Wikipedia eta itzulpen automatikoa: “harri batez bizpalau xori”

Iñaki Alegria*, Unai Cabezón*, Unai Fernandez de Betoño**, Gorka Labaka*,
Aingeru Mayor*, Kepa Sarasola*, Arkaitz Zubiaga**

*Ixa Taldea, https://ixa.si.ehu.es
**Euskal Wikipedia, http://eu.wikipedia.org

Laburpena

Artikulu honetan elkarlanean egindako proiektu bat aurkezten dugu. Boluntario talde bat bildu dugu espainierazko Wikipediako hainbat artikulu euskarara itzultzeko, baina boluntarioen lana errazteko, Matxin itzultzaile automatikoa erabili dugu aurre-itzulpenak sortzeko, eta horrela boluntarioen lana errore eta akatsak dituzten itzulpen automatiko horiek aztertu eta zuzentzea izan da. Lan honekin, batetik, Euskal Wikipedia aberastu dugu, 50.000 hitz berri gehituz. Beste alde batetik, sistema automatikoaren itzulpenak eta posteditatutako bertsio zuzenduekin corpus bat sortu dugu. Corpus hori erabili dugu posteditore estatistiko bat sortzeko, Matxin itzulpen automatikoko sistemaren irteeraren doitasuna %10ean hobetuz.

Abstract

In this paper we define a collaboration framework that was tested with editors of Basque Wikipedia. Their post-editing of Computer Science articles has been used to improve the output of a Spanish to Basque MT system called Matxin. For the collaboration between editors and researchers, we selected a set of 100 articles from the Spanish Wikipedia. These articles would then be used as the source texts to be translated into Basque using the MT engine. A group of volunteers from Basque Wikipedia reviewed and corrected the raw MT translations. This collaboration ultimately produced two main benefits: (i) the change logs that would potentially help improve the MT engine by using an automated statistical post-editing system, and (ii) the growth of Basque Wikipedia. The results show that this process can improve the accuracy of an Rule Based Machine Translation system in nearly 10% benefiting from the post-edition of 50,000 words in the Computer Science domain.

1 Sarrera

Artikulu honetan elkarlanaren onuraz arituko gara. Lan asko egin behar denean eta baliabideak urriak direnean, indarrak biltzea izaten da gakoa. Eta lan bera erabiliz behar ezberdinak asetzen baditugu, helburu desberdinak betetzen baditugu, askoz hobe.

Gurean bi behar, bi helburu, izan ditugu eskuartean: itzulpen-automatikoaren garapena [1] batetik eta Euskal Wikipediaren[1] aberasketa bestetik.

Euskal Herriko Unibertsitateko IXA taldean[2] Hizkuntzaren Prozesamenduaren arloan ikerketa burutzen dugu [2, 3], besteren artean itzulpengintza automatikoan. Euskararekin lan egiten duen eta publikoki erabilgarria izan zen lehenengo itzulpen automatikoko sistema, Matxin, eraiki ondoren, erregeletan oinarritutako sistema horren irteera hobetzeko posteditore estatistiko bat eraikitzea

1 http://eu.wikipedia.org
2 http://ixa.si.ehu.es

erabaki genuen. Baina horretarako automatikoki itzulitako esaldiak eta horien eskuzko postedizioak bilduko lituzkeen corpus bat beharrezkoa genuen.

Lankidetzaz sortutako Wikipedia entziklopedia eleanitza euskaraz ere badugu, baita osasun osoz eduki ere [4, 5]. Baina euskara bezalako baliabide urrietako hizkuntzetako Wikipediak aberastea ez da erraza, sarrera berriak idatziko dituzten boluntario kopurua txikia delako, eta beraz, boluntario bakoitzak eginbeharreko esfortzua handia delako. Editore kopuru txikia duten hizkuntzek ezin dute lehiatu ingelesa edo espainiera bezalako hizkuntzetan ematen den Wikipediaren hazkuntzarekin. Baina alde positibotik ikusita, Wikipedia txikietako editoreak profitatu ahal izango dira hizkuntza handietan sortutako eduki kopuru handiez, eduki horiek haien hizkuntzara itzuliz, bide hau eduki berriak sortzea baino askoz ere merkeagoa izanik [6].

Itzulpen automatikoa (IA) laguntza ederra izan daiteke itzulpen-prozesu hori errazte aldera. Gaur egungo sistema automatikoek ematen dituzten itzulpenak okerrak eta akatsez betetakoak badira ere, gure hipotesiaren arabera itzultzaile ez profesionalen kasuan sistema automatikoaren irteera zuzentzea hutsetik itzultzea baino errazagoa da.

Hona hemen gure ideia: Euskal Wikipediako sarrera berriak idaztea espainierazko Wikipediako artikuluen itzulpen automatikoa zuzenduz. Horrela, alde batetik, editoreen lana erraztu dezakegu Euskal Wikipedia aberasteko eta, bestetik, itzulpen automatikoaren irteera okerrak eta eskuz zuzendutako postedizioak bildu ditzakegu corpus batean, eta postediziozko corpus horren bitartez posteditore estatistikoa eraiki gero [7, 8]. Esaera zaharrak dioen bezala: “*harri batez bizpalau xori*”.

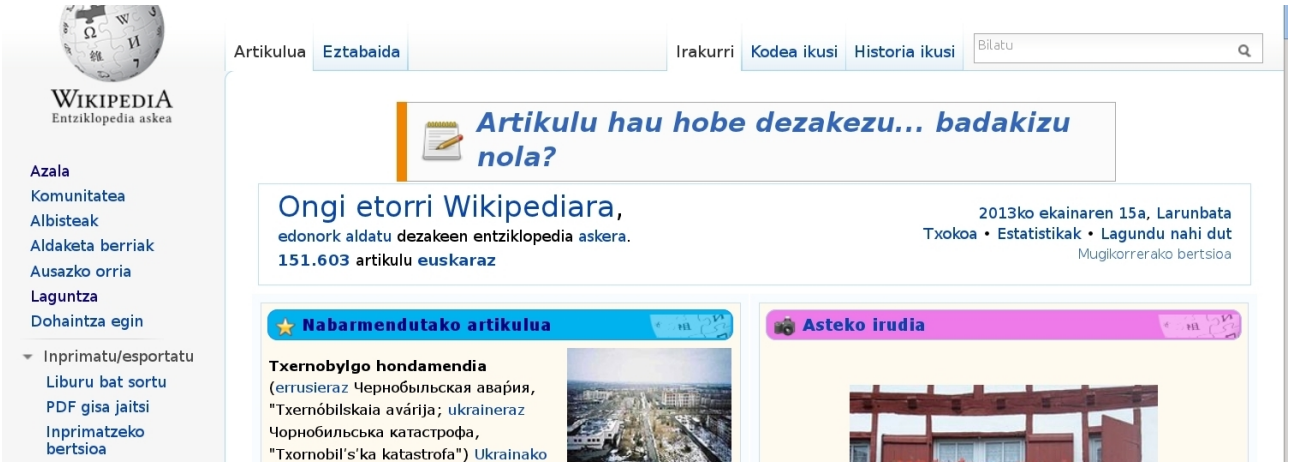
2 Aurkezpena

2.1 Wikipedia: Entziklopedia askea

Wikipedia Interneten argitaratutako eduki askeko entziklopedia eleanitza da. Lankidetzaz editatua, mundu osoko boluntarioek idazten dute. Gaur egun, Interneten dagoen kontsultarako tresnarik handiena eta zabalena da, zalantzarik gabe, eta baita Web 2.0 eta haren parte hartzearen filosofiaren eredu argienetakoa ere. Etengabe eguneratzen, handitzen eta zuzentzen da. Wikipediak ezagutza librea sustatzen du. *Libre* izatea ez da doakoa delako bakarrik, baita bildutako jakintzaren berrerabiltzea baimentzen duelako ere; jakintzaren eboluzioa sustatzeko oinarri sendoa izan da hori.

2001. urtean sortu zen eta hamar urtebete eta gutxira, 2012ko azaroaren lehenean, 285 hizkuntzatako edizioak zituen, guztira 23 milioi artikulu, horietatik lau milioitik gora ingelesez. Milaka boluntario, profesional gutxi batzuk baino etekin hobea lortzen ari dira, kantitatea kalitate bihurtuz. Euskal Wikipedia, 150.000 artikulurekin, osasun onean dagoela esan genezake. Wikipediako hizkuntzen zerrendan euskara 35. posizioan agertzen da. Hori bai, tamalez, bere tamaina oraindik txikia da Hizkuntzaren Prozesamenduko aplikazio aurreratuetan corpus gisa erabilia izateko.

Esan beharra dago euskaraz sekula ez dela izan Wikipedia baino entziklopedia zabalagorik. Euskal Wikipediako komunitatea lanean gogoz dabil, bai euskarazko artikulu berriak sortzen, bai editore berriak bilatzen. Eta lan honetako helburu bat horixe izan da: Euskal Wikipedia aberastea.



1. irudia: *Euskal Wikipedia*

2.2 Matxin: Itzultzaile automatikoa

Matxin [9, 10] erabilera publikoko euskararako lehenengo itzulpen automatikoko sistema izan zen, Euskal Herriko Unibertsitateko IXA Taldearen eta Elhuyarren artean garatua. Software librea da, GNU GPL lizentziapean argitaratutakoa.

Matxin erregeletan oinarritutako itzulpen automatikoko sistema bat da. Hiru fasetan egiten du itzulpena: analisia (espainierarako Freeling [11] software libreko paketea erabiltzen du), transferentzia (lexikala eta estrukturala) eta sorkuntza (sintaktikoa eta morfologikoa).

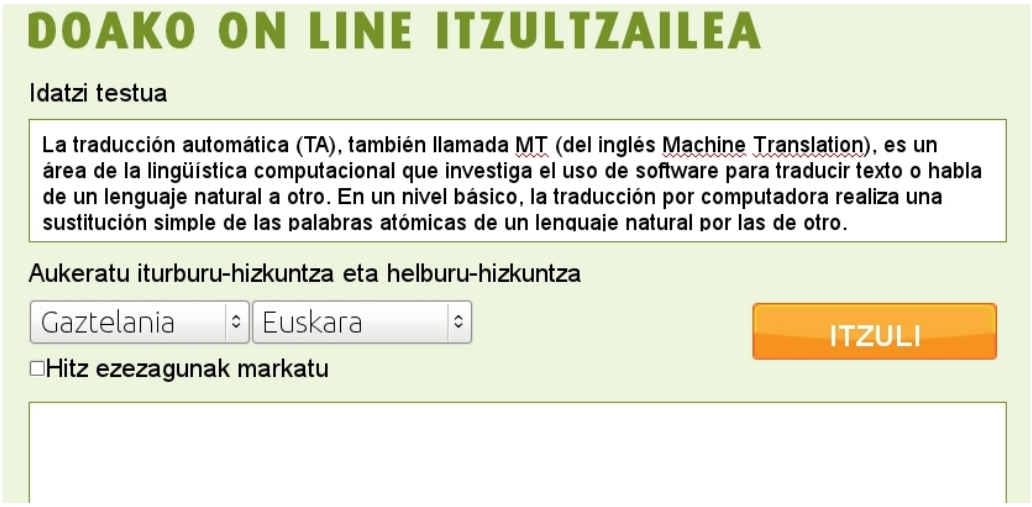
Ezberdintasun lexikal eta sintaktiko handiak aurkezten dituzten hizkuntzen arteko itzulpenak burutzeko eta abiapuntu-hizkuntzatik independentea izateko diseinatuta dago. Matxin 2.0 prototipoak espainieratik euskarara itzultzen du eta erabilera orokorrekoa da. Sarean erabil daiteke bere webgunean[3] eta OpenTrad proiektuko webgunean[4] proba daiteke, eta Sourceforge-n kode irekiko software libre gisa banatzen da[5].

Gaur egun espainiera-euskara sistema hobetzen jarraitzeaz gain, ingelesetik euskarara itzultzen duen prototipoa garatzen ari da IXA taldea. Gainera, estatistikan oinarritutako sistemetan eta sistema hibridoetan ere ikerketa burutzen ari da [12].

IAko sistemak hobetzeko modu bat postedizio-lanaz baliatzea da. Postedizioa IAko sistemak ematen dituen itzulpenen zuzenketa da. Postedizioa gizakiek egin ohi dute, sistema automatikoaren kalitatea ona denean itzulpenak egiteko lana murriztu daiteke horrela.

Posteditore automatikoak ere eraiki daitezke. Horretarako itzultzaile automatikoaren emaitzak eta emaitza horien eskuz posteditatutako irteerak erabiltzen dira, horiekin ikasketa automatikoa aplikatu ahal izateko [13]. Hauxe da, hain zuzen, elkarlan hau abiarazi zuen asmoa: Matxin erregeletan oinarritutako sistemaren irteera hobetuko lukeen posteditore estatistiko bat eraikitzea.

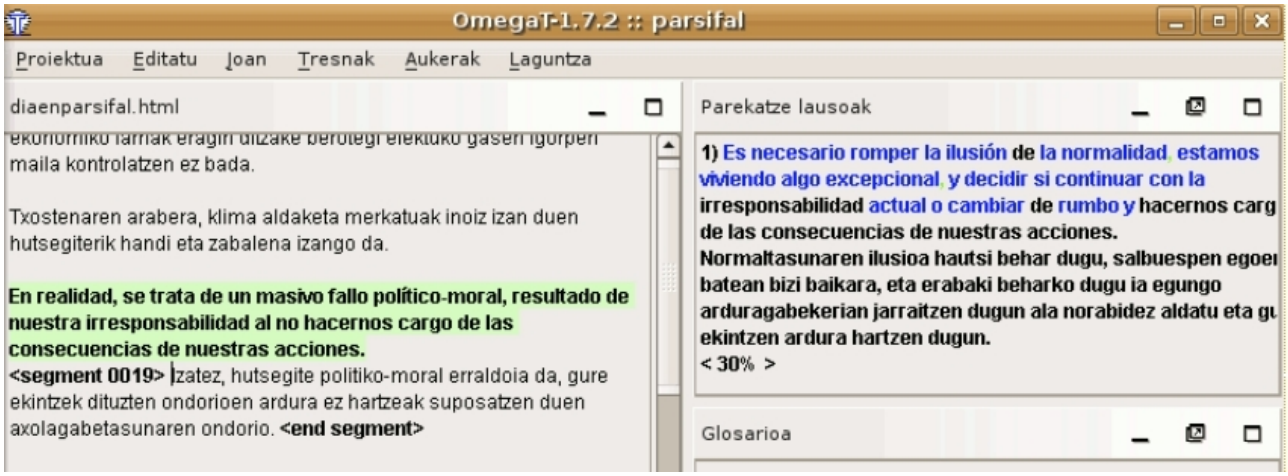
3 http://matxin.elhuyar.org/
4 http://www.opentrad.com/
5 http://matxin.sourceforge.net/



2. irudia: *Matxin itzulpen automatikoko sistema on-line*

2.3 OmegaT: Ordenagailuz lagundutako itzulpen-tresna

OmegaT[6] aplikazioa ordenagailuz lagundutako itzulpen-tresna bat da, giza-itzultzaileei laguntzen diena, besteak beste, itzulpen-memoriekin. OmegaT-k, aurretik egindako hainbat itzulpen gordeta dituenez, itzultzaileak esaldi berri bat itzuli behar duenean, aurretik itzulitakoen artean antzekoenak bilatu eta haientzat zeuzkan gordetako itzulpenak proposatzen ditu, baten bat egokia izanez gero giza-itzultzaileak berrerabil dezan. Itzulpen-memoriak oso erabilgarriak izan daitezke itzuli beharreko testuen artean antzeko atalak dituzten testu-zatiak agertzen badira.


OmegaT-1.7.2 :: parsifal
Proiektua Editatu Joan Tresnak Aukerak Laguntza
diaenparsifal.html
maila kontrolatzen ez bada.
Txostenaren arabera, klima aldaketa merkatuak inoiz izan duen hutsegiterik handi eta zabalena izango da.
En realidad, se trata de un masivo fallo político-moral, resultado de nuestra irresponsabilidad al no hacernos cargo de las consecuencias de nuestras acciones.
<segment 0019> Izatez, hutsegite politiko-moral erraldoia da, gure ekintzek dituzten ondorioen ardura ez hartzeak suposatzen duen axolagabetasunaren ondorio. <end segment>
Parekatze lausoak
1) Es necesario romper la ilusión de la normalidad, estamos viviendo algo excepcional, y decidir si continuar con la irresponsabilidad actual o cambiar de rumbo y hacernos carg de las consecuencias de nuestras acciones.
Normaltasunaren ilusioa hautsi behar dugu, salbuespen egoer batean bizi baikara, eta erabaki beharko dugu ia egungo arduragabekerian jarraitzen dugun ala norabidez aldatu eta gu ekintzen ardura hartzen dugun.
< 30% >
Glosarioa

3. irudia: *OmegaT tresnaren interfazea*

Itzulpen-memoriaz gain, OmegaT-k beste hainbat funtzionalitate eskaintzen ditu:

- Internetetik zuzenean hainbat itzultzaile automatiko erabiltzeko aukera eskaintzen du (adibidez, Google Translate, Apertium edo Belazar), jatorrizko esaldien oinarrizko itzulpen bat eskuratzeko.
- Glosario eta hiztegien sorkuntza eta inportazioa. Terminologiaren kudeaketarako lagntza da hori, itzulpenen koherentzia lortzeko. Glosarioetan domeinu berezi baterako hitz bakanak edo hitz anitzeko unitate lexikalak gordetzen dira. OmegaT-k itzuli behar dugun uneko segmentuko hitzen ordainak erakusten ditu, hiztegietan edo glosarioetan baldin badaude.

6 http://www.omegat.org

- Lematizatzaileen erabilera, hitzen lemak hobeto erauzteko. Modu horretan asko hobetzen da parekatze lausoen bilaketa eta glosarioen erabilera. Ezaugarri hau euskara bezalako hizkuntza deklinatuentzako bereziki interesgarria da.

OmegaT-k hainbat formatutako fitxategiak onartzen ditu (HTML, MS Office eta OpenOffice bulegotika-aplikazioen formatuak, DocBook, PO, etab.) eta fitxategiak iragazteko dituen arauei jarraituta formatu-markak identifikatu eta itzuli behar den testu soila lortzen du. Testu gordin hori segmentutan banatu eta segmentuak banaka erakusten dizkio giza itzultzaileari beronek lehen aipatutako laguntzekin itzul ditzan.

OmegaT kode irekiko plataforma da eta erabiltzaileen eta garatzaileen komunitate aktibo batek sostengatzen du.

OmegaT itzulpen-ingurunea aukeratu dugu Wikipediako espainierazko artikuluen itzulpen-automatikoen postedizio lanetarako, software librea izatean, aukera emango digulako sistema gure beharretara egokitzeko, besteak beste funtzionalitate berriak integratzen.

Proiektuaren hasieran itzulpenean laguntzeko beste sistema posible batzuk aztertu genituen: (1) World Wide Lexicon Translator (WWL3)[7], Firefox-erako gehigarri bat webguneak itzulita ikusi ahal izateko, konbinatzen zuen giza-itzulpena eta itzulpen automatikoa, baina posteditatzeko intefazea ez zebilen oso ondo; (2) Google Translation Toolkit[8] tresnak Wikipediako sarrerak itzultzeko laguntza eskaintzen du baina software libre eta irekia ez denez ezin izan genuen moldatu gure beharretara. OmegaT aukeratu genuen, software librea izanda moldatzeko aukera ematen zigun eta.

3 Elkarlan proiektua

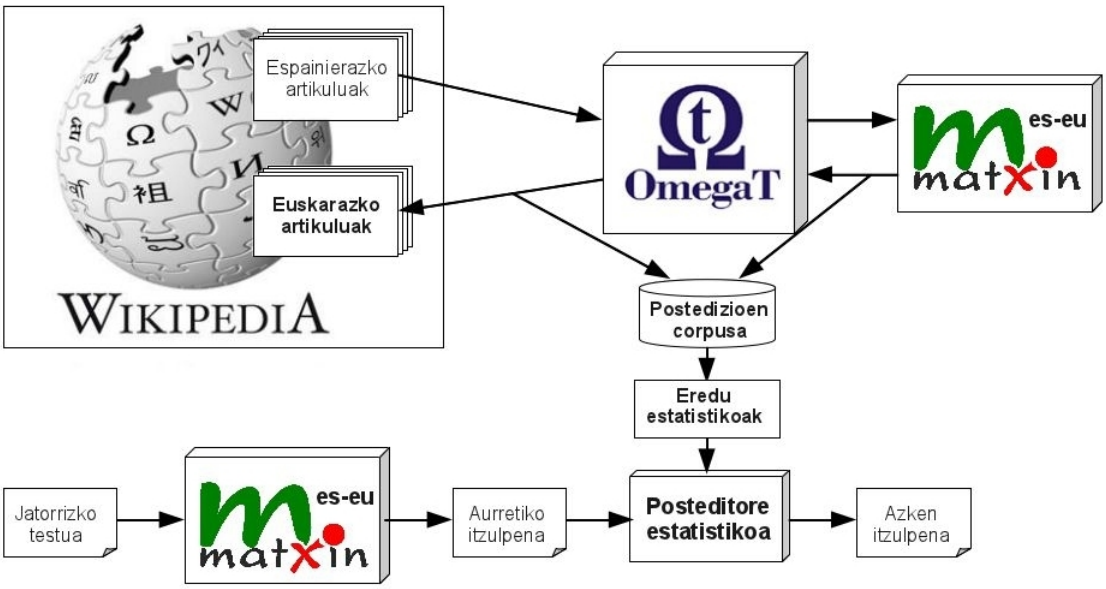
3.1 Diseinua

5. irudia: *Wikitzul saioa (2012/10/12)*

Gure proiektua boluntarioen lankidetzan oinarritu da. OmegaT plataforma egokitua erabiliz boluntarioek espainierazko Wikipediako artikuluen itzulpen-zirriborroa lortzen zuten Matxin itzultzaile automatikoa erabiliz eta, ondoren, itzulpen gordin hori aztertu eta zuzentzen zuten, Postedizio-lan horretan, batzuetan, itzulpen automatikoaren emaitza oso txarra zenean, itzulpen hori ez zen erabiltzen eta esaldi osoa berriro itzultzen zuen boluntarioak bere kabuz. Baina, gehienetan,

7 http://wordpress.org/plugins/speaklike-worldwide-lexicon-translator/
8 http://translate.google.com/toolkit

automatikoki sortutako itzulpeneko akatsak zuzenduz lan asko aurrezten zen.



4.irudia: *Prozesu osoaren ikuspegia*

Matxin itzulpen automatikoko sistemaren emaitzak hobeak izateko asmoz ezagutza-eremu zehatz baterako egokitzea erabaki genuen. Aukeratu genuen eremua informatikarena izan zen, batetik eremu teknikoa izatean faktore kulturalen menpekotasuna txikiagoa delako eta bestetik gure gertuko komunitatearentzat gai ezaguna delako. Matxin sistemaren lexikoiak eremu horretarako egokitu genituen eta, noski, *Informatika* kategoriakoak izan ziren gure proiektuan itzultzeko aukeratuko genituen Wikipediako artikuluak.

Proiektua publiko egin aurretik eta boluntarioak biltzen hasi aurretik, probazko saiakera batzuk egin genituen diseinatzen ari ginen prozesua aztertzeko, eta artikulu luzeen postedizioa egitea esfortzu handiko lana zela ikusi genuen (artikulu luzeak posteditatzeko 8 ordu baino gehiago behar izaten ziren). Horrek boluntarioak lortzea zaildu zezakeenez eta lana bukatu gabe uzteko posibilitatea handitzen zuenez, artikulu luzeak itzuli beharrean neurri ertaineko artikuluak proposatzea erabaki genuen.

Boluntario bakoitzari hasieran 2-3 lerrotako artikulu motz batekin probatzen hastea eskatu zitzaion, OmegaT plataforma erabiltzeko prozesu osoan trebatzeko: artikulua Wikipediatik jaistea, itzulpen automatikoa lortzea, itzulpena posteditatzea eta azkenik emaitza Euskal Wikipediara igotzea. Ondoren 10-20 lerroetako artikulu bat itzultzeko eskatzen zitzaion, eta gustura aritu ziren batzuk artikulu bat baino gehiago ere itzuli zuten.

Wikiproiektu bat martxan jarri genuen[9]. Bertan OmegaT-ren bertsio egokitua jaisteko loturarekin batera, Espainerazko Wikipediako hainbat artikuluren zerrenda jarriz, garai horretan euskarazko itzulpenik ez zutenak, boluntario bakoitzak artikulu horien artean gogokoena(k) aukeratu ahal izateko.

Lankidetza-kanpaina publikoa zortzi hilabetez egon zen martxan, 2011ko uztailetik 2012ko otsailera. Boluntarioekin aritzeko trebatze-saio pare bat antolatu genituen, artikulu motzekin lortu behar zen trebakuntza hori taldean eta lagunduta landu ahal izateko. Gainera, “Wikitzul” deituriko elkarlan-saioetara boluntario gehiago erakartzeko bi deialdi egin ziren[10].

Guztira, prozesu osoan, 36 boluntariok parte hartu zuten eta, 100 artikulu itzuli ondoren, euskarazko 50.204 hitz gehitu dira Euskal Wikipediara.

9 http://eu.wikipedia.org/wiki/Wikiproiektu:OpenMT2_eta_Euskal_Wikipedia
10 http://www.unibertsitatea.net/blogak/ixa/2011/12/05/euskal-wikipediaren-edizio-maratoia-durangoko-azokan/

Eskuz posteditatutako artikulu horiekin bi helburu lortu ditugu. Batetik, Euskal Wikipedia zabaltzea eta, bestetik, automatikoki itzulitako testuen eta haien eskuzko postedizioen corpusa osatzea, posteditore estatistiko bat eraikitzeko erabili dena. Gainera, prozesu honetan, bai Matxin sisteman bai OmegaT-n hainbat hobekuntza egin dira.

3.2 Wikipediako artikuluen aukeraketa

Boluntarioak gureganatu ahal izateko, trebakuntzarako artikulu labur batzuk, baina batez ere luzera ertaineko artikulu asko identifikatu behar genituen. Azkenean luzera ertaineko artikuluen itzulpenen postedizioa proposatzea erabaki genuen, hasierako kontaktuetan argi ikusi baikenuen proposatutako artikuluak handiegia izanda ez zela boluntariorik hurbilduko.

Wikipediako kategoria batean dauden artikuluen zerrenda ematen duen programa prestatu genuen informatika eremuko artikulu ertainak eskuratzeko. Artikulu bakoitzean ea beste hizkuntzetan baliokideak ba ote zeuden eta horien luzera ere eskuratu genituen programa horrekin.

Katalanezko Wikipediaren tamaina (378.408 artikulu) espainierazko (902.113 artikulu) eta euskarazko (135.273 artikulu) Wikipedien tamainen artean kokatzen zenez, gure ustez katalanezko Wikipedian dagoen eta euskarazkoan ez dagoen artikulu bat lehenago gehitu beharko litzateke Euskal Wikipedian, katalanezko Wikipedian ez dagoen artikulu bat baino.

Beraz, Euskal Wikipediaren hutsuneak identifikatzeko orduan gure programaren emaitza erabili genuen irispide honekin: Katalanezko Wikipedian *Informàtica* kategorian eta Espainierazko Wikipedian izanik, euskarazkoan existitu gabe espainierazkoan 10-20 lerro arteko luzera zuten artikuluak identifikatzeko.

Horrela 140 artikulu identifikatu genituen, Euskal Wikipedian sartzeko artikuluen proposamen-zerrendan boluntarioei eskaintzeko.

3.3 Matxinen egokitzapenak

Informatika eremurako egokitu dugu Matxin itzulpen-sistema, bere lexikoi elebiduna bi modutara aberastuta [14]:

- Lexikoiaren egokitzapena hiztegi elektronikoak erabiliz. Hainbat espainiera-euskara on-line hiztegitan informatika eremurako hitzen ordainen bilaketa sistematikoa burutu dugu. Horrela 1.623 sarrera berri gehitu ziren Matxinen lexikoian. Termino gehienak hitz anitzekoak ziren, adibidez “base de datos” (datu-base) edo “lenguaje de programación” (programazio-lengoaia). Hitz bakarreko termino batzuk ere lortu ziren, adibidez “iterativo” (iteratibo), “ejecutable” (exekutagarri) edo “ensamblador” (mihiztatzaile). Gainera, hautapen lexikalerako ordainen ordena aldatu zen 184 hitzetan; esate baterako, erdarazko “rutina” sarreraren ordainen ordena aldatu genuen, “errutina” ordaina jarri genuen lehenengo ordain modura, ordura arte lehenesten zen “ohitura” ordaina atzerago eramanda.

- Lexikoiaren egokitzapena corpus paralelo bat erabiliz. Informatikaren eremuko espainiera-euskara corpus paralelo bat bildu genuen, Mozilla software librearen lokalizazioan sortutakoa (138.000 segmentu, 600.000 hitz espainieraz eta 440.000 euskaraz). Corpus hau itzulpen automatiko estatistikorako egokia ez bada ere, erabilgarria izan daiteke erlazio lexikalak erauzteko. Giza++ lerrokatzeetan oinarrituta, corpuseko sarrera bakoitzerako itzulpen posibleen zerrenda erauzi genuen, itzulpen posible bakoitzaren probabilitatearekin. Doitasunagatik, zerrenda hauek hautapen lexikalerako soilik erabili genituen. Ordainen ordena aldatu zen lexikoiko 444 sarreratan. Adibidez ”dirección” sarrerarako “helbide” hobetsi zen, “norabide” beharrean, egokiagoa izango zelakoan.

3.4 OmegaTren egokitzapenak

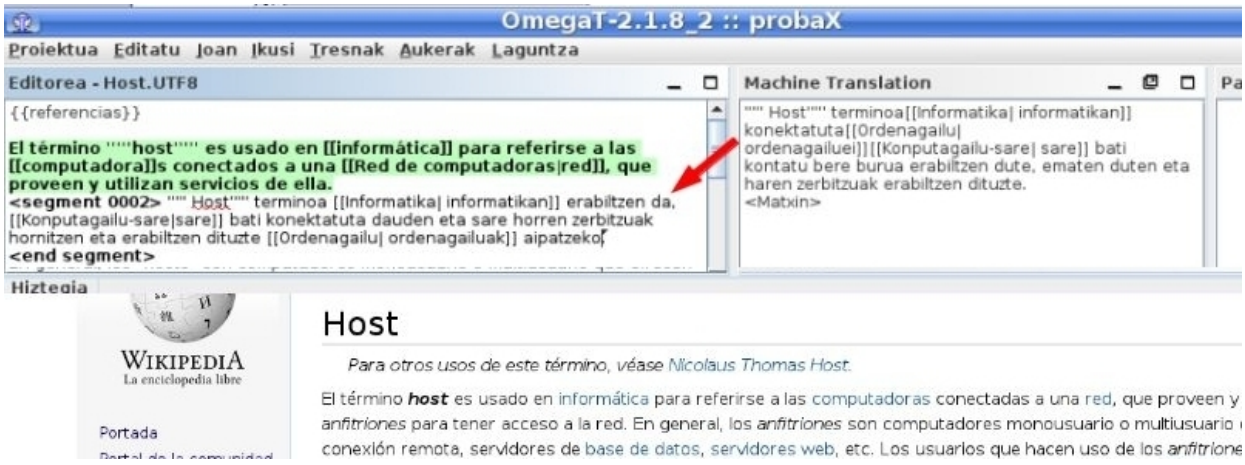
Posteditoreentzat erabilerrazago egiteko, OmegaT egokitu dugu hainbat funtzionalitate gehituz:

- Matxin es-eu itzultzaile automatikoaren integrazioa. OmegaT-k badu itzulpen automatikoko zerbitzuak konektatzeko klase bat, zerbitzu berriak erraz gehitu ahal izateko. Matxin OmegaT-ren barruan integratzeko erabili genuen guk klase hori. Integrazio-lana errazteko, Matxin sistema egokitu genuen web zerbitzu gisa inplementatuz, SOAP bidez API deiekin atzigarria izateko. Horrela OmegaT barrutik erraz erabil daiteke Matxin sistema.

- Inportatu/esportatu Wikipediako artikuluak OmegaT-n. OmegaT-k eskaintzen dituen MediaWiki dokumentuen inportaziorako aukerei gehituz, ezaugarri berri bat inplementatu genuen espainierazko Wikipediako artikuluak inportatzeko eta beste bat posteditatutako artikuluak Euskal Wikipediara igotzeko. Horretarako *login* egiteko modulu berri bat inplementatu behar izan genuen. Artikulu berria Euskal Wikipediara igotzean, gainera, Matxinek emandako itzulpen automatikoa eta boluntarioak posteditatutako itzulpena geure zerbitzari batera bidaltzen dira, lortu nahi dugun postedizio-corpusa osatzen joateko.

- Euskarazko zuzentzaile ortografikoaren integrazioa, postedizio-lanak errazteko.

- Wikipediako estekak itzultzeko programa, Wikipediako metadatuen informazioa erabiliz. Ikus dezagun, adibidez, [[gravedad | gravedad]] Wikipediako barne loturaren itzulpena. Lotura horren lehenengo terminoak Wikipediako sarrera bati egiten dio erreferentzia, eta bigarren terminoa lotura horretan erakutsiko den testua da. Wikipediako artikulu bati erreferentzia egiten dion lehenengo termino hori itzultzeko, gure programak Wikipedia barruko informazioa erabiltzen du (zein beste hizkuntzatan dagoen artikulua eta zein den bere sarrera hizkuntza horietan) espainierazko artikulu horri dagokion euskarazko artikulua eskuratzeko, kasu honetan “grabitazio”. Bigarren terminoa, lotura hori adierazteko erakutsiko den testua, Matxin itzultzaile automatikoa erabiliz itzuliko da, “larritasuna” lortuz. Horrela, emandako barne loturaren itzulpena [[grabitazio | larritasuna]] izango da. Posteditorearentzat laguntza ederra da automatikoki lortzea Euskal Wikipediako lotura (*grabitazio*), berak ezer bilatu beharrik gabe. Gainera, euskaraz dagokion sarrera ikusita erraz zuzendu dezake Matxin-ek aukeratu duen ordaina, testuinguru horretan egokia ez bada (*larritasuna-->grabitazioa*). Ordezkapen hori beti automatikoki egitea lagungarria izan daiteke, baina ez beti; epe erdirako eginkizunen artean jarri dugu funtzionaliatate hau aztertu eta inplementatzea.


OmegaT-2.1.8_2 :: probaX
Proiektua Editatu Joan Ikusi Tresnak Aukerak Laguntza
Editorea - Host.UTF8
{{referencias}}
El término """host""" es usado en [[informática]] para referirse a las [[computadora]]s conectados a una [[Red de computadoras|red]], que proveen y utilizan servicios de ella.
<segment 0002> """ Host""" terminoa [[Informatika| informatikan]] erabiltzen da, [[Konputagailu-sare|sare]] bati konektatuta dauden eta sare horren zerbitzuak hornitzen eta erabiltzen dituzte [[Ordenagailu| ordenagailuak]] aipatzeko
<end segment>
Machine Translation
""" Host""" terminoa[[Informatika| informatikan]] konektatuta[[Ordenagailu| ordenagailuei]][[Konputagailu-sare| sare]] bati kontatu bere burua erabiltzen dute, ematen duten eta haren zerbitzuak erabiltzen dituzte.
<Matxin>
Hiztegia
WIKIPEDIA
La enciclopedia libre
Portada
Host
Para otros usos de este término, véase Nicolaus Thomas Host.
El término host es usado en informática para referirse a las computadoras conectadas a una red, que proveen y anfitriones para tener acceso a la red. En general, los anfitriones son computadores monousuario o multiusuario
conexión remota, servidores de base de datos, servidores web, etc. Los usuarios que hacen uso de los anfitriones

6. irudia: *Wikipediako esteken itzulpena. Espainerazko Wikipediako “Host” artikuluan “red” hitza esteka moduan agertzen da. Esteka hori [[Red de computadoras | red]] moduan adierazten da. OmegaT eta Matxinek [[Konputagailu-sare|sare]] eskaintzen dute.*

4 Emaitzak eta hobekuntzak

Guztira hauek dira sortu eta modu irekian plazaratu ditugun produktuak:

- Sortutako Corpusak
  - Espainiera/euskara corpus paralelo bat.[11] Mozilla softwarearen lokalizazioan sortu denaren bertsio berria (138.000 segmentu, 600.000 hitz espainieraz eta 440.000 euskaraz).
  - Testu itzuli eta zuzenduen corpus bat.[12] Espainierazko Wikipediako 100 artikuluak Matxin itzultzaile automatikoak sortutako itzulpenekin eta boluntarioek eskuz posteditatutako itzulpen zuzenduekin. Corpus horren euskarazko aldeak 50.204 hitz dauzka.
- Wikipedia
  - Euskarazko artikulu berriak: 100 artikulu berri gehitu dira Euskal Wikipedian, guztira 50.204 hitz.[13]
  - Artikuluak bilatzeko programatxoa (wikigaiak4koa.pl). Tresna hau, perl-ez inplementatua, Wikipediaren edozein hizkuntzatako kategoria baten edukia aztertzeko erabil daiteke. Kategoria bat eta lau hizkuntza emanda, lehenengo hizkuntzaren kategoria horretako artikuluen zerrenda ematen digu, artikulu bakoitzaren beste hiru hizkuntza horietako baliokideekin eta beren luzerekin.[14]
- Matxin
  - Matxinen bertsio egokitua informatika arlorako. Sistemaren lexikoia eremu zehatz horretara egokitu dugu eta SOAP[15] zerbitzu moduan inplementatu dugu. Jatorrizko Matxin sistema eta informatika arlorako egokitutakoa automatikoki ebaluatu ditugu. Espainierazko Wikipediatik aukeratutako artikuluak erabili dira, boluntarioek postedizio bidez sortutako itzulpen zuzenduak erreferentzia gisa hartuta. Erabilitako metrika automatiko [15] guztietarako (MBLEU, BLEU, NIST, METEOR, TER, WER eta PER) egokitutako sistemak emaitza hobeak ematen ditu: hobekuntza erlatibo handienak BLEUk eta MBLEUk adierazten dute (%15) eta hobekuntza txikienak WER metrikak (%3,5)
  - Postedizio automatikorako modulua. Automatikoki lortutako itzulpenak eta itzulpen horien eskuz posteditatutakoak biltzen dituen 50.000 hitzeko corpusa baliatuta eta teknika estatistikoak erabiliz Matxin sistemarekin lortutako itzulpenak automatikoki posteditatzen dituen programa berri bat sortu dugu. Programa honek Matxin sistemaren irteera jaso eta postedizio automatikoa egiten du. Ebaluazio automatikoak erakutsi du sistema berri honek %10eko hobekuntza lortzen duela, Matxin sistema soilarekin konparatuz gero (Alegria et al., 2013). Gurean erabilitako corpusaren tamaina postedizio estatistikoari buruzko nazioarteko esperimentu nagusietan erabili dituztenak baino txikiagoa da (adibidez Simard et al. (2007) 100.000 hitzetako corpusa erabiltzen dute). Beraz postedizio corpus handiago bat lortuz hobetzeko bidea badagoela aurreikusten dugu.
- OmegatT
  - Matxin itzultzailea eta euskarazko zuzentzailea OmegaT-n erabiltzeko aukera gehitu dugu.
  - Wikipediako artikuluak inportatzeko eta esportatzeko funtzionalitateak gehitu ditugu.

11 http://ixa2.si.ehu.es/glabaka/lokalizazioa.tmx
Eskerrak Elhuyarri eta Julen Ruiz-i

12 http://ixa2.si.ehu.es/glabaka/OmegaT/OpenMT-OmegaT-CS-TM.zip

13 http://eu.wikipedia.org/w/index.php?title=Berezi:ZerkLotzenDuHona/Txantiloi:OpenMT-2&limit=250

14 http://www.unibertsitatea.net/blogak/testuak− lantzen/2011/11/22/wikigaiak4koa

15 http://eu.wikipedia.org/wiki/SOAP

Ezaugarri hau hizkuntzarekiko independentea da, eta euskara ez beste hizkuntzentzat ere erabil daiteke (ezaugarri hau ez da oraindik probatu beste karaktere multzo bat erabiltzen duten hizkuntzentzat, arabiera adibidez).

- Wikipediako estekak itzultzeko programa inplementatu dugu, Wikipedia barruko informazioa erabiliz espainierazko artikulu bati dagokion euskarazkoa zein den lortzen duena.
- OmegaT deskargatzeko, instalatzeko eta erabiltzeko eskuliburua sortu dugu,[16] Wikipediako artikuluak posteditatzeko zehaztasunekin.

5 Ondorioak eta etorkizuneko lana

Boluntario-lana lortzea zaila izan da gurean, eginbeharreko lana burutzeko komunitatea sortzea eta koordinatzea esfortzu handia suposatu baitigu. Bagenekien hasieratik euskara bezalako hiztun gutxiko hizkuntza baten kasuan hala izango zela, eta hala ere gure helburua lortu dugu.[17] 36 boluntariok parte hartu dute proiektuan, horietako 20k luzera ertaineko artikulu bat amaitu dutelarik. Argi gelditu zaigu Wikipediako artikulu motzak aukeratzea egokiagoa dela boluntarioak erakartzeko proiektura, gehienetan motibazioa ez delako nahiko handia gehiegizko esfortzua inbertitzeko.

Itzulpen automatikoa itzultzaile profesionalentzat oso lagungarria ez bazirudien ere, itzultzaile amateurrentzat erabilgarria izan zitekeela zen gure hasierako hipotesia. Proiektuan sortutako postedizioak aztertu ondoren, gure hipotesia baiezta dezakegu, itzulpen automatikoko sistemaren irteeraren kalitatea oso ona ez bada ere, nahikoa delako editoreei laguntzeko, eginbeharreko esfortzua gutxituta.

OmegaT lokalki instalatu eta konfiguratu behar izatea eragozpen bat izan da, posteditoreen komunitate handi bat erabiltzea nahi genuela kontuan harturik. Gure proiekturako eskertzekoa izango zen Google Translation Toolkit bezalako on-line elkarlan plataforma bat erabili ahal izatea. Hau dela eta, etorkizuneko proiektuetan erabiltzeko on-line plataforma egokiago bat egokitzea edo sortzea planeatzen ari gara.

Erabiltzaile berri batek OmegaT tresna erabiltzeko izan ditzakeen zailtasunak ere kontuan hartu behar dira, plataformak aukera eta funtzionalitate asko eskaintzen dituelako, eta horien artean norberak behar duen azpimultzoa soilik erabiltzea konplexua egiten delako. Zorionez, dokumentazio asko existitzen da horretan laguntzeko; gainera, guk geure proiektuan erabiltzen diren funtzionalitateak argitzen dituen eskuliburu bat sortu dugu posteditoreen lana gidatzeko. Gure proiektuan ikusi ahal izan dugu, gida hauekin, erabiltzaileek hasierako zailtasunak gainditzen dituztela eta berehala lortzen dutela OmegaT-rekin bere kabuz lan egiteko trebetasuna.

Wikipediako artikuluen metadatuen tratamendua erronka bat da, bai itzulpen automatikoko sistemarentzat eta bai giza itzultzaileentzat. Honetan lagungarria izan da guk inplementatutako programa, artikuluetan agertzen diren loturen baliokidetzak lortzen dituena, Wikipediaren hizkuntzarteko loturen informazioa erabiliz. Etorkizunean Wikipediako esteketan eta barne antolakuntzan dagoen informazioa era sakonagoan erabiltzea aurreikusten dugu.

Buruan dugu itzulpen-sistemaren lexikoia aberastea ere, domeinuaren arabera ordain egokiagoak hautatzeko. Informatika ez den beste arlo batera ere zabal genezake gure sistema.

Etorkizuneko proiektuetan boluntarioak erakartzeko estrategiak ere findu beharko genituzke, boluntario-nitxo berriak identifikatuz edo: unibertsitateak, hizkuntz-eskolak, euskaltegiak...

16 http://siuc01.si.ehu.es/~jipsagak/OpenMT_Wiki/Eskuliburua_Euwikipedia+Omegat+Matxin.pdf

17 Boluntarioei gure eskerrik beroena, eurak izan baitira proiektuaren emaitza arrakastatsua lortzea ahalbidetu dutenak

Eskerrak

Ikerketa hau bi erakundek finantzatu dute: Eusko Jaurlaritza (Berbatek proiektua, Etortek deialdiko IE09–262) eta Espainiako Hezkuntza eta Zientzia Ministerioa (OpenMT2 proiektua, TIN2009-14675-C03-01). Elhuyarrek eta Julen Ruizek lagundu ziguten baliabideak biltzen Matxin itzultzailea informatika arlora egokitzen. Eta gure eskerrik beroena 36 boluntario kolaboratzaileei.

Erreferentziak

[1] HUTCHINS W.J. eta SOMERS H. 1992. «An introduction to machine translation». London: Academic Press. http://goo.gl/U0IbV

[2] ADURIZ I., ALEGRIA I., ARTOLA X., DÍAZ DE ILARRAZA A., SARASOLA K. 2011. «Teknologia garatzeko estrategiak baliabide urriko hizkuntzetarako: euskararen eta Ixa taldearen adibidea». Linguamatica — ISSN: 1647–0818, Vol. 3 Núm. 1 - Junho 2011 - Pág. 13–31

[3] HERNÁEZ I., NAVAS E. , ODRIOZOLA I., SARASOLA K., DIAZ DE ILARRAZA A., LETURIA I., DIAZ DE LEZANA A., OIHARTZABAL B. , SALABERRIA J. 2012. «The Basque language in the digital age / Euskara aro digitalean». METANET White Paper Series. Georg Rehm, Hans Uszkoreit (editors). Springer. ISBN 978-3-642-30795-9; e-book ISBN 978-3-642-30796-6. http://goo.gl/Fm9v7

[4] FERNANDEZ DE BETOÑO U. 2011. «Hamar urte jakintza librea zabaltzen». Gaur8.info. http://goo.gl/TO1to

[5] MUJIKA A. 2011. «Wikipedia: milioika artikulu ez ezik, zaleak eta kritikoak pilatu dituen kaxa erraldoia». Gaur8.info. http://goo.gl/TO1to

[6] WAY A. 2010. «Machine translation. In: A. Clark, C. Fox, and S. Lappin (eds.) . The Handbook of Computational Linguistics and Natural Language Processing», pp. 531–573. Oxford: Wiley-Blackwell.

[7] ALEGRIA I., CABEZÓN U. , FERNANDEZ DE BETOÑO U., GONZALEZ G., ITURBE M., LABAKA G., MAYOR A., SARASOLA K., ZUBIAGA A. 2013. «OpenMT2 eta Euskal Wikipedia wikiproiektuaren emaitzak». IX. Informatika Euskaldunen Bilkura, IEB2013. Udako Euskal Unibertsitatea. Donostia

[8] ALEGRIA I., CABEZÓN U. , FERNANDEZ DE BETOÑO U., GONZALEZ G., ITURBE M., LABAKA G., MAYOR A., SARASOLA K. 2013. «Reciprocal Enrichment between Basque Wikipedia and Machine Translators». In: The People's Web Meets NLP: Collaboratively Constructed Language Resources. Springer. http://goo.gl/JHpwJ

[9] MAYOR A. 2007. «Matxin: Erregeletan oinarritutako itzulpen automatikoko sistema baten eraikuntza estaldura handiko baliabide linguistikoak berrerabiliz». Tesi-lana. LSI Saila (EHU). Donostia. http://goo.gl/j5aBI

[10] MAYOR A, DIAZ DE ILARRAZA A, LABAKA G, LERSUNDI M, SARASOLA K. 2011 «Matxin, an open-source rule-based machine translation system for Basque». Machine Translation Journal 25, 1:53–82.

[11] ATSERIAS J., CASAS B., COMELLES E., GONZÁLEZ M., PADRÓ L., PADRÓ M. 2006. «FreeLing 1.3: Syntactic and semantic services in an open-source NLP library». In: Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'06). Genoa, Italy, pp. 48–55.

[12] LABAKA G. 2010. «EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language». Its use in SMT-RBMT-EBMT hybridation. Tesi-lana. LSI Saila (EHU). Donostia. http://goo.gl/b0Rga

[13] SIMARD M., UEFFING N, ISABELLE P, eta KUHN R 2007. «Rule-based translation with statistical phrase-based post-editing». In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 203–206. Prague.

[14] ALEGRIA I., DIAZ DE ILARRAZA A, LABAKA G, LERSUNDI M, MAYOR A, eta SARASOLA K. 2011. «Matxin-Informatika: Versión del traductor Matxin adaptada al dominio de la informática». In: Proceedings of the XXVII Congreso SEPLN, pp. 321–322. Huelva, Spain

[15] SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L., eta MAKHOUL J. 2007. «A study of translation edit rate with targeted human annotation». In: Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA), pp. 223–231 Cambridge, Massachusetts, USA.