel corpora using semantic machine translation evaluation metric: The nrc supervised submissions to the parallel corpus filtering task. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 908–916, 2018.

Lu J., Ge X., Shi Y., eta Zhang Y. Alibaba submission to the wmt20 parallel corpus filtering task. *Proceedings of the Fifth Conference on Machine Translation*, 979–984, 2020.

Lü Y., Huang J., eta Liu Q. Improving statistical machine translation performance by training data selection and optimization. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 343–350, 2007.

Ma X. eta Liberman M. BITS: A Method for Bilingual Text Search over the Web. *Machine Translation Summit VII*, 538–542, Singapore, 1999.

Mahalanobis P.C. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, number 2, 49—-55, 1936.

Mahata S., Das D., eta Bandyopadhyay S. Bucc2017: A hybrid approach for identifying parallel sentences in comparable corpora. *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 56–59, 2017.

Mahata S., Das D., eta Pal S. WMT2016: A Hybrid Approach to Bilingual Document Alignment. *Proceedings of the First Conference on Machine Translation*, 724–727, Berlin, Germany, 2016.

Mahata S.K., Mandal S., Das D., eta Bandyopadhyay S. Smt vs nmt: a comparison over hindi & bengali simple sentences. *arXiv preprint arXiv:1812.04898*, 2018.

Mansour S., Wuebker J., eta Ney H. Combining translation and language model scoring for domain-specific data filtering. *International Workshop on Spoken Language Translation (IWSLT) 2011*, 2011.

Matsoukas S., Rosti A.V., eta Zhang B. Discriminative corpus weight estimation for machine translation. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 708–717, 2009.

Mediani M., Winebarger J., eta Waibel A. Improving in-domain data selection for small in-domain sets. *Proceedings of IWSLT*, 47 lib., 2014.

Medveď M., Jakubícek M., eta Kovár V. English-French Document Alignment based on Keywords and Statistical Translation. *Proceedings of the First Conference on Machine Translation*, 728–732, Berlin, Germany, 2016.

Mikolov T., Sutskever I., Chen K., Corrado G.S., eta Dean J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111–3119, 2013.

Miller G.A. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

Miller G.A. *WordNet: An electronic lexical database*. MIT press, 1998.

Moon T.K. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

Moore R.C. eta Lewis W. Intelligent selection of language model training data. 2010.

Morin E., Hazem A., Boudin F., eta Clouet E.L. Lina: Identifying comparable documents from wikipedia. 2015.

Morrison D.R. Patricia—practical algorithm to retrieve information coded in alphanumeric. *Journal of the ACM (JACM)*, 15(4):514–534, 1968.

Munteanu D.S. eta Marcu D. Processing comparable corpora with bilingual suffix trees. *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, 289–295, 2002.

Munteanu D.S. eta Marcu D. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 2005.

Nigam K., Lafferty J., eta McCallum A. Using maximum entropy for text classification. *IJCAI-99 workshop on machine learning for information filtering*, 1 lib., 61–67. Stockholom, Sweden, 1999.

Och F.J. eta Ney H. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51, 2003.

Papavassiliou V., Prokopidis P., eta Piperidis S. The ILSP/ARC submission to the WMT 2016 Bilingual Document Alignment Shared Task. *Proceedings of the First Conference on Machine Translation*, 733–739, Berlin, Germany, 2016.

Papineni K., Roukos S., Ward T., eta Zhu W.J. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318, 2002.

Paramita M.L., Guthrie D., Kanoulas E., Gaizauskas R., Clough P., eta Sanderson M. Methods for collection and evaluation of comparable documents. *Building and using comparable corpora*, 93–112. Springer, 2013.

Patry A. eta Langlais P. Automatic Identification of Parallel Documents with Light or Without Linguistic Resources. *Proceedings of the 18th Canadian Society Conference on Advances in Artificial Intelligence*, 354–365, Victoria, Canada, 2005.

Patry A. eta Langlais P. Identifying Parallel Documents from a Large Bilingual Collection of Texts: Application to Parallel Article Extraction in Wikipedia. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 87–95, Portland, Oregon, 2011a.

Patry A. eta Langlais P. Identifying parallel documents from a large bilingual collection of texts: Application to parallel article extraction in wikipedia. *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, 87–95, 2011b.

Peris Á., Chinea-Ríos M., eta Casacuberta F. Neural networks classifier for data selection in statistical machine translation. *arXiv preprint arXiv:1612.05555*, 2016.

Peris Á., Chinea-Ríos M., eta Casacuberta F. Neural networks classifier for data selection in statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108(1):283–294, 2017.

Pierre Zweigenbaum S.S. eta Rapp R. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In Rapp R., Zweigenbaum P., eta Sharoff S., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-07-8.

Pilehvar M.T., Jurgens D., eta Navigli R. Align, disambiguate and walk: A unified approach for measuring semantic similarity. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1341–1351, 2013.

Pinnis M., Ion R., Stefanescu D., Su F., Skadiņa I., Vasiļjevs A., eta Babych B. Accurat toolkit for multi-level alignment and information extraction from comparable corpora. *Proceedings of the ACL 2012 System Demonstrations*, 91–96, 2012.

Poncelas A., Maillette de Buy Wenniger G., eta Way A. Feature decay algorithms for neural machine translation. 2018.

Radford A., Narasimhan K., Salimans T., eta Sutskever I. Improving language understanding by generative pre-training. 2018.

Radford A., Wu J., Child R., Luan D., Amodei D., eta Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ramponi A. eta Plank B. Neural unsupervised domain adaptation in nlp—a survey. *arXiv preprint arXiv:2006.00632*, 2020.

Rapp R. Identifying word translations in non-parallel texts. *arXiv preprint cmp-lg/9505037*, 1995.

Rauf S.A. eta Schwenk H. On the use of comparable corpora to improve smt performance. *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, 16–23, 2009.

Rauf S.A. eta Schwenk H. Parallel sentence generation from comparable corpora for improved smt. *Machine translation*, 25(4):341–375, 2011.

Reimers N. eta Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Resnik P. eta Smith N.A. The Web as a Parallel Corpus. *Computational Linguistics*, 29(3):349–380, 2003.

Rossenbach N., Rosendahl J., Kim Y., Graça M., Gokrani A., eta Ney H. The rwth aachen university filtering system for the wmt 2018 parallel corpus filtering task. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 946–954, 2018.

Rousseau A. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, (100):73–82, 2013.

Salton G. eta McGill M.J. Introduction to modern information retrieval. 1986.

Sánchez-Cartagena V.M., Bañón M., Ortiz-Rojas S., eta Ramírez-Sánchez G. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium, October 2018. Association for Computational Linguistics.

Sánchez-Cartagena V.M., Banón M., Ortiz-Rojas S., eta Ramírez-Sánchez G. Prompsit's submission to wmt 2018 parallel corpus filtering shared task. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 955–962, 2018.

Sanh V., Debut L., Chaumond J., eta Wolf T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Sarikaya R., Maskey S., Zhang R., Jan E.E., Wang D., Ramabhadran B., eta Roukos S. Iterative sentence-pair extraction from quasi-parallel corpora for machine translation. *Tenth Annual Conference of the International Speech Communication Association*, 2009.

Schwenk H. Filtering and mining parallel data in a joint multilingual space. *arXiv preprint arXiv:1805.09822*, 2018.

Schwenk H., Chaudhary V., Sun S., Gong H., eta Guzmán F. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*, 2019.

Sen S., Ekbal A., eta Bhattacharyya P. Parallel corpus filtering based on fuzzy string matching. *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, 289–293, 2019.

Sennrich R. eta Volk M. Mt-based sentence alignment for ocr-generated parallel texts. 2010.

Sethy A., Georgiou P.G., Ramabhadran B., eta Narayanan S. An iterative relative entropy minimization-based data selection approach for n-gram model adaptation. *IEEE transactions on audio, speech, and language processing*, 17(1):13–23, 2009.

Sharoff S., Rapp R., eta Zweigenbaum P. Overviewing important aspects of the last twenty years of research in comparable corpora. *Building and Using Comparable Corpora*, 1–17. Springer, 2013.

Sharoff S., Zweigenbaum P., eta Rapp R. Bucc shared task: Cross-language document similarity. *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, 74–78, 2015.

Skadiņa I., Aker A., Mastropavlos N., Su F., Tufis D., Verlic M., Vasiļjevs A., Babych B., Clough P., Gaizauskas R., *et al.*. Collecting and using comparable corpora for statistical machine translation. *Proceedings of the 8th international conference on language resources and evaluation (LREC), Istanbul, Turkey*, 2012.

Smith J., Quirk C., eta Toutanova K. Extracting parallel sentences from comparable corpora using document level alignment. *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, 403–411, 2010.

Snover M., Dorr B., Schwartz R., Micciulla L., eta Makhoul J. A study of translation edit rate with targeted human annotation. *Proceedings of association for machine translation in the Americas*, 200 lib. Citeseer, 2006.

Sun B. eta Saenko K. Deep coral: Correlation alignment for deep domain adaptation. *European conference on computer vision*, 443–450. Springer, 2016.

Taghipour K., Khadivi S., eta Xu J. Parallel corpus refinement as an outlier detection algorithm. *Proceedings of the 13th Machine Translation Summit (MT Summit XIII)*, 414–421, 2011.

Thompson B. eta Koehn P. Vecalign: Improved sentence alignment in linear time and space. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1342–1348, Hong Kong, China, 2019. Association for Computational Linguistics.

Tiedemann J. *Bitext Alignment*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, Williston, VT, 2011.

Tiedemann J. Parallel data, tools and interfaces in opus. *Lrec*, 2012 lib., 2214–2218, 2012.

Tseng H., Chang P.C., Andrew G., Jurafsky D., eta Manning C.D. A conditional random field word segmenter for sighan bakeoff 2005. *Proceedings of the fourth SIGHAN workshop on Chinese language Processing*, 2005.

Tufiş D., Ion R., Ceauşu A., eta Stefanescu D. Improved lexical alignment by combining multiple reified alignments. *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

Üstün A., van der Goot R., Bouma G., eta van Noord G. Multi-team: A multi-attention, multi-decoder approach to morphological analysis. *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 35–49, 2019.

Uszkoreit J., Ponte J.M., Popat A.C., eta Dubiner M. Large scale parallel document mining for machine translation. *Proceedings of the 23rd International Conference on Computational Linguistics*, 1101–1109, Beijing, China, 2010.

Varga D., Halácsy P., Kornai A., Nagy V., Németh L., eta Trón V. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247, 2007.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., eta Polosukhin I. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Veliz C.M., De Clercq O., eta Hoste V. Is neural always better? smt versus nmt for dutch text normalization. *Expert Systems with Applications*, 170: 114500, 2021.

Vogel S., Ney H., eta Tillmann C. Hmm-based word alignment in statistical translation. *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*, 1996.

Wong D.F., Lu Y., eta Chao L.S. Bilingual recursive neural network based data selection for statistical machine translation. *Knowledge-Based Systems*, 108:15–24, 2016.

Xu H. eta Koehn P. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2945–2950, 2017.

Yang L., Zhang M., Li C., Bendersky M., eta Najork M. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 1725–1734, 2020.

Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., eta Le Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*, 2019.

Ye T., Wang T., McGuinness K., Guo Y., eta Gurrin C. Learning multiple views with orthogonal denoising autoencoders. *International Conference on Multimedia Modeling*, 313–324. Springer, 2016.

Young T., Hazarika D., Poria S., eta Cambria E. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

Zafarian A., Sadeghi A.P.A., Azadi F., Ghiasifard S., Panahloo Z.A., Bakhshaei S., eta Ziabary S.M.M. Aut document alignment framework for bucc workshop shared task. *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, 79–87, 2015.

Zhang Z. eta Zweigenbaum P. znlp: Identifying parallel sentences in chinese-english comparable corpora. *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 51–55, 2017.

Zhao B. eta Vogel S. Adaptive parallel sentences mining from web bilingual news collection. *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, 745–748. IEEE, 2002.

Zhou X., Pappas N., eta Smith N.A. Multilevel text alignment with cross-document attention. *arXiv preprint arXiv:2010.01263*, 2020.

Zweigenbaum P., Sharoff S., eta Rapp R. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, 60–67, 2017.

Ștefănescu D., Ion R., eta Hunsicker S. Hybrid parallel sentence mining from comparable corpora. *Proceedings of the 16th conference of the European Association for Machine Translation*, 137–144, 2012.