



IKER
GAZTE
NAZIOARTEKO
IKERKETA EUSKARAZ

II. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2017ko maiatzaren 10, 11 eta 12
Iruñea, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

**Osasun-zientzietako
terminologiaren euskaratze
automatikoaren ebaluazioa,
osasungintzako euskal
komunitatea inplikatur**

*Olatz Perez-De-Viñaspre eta
Maite Oronoz Anchordoqui*

141-148 or.

<https://dx.doi.org/10.26876/ikergazte.ii.03.20>

ANTOLATZAILEA:



ELKARLANEAN:



LAGUNTZAILEAK:



UDALBILTZA



Osasun-zientzietako terminologiaren euskaratze automatikoaren ebaluazioa, osasungintzako euskal komunitatea inplikatur

Olatz Perez-de-Viñaspre eta Maite Oronoz

IXA taldea (UPV/EHU). Manuel Lardizabal Pasealekua 1. 20018 Donostia

Laburpena

Osasun-arloko terminologia euskaraz lortzeko bidean, SNOMED CT nazioarteko terminologia euskaratzen ari gara. Horretarako KabiTerm eta MatxinMed sistemak aurkeztu ditugu artikulu honetan. KabiTermek termino habiaratuaren egitura baliatzen du ingelesezko termino konplexuak euskaratzeko. MatxinMed, Matxin Itzultzaile Automatikoaren osasun-zientzien domeinurako egokitzapena da. Horien ebaluaziorako Medbaluatoia kanpaina sortu dugu zeinak osasun-zientzien euskal komunitatea inplikatzeko duen. Jendearen inplikazioa oso positiboa izan da eta lortutako emaitzek KabiTermen nagusitasuna erakutsi dute.

Hitz gakoak: Osasun-zientzietako terminologia, SNOMED CT, Komunitate-ebaluazioa

Abstract

In order to obtain SNOMED CT terminology in Basque, we developed KabiTerm and MatxinMed systems. KabiTerm uses the nested structure of the English complex terms, to produce Basque equivalences. MatxinMed is the health domain adaptation of the Machine Translation system Matxin. We evaluate both systems by means of the Medbaluatoia campaign that implies the Basque health community. The implication of the people has been very positive and the results show the superiority of KabiTerm.

Keywords: Health terminology, SNOMED CT, Community-evaluation

1 Sarrera eta motibazioa

Osasun-arloan euskaraz aritu nahi badugu, ezinbestekoa da terminologia finkatzea. Osasun langilearen gana hurbildu eta “*tengo rotura del tendón del músculo esplenio del cuello*” esan ordez “lepoko esplenio giharraren tendoiak hautsia dut” esateko gai bagara, euskararen normalizazio prozesuan aurrerapauso izugarria egingo dugu.

SNOMED CT da ingelesez, gazteleraz eta beste hainbat hizkuntzatan, osasun alorreko terminologia jasotzen duen baliabide zabalduena, eta era berean, orain arte sortu den terminologia eleanitz osatutzat hartzen da. Nolabait esateko, hizkuntza eta sistema desberdinen arteko osasun-txostenen adierazpen eta interpretazio automatikoa eta anbiguotasunik gabea ahalbidetuko duen hiztegi normalizatua da, eta hiztegi-sarreraren arteko harremanak zehaztuta ditu. Osasun-txostenetan agertzen diren kontzeptuak, deskribapenak eta erlazioak barnebiltzen ditu. Jatorrizko hizkuntza duen ingelesezko 300.000 kontzeptu baino gehiago ditu definituak, baita horiek deskribatzeko 1.000.000 termino baino gehiago ere.

Osasun kontuetan euskaraz aritzeko, baliabide hau euskaraz behar dugu. SNOMED CTko terminoentzako euskarazko ordainak automatikoki lortzen ari gara azken urteotan (Perez-de Viñaspre eta Oronoz, 2015) eta euskarazko SNOMED CTren bertsio ofiziala ere diseinu eta garapen fasean da.

Euskarazko osasun-terminoak itzulpen tekniken bidez automatikoki lortzeko bidean, hainbat urratsetan egin dugu lan: dagoeneko sortutako baliabide lexikalak erabiliz lehenengo, hitzen morfosemantika edo lexema-osaera erabiliz geroago, eta terminoen luzera hitzetan handitzen joan den heinean, bi modu erabili ditugu: i) termino batzutan beste termino batzuk habiaratuak daudela baliatu dugu KabiTerm tresnan eta ii) erregetan oinarritutako Matxin itzultzaile automatikoa egokitu dugu, MatxinMed sortuaz. Termino konplexuen, hau da, hitz bat baino gehiago dituzten terminoen, euskarazko ordaina lortzea dute helburu KabiTermek eta MatxinMedek. Aipatu berri ditugun bi tresna horiek sortutako termino konplexuen ebaluazioa nola egin dugun azaltzea du helburu lan honek.

Sortutako euskarazko termino konplexuen ebaluazio automatikoa egiteko, ez dago hizteirik edo baliabiderik. Aukera bakarra ebaluazioa eskuz adituek (medikuek eta erizainek, besteak beste) egitea da, eta hori oso garestia da bai ekonomikoki baita denbora aldetik ere. Osasun-zientzietan euskaraz aritzen den komunitatea, har dezagun Osasungoa Euskalduntzeko Erakundea (OEE)¹ adibide moduan, oso komunitate aktiboa da. Hori aprobetxatu nahi izan dugu, termino konplexuen ebaluazioa komunitatean egiteko. Medbaluatoia izena eman diogu, osasun-zientzietako terminoen komunitatean egindako ebaluazioari.

Artikuluari aurkeztuko dugun lana atal honetan kokatu ondoren, 2 atalean arloko egoaren inguruan arituko gara. Jarraian dagoen 3 atalean KabiTerm eta MatxinMed sistemak deskribatuko ditugu, sistema hauen komunitate-ebaluazioa, Medbaluatoia, azalduko dugu eta emaitzak erakutsiko ditugu. Bukatzeko, 4 atalean lan honekin ateratako ondorioak laburbilduko ditugu eta 5 atalean etorkizunean egin nahiko genituzkeen lanak zerrendatuko ditugu.

2 Arloko egoera eta ikerketaren helburuak

Gaur egun SNOMED CT atzigarri dago AEBtako ingelesez, Britainia Handiko ingelesez, gazteleraz, danieraz eta suedieraz. Itzulpenak egiten ari dira, frantsesera, lituanierara eta beste hainbat hizkuntzetara². Itzulpenak egiteko teknika ezberdinak erabili dira: batzuk eskuz egin dira (danieraren kasua (Petersen, 2011)), itzulpen automatikoa eta eskuzko lana konbinatuaz (txinerarako adibidez (Zhu *et al.*, 2012)) edo itzulpen automatikoan laguntzeko sistema bat erabiliz (hau da frantsesaren kasua (Abdoune *et al.*, 2011)).

Itzulpen Automatikoaren (IA) helburua ordenagailuen bidez hizkuntza batetik bestera itzulpenak egitea da. Itzulpen Automatikoari aurre egiteko hurbilpenak bi multzo nagusitan sailkatuak izan dira: erregeletan oinarritutako itzulpen automatikoa eta corpusetan oinarritutako itzulpen automatikoa. Erregeletan oinarritutakoek, hizkuntzekiko dugun jakintza linguistikoa erabiltzen dute oinarri gisa eta corpusean oinarritutakoek aurretik egindako itzulpenak oinarri hartuta, hurbilpen enpirikoak garatzen dituzte.

Gaur egunean ingelesa-euskara parerako aurkitzen ditugun sistema gehienak erregeletan oinarritutako sistemak dira, hizkuntzen urruntasunak eta corpus-paralelo erraldioen faltak baldintzatuta ziurrenik. Erregeletan oinarritutako sistema bat **Matxin** (Mayor *et al.*, 2011) dugu IXA taldean garaturikoa eta kode irekikoa. Jatorrian gaztelania-euskara hizkuntza parerako garatu bazen ere, ingelesa-euskara parerako ere egokitua izan da (Aranberri *et al.*, 2015). **TectoMT** erregeletan oinarritutako itzulpen sistema bat da (Popel eta Žabokrtský, 2010) teknika estatistikoak baliatzen dituena itzulpen prozesuaren hainbat modulutan. Jatorrian ingelesetik txekierara itzultzeko sistema bada ere, berriki QTLeap³ proiektuaren baitan sistema ingelesa-euskara parerako egokitu da (Aranberri *et al.*, 2016b). Google enpresa erraldoiaren **Google Translate** dohaineko itzulpen zerbitzua 2001. urtean argitaratu zutenean Google Translate Systran deituriko erregeletan oinarritutako sistemari oinarritzen zen. Geroago, 2005. urtetik aurrera, sistema estatistikoak erabiltzen hasi ziren eta 2010etik aurrera, euskararentzako alpha bertsioa argitaratu zuten. Eusko Jaurlaritzaren bitartez **Lucy** erregeletan oinarritutako sistema komertziala (Gieselmann, 2008) garatua izan da baina tamalez informazio gutxi dago honi buruz. **EuSMT** IXA taldean garatutako sistema estatistiko bat da Mosesen (Koehn *et al.*, 2007) oinarritutakoa eta bi hurbilpen dituena ingelesa-euskara hizkuntza parerako: **EusSMT** oinarria eta segmentazioa gehitua duen **EuSMTs**. **SMatxinT** sistema EuSMTren bi hurbilpenak eta Matxin hibridatuz IXA taldean sortutako sistema da (Labaka *et al.*, 2014). Ikusi dugunez, hainbat sistema sortu dira ingelesa-euskara parearen itzulpenarako. Lan honetan Matxinen medikuntzarako egokitzapena azalduko dugu 3.2 azpiatalean.

Eskuzko ebaluazioa garestia dela esan dugu dagoeneko baita testu bildumen sortzea eta etiketatzea ere. Gaur egun, Internetaren hedapen izugarria dela-eta, pertsona asko dago prest honelako baliabideak sortzen laguntzeko. Adibidez, Bontcheva *et al.* (2013) lanean GATE Teamware tresna sortu zuten testu-anozio kolaboratiborako. Ebaluaziorako ere erabili izan da Internet komunitatea, orain arte aipatu ditugun hainbat ingelesa-euskara itzultzaileen ebaluaziorako, hain zuzen ere (Aranberri *et al.*, 2016a). Azken lan honetan du oinarria, 3.3 atalean azalduko dugun Medbaluatoiak.

Arloa aztertuta, ikerketa-lan honen helburu nagusia zehaztuko dugu. Helburu nagusia, automatikoki sortutako osasun-zientzietako terminologia eta bide batez, haiek sortzeko sistemak, ebaluatzea da,

¹<http://www.oegunea.eus/>

²Supporting Different Languages, IHTSDO. <http://www.ihtsdo.org/snomed-ct/snomed-ct0/different-languages>

³<http://qt leap.eu>

horretarako komunitatean egindako ebaluazioa baliatuz eta Medbaluatoia kanpaina abiatuz.

3 Ikerketaren muina

Ikerketa honetan, osasun-zientzien domeinuko terminologia euskaratzeko bi sistema garatu ditugu. Lehenengo, termino habiaratuetan oinarritzen den KabiTerm sistema da. Sistema horrek, termino konplexuen barruan agertzen diren beste terminoak oinarri harturik, euskaratze patroia batzuen bitartez euskarazko ordainak sortzen ditu. Bigarrenak, MatxinMedek, Matxin deritzon Itzultzaile Automatiko du oinarri (Mayor *et al.*, 2011) eta bere osasun-zientzien domeinurako egokitzapena da.

3.1 KabiTerm: Termino habiaratuen bidezko sorkuntza

KabiTerm sistema, termino habiaratuetan oinarritzen da transduktoreen bitartez termino konplexuentzat ordainak sortzeko. Lan honetan transduktoreak ingelesa-euskara hizkuntza parearentzat implementatu baditugu ere, modu errazean egokitu daitezke edozein hizkuntza parerako.

Idea nagusia zera du: termino konplexuen barruan maiz beste termino batzuk agertu ohi dira kabiari. Habiaratutako terminoa, barrukoa, euskaratuta badago, euskaratze-patroiak definitu ahal dira termino konplexu osoaren ordainak lortzeko.

Euskaratze-patroi horiek definitzeko, bi adituei ingelesezko terminoen lagin bat euskaratzeko eskatu diegu. Lagina aukeratzeko, SNOMED CTko terminoak eta haien egiturak aztertu ditugu. Egituren azterketa egiteko AnaMed deituriko analizatzaile bat sortu dugu. Jatorrian ingelesezko terminoak ditugunez, AnaMedek Stanforderko CoreNLP tresna (Manning *et al.*, 2014) hartzen du oinarrian, eponimoen zerrenda bat jasotzen du eta SNOMED CTko terminoen identifikazioa egiten du. Horrela, informazio linguistikoa, informazio terminologikoa gehitu diegu.

AnaMedi esker, SNOMED CTko terminoak termino habiaratuen egituraren arabera sailkatu ditugu. Hurrengo taulan (1 taula), egitura horien adibide batzuk ikus ditzakegu. Adibidez, *malignant neoplasm of renal calyx* termino konplexuan bi termino nagusi aurkitu ditugu *malignant neoplasm* nahasmenduen (*disorder*) hierarkiako bata eta *renal calyx* gorputz-egituren (*body structure*) hierarkiakoak besteak. Aipatu beharra dugu, habiaratutako termino gehiago ere aurkitzen dituela AnaMedek, hala nola *calyx* gorputz-egitura, *malignant* kalifikatzailea (*qualifier*) edo *neoplasm* nahasmendua.

Terminoak	Multzokatzea	Egitura
<i>structure of radial tuberosity</i>	<i>structure of radial_tuberosity</i>	structure+of+BODYSTRUCTURE
<i>Baelz's disease</i>	<i>Baelz's disease</i>	EPONYM+'s+DISORDER
<i>maximal gastric acid output measurement</i>	<i>maximal_gastric_acid_output_measurement</i>	PROCEDURE+PROCEDURE
<i>malignant neoplasm of renal calyx</i>	<i>malignant_neoplasm of renal_calyx</i>	DISORDER+of+BODYSTRUCTURE

1 Taula: AnaMeden bidez lortutako egiturak eta multzokatzeak.

Habiaratutako termino guztiekin sortu ditugu egiturak (ikusi 1 taulan azken zutabea), eta agerpen kopuruaren eta dependentsia kopuruaren arabera sailkatu ditugu. Hau da, egitura zenbat terminotan ematen den zenbatu dugu, eta termino hori zenbat terminotan agertzen den habiaratuta kontuan hartu dugu. Horrela, lehenengo faserako 41 egitura aukeratu ditugu eta bakoitzeko gutxienez ausaz erauzitako 3 adibide. Aditu bakoitzari 28 egitura eman dizkiogu, zeinetatik 14 egitura komunak diren (denera 100 bat adibide, 59 termino komunean dituztelarik). Komunean dituzten adibideen kasuan, adostasun altua neurtu dugu, eta desadostasun kasuekin, beraien arteko adostasuna bilatu dugu, irizpide berdina erabil ditzaten. Oinarriko irizpideak argi izanik, bigarren faserako, 25na egitura prestatu ditugu, aditu bakoitzak denera 100 adibide inguru dituelarik. Bi faseak elkaturik, 340 adibide inguru izan ditugu euskaratze-patroiak definitzeko eta horietatik abiatuta 53 euskaratze-patroi definitu ditugu.

Hurrengo irudian (1 irudia), KabiTermen funtzionamendua irudikatu dugu. Aurrena sarrera terminoan termino habiaratuak identifikatzen dira: *fracture* nahasmendu bat da eta *elbow* gorputz-egitura bat. Hori horrela izanik, euskaratze-patroia aplikatzen zaio (egitura baliokidea), eta horren arabera euskarazko egitura sortzen da, termino habiaratuen euskarazko ordainekin (“ukondo+aren haustura”). Bukatzeko, euskarazko morfologia arauak errespetatuz, deklinabideak gehitzen zaizkio horrela etiketatuta dauden hitzei, euskarazko ordaina sortuz (“ukondoaren haustura”).

Jatorri-terminoa: *fracture of elbow*

Sorkuntza-prozesua:

Identifikatutako egitura: GAIXOTASUN + of + GORPUTZ_EGITURA

Egitura baliokidea: GORPUTZ_EGITURA+aren + GAIXOTASUN

Euskaratze-patroiaren irteera: ukondo+aren haustura

Sortutako ordaina: ukondoaren haustura

1 Irudia: KabiTermen bidez lortutako ordainen adibidea.

3.2 MatxinMed: osasun-zientzietarako Itzultzaile Automatikoa

Atal honetan Matxin itzultzaile automatikoaren osasun-zientzien domeinurako egokitzapena azalduko dugu. Azken urteetan itzultzaile automatikoak domeinu zehatz batera egokitzeko joera zabaldu da, frogatu baita domeinu batera mugatzean, domeinu horretako itzulpenen kalitatea hobetzen dela. Erregeletan oinarritutako sistemak domeinu zehatz batera egokitzeko, hiztegiak eta gramatikak egokitu ohi diren arren, horretarako eskuzko lan handia egin behar denez, ohikoa da egokitzapena hiztegien aberastera mugatzea (Weijnitz *et al.*, 2004). Itzulpen Automatiko Estatistikoak aldiz, domeinu zehaztutara egokitzeko, domeinu horretako corpus-paraleloak behar dira.

Matxin aukeratu dugu osasun-zientzien domeinura egokitzeko sistema moduan, ingelesa-euskara parerako ditugun baliabideak kontuan izanik, erregeletan oinarritutako sistema bat baita bideragarriena. Euskararako osasun-arloko corpus-paraleloen urritasunak, Itzultzaile Automatiko Estatistiko bat egokitzea oraingoz oso zaila egiten du. Gainera, Matxin IXA taldean bertan garatua egonik, egokitzapenean lagundu ahal gaitun ikerlari-taldea hurbil dugu. Horretaz gain, hiztegiak modu errazean zabal daitezke, eta inbertsio txikiarekin, sistema domeinura egokitu dezakegu.

Lehenik eta behin, Matxini domeinua aukeratzeko funtzionalitatea gehitu diogu. Horretarako, hiztegiei ezaugarri berri bat gehitu diegu, eta lexikoaren transferentzia egiteko momentuan, Matxinek ordainak domeinuaren arabera aukeratzeko dituen. Bigarrenik, ordainen arteko ordena zehazteko Hizkuntza Eredu bat definitu dugu. Azkenik, termino konplexuak lantzeko erregelak sortu ditugu.

Lexikoa aukeratzeko baliabide nagusia hiztegia da, Matxinek formatu zehatz batean gordeta duena. Hiztegi horrek, transferentzia lexikoa egiteko beharrezko informazio guztia jasotzen du, hitzen itzulpenak eta hitzen kategoria gramatikala, besteak beste. Hiztegiari ezaugarri berri bat gehitu diogu domeinua adierazten duena. Matxinek hitz baten ordaina aukeratu behar duenean, hitz horren parekatze guztien artean zehaztutako domeinukoak aukeratu ditu. Aukera dago domeinu bat baino gehiago zehazteko, eta haien artean hierarkia bat sortzeko. Hitzaren ordaina bilatzean, lehen domeinuan ordainik ez balego, bigarren domeinukoak bilatuko ditu, gero hirugarrenekoak e.a. Adibidez, jatorrizko hiztegiaren *colon* ingelesezko terminoa “bi puntu” gisa euskaratzen den bitartean, medikuntzaren domeinuan “kolon” da ordain egokia. Transferentzia lexikalean egindako egokitzapen honi esker, 2 adibidean ikus dezakegun bezala, MatxinMed “*He has colon cancer*” bezalako esaldiak ondo itzultzeko gai da, transferentzia lexikala egiterakoan parekatze hobea aukeratzeko baliabideak dituelako.

Jatorrizko esaldia	<i>He has colon cancer.</i>
Matxinen itzulpena	Hark bi puntu minbizia dauka.
MatxinMeden itzulpena	Hark kolon minbizia dauka.

2 Taula: Matxin eta MatxinMeden itzulpenen adibidea.

Bestalde, ingeleseko hainbat terminok, euskarazko ordain bat baino gehiago dute. Matxinen une honetako inplementazioak, ordainen gaineko desanbiguaziorako tresnak nahikoa garatu gabe ditu, eta hortaz ordainen arteko ordena bat ezartzea ezinbestekoa da. Oraingoz, lehenengo ordaina soilik erabiliko bada ere, interesgarria iruditzen zaigu ordain guztiak hiztegiaren gehitzea, etorkizunean desanbiguatzeak aukera izateko.

Ordainen arteko ordena zehazteko, Hizkuntza Eredu bat osatu dugu. Hizkuntza Ereduak Itzultzaile Automatiko Estatistikoetan erabiltzen diren eredu probabilistikoak dira, helburu-hizkuntzarako sortzen direnak. Eredu horiek, testu-kate batek helburu-hizkuntzako baliozko esaldia izateko duen probabilitatea adierazten dute, eta helburu nagusia itzulpena helburu-hizkuntzara ahalik eta antzekoena dela

bermatzea da. Hizkuntza eredu bat sortzeko, helburu-hizkuntzaren corpus bat beharrezkoa da. Gure kasuan, euskarazko corpora behar dugu eta gainera, osasun-zientzien domeinukoa. Domeinuko corpora osatzeko, Udako Euskal Unibertsitatearen osasun-zientzietako liburuak, Elhuyar Fundazioaren itzulpen-memoriak, Euskal Herriko Unibertsitatearen medikuntzako ikasleen apunteak, Osasungoa Euskalduntzeko Erakundearen osasun biltzarren txostenak eta Osakidetzaaren zabalkunderako txostenak elkartu ditugu, eta horrela 3.533.732 hitzetako corpora osatu dugu.

Azkenik, Matxinek Hitz Anitzeko Unitate Lexikalak (HAULak) ezagutzeko duen modulua erabili dugu termino konplexuen identifikazioa egiteko. Identifikaziorako sortutako erregelekin, MatxinMed HAUL berriak, gure kasuan termino konplexuak, identifikatzeko gai da. Erregela horiei esker, 3 taulan erakusten dugun bezala, MatxinMedek hobekuntza nabaria erakusten du termino konplexuak dituzten itzulpenetan.

Jatorrizko esaldia	<i>He has Osler syndrome.</i>
Matxinen itzulpena	Hark sindrome Osler dauka.
MatxinMeden itzulpena	Hark Osler-en sindromea dauka.

3 Taula: Matxin eta MatxinMeden itzulpenen adibidea.

3.3 Medbaluatoia: komunitatean oinarritutako ebaluazioa

KabiTerm eta MatxinMed sistemak ebaluatzeko, osasun-zientzien euskal komunitatea inplikatzeko duen ebaluazio bati heldu diogu: Ebaluatoiarenekin (Aranberri *et al.*, 2016a) moldaera moduan Medbaluatoia deritzon kanpaina diseinatu eta martxan jarri dugu. Ebaluatoia, 2015ean Itzultzaile Automatikoak sailkatzeko erabili zen euskal hitzunen komunitatea inplikatzeko, eta oso emaitza interesgarriak jasoaz.

Itzultzaile Automatikoaren ebaluazioan ohikoena neurri automatikoak ateratzea bada ere, BLEU (Papi-*neni et al.*, 2002) izanik zabalduena, pertsonen ebaluazioa ekidinezina da tresna pertsonen erabiltzea nahi baldin badugu. Hori horrela izanik, IA ebaluazio gehienetan ebaluatzaile talde txiki batek parte hartzen du, nagusiki ikerketa taldeko bertako ikerlariak izanik eta oso lagin txikia ebaluatuz. Ebaluatoiak aldiz, komunitatean oinarritutako ebaluazioa proposatzen du. Euskaldunon komunitatea kontzientziatua eta dinamikoa izanik, inbertsio handiegirik egin gabe, konfiantza maila nahikoa duten ebaluazioak jaso ahal dira modu azkarrean. Jakina den bezala, hizkuntza gutxituen komunitateek, haien hizkuntzek biziraun dezaten lagungarriak izan daitezkeen ekitaldietan zintzoki parte-hartzeko prestutasuna erakutsi ohi dute.

Guk Medbaluatoian Ebaluatoian garatutako metodologia erabili dugu terminoen ordain hautagaiak sailkatzeko. Horrela, esaldiak beharrezko erabiltzaileek termino konplexuak (2 eta 8 token artekoak) ebaluatu dituzte. Erabiltzaileari jatorrizko ingelesezko terminoa aurkezten zaio, honi emandako bi euskarazko ordain eta erabiltzailearen eginbeharrekoa da, bi aukeren artean zein den egokiagoa erabakitzea. Bi terminoak izan daitezke zuzenak, ala biak okerrak, beraz, Medbaluatoian hori adierazteko funtzionalitatea gehitu diogu jatorrizko Ebaluatoia sistemari.

Medbaluatoia KabiTerm eta MatxinMed ebaluatzeko erabili dugu, hortaz bi ebaluaziorako multzo lortu ditugu. Alde batetik, KabiTerm ebaluatzeko, hiru sistema konparatu ditugu: Google Translate oinarri-lerroko sistema gisa, KabiTerm sistema eta MatxinMed-1 (KabiTermekin sortutako terminorik gabe). Multzo honi KabiTermen multzoa deituko diogu. Bestetik, MatxinMed bera ebaluatuko dugu. Horretarako, oinarri-lerroko sistema berdina erabiliko dugu, Google Translate, eta MatxinMeden bi aldaera erabiliko ditugu. Lehenengoa, MatxinMed-1, KabiTermen ebaluazioan erabiliko dugun berdina, eta bigarrenagoari, MatxinMed-2, KabiTerm sistemak sortutako ordainak ere integratu dizkiogu. Multzo horri MatxinMeden multzoa deritzogu. Hurrengo taulan (4 taula), bi ebaluazioetan parte hartuko duten sistemak erakusten ditugu.

	KabiTerm	Google	MatxinMed-1	MatxinMed-2
KabiTermen multzoa	✓	✓	✓	
MatxinMeden multzoa		✓	✓	✓

4 Taula: Medbaluatoiarenekin bidez ebaluatutako sistemak.

Medbaluatoian 1.000 jatorri-termino ebaluatu ditugu. Kontuan izanik bi ebaluazio egingo ditugula, ebaluazio bakoitzerako 500 termino aukeratu ditugu SNOMED CTren termino konplexuetatik, eta multzo horiek token kopuruaren eta hierarkiaren arabera estratifikatu ditugu.

KabiTermen ebaluazio multzorako, ausaz aukeratu ditugu 500 termino konplexu, beti ere, KabiTerm jatorri-terminoen ordainak sortzeko gai bada. MatxinMeden multzoa sortzeko aldiz, Kabitermek itzuli ezin dituen jatorri terminoetatik 500 jaso ditugu. Izan ere, MatxinMed soilik KabiTermek euskaratzen ez dituen terminoak euskaratzeko erabiliko dugu, eta horrela egin dugu ebaluaziorako lagina lortzeko ere.

KabiTermen ebaluaziorako terminoen gainsorkuntza kontrolatzeko beharra izan dugu. Izan ere, gainerako sistemek ordain bakarra ematen duten bitartean, KabiTermek hainbat ordain emateko joera dauka, eta sistemak ez leudeke baldintza berdinetan lehiatzen. Lehia orekatzeko gure osasun-zientzien euskarazko Hizkuntza Eredua erabili dugu. Modu horretan, KabiTermek proposaturiko ordain guztien artean, Hizkuntza Ereduan probabilitate altuenekoa aukeratu dugu. Tamalez, horrek ez du ziurtatzen ordain onena izango denik baina metodo automatikoen artean ezagutzen dugun metodo ziurgarriena da.

Denera 1.000 termino ebaluatu ditugu, eta termino bakoitzarekin hiru sistema-bikote (hiru sistema). Horrela, 3.000 ebaluazio behar izan ditugu. Subjektibotasunari aurre egiteko, ebaluazio bakoitza 5 erabiltzailek egitea erabaki dugu. Horrela, denera 15.000 ebaluazio behar izan ditugu Medbaluatoia kanpaina bukatzeko. Gainera, erabiltzailearen arreta neurtzeko, kontrolerako ebaluazioak gehitu ditugu. Kontrolerako ebaluazio horiek, eskuz emandako ordain zuzenez, eta oso nabariak diren ordain okerrez osatuta daude. Kontrolerako ebaluazioetatik herena gaizki eginez gero, parte-hartzailea kanporatu egiten da.

Medbaluatoiarene zabalkunderako, Medikuntzako Fakultateko euskara taldeko klase guztietatik pasa ginen kanpainaren hasiera egunean. Gainera, mezu elektronikoko bat zabaldu genuen, profesionalen artean parte-hartzea lortzeko. Zabalkundean gakoak izan dira OEE eta UPV/EHUko Euskara Errektoreordetza.

3.4 Emaitzak

Medbaluatoiarene kanpaina oso arrakastatsua izan dela azpimarratu beharrea gaude. Hiru astetarako aurreikusitako lana, aste bakarrean bukatu dute parte-hartzaileek. Denera 218 parte-hartzaile izan ditugu kanpainan, eta euretako 13 baztertuak izan dira kontrol-ebaluazioetan kale eginagatik. Ebaluatzaileen artean Cohenen $kappa$ (Cohen, 1960) neurtu dugu, beste IA ebaluazio batzuetan egin izan den bezala (Aranberri *et al.*, 2016a). Lortu ditugun $kappa$ balioak onargarriak dira: 0,30 eta 0,48 artekoak.

Medbaluatoian zehar, parte-hartzaileei jatorrizko terminoa, eta bi euskarazko ordain erakutsi dizkiegu, eta bien arteko onena aukeratu behar izan dute, horrela, ebaluazio bat eginez. Ez dugu onena aukeratzeko irizpiderik eman, bakoitzaren irizpideen arabera onena iruditzen zaiena jaso nahi dugulako.

Jatorrizko termino bakoitzarentzat eta sistema pare bakoitzeko bost ebaluazio jaso nahi izan ditugu. Hala ere, web-aplikazioaren konfigurazioa dela-eta, batzuentzat 7 ebaluazio ere lortu ditugu. Ebaluazio gehigarri horiek ez dira akasduak eta erabiltzea erabaki dugu. Hurrengo taulan (5 taula), sistema bakoitzarentzat lortutako ebaluazio kopuruak erakusten ditugu.

	Google KabiTerm	Google MatxinMed-1	KabiTerm MatxinMed-1	Google MatxinMed-2	Google MatxinMed-1	MatxinMed-2 MatxinMed-1
Ebaluazioak denera	2.529	2.523	2.527	2.540	2.535	2.535

5 Taula: Sistema pare bakoitzak denera lortutako ebaluazioak.

Ebaluazioetatik sistema onena aukeratzeko honako estrategia jarraitu dugu: bi sistemen arteko bozka diferentzia bi baino handiagoa bada, zalantzarik gabeko irabazlea da (taulan “X.sistema++” moduan kodetu dugu). Sistemen arteko diferentzia 1 edo 2 bada, sistema irabazlea dela ere deritzogu (taulan “X.sistema+” bezala kodetu dugu). Bi sistemek bozka kopuru berdina lortzen badute, orduan berdintza dagoela deritzogu. Adibidez, termino baten ebaluazioan, KabiTermek 4 bozka jaso baditu, eta Google-ek 1, bien arteko ezberdintasuna 3 denez, KabiTermen alde, “2.sistema++”-n bozka bat gehituko da.

Emaitzetan ikus dezakegunez, KabiTermen multzoan (lehenengo 3 zutabeak), KabiTerm sistema izan da emaitza onenak lortu dituen alde handiarekin. Bereziki handia izan da Google-en itzultzailearekin alderatuz, % 86,6 kasutan izan baita hobea (% 73,2 gehi % 13,4). MatxinMed-1ekin konparazioan alde txikiagoa bada ere, nabarmenki jaso ditu emaitza hobek (% 46,4 kasutan nabarmen irabazi du, eta MatxinMed-1ek aldiz % 21,0 kasutan).

MatxinMeden multzoari dagokionean, emaitzak horren nabarmenak ez badira ere, Matxinen bi bertsiok Google Translatek baino emaitza hobek lortu dituzte. Izan ere, irabazten duten portzentaiak

	Google - KabiTerm	Google - MatxinMed-1	KabiTerm - MatxinMed-1	Google - MatxinMed-2	Google - MatxinMed-1	MatxinMed-2 - MatxinMed-1
1.sistema++	6,8 (34)	13,2 (66)	46,4(232)	19,4 (97)	21,6(108)	6,6 (33)
1.sistema+ berdin	3,2 (16)	9,2 (46)	14,4 (72)	12,0 (60)	16,4 (82)	13,0 (65)
2.sistema+	3,2 (16)	7,8 (39)	5,6 (28)	7,0 (35)	8,8 (44)	62,4(312)
2.sistema++	13,4 (67)	15,2 (76)	12,4 (62)	22,0(110)	19,4 (97)	14,6 (73)
	73,2(366)	54,4(272)	21,0(105)	39,8(199)	34,0(170)	3,6 (18)

6 Taula: Medbaluatoia kanpainaren sistemen ebaluazioaren emaitzak.

gehituta, % 61,8 kasutan MatxinMed-2k irabazi du (% 22,0 gehi % 39,8), eta Google-ek aldiz % 31,4 kasutan (% 19,4 gehi % 12,0). Dena dela, MatxinMeden bi bertsioak konparatzen ditugunean, emaitzek ez dute sistema baten edo beste baten alde egiten. Izan ere, kasuen % 62,4tan sistemek kalitate berdina erakutsi dute, % 19,6tan MatxinMed-2k emaitza hobea (portzentaia gehituta) eta % 18,2tan MatxinMed-1ek hobea. Ezin dugu ahaztu, bi sistemen alde KabiTermek sortutako terminoetan datzala. Horrela, MatxinMed-2k KabiTermek euskaratutako termino habiaratuak aurkitzen dituenen bakarrak sortuko ditu ordain ezberdinak, eta esku artean izan dugun laginean kasu gutxitan gertatu da hori. Hala ere, ezin dugu kasu horietarako zein sistema den hobea ondorioztatu, emaitza oso antzekoak lortu baitituzte bi sistemek. Azpimarratzekoa da, KabiTermen multzoan agertzen diren terminoen egitura mugatua dela, eta egitura horretan MatxinMedek emaitza hobea lortzen dituela Google-ekin alderatuz.

Termino zuzena aukeratzeko funtzionalitateak ez du arrakastarik izan. Uste dugu, ordain “ona” eta ordain “hobea” kontzeptuen arteko bereizketa ez dela garbi geratu, parte-hartzaileen arabera oso emaitza ezberdinak lortu baititugu lagin berarekin. Emaitza esanguratsuak ez direnez, ez ditugu erakutsiko.

4 Ondorioak

Artikulu honetan osasun-zientzien domeinuko terminologia euskaratzeko bi sistema aurkeztu ditugu eta horien ebaluazioa egin dugu euskal komunitatea inplikatur. Lehenengoak, KabiTermek, termino konplexuetan dauden termino habiaratuak baliatzen ditu euskaratzea egiteko. Bigarrena aldiz, MatxinMed, erregeletan oinarritutako itzultzaile baten osasun-zientzien domeinura egindako egokitzapen bat da.

Bi teknikak Medbaluatoia deritzon kanpainan ebaluatu ditugu, eta artearen egoerako Google Translate tresnarekin konparatu ditugu. Emaitza oso positiboak izan dira, bai parte-hartzeari dagokionean, bai eta gure sistemen kalitatearen inguruan Google-ekin alderatzean. Bereziki azpimagarriak dira KabiTermek lortutako emaitzak, gainerako sistemei nabarmenki irabazten baitie.

5 Etorkizunerako planteatzen den norabidea

Etorkizunerako, SNOMED CT euskaratzeko erabili ditugun patroiak beste terminologia batzuk euskaratzeko egokitu nahi ditugu. Izan ere, Osakidetza Euskara Zerbitzuak interes berezia dauka “Gaixotasunen Nazioarteko Sailkapenaren 10.en bertsioa” (GNS-10) euskaratzeko, eta proiektu bat abiatu dugu gure tresna egokitzeko. Patroi berriak definitu beharko ditugu horretarako, GNS-10k dituen terminoen egitura oso berezia delako, sailkapenera bideratuta baitago. Adibidez “*Salmonella infection, unspecified*” moduko deskribapenak oso maiz topa ditzakegu bertan.

Horretaz gain, Medbaluatoia SNOMED CTren terminologia baliozkotzeko egokitu nahi dugu. Izan ere, kanpainan zehar hainbat parte-hartzailek zuzenketak egiteko aukera eskatu digute, eta gehiagorako gogoarekin geratu dira. Jendearen prestutasuna kontuan izanik, euskarazko SNOMED CT osoaren zuzenketa egin nahiko genuke horrelako kanpaina batekin. Horrela, jende gehiagoren oniritziarekin lortuko genituzke SNOMED CTren sinonimoak. Kontzeptu baten deskribapenen artean hobetsia zein den aukeratzeko ataza aditu talde txiki baten esku utziko genuke, terminologoak eta osasun-langileak barnebilduz.

Erreferentziak

ABDOUNE, HOCINE, TAYEB MERABTI, STÉFAN J. DARMONI, eta MICHEL JOUBERT. 2011. Assisting the Translation of the CORE Subset of SNOMED CT Into French. In *Studies in Health Technology and Informatics*, volume 169, 819–823.

- ARANBERRI, NORA. 2016. Ebaluatoia: crowd evaluation of English-Basque machine translation.
- , GORKA LABAKA, A DIAZ DE ILARRAZA, eta KEPA SARASOLA. 2015. Exploiting portability to build an RBMT prototype for a new source language. In *Proceedings of EAMT*.
- , GORKA LABAKA, ARANTZA DÍAZ DE ILARRAZA, eta KEPA SARASOLA. 2016a. Ebaluatoia: crowd evaluation for English-Basque machine translation. *Language Resources and Evaluation* 1–32.
- , GORKA LABAKA INTXAUSPE, ONEKA JAUREGI, ARANTZA DÍAZ DE ILARRAZA SÁNCHEZ, IÑAKI ALEGRÍA LOINAZ, eta ENEKO AGIRRE BENGEOA. 2016b. Tectogrammar-based machine translation for English-Spanish and English-Basque. *Procesamiento del Lenguaje Natural* 73–80.
- BONTCHEVA, KALINA, HAMISH CUNNINGHAM, IAN ROBERTS, ANGUS ROBERTS, VALENTIN TABLAN, NIRAJ ASWANI, eta GENEVIEVE GORRELL. 2013. Gate teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation* 47.1007–1029.
- COHEN, JACOB. 1960. "A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement* 20.
- GIESELMANN, P. 2008. Architecture of the Lucy translation system. *Second machine translation marathon, Wandlitz, Berlin* 28.
- KOEHN, PHILIPP, HIEU HOANG, ALEXANDRA BIRCH, CHRIS CALLISON-BURCH, MARCELLO FEDERICO, NICOLA BERTOLDI, BROOKE COWAN, WADE SHEN, CHRISTINE MORAN, RICHARD ZENS, eta OTHERS. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 177–180. Association for Computational Linguistics.
- LABAKA, GORKA, CRISTINA ESPAÑA-BONET, LLUÍS MÀRQUEZ, eta KEPA SARASOLA. 2014. A Hybrid Machine Translation Architecture Guided by Syntax. *Machine translation* 28.91–125.
- MANNING, CHRISTOPHER D., MIHAI SURDEANU, JOHN BAUER, JENNY FINKEL, STEVEN J. BETHARD, eta DAVID MCCLOSKEY. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*, 55–60.
- MAYOR, AINGERU, INAKI ALEGRIA, ARANTZA DIAZ DE ILARRAZA, GORKA LABAKA, MIKEL LERSUNDI, eta KEPA SARASOLA. 2011. Matxin, an Open-source Rule-based Machine Translation System for Basque. *Machine Translation* 25.53–82. 10.1007/s10590-011-9092-y.
- PAPINENI, KISHORE, SALIM ROUKOS, TODD WARD, eta WEI-JING ZHU. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on ACL*, 311–318. Association for Computational Linguistics.
- PEREZ-DE VIÑASPRE, OLATZ, eta MAITE OROÑOZ. 2015. SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. *BMC medical informatics and decision making* 15.S5.
- PETERSEN, PALLE G., 2011. How to Manage the Translation of a Terminology. Presentation at the IHTSDO October 2011 Conference and Showcase.
- POPEL, MARTIN, eta ZDENĚK ŽABOKRTSKÝ. 2010. TectoMT: modular NLP framework. In *International Conference on Natural Language Processing*, 293–304. Springer.
- WEIJNITZ, PER, EVA FÖRSBOM, EBBA GUSTAVII, EVA PETTERSSON, eta JÖRG TIEDEMANN. 2004. MT Goes Farming: Comparing Two Machine Translation Approaches on a New Domain. In *LREC*.
- ZHU, YANHUI, HUITING PAN, LEI ZHOU, WEI ZHAO, ANA CHEN, ULRICH ANDERSEN, SHUXIANG PAN, LIXIN TIAN, eta JIANBO LEI. 2012. Translation and Localization of SNOMED CT in China: A pilot study. *Artificial Intelligence in Medicine* 54.147–149.

6 Eskerrak eta oharrak

Eskerrak lan honetan lagundu diguten iXakide guztiei, bereziki Nora Aranberri eta Gorka Labakari, MatxinMed eta Medbaluatoia ez baitziren posible izango beraien aholku eta laguntzarik gabe.

Eskerrak gainera, Medbaluatoia kanpainen parte-hartu duten osasun-zientzietako ikasle zein langileei. Eta nola ez Felix Zubiari eta Karlos Iburgureni, KabiTermen patroietan horrenbeste laguntzeagatik.

Artikulu hau idazte fasean dagoen tesiaren kapitulu baten laburpena da. Tesi hori, Eusko Jaurlaritzak emandako beka bati esker garatu da: BFI-2011-389 doktoretza-aurreko programakoa hain zuzen ere.