Lengoaia eta Sistema Informatikoak Saila

Informatika Fakultatea

CLIR Teknikak
Baliabide Urriko Hizkuntzetarako

Xabier Saralegi Urizarrek

Informatikan Doktore titulua eskuratzeko aurkeztutako

TESI-TXOSTENA

Donostia, 2017ko apirila

Lengoaia eta Sistema Informatikoak Saila

Informatika Fakultatea

CLIR Teknikak
Baliabide Urriko Hizkuntzetarako

Xabier Saralegi Urizarrek Eneko Agirre eta Iñaki Alegriaren zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua

Donostia, 2017ko apirila

Eskerrak

- EHUko Iñaki Alegria eta Eneko Agirre zuzendariak: Nire tesi-lana maisuki zuzendu izanagatik, tesi-lanean zehar uneoro laguntzeko erakutsitako prestutasunagatik eta, batez ere, honelako lan luze bati heltzeko behar den etengabeko motibazioa bizirik mantentzea lortzeagatik.

- Elhuyar I+Gko Maddalen Lopez de Lacalle: Tesi-lan honetako ikergai nagusia den hizkuntza arteko informazioaren berreskurapenari buruzko esperimentuetako zenbait lanetan laguntzeagatik. Nabarmentzekoa eta benetan eskertzekoa izan da datu-multzoak prestatzen eta Indri tresnaz egindako esperimentuetan Maddalenen aldetik jaso dudan laguntza.

- Elhuyar I+Gko Iker Manterola: Pibotaje-tekniken bidezko hiztegien sorkuntzaren inguruko esperimentuetan bere aldetik jasotako laguntzagatik, batez ere, emaitzen ebaluazioan eta zenbait algoritmoren inplementazioan.

- Elhuyar I+Gko Iñaki San Vicente: Pibotaje bidezko hiztegien sorkuntzari buruzko esperimentuetako zenbait lanetan, tesi-lan hau osatzen duten artikulu zientifikoen ingelesezko erredakzioan eta LaTex-en maketatzen laguntzeagatik.

- Universidade de Santiago de Compostela-ko Pablo Gamallo: Amarauneko bilatzaileen bidez bildutako testuinguruen azterketa linguistikoa lantzean jasotako laguntzagatik.

- Elhuyar I+Gko Igor Leturia: Tesi-lan honen erredakzioa orrazten laguntzeagatik, eta Elhuyarreko I+G taldeko arduraduna zela hizkuntza arteko berreskurapenaren ildoaren aldeko apustua egin izanagatik.

- Elhuyar I+Gko Antton Gurrutxaga: Tesi-lan honen erredakzioan erabilitako terminologiaren inguruko zalantzak argitzeagatik, eta Elhuyarreko I+G taldeko arduraduna zela hizkuntza arteko berreskurapenaren ildoa martxan jartzea sustatzeagatik.
- Elhuyar I+Gko Josu Aztiria: Elhuyarreko Hizkuntza eta Teknologia saileko arduraduna zelarik, hizkuntza teknologien garapenean erakutsitako konfiantzagatik.
- Ezin utzi aipatu gabe une puntualetan laguntza eman edo eskaini didaten Elhuyar Fundazioko eta IXA taldeko beste hainbat kide.

Gaien Aurkibidea

I. KAPITULUA

Sarrera

I.1 Motibazioa eta eszenatokia

Gaur egungo jendartean informazio-trukaketak garrantzia handia du eremu eta alor askotan. Informazio digitala gero eta ugariagoa da, eta horren ondorioz, informazio hau ustiatzeko metodo automatikoen garapena ezinbestekoa da.

Automatizatu beharreko egitekoen artean bilaketa eta, harekin lotuta, informazioaren berreskurapena (*Information Retrieval* edo *IR* ingelesez) dugu. Informazioaren berreskurapenean sistema automatiko batek erabiltzaileak duen informazio-behar bati erantzun behar dio, erabiltzailearen informazio-beharrarekiko adierazgarria den informazioa itzuliz.

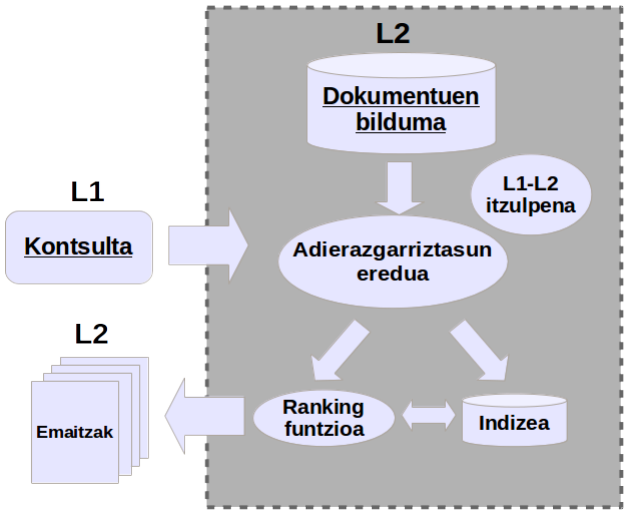
Hainbat testuingurutan (adibidez, denda birtualak, posta elektronikoa, eztabaida-foroak, kazetaritzako dokumentazio-lanak, amaraunean egiten diren kontsulta informazionalak) bilatu nahi den informazio adierazgarria hainbat hizkuntzatan egon daiteke. Baina informazioren berreskurapenerako sistema klasiko batean kontsultak eta kontsultatu beharreko bilduma hizkuntza berekoak izan ohi dira. Beraz, horrelako sistema klasiko bat ez litzateke egokia izango ingurune eleaniztunetan, non kontsultak formulatzeko hizkuntzak eta kontsultatu beharreko bildumaren hizkuntzak ez duten zertan berdinak izan behar.

Hizkuntza arteko informazioaren berreskurapena (*Cross Lingual Information Retrieval* edo CLIR ingelesez) da, hain zuzen ere, aipatutako eszenatoki eleaniztunetan egokia den prozesua. Prozesu hau informazioaren

berreskurapenaren aldaera bat da, non kontsultaren hizkuntza (*L1*) eta helburu-bildumarena (*L2*,*L3*,...) desberdinak diren (ikusi I.1 irudia).

Informazioa bilatzen duen erabiltzailea eleaniztuna izango balitz, edo bere hizkuntzara itzultzen duen itzulpen automatikorako sistema egoki bat eskuragarri izango balu, emaitza aberatsagoak lor litzake CLIR sistema baten bidez IR sistema elebakar batez baino. Benetan aberatsagoak izango lirateke baldin eta hizkuntza guztietatik jasotako informazio-zatiak osagarriak izango balira (adibidez, herrialde ezberdinetako egunkari digitalak).

Hortaz, ondoriozta daiteke CLIR sistema bat guztiz beharrezkoa izan dadin, bi baldintza bete beharko lituzkeela eszenatokiak: alde batetik, erabiltzaileak hizkuntza bat baino gehiago ulertzea, eta bestetik hizkuntza horien informazio adierazgarria osagarria izatea.



I.1 irudia: CLIR ($L1 \rightarrow L2$) sistema baten arkitekturaren grafikoa.

Kontrakoa pentsa daitekeen arren, ingurune eleaniztunak ugariak dira munduan, eta munduko biztanle gehienak elebidunak edota eleaniztunak dira. David-en [1997] arabera munduko umeen bi herenak ingurune elebidunetan hazten dira. [Tabouret-Keller, 2004] lanaren arabera Europako eleaniztasun-tasa populazioaren %50ekoa da. Hortaz, esan liteke oso fenomeno naturala dela gaur egungo gizarteetan, bai herrialde garatuetan baita garapenean dauden herrialdetan ere.

Munduan 6.000 hizkuntza inguru existitzen dira, eta gehienak hiztun

gutxiko hizkuntzak dira. Izan ere, 60 hizkuntzak soilik dauzkate 10 milioitik gora hiztun. Hauen artean, ingelesa nagusitu da nazioarteko esparru askotan, mundu mailako *lingua franca* bilakatuz. Guztira 510 milioi pertsonak daki ingelesez. Pentsa liteke, hortaz, hizkuntza bakar bat nagusitzen ari denez, hizkuntza arteko informazioaren berreskurapenerako teknologia ez dela beharrezkoa. Halere, hiztunen kopuruak sakonago aztertzen baditugu, ingelesezko hiztunak mundu osoko populazioaren %8ra ez direla iristen ikusten dugu. Alegia, populazioaren %92ak ingelesez ez daki. Horrez gain, 510 milioi horietatik 340 baino ez dira jatorrizko hiztunak. Datu hau garrantzitsua da jarraian ikusiko dugun bezala.

Pertsonak edozein ulermen- edo adierazpen-prozesutan seguruago sentitzen dira beren ama-hizkuntza erabiltzen dutenean. Informazioaren berreskurapenerako sistemen erabileran ere hori betetzen dela frogatu da zenbait ikerketatan [Zazo et al., 2005, Vundavalli, 2008]. Europako Batzordeko azterketa bat [Eurobarometer, 2011] antzeko ondorioetara iritsi zen. Azterketa honen arabera amarauneko erabiltzaileen %90ek bere hizkuntzan nabigatu nahiago izaten du, eta erabiltzaileen %44k uste du informazioa galtzen duela bere hizkuntzan bakarrik nabigatzen duenean.

Hizkuntza arteko informazioaren berreskurapenak irtenbidea ematen dio aurreko paragrafoan aipatutako problematikari [Gonzalo, 2002, Marlow et al., 2008]. Halere, hizkuntza arteko informazioaren berreskurapenerako sistemen garapena ez da tribiala, batez ere, garapen digital txikiko hizkuntzen (Baliabide Urriko Hizkuntzak izendatuko ditugunak) kasuan, non hizkuntza handietarako eskura dauden zenbait baliabide askoz urriagoak diren. Baliabide Urriko Hizkuntzak testuinguru eleaniztun gehienetan aurkituko ditugu.

Tesi honetan hizkuntza arteko informazioaren berreskurapenerako zenbait teknika ikertu ditugu, baliabide urriko hizkuntzetarako egokiak direnak. Azken helburua baliabide urriko hizkuntzetarako CLIR sistema eraginkor bat nola gara daitekeen ikertzea da. Baliabide urriko hizkuntzetarako eszenatoki honek corpus paralelo eta itzulpen automatikorako sistemak ez erabiltzera mugatuko gaitu.

CLIR sistemetan bi estrategia nagusi daude hizkuntzaren arteko berreskurapenari aurre egiteko: kontsulta bildumaren hizkuntzara itzuli edo bilduma kontsultaren hizkuntzara. Artearen egoerari buruzko kapituluan estrategia guztiei eta beraiek inplementatzeko zenbait teknikari buruz hitz egingo badugu ere, aurreratu dezakegu kontsulta itzultzea dela gehien erabilitako estrategia, batez ere, estrategia eskalagarriagoa delako.

Kontsultaren itzulpena egiteko zenbait teknika proposatu dira literaturan, lau talde nagusitan sailkatu daitezkeenak, bakoitzak eskatutako

baliabideari erreparatzen badiogu:

1. Itzulpen automatikoan oinarritutakoak.
2. Corpus paraleloetan oinarritutakoak.
3. Hiztegi elebidunetan oinarritutakoak.
4. Corpus konparagarrietan oinarritutakoak.

Baliabide hauek eskuragarritasunaren edo eskuratzeko zailtasunaren arabera sailkatu behar izango bagenitu, itzulpen automatikoak eta corpus paraleloak izango lirateke lehenengoak, corpus konparagarriak eta hiztegi elebidunak errazago lortu daitezkeelarik.

Lan honetan, hain zuzen, azken baliabide horietan oinarritutako CLIR teknikak landu ditugu, lanaren helburua baliabide urriko hizkuntzetarako CLIR teknikak garatzea baitugu. Halaber, kontsulta-saioak, IR sistema guztietan lortu daitekeen baliabidea, ustiatu nahi ditugu kontsulta itzultzeko prozesua hobetze aldera. Teknika hauen bidez CLIR sistema gai izango da erabiltzaileak formulatutako kontsulta bildumaren hizkuntzara itzultzeko.

Itzulpen-prozesu horretan aurre egin beharreko arazo nagusiak honako biak dira:

- Kontsultako hitz bakoitzeko itzulpen-hautagaiak lortzea.
- Itzulpen zuzenak hautatzea.

Halaber, CLIR teknika hauetarako baliabide nagusia den hiztegi elebiduna modu automatikoan sortzeko teknikak ere landu ditugu tesi-lan honetan. Kasu honetan ere hiztegietan eta corpus konparagarrietan soilik oinarritutako teknikak aztertu eta garatu ditugu. Bi hiztegi elebidun (D_{a-b} eta D_{b-c}) gurutzatuz eta hizkuntza bat pibote erabiliz hiztegi elebidun berriak (D_{a-c}) sortzeko estrategia aztertu eta landu dugu.

I.2 Helburuak

Dokumentuen berreskurapenerako sistema baten helburua erabiltzaileak egindako kontsultarekiko adierazgarriak diren dokumentuak berreskuratzea da. Dokumentuak eta kontsulta hizkuntza ezberdinetakoak baldin badira, hizkuntza arteko sistema bati buruz ari gara.

Hizkuntza arteko informazioaren berreskurapenerako sistema bat garatzean kontsulta itzultzea da hizkuntzaren mugari aurre egiteko

hurbilpenik erabiliena. Kontsulta itzultzeko estrategia arrakastatsuenak itzulpen automatikoko sistema edo corpus paraleloetan oinarritzen dira, baina baliabide hauek urriak dira aurreko atalean aipatutako baliabide urriko hizkuntzen eszenatokietan. Horrelako egoeretan egokiagoa litzateke eskuragarriago diren baliabideetan oinarritutako kontsulta itzultzeko estrategia bat. Tesi honetan frogatu nahi dugu baliabide nagusi horiek hiztegi elebiduna eta horren osagarri diren corpus konparagarriak eta kontsulta-saioak izan daitezkeela. Hipotesi hori baieztatzeko honako azpihelburuak finkatu ditugu:

- Kontsultak itzultzeko hiztegietan oinarritutako tekniken azterketa eta konparaketa, arreta berezia itzulpen-hautapenaren arazoan jarriz.
- Amarauna corpus konparagarri gisa erabiltzea kontsultaren itzulpen-prozesua hobetzeko.
- Kontsulta-saioak emandako testuinguru zabalaren informazioa erabiltzea kontsulten itzulpen-prozesua hobetzeko.
- Pibotaje bidezko metodoen azterketa hiztegi elebidunak sortzeko baliabide urriko hizkuntzetarako, hizkuntza pare batentzat hiztegirik ez dagoenean.

Helburu hauek lortzeko metodologia eta emaitzak hainbat argitalpen zientifikotan argitaratu dira. Hortaz, tesi hau artikulu-bilduma bezala aurkeztu dugu, non artikuluak aurreko azpihelburuekin bat datozen. Horrez gain, azterketa bibliografiko bat egin da gaiaren gaur egungo egoera islatzeko asmoz, eta tesiaren garapenean jorratu diren ekarpenak bildu dira bukaeran.

I.3 Tesiaren Egitura

Esan bezala, tesi honen erredakzioak artikulu bildumaren formula jarraitzen du. Artikuluak III. IV. V. eta VI. kapituluetan aurkezten dira. II. kapituluan CLIR ataza burutzeko tekniken inguruko azterketa bibliografikoa eta tesi-lan honek biltzen dituen esperimentuetan erabilitako baliabideak deskribatzen ditugu. Tesi honetan egindako lanen ondorio nagusiak VII. kapituluan aurkezten ditugu.

Artikuluak lau kapitulutan antolatu ditugu erreferentzia gisa azpihelburuak erabiliz:

- Kontsultak itzultzeko hiztegietan oinarritutako metodoen azterketa eta konparaketa (III. kap.):

  - Xabier Saralegi and Maddalen Lopez de Lacalle. Comparing different approaches to treat translation ambiguity in CLIR: Structured queries vs. target co-occurrence based selection. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 398–404. IEEE, 2009
  - Xabier Saralegi and Maddalen Lopez de Lacalle. Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010a

- Amarauna corpus konparagarri gisa erabiltzea kontsultaren itzulpena hobetzeko (IV. kap.):
  - Xabier Saralegi and Maddalen Lopez de Lacalle. Estimating translation probabilities from the web for structured queries on CLIR. In *European Conference on Information Retrieval*, pages 586–589. Springer, 2010b
  - Xabier Saralegi and Pablo Gamallo. Analyzing the sense distribution of concordances obtained by web as corpus approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 355–367. Springer, 2013

- Saioen informazioa baliatzea kontsultaren itzulpen-prozesua hobetzeko (V. kap.):
  - Xabier Saralegi, Eneko Agirre, and Iñaki Alegria. Evaluating Translation Quality and CLIR Performance of Query Sessions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016

- Hiztegi elebidunen sorkuntza pibotaje-tekniken bidez (VI. kap.):
  - Xabier Saralegi, Iker Manterola, and Inaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011

I.4 Tesiaren Laburpena

CLIR sistema batean kontsultaren itzulpen-prozesuan aurre egin beharreko arazo nagusiak honakoak dira:

- Kontsultako hitz bakoitzeko itzulpen-hautagaiak lortzea.
- Itzulpen zuzenak hautatzea.

III. kapituluan bi arazo horiek tratatzeko zenbait teknikaren azterketa eta konparaketa azaltzen dira. Aztertutako tekniketan itzulpen-hautagaiak hiztegi elektroniko batetik eskuratzen dira. Itzulpen anbiguoak ebazteko bi metodo aztertu eta alderatu ditugu: Pirkolaren metodoa [Pirkola, 1998], eta agerkidetzan oinarritutako algoritmo iteratibo bat [Monz and Dorr, 2005]. Pirkolaren metodoak hautapen espliziturik ez du egiten, berreskurapen-prozesuan hitz bakoitzeko hiztegitik eskuratutako itzulpen-hautagai guztiak hartzen baititu. Ranking-funtzioa kalkulatzean jatorrizko hitz bereko itzulpen-hautagaien estatistikak modu berezi batean kalkulatzen dira haien maiztasunak bateratze aldera.

Agerkidetzan oinarritutako algoritmoak, aldiz, esplizituki egiten du itzulpen-hautapena, eta berreskurapen-algoritmoa aplikatu aurretik. Agerkidetzak, helburu-bildumatik hartuak, helburu-hizkuntzan probabilitate handieneko itzulpen-hautagai multzoa kalkulatzeko erabiltzen dira, algoritmo iteratibo baten bidez.

Ebaluazioaren arabera bi metodoek nabarmen gainditzen dute itzulpen-hautapenerako hiztegitik lehen itzulpena hartzen duen *baseline*a edo oinarri-lerroa (%10-15 hobea). Kontsulta laburrak (2-3 hitzekoak) prozesatu behar direnean portaera antzekoa lortzen da bi metodoekin; kontsulta luzeekin, aldiz, Pirkolaren metodoak portaera eraginkorragoa eskaintzen du.

III. kapituluan, itzulpen-hautapenerako teknikak lantzeaz gain, hiztegien bidezko itzulpen-estrategian gertatzen diren erroreen azterketa ere egin dugu. Errore bakoitzak berreskurapen-prozesuaren eraginkortasunean duen eragina kuantifikatzeaz gain, erroreak tratatzeko metodo bana proposatu eta ebaluatu dira. Guztira hiru errore-iturri aztertu ditugu:

- Hiztegitik kanpoko -edo *Out of Vocabulary* (OOV)- hitzak.
- Hitz anitzeko terminoak.
- Itzulpen anbiguoak.

Itzulpen-anbiguotasuna eta hitz anitzeko unitateen presentzia dira CLIR prozesuan kalitate-jaitsiera handiena eragiten duten faktoreak. Hitz anitzeko unitateen presentziak %20 eta %9 inguruko beherakadak eragiten ditu kalitate aldetik, kontsulta laburren eta luzeen berreskurapen-prozesuan hurrenez hurren. Itzulpen-anbiguotasunaren kasuan %32ra eta %17ra iristen

dira jaitsierak, kontsulta laburretan eta luzeetan. OOV hitzen presentziak %12ko jaitsiera eragiten du kontsulta laburren berreskurapenean, eta %4koa kontsulta luzeen kasuan.

Itzulpen-anbiguotasuna tratatzeko agerkidetzan oinarritutako metodoaren zenbait aldaera aztertu dugu, kontsulta laburren kasuan %23ko hobekuntza lortuz hiztegitik lehen itzulpena hartzen duen *baseline*aren gainean. Hitz anitzeko terminoak itzultzeko hiztegi terminologikoan oinarritutako parekatze-metodo bat erabili da, *baseline*aren gainean %5eko hobekuntza lortzen duena. OOV hitzen presentziari aurre egiteko kognatuen detekzioan oinarritako metodo bat baliatu da, *baseline*aren gainean %9ko hobekuntzara heltzen dena.

IV. kapituluan, corpus konparagarri elebidunetan ezkutuan dagoen itzulpen-ezagutza CLIR prozesuan baliatzeko modua aztertu dugu, zehazki itzulpen-hautapenean aplikatuta.

Horretarako Pirkolaren metodoaren aldaera bat hartu dugu oinarri, [Darwish and Oard, 2003] artikuluan aurkeztutakoaren ildotik. Pirkolaren metodoaren aldaera honek jatorrizko hitz baterako hiztegitik hartutako itzulpen-hautagaiei pisuak esleitzea ahalbidetzen du. Pisu horiek antzekotasun distribuzionala konputatuz estimatu dira, corpus konparagarri elebidun batetik abiatuta.

Antzekotasun distribuzionala konputatzeko behar besteko testuinguruen kopurua bermatzeko amarauna corpus konparagarri erraldoi gisa nola erabil daitekeen aztertu dugu. Modu horretan lortutako hizkuntza arteko pisuak berreskurapen-prozesuan erabilita Pirkolaren metodoan oinarrituako *baseline*aren gainean %2ko hobekuntza lortzen da.

Esperimentuak gaztelania-ingelesa CLIR ataza baten barruan burutu dira, testuinguruak berreskuratzeko amaraunerako zenbait bilatzaile (*WebCorp*, *Google Blog Search* eta *Google News Archive*) aztertu nahi izan zirelako. Dena dela, emaitzak edozein hizkuntza-bikotetara estrapolatu daitezke.

Bilatzaileen erabilera, testuinguruak eskuratze aldera, sakonago aztertu nahi izan dugu beste esperimentu batean. Bilatzaileak ez daude pentsatuta datu linguistikoak biltzeko, baizik eta informazio adierazgarria aurkitzeko. Esperimentuan, hain zuzen ere, horixe aztertu nahi izan dugu, zenbateraino diren linguistikoki adierazgarriak bilatzaileen bidez bildutako testuinguruak. Azterketa horren motibazioa aurreko esperimentuan aurkitzen dugu, non antzekotasun distribuzionala konputatzeko testuinguruak bilatzaileen bidez lortu ditugun. Horretarako SemCor[1] (adierekin etiketatutako

[1] http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor

corpusa) eta bilatzaileen bidezko testuinguruen arteko aldea neurtu dugu, adiera-banaketari dagokionez. Adiera-banaketaz gain adiera-aniztasuna eta koherentzia linguistikoa ere neurtu ditugu.

Adiera-banaketari dagokionez, korrelazio apur bat dago SemCor eta bilatzaileen artean. Korrelazio baxuena WebCorp-ek ematen du, ziur asko azpian dituen amaraun osoko bilatzaileetan kontsulta nabigazionalei eta popularitateari lotutako faktoreei ematen zaien pisuarengatik.

V. kapituluan kontsultaz gaindiko testuinguruak baliatu nahi izan ditugu hiztegian oinarritutako CLIR prozesuaren itzulpen-hautapena burutzeko. Zehazki, kontsulta-saioetatik erauzitako testuinguruaren erabilera aztertu nahi izan dugu. Itzulpen-hautapena egiteko problema nagusietako bat testuinguru-informazioaren falta izaten da, CLIR sistema batean kontsulta itzuli nahi denean. Kontsulta asko laburrak direnez, kontsultako hitzek adierazitako testuingurua ez da nahikoa itzulpen-hautapena eraginkortasunez gidatzeko.

Esperimentuak birformulazio motako saioetan zentratu dira: a) zehazte, b) dibagazio, eta c) orokortze birformulazioetan zehazki. Birformulatutako kontsulta itzultzeko aurreko kontsultak adierazitako testuinguruak zenbateraino laguntzen duen neurtu da ebaluazioan. Hobekuntza lortzen da zehazte eta orokortze birformulazioen kasuan.

Hiztegietan oinarritutako CLIR sistemek orokorrean emaitza onargarriak eskaintzen dituzte. Hiztegi elebiduna da, beraz, funtsezko baliabidea. Baina zenbait hizkuntza-bikoteren kasuan hiztegi elebidunik ez da existitzen, edo ez da eskuragarri egoten. Hiztegi elebidunak modu automatikoan sortzeko estrategiak aztertu ditugu VI. kapituluan, hiztegi elebidunen gurutzaketan oinarritutako teknikak erabiliz.

Gurutzaketaren bidez D_{a-c} hiztegi elebiduna sortzen dugu D_{a-b} eta D_{b-c} hiztegietatik abiatuta, hau da, b hizkuntza zubi edo pibote baten modura erabiliz. Gurutzaketa hutsaren bidez lortutako hiztegitik itzulpen okerrak kimatzeko bi teknika eta beren konbinaketa bat aztertu ditugu. Bata, alderantzizko kontsulta (*Inverse Consultation*, IC ingelesez) [Tanaka and Umemura, 1994], hiztegien egituraz soilik baliatzen da. Bestea [Kaji et al., 2008, Gamallo and Pichel, 2010], corpus konparagarri elebidunetatik kalkulatutako antzekotasun distribuzionalean dago oinarrituta.

Esperimentuetan D_{eu-en} (euskara-ingelesa) eta D_{en-es} (ingeles-espainiera) hiztegiak gurutzatu ziren D_{eu-es} (euskara-espainiera) hiztegi bat sortzeko. Zenbait eszenatoki simulatzen zituen ebaluazio baten arabera, IC teknikak doitasun handiko hiztegiak bermatzen ditu. Antzekotasun distribuzionalak, aldiz, estaldura handiagoa ahalbidetzen du. Bi teknikak konbinatuz doitasun eta estalduraren arteko orekarik onena

lortzen da.

I.5 Argitaratutako artikuluak

Aipatu bezala tesia artikulu-bilduma bezala egituratu da. Tesiak biltzen dituen artikuluak honakoak dira:

- Xabier Saralegi and Maddalen Lopez de Lacalle. Comparing different approaches to treat translation ambiguity in CLIR: Structured queries vs. target co-occurrence based selection. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 398–404. IEEE, 2009
- Xabier Saralegi and Maddalen Lopez de Lacalle. Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010a
- Xabier Saralegi and Maddalen Lopez de Lacalle. Estimating translation probabilities from the web for structured queries on CLIR. In *European Conference on Information Retrieval*, pages 586–589. Springer, 2010b
- Xabier Saralegi and Pablo Gamallo. Analyzing the sense distribution of concordances obtained by web as corpus approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 355–367. Springer, 2013
- Xabier Saralegi, Eneko Agirre, and Iñaki Alegria. Evaluating Translation Quality and CLIR Performance of Query Sessions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016
- Xabier Saralegi, Iker Manterola, and Inaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011

Tesi honetatik kanpo badaude ere, jarraian zerrendatzen dira hizkuntzaren prozesamenduaren alorreko terminologia-erauzketa eta iritzien erauzketa ikergaietan elkarlanean argiratu ditudan beste artikulu adierazgarri batzuk.

Terminologia-erauzketan:

- Iñaki Alegria, Antton Gurrutxaga, Pili Lizaso, Xabier Saralegi, Sahats Ugartetxea, and Ruben Urizar. A XML-Based Term Extraction Tool for Basque. In *LREC*, 2004

- Antton Gurrutxaga, Xabier Saralegi, Sahats Ugartetxea, and Iñaki Alegria. Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. In *Atti del XII Congresso Internazionale di Lessicografia: Torino, 6-9 settembre 2006*, pages 159–165, 2006

- Xabier Saralegi, Iñaki San Vicente, and Antton Gurrutxaga. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of Building and using Comparable Corpora workshop*, pages 27–32, 2008

Iritzien erauzketan:

- Xabier Saralegi, Iñaki San Vicente, and Irati Ugarteburu. Cross-Lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 96–108. Springer, 2013

- Iñaki San Vicente and Xabier Saralegi. Polarity Lexicon Building: to what Extent Is the Manual Effort Worth? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA)

I.6 Sortutako baliabideak

Tesi-lan honen testuinguruan hiztegietan oinarritutako hizkuntza arteko bilaketa inplementatzen duen kode irekiko pakete bat, Bilakit[2] izenekoa, garatu da. Paketeak Solr/Lucene bilaketa-tresnaren gainean funtzionatzen du eta erraz integratu daiteke edozein garapen-proiektutan. Bilakitek inplementatzen dituen funtzio nagusiak honakoak dira:

- Hizkuntza arteko informazioaren berreskurapena.

[2]https://github.com/Elhuyar/Bilakit

- Lematizazio eleaniztuna.
- Entitateen erauzketa eta sailkapena.
- Hizkuntza gehiagotara hedatzeko erraza.

Bilakit paketea Elhuyarrek merkaturatzen duen Xenda[3] izeneko soluzioaren oinarri teknologikoa da. Xendak adimena, eleaniztasuna eta zehaztasuna eransten dizkie informazio-sistemei ondoko bilaketa- eta nabigazio-funtzio aurreratuen bidez:

- Hizkuntza batean bilaketa bat egin, eta hizkuntza batean baino gehiagotan erakusten ditu emaitzak.
- Bilaketetarako laguntza-funtzioak: auto-osaketa eta zuzenketa ortografikoa.
- Emaitzak ordenatzeko irizpide bat baino gehiago, hainbat eremutan aplikagarriak.
- Lematizazioa.
- Erlazionatutako edukiak, barrukoak zein kanpoko iturburuetakoak.
- Edukien aberaste semantikoa izendun entitateen markaketaren bidez (pertsonak, tokiak eta erakundeak).

Bestalde, zenbait hizkuntzatarako hiztegi elebidunak ere sortu ditugu, kasu batzuetan pibotaje-teknikak baliatuz[4].

[3] http://xenda.elhuyar.eus/
[4] http://hiztegiautomatikoak.elhuyar.eus/

II. KAPITULUA

Artearen egoera eta erabilitako baliabideak

Kapitulu honetako lehenengo lau ataletan CLIR ataza aurrera eramateko literaturan proposatutako hurbilketak eta teknikak azalduko ditugu. IR eta CLIR atazetan erabilitako estrategia nagusiak aurkezteaz gain, tesi-lan honetan landuko ditugun hiru gaitan jarriko dugu arreta berezia: a) kontsulten itzulpenerako metodoak, b) corpus konparagarrien aplikazioa CLIRen, eta c) kontsulta-saioen ustiaketa. Bosgarren atalean kontsultak itzultzeko zenbait hurbilketetan ezinbestekoak diren hiztegi elebidunak modu automatikoan sortzeko pibotaje-tekniken azterketa bibliografikoa aurkeztuko dugu. IR eta CLIR sistemak ebaluatzeko literaturan proposatzen diren metriken laburpena seigarren atalean azalduko dugu. Zazpigarren atalean IR eta CLIR alorretan euskararen gainean burutu diren ikerlanen laburpena azalduko dugu. Bukatzeko, zortzigarren atalean, tesi-lan honetako esperimentuetan erabilitako baliabideak, datu-bildumak, eta tresnak deskribatuko ditugu.

II.1 IR eta CLIR inguruneak

Informazio digitala gero eta ugariagoa da. Hori dela eta, gizakiak horren kopuru handiak kontsultatzeko bilaketa edo informazioaren berreskurapenerako (*Information Retrieval* edo IR) sistemak behar ditu [Salton and McGill, 1986, Baeza-Yates et al., 1999, Manning et al., 2008]. Mota honetako sistemek bultzada nabarmena izan zuten amaraunaren eztandarekin, horren biltegi handian informazio adierazgarria modu azkarrean eta eraginkorrean aurkitzea premiazko bilakatu baitzen. Adibide

adierazgarriena Google bilatzailean dugu, gaur egun pertsona gehienentzako eguneroko tresna bilakatu dena.

IRko ikergaien multzoa zabala bada ere, ebatzi beharreko ataza oinarrizkoena testu-bilduma batean gai zehatz bati buruzko testuak lortzea da. Horrela, gaia, edo informazio-beharra, zenbait hitzetako kontsulta (*query* ingelesez) baten bidez $q = (w_0, .., w_n)$, adierazten du erabiltzaileak, adibidez, $q_0 = (ibilgailu, elektrikoen, teknologia)$. Sistemaren emaitza kontsulta horrekiko adierazgarriak diren edukien zerrenda edo ranking bat da, $D = (d_0, .., d_n)$, adierazgarritasun-mailarako irizpidea gai-antzekotasuna (*topic similarity* ingelesez) izanik. Zerrenda hau adierazgarritasun-mailaren arabera ordenatuta egon daiteke. Zenbat eta eduki adierazgarri gehiago zerrenda-hasieran, orduan eta arrakasta-maila hobearekin betetzen da ataza. Ebaluazio-metodoei buruz 7. atalean hitz egingo dugu.

Literaturako IR sistemak honako ereduen arabera sailkatzen dira:

- Eredu boolearra [Lancaster and Gallup, 1973]: Multzoen teorian eta boolear aljebran oinarritutako eredua da. Dokumentuak multzoen bidez adierazten dira, eta kontsultak adierazpen logikoen bidez. Kontsultari dagokion adierazpen logikoa dokumentu baten hitz multzoari aplikatuz lortzen da bigarren honen adierazgarritasun-maila.

- Bektore-espazioan oinarritutako eredua [Salton et al., 1975]: Dokumentuak eta kontsultak bektoreen bidez adierazten dira. Adierazgarritasun-maila bi bektoreak konparatuz lortzen da.

- Eredu probabilistikoa: Dokumentuak eredu probabilistikoen bidez adierazten dira. Kontsulta bat emanda, harekiko adierazgarriak izateko probabilitatearen arabera ordenatzen dira dokumentuak. Bi eredu probabilistiko mota ezberdin bereizi daitezke; eredu klasikoak [Robertson et al., 1980] eta hizkuntza ereduak [Ponte and Croft, 1998].

- Ikasketa automatikoan oinarritutako ereduak [Cao et al., 2006] edo *Learning to Rank* (LTR) hurbilketa: Ikasketa automatikoaren bidez trebatzen dira ereduak erreferentziazko rankingez osatutako entrenamendu-datuekin.

Eredu hauek guztiak *bag-of-words* delako (hitzen poltsa) paradigman oinarritzen dira, hau da, dokumentuen errepresentazioa hitzen bidez egiten da.

Berreskurapen-atazaren adibidera (*q=(ibilgailu, elektrikoen, teknologia)*) itzuliz, sar ditzagun bi aldagai prozesuan; kontsultaren eta edukien hizkuntzak. Eszenatoki sinpleenean biak hizkuntza bera lirateke

(euskara), baina bi hizkuntza horiek ezberdinak direneko eszenatokia (adibidez, euskara eta ingelesa) ere gerta daiteke. Orduan, informazioaren berreskurapen elebakarra dena eszenatoki eleaniztun batean hizkuntza arteko informazioren berreskurapen ataza bihurtzen da.

Hizkuntzen arteko langa gainditzeko estrategia desberdinak proposatzen dira itzultzen den elementuaren arabera: kontsulta itzultzea, dokumentua itzultzea edo biak. Egile gehienek kontsultak itzultzeko estrategia lantzen dute, batez ere, estrategia hau oso arina delako memoria- eta prozesu-eskakizunei dagokienez [Hull and Grefenstette, 1996]. Kontsultak motzak dira, dokumentu-bildumaren tamainarekin alderatuta.

Estrategia bakoitzarekin lortutako emaitzei erreparatuz gero, dokumentuak itzuliz eraginkortasun onena lortzen da. Arrazoia honakoa da; dokumentuek kontsultek baino testuinguru zabalagoa dute. Hori dela eta, itzulpen-hautagai zuzena aukeratzea errazagoa da kasu horretan. Horrez gain, dokumentu batean kontsulta batean baino esaldi edo adibide gehiago daude. Hitz bat gaizki itzuli daiteke esaldi batean, baina askoz zailagoa da esaldi guztietan gaizki itzultzea. Kontsultan, ordea, hitza ondo itzultzeko aukera bakarra dago. [Oard, 1998] lanean erakutsi dute itzulpenaren kalitatea eta berreskurapen-prozesuaren eraginkortasuna hobetzen direla bildumak itzultzen direnean. Beste ikerlari batzuek [McCarley, 1999, Chen and Gey, 2003] oraindik emaitza hobeak lortu dituzte kontsultak eta bildumak itzuliz. Ranking bana sortzen dituzte kontsultak eta bildumak abiapuntutzat hartuta. Ondoren, bi eratan lortutako rankingak konbinatzen dituzte.

Bestalde, itzulpena nola egingo den erabaki behar da. Itzulpena burutzeko teknikak lau multzo nagusitan banatu daitezke erabilitako ezagutza-iturriaren arabera:

- Itzulpen automatikoa (*Machine Translation* edo MT).
- Corpus paraleloak.
- Hiztegi elebidunak.
- Corpus konparagarriak.

Hiztegi elebidunetan eta corpus paraleloetan oinarritako tekniketarako familia estatistiko diferenteak proposatu dira literaturan; hizkuntza arteko eredu probabilistikoak, eta hizkuntza arteko hizkuntza-ereduak. Lehenengoa hiztegi elebidunekin konbinatzera zuzenduta dago, itzulpen anbiguoak tratatzeko eragile bereziak eskainiz. Bigarrenak corpus paralelo batetik

lortutako itzulpen-probabilitateak eredu formalago batean integratzen ditu [Hiemstra, 2001].

Bi eredu hauen eraginkortasuna erabilitako baliabide motaren araberakoa bada ere, kasu gehienetan corpus paraleloekin trebatutako hizkuntzen arteko hizkuntza-ereduek emaitza hobeak lortzen dituzte [Xu et al., 2001]. Halere, lehen esan bezala, eredu honek corpus paraleloak eskatzen ditu, eta baliabide hau oso urria izaten da hizkuntza bikote gehienetarako. Hiztegi elebidunak, ordea, askoz ugariagoak dira, baina zoritxarrez ez dituzte itzulpen-probabilitateak ematen. Hortaz, itzulpen anbiguoak tratatzeko bestelako metodoak inplementatu behar dira, hurrengo atalean azalduko ditugunak.

MT sistemak ere erabil daitezke CLIR ataza bat burutzeko. Estrategiarik sinplenean kutxa beltz bat bezala erabiliko litzateke MT sistema, kontsulta edo dokumentuak itzultzeko [Jones et al., 1999, Wu et al., 2008]. Beste egile batzuek MT sistema CLIR atazara egokitzeko teknikak ikertu dituzte [Magdy and Jones, 2014, Ture et al., 2012, Sokolov et al., 2014].

Egile batzuk itzulpen automatiko estatistikoko (*Statistical Machine Translation* edo SMT) ereduak baliatzen dituzte CLIR atazari heltzeko. [Azarbonyad et al., 2012] ikerlanean corpus paralelo batetik ateratako itzulpen-ezagutza *Learning to Rank* eredu batean erabiltzen dute CLIR sistema bat garatzeko. Kontsulta-eduki bikoteetarako ateratako hainbat ezaugarritatik trebatzen dute ranking-eredua. Hizkuntzen arteko ezaugarrien mapaketa egiteko corpus paralelo batetik lortutako itzulpen-probabilitateak baliatzen dituzte. Hiztegietan eta corpus konparagarrietan oinarritutako metodoekin alderatuta hobekuntza nabarmena lortzen dute.

Tesi-lan honetan galderen itzulpena, eta hori aurrera eramateko MT sistema eta corpus paralelorik gabeko metodoak aztertu ditugu, baliabide urriko hizkuntzen esparrua baita gure eszenatokia.

Hurrengo ataletan CLIR arloko literaturaren azterketa aurkeztuko dugu honako gaiei bereziki erreparatuz: kontsultak itzultzeko metodo orokorrak, corpus konparagarrien erabilera, kontsulta-saioen ustiapena, hiztegien eraikuntza pibotaje bidez, eta gai hauetan euskararen gainean egindako lana.

II.2 Kontsulta itzultzeko metodoak

Hiztegietan oinarritutako itzulpen-metodoen aldetik [Pirkola, 1998] artikulua klasiko bat da. Bertan itzulpen-hautapena lantze aldera eredu

probabilistikoetan **kontsulta egituratuak** integratzea proposatu zen. Kontsulta egituratuek kontsulta bateko esanahi berdineko hitzak eragile baten bidez (*syn*) multzokatzeko aukera ematen dute (ikusi adibidea II.1). CLIR sistema baten kasuan, jatorrizko hitz bereko itzulpen-hautagai guztiak *syn* multzo berean sartuko genituzke, adibidean ikusten dugun bezala:

Kontsulta	Galdera egituratua
"kutsatu odol epai"	"#syn(pollute impregnate infect) #syn(blood kinship) #syn(sentence judgment scratch cut)"
s_1="kutsatu" s_2="odol" s_3="epai"	tr(s_1)={"pollute","impregnate","infect"} tr(s_2)={"blood","kinship"} tr(s_3)={"sentence","judgment","scratch","cut"}

II.1 taula: Kontsultaren itzulpena galdera egituraruan.

Multzo bereko (*syn* operadoreaz adierazita) itzulpen guztiak hitz bera izango balira bezala tratatzen dira TF_j (terminoaren maiztasuna edo *term frequency* D_j dokumentuan) eta DF (dokumentu-maiztasuna edo *document frequency*, bildumako zenbat dokumentutan agertzen den terminoa) estatistikoak kalkulatzen direnean, berreskurapen-garaian. Jatorrizko s_i hitz bakoitzeko itzulpen-hautagaien multzoaren ($tr(s_i)$) TF eta DF balioak itzulpen-hautagaien TF eta DF balioak batuz kalkulatzen dira III.1 eta III.2 formuletan azaltzen den bezala.

$$TF_j(s_i) = \sum_{\{t \in tr(s_i)\}} TF_j(t) \tag{II.1}$$

$$DF(s_i) = | \bigcup_{\{t \in tr(s_i)\}} \{d | t \in d\}| \tag{II.2}$$

Esan daiteke *syn* operadorearen araberako TF eta DF estatistikoen kalkulua kontserbadorea dela, itzulpen-hautagaiei pisu handiegia eragitea galarazten baitzaie, zuzenak direlarik ere. Horrela, hautagairen bat hitz orokorra baldin bada (DF balio handikoa), multzoari, edo jatorrizko hitzari, garrantzia kenduko zaio kontsultako gainerako hitzen aldean. Adibidez, *"aurkitu"* hitzerako *"find"*, *"discover"*, *"be"*, eta *"feel"* ingelesezko itzulpenak lortzen ditugu hiztegi elebidunetik. *"be"* hitzaren DF balioa oso

altua da, *"aurkitu"* hitzaren DF balioaren jaitsiera nabarmen eraginez. Hautagai orokor hori itzulpen zuzena baldin bada, ez dago problemarik. Baina itzulpen okerra balitz, jatorrizko hitzak pisua galduko luke eduki adierazgarrien rankinga kalkulatzen denean. Itzulpen-hautagaien artean DF baxuena duena aukeratuko balitz, arriskua legoke itzulpen-hautagai hori okerra izateko, eta ondorioz, behar baino piso gehiago emateko jatorrizko hitzari. Okerreko aukeraketa honek kalkulu kontserbadorearen bidez lortzen denak baino jaitsiera handiagoa eragin dezake eduki adierazgarrien rankingean.

Ildo horretan, autore batzuek [Darwish and Oard, 2003] **kontsulta egituratu probabilistikoak** proposatzen dituzte, non itzulpen-hautagaiei pisuak ematen zaizkien. Era horretan TF eta DF estatistikoak pisu horien arabera kalkulatzen dira modu orekatuago baten. Arestian aipatutako adibidean *"be"* itzulpen-hautagaiari pisu baxuena esleitu ahal zitzaion, suposatuz probabilitate gutxieneko hautagaia dela.

Kontsulta egituratuak ez ezik, **agerkidetzetan oinarritutako metodoak** proposatzen dira [Monz and Dorr, 2005, Ballesteros and Croft, 1998, Gao et al., 2001] itzulpen-hautapenari aurre egiteko. Metodo hauek helburu-bilduma hizkuntza-eredu bat izango balitz bezala erabiltzen dute itzulpen-hautapena bideratzeko. Literaturan proposatutako algoritmoek helburu-bilduman elkartze-mailarik handiena duten hautagaiak aukeratzen dituzte itzulpen zuzen moduan. Elkartze-maila orokor hori kalkulatzeko estrategia ezberdinak landu dira literaturan [Monz and Dorr, 2005, Gao et al., 2001, 2002, Liu et al., 2005].

Itzulpenaren anbiguotasuna ez ezik, lehen aipatu bezala, badaude beste arazo batzuk kontsultaren itzulpen-prozesuan gertatzen direnak: hiztegitik kanpoko hitzen presentzia eta hitz anitzeko unitateen itzulpena. Hiztegitik kanpoko hitzen presentzia itzulpen-prozesuan erabilitako itzulpen-baliabidearen estaldurari lotutako problema bat da. Irtenbiderik erabiliena helburu-bilduman hiztegitik kanpoko hitzaren kognatu (*cognate* ingelesez, antzeko hitza) bat aurkitzea da (Adib., "*banku*" $\rightarrow$ "*bank*") [Knight and Graehl, 1997]. Kognatuak harrapatzeko antzekotasun ortografikoa kalkulatzeko neurriak aplikatzen dira. Hitz anitzeko unitateak itzultzeko lehen aipatutako agerkidetzetan oinarritutako metodoak [Monz and Dorr, 2005, Ballesteros and Croft, 1998, Gao et al., 2001] eta hitz anitzeko unitateen zerrendak erabiltzen dira [Ballesteros and Croft, 1997].

Tesiaren esparrutik kanpo egon arren, MT sistemetan oinarritutako kontsulta itzultzeko hainbat metodo proposatu dira azken urteotan. [Magdy and Jones, 2014] lanean CLIR atazaren testuinguruan MT prozesua arintzeko teknikak proposatzen dituzte. MT sistema entrenatzeko

corpusa aurreprozesatzen dute hitz funtzionalak (*stop-word* izenarekin ezagutzen direnak) kenduz eta hitzen erroak utziz lematizazioaren bitartez. Aurreprozesu hau eginez MT prozesua nabarmen arintzen dute berreskurapen-prozesuan eragin gabe. [Ture et al., 2012] lanean SCFG (*Synchronous Context-Free Grammar*) sistema baten bidez burutzen dute kontsultaren itzulpena. Kontsultako hitz bakoitzeko n itzulpen onenak eraikitzen dituzte *reading* (kontsulten itzulpen-hautagai) bakoitza tokenizatuz eta itzulpenen probabilitateak metatuz. N itzulpen onenak hartuz itzulpen onenean oinarritutako oinarri-lerroa hobetzen dute.

[Sokolov et al., 2014] lanean SMT sistema optimizatzen dute CLIR atazaren ebaluazioaren arabera. Horretarako SMTrako entrenamendu diskriminatzailerako teknikak eta *linked data*tik modu automatikoan bildutako adierazgarritasun-epaiak (ingelesez *relevance judgments*) erabiltzen dituzte. Proposatutako estrategiak itzulpen- eta berreskurapen-prozesuei lotutako informazioa integratzen du CLIRera zuzendutako eredu bakarrean. Itzulpenak corpus paraleloen bidez estimatzen dira. Ezaugarri lexikaletarako parametroen estimazioa adierazgarritasun-epaien arabera egiten da. Hieber and Riezler [2015] artikuluan berreskurapen-prozesuari lotutako ezaugarriekin aberasten dute SMT eredu bat. Ezaugarri hauek itzulpen-ezaugarriekin batera optimizatzen dira ranking-helburu batekiko.

Tesi honetan baliabide urriko hizkuntzetarako metodoak garatu nahi ditugunez, **hiztegi elebidunetan oinarritutako metodoak** landu ditugu. Artearen egoerako zenbait metodo [Monz and Dorr, 2005, Pirkola, 1998, Darwish and Oard, 2003] ebaluatu eta alderatu ditugu, beraiek konbinatzen dituzten metodo hibrido berriak eta bestelako egokitzapenak ere landuz. Euskara aintzat hartzen duten mota honetako lehenengo esperimentuak dira literaturan. Esperimentu hauek III. kapituluan azaltzen ditugu.

II.3 Corpus konparagarriak CLIRen

Zenbait egilek corpus konparagarrietan oinarritutako CLIR estrategiak proposatzen dituzte. Kontsulta itzultzeko prozesua hobetzeko baliatzen dituzte mota honetako corpusak.

[Sadat et al., 2003] artikuluan corpus konparagarrietatik erauzten dituzte kontsulta itzultzeko itzulpen-hautagaiak. Antzekotasun distribuzionala eta POS informazioaren araberako murriztapenak baliatuz erauzten dituzte itzulpen-hautagaiak. Itzulpen-hautagaien artetik amaraunean edo corpus batean maiztasun handiena duena aukeratzen da.

Talvensaari et al.-ek [2007] corpus konparagarrietatik ateratako itzulpen-ezagutza baliatzen dute kontsultak itzultzeko. Zehazki, antzekotasun-thesaurus modura erabiltzen dute beraiek eraikitako corpus konparagarri bat. Hiztegian oinarritutako itzulpen-sistema eta thesaurus-ean oinarritutakoa konbinatuz emaitzak hobeak lortzen dituzte banaka erabiliz baino.

Rahimi eta Shakery-k [2013] informazioaren berreskurapenean erabilitako hizkuntza-ereduen hurbilketa hartzen dute kontsulta itzultzeko. Lan honetan corpus konparagarria da itzulpenak lortzeko iturburu bakarra. Hurbilketaren hipotesia da itzulpen-baliokideek antzeko ekarpenak dituztela corpus konparagarritik lerrokatutako dokumentuen hizkuntza-ereduetan. Antzekotasuna jatorrizko eta helburu-hitzen ereduen arteko KL-dibergentzian oinarrituta dago. Berreskurapen elebakarraren %43.32ko eraginkortasuna lortzen dute, hiztegian oinarritutako oinarri-lerroa (*baseline*) gaindituz.

Corpus konparagarriak biltzeko prozesuaren inguruan hainbat artikulu aurki daitezke bibliografian. Baliabide hauek zenbait hizkuntzaren prozesamenduko atazatan ustiatzen dira. Artikulu gehienetan amarauna corpus-iturburu gisa baliatzea proposatzen da [Talvensaari et al., 2008, Klementiev and Roth, 2006].

Tesi-lan honetan kontsultaren itzulpen-prozesua hobetzeko ustiatu ditugu corpus konparagarriak. Zehazki, amarauneko bilatzaileen bidez bildutako testuinguru konparagarrietatik itzulpen-probabilitateak estimatzeko estrategia berri bat proposatzen dugu. [Sadat et al., 2003] eta [Talvensaari et al., 2007] lanetan bezala, hizkuntza arteko hitzen distantziak kalkulatzen ditugu corpus konparagarrietatik, itzulpen-probabilitateak estimatzeko soilik gure kasuan. Amaraunetik bildutako hizkuntza arteko testuinguru konparagarriak baliatzea izango litzateke lan horiekiko diferentzia nabarmenena. Halaber, amarauneko bilatzaileetatik bildutako testuinguruen adierazgarritasun linguistikoaren azterketa berri bat landu dugu. Esperimentu hauek IV. kapituluan azalduko ditugu.

II.4 Kontsulta-saioen informazioa ustiatzen

Azkeneko urteotan kontsulta-saioak ranking hobeak lortzeko baliatu daitezkeela defendatu dute egile batzuek [Carterette et al., 2011]. Egile hauen ustez erabiltzaileek kontsulta bat baino gehiago behar izaten dituzte beren informazio-beharra asetzeko. Informazio-behar bati lotutako prozesuak kontsulta bat baino gehiago hartzen dituela diote.

Horrela, hasierako kontsulta q_i sartu ondoren, erabiltzaileek kontsulta hau birformulatzeko q_r joera izaten dute, hainbat modutan:

- Zehaztea: Hasierako kontsulta zehaztea (adib., q_i=*"Eszitia"* q_r=*"Eszitia mitologian"*).
- Orokortzea: Hasierako kontsulta orokortzea (adib., q_i=*"ordenadore harra"* q_r=*"malware"*).
- Dibagazioa: Hasierako kontsultarekin zerikusia duen beste kontsulta bat formulatzea (adib., q_i=*"eguzki-orbanen aktibitatea"* q_r=*"eguzki orbanak eta lurrikarak"*).

Bilatzaileen *log* fitxategien gainean egindako zenbait azterketak frogatu dute erabiltzaileen erdiak birformulatzen dutela hasierako kontsulta. Adibidez, 1997ko eta 2001eko Excite datu-multzoetan erabiltzaileen %52k eta %45ak egina zituzten birformulazioak [Wolfram et al., 2001].

Kontsulta-saioetan oinarritutako informazioaren berreskurapena lantzeko helburuarekin abian jarri zuten 2010ean *Session Track*[1] delako ataza TREC lehiaketaren barruan. 2010eko edizioan [Kanoulas et al., 2010] bi kontsultako saioen ustiaketa aztertu zen. Lehiakideek birformulazioen berreskurapen-prozesua hobetu behar zuten hasierako kontsultak ustiatuz. Orokorrean, lehiaketan aurkeztutako sistemen eraginkortasuna hobea zen orokortze eta dibagazio motako birfomulazioak ustiatuz zehazte motakoak ustiatuz baino. Dena dela, partaide batek soilik lortu zuen hobekuntza estatistikoki esanguratsua, hasierako kontsultak ustiatu gabeko sistema baten gainean. Hurrengo edizioetarako saio luzeagoak eraiki ziren atazarako. 2011 eta 2012 urteetako edizioetan partaideen erdiak lortu zuten hobekuntza esangursatsua saioen informazio erabilita, eta 2013 eta 2014ko edizioetan partaide gehientsuek lortu zuten hobekuntza esanguratsua.

Kontsulta-saioek emandako informazioaren erabilera CLIR ataza hobetzeko oso artikulu gutxitan aztertu da. Jarraian literaturan aurkitu ditugun lanak aipatuko ditugu.

Gao et al-ek [2007], adibidez, kontsulta-logak baliatzen dituzte hizkuntza arteko kontsultak proposatzeko (*query suggestion*) atazan. Jatorrizko hizkuntzako kontsulta eta helburu-hizkuntzako kontsultaren arteko antzekotasuna kalkulatzeko metodo berri bat proposatzen dute. Metodo honek, itzulpen-informazioaz gain, termino agerkidetzak, kontsulta-logak eta kliken datuak ustiatzen ditu. Hizkuntza arteko kontsulta-antzekotasunerako eredu diskriminatzaile baten bidez ikasten da eskuz itzulitako kontsultetatik.

[1] http://trec.nist.gov/data/session.html

[Hu et al., 2008] lanean kliken datuetan ezkutuan dagoen ezagutzatik kontsulten itzulpenak erauzteko metodo bat proposatzen dute. Lehenengo urrats batean kliken datuetan URLen patroi elebidunak aurkitzen dituzte. Ondoren, erabiltzaileen klikek duten portaeraren araberako kontsulten itzulpen-bikoteak topatzen dituzte. Kontsulten itzulpenak kliken datuetako agerkidetzak ustiatuz finkatzen dituzte.

Tesi-lan honetan Carterette et al.-ek [2011] zehaztutako birformulazio-motak ustiatu ditugu kontsultaren itzulpena hobetzeko. Zehazki, birformulatutako kontsulta baten itzulpen-prozesuan aurreko kontsulten informazioa baliatzea aztertu dugu. Denbora errealean, erabiltzaileak kontsultak egin ahala, lortu daiteken informazioa da. Hortaz, ez ditugu behar lehendik bildutako kliken datuak eta kontsulta-logak [Gao et al., 2007, Hu et al., 2008] lanetako metodoetan bezala. Esperimentuak V. kapituluan azalduko ditugu.

II.5 Pibotaje bidezko hiztegi elebidunen sorkuntza

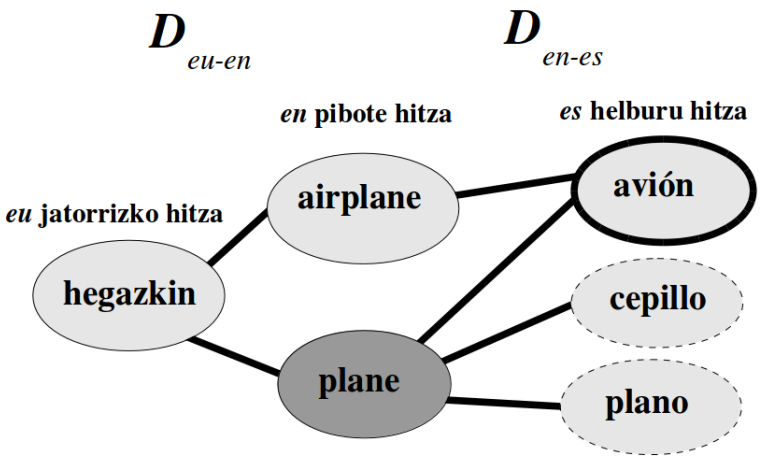
Hiztegi elebidunen sorkuntza oso prozesu luzea eta konplexua da, eskuz eginez gero. Baliabide urriko hizkuntzen kasuan bestelako hizkuntzetan baino hiztegi elebidun gutxiago daude eskuragarri. Hori dela eta, mota honetako hizkuntzek etekin handia atera diezaiekete hiztegien sorkuntza automatizatzeko metodoei.

Amaraunaren hazkundearen ondorioz Wikipedia [2] bezalako baliabideak lexiko elebidun berria erauzteko aukera ona dirudien arren [Erdmann et al., 2008], hiztegi baten eta entziklopedia baten artean alde handia dago. Wiktionary [3] lexikografiara zuzendutako etorkizun handiko baliabidea bada ere, baliabide urriko hizkuntzen estaldura mugatua eskaintzen du gaur egun.

Zabaldunke handiko hizkuntzak zubi edo pibote gisa erabiliz, hiztegi elebidunak sortzeko hainbat metodo proposatu dira literaturan. Pibotaje bidezko teknika hauek hiztegi elebidunak modu automatikoan sortzea ahalbidetzen dute. Ataza ez da hutsala, hitzen arteko itzulpen-prozesua ez baita guztiz trantsitiboa, hitz anbiguoen presentziak okerreko itzulpen-hautagaiak eragin baititzake, II.1 irudiko adibidean ikus dezakegunez.

Literaturan zenbait metodo proposatu dira itzulpen anbiguoen ondorioz sortutako okerreko itzulpenak kimatzeko. [Tanaka and Umemura, 1994] artikuluan pibotaje-prozesuan erabilitako hiztegien egituran soilik

[2] https://www.wikipedia.org/
[3] https://www.wiktionary.org/



II.1 irudia: D_{eu-en} eta D_{en-es} hiztegiak gurutzatuz eta *en* pibote erabiliz *"hegazkin"* hitzerako lortutako itzulpen-hautagaiak. Itzulpen okerrak ertz marradununeko borobiletan erakutsita.

oinarritutako metodo bat proposatzen dute, Alderantzizko Kontsulta (*Inverse Consultation* ingelesez) izenekoa. Alderantziko kontsulten bidez hitzen arteko distantzia semantikoa neurtzen dute konpartitzen dituzten pibote-hitzak kontatuz. Metodo honen hedapenak proposatu dira, klase semantikoak eta POS informazioa ere integratzen dituztenak [Bond et al., 2001, Bond and Ogura, 2008]. Sjöbergh eta al-ek [2005] definizioen konparaketaren ustiaketa aztertu zuten adiera bereko hitzak topatzeko. Zoritxarrez, klase semantikoak edo definizioak dituzten hiztegi elebidunak gutxi dira.

[István and Shoichi, 2009] artikuluan pibote-hizkuntzari dagokion WordNet [Miller, 1995] erabiltzea proposatzen dute. WordNet-ek eman dezakeen informazio semantikoa baliatzen dute pibotaje-prozesuaren doitasuna hobetzeko. Soderland et al. [2009] lanean pibote gisa hizkuntza bat baino gehiago batera erabiltzea aztertu zuten. Beraien hipotesia da zenbat eta pibote-hizkuntza gehiago erabili orduan eta ebidentzia gehiago izango dituztela itzulpen-baliokideak topatzeko.

[Tsunakawa et al., 2008] artikuluan corpus paraleloak erabili zituzten itzulpen-baliokide hautagaien arteko probabilitateak estimatzeko. Gutxieneko probabilitate batera iristen direnak hiztegian sartzeko moduko itzulpen zuzentzat hartzen dira.

Corpus paraleloak urriak dira planteatu dugun lan-eszenatokian, eta alternatiba moduan egile batzuk [Kaji et al., 2008, Gamallo and Pichel, 2010] corpus konparagarrietan oinarritutako metodoak proposatu dituzte. Corpus konparagarrietatik estimatutako hizkuntza arteko antzekotasun distribuzionala baliatzen dute okerreko itzulpen-baliokideak baztertzeko. [Shezaf and Rappoport, 2010] artikuluan antzekotasun distribuzionalaren aldaera bat proposatzen dute, haien esperimentuetan Alderantzizko Kontsultaren metodoa baino emaitzak hobeak ematen dituena.

[Wushouer et al., 2013] artikuluan familia bereko hizkuntzetarako heuristikoetan oinarritutako estrategia bat proposatzen dute. Probabilitatea, distantzia semantikoa eta antzekotasun ortografikoa erabiltzen dira heuristiko gisa pibotaje bidez lortutako itzulpen-hautagaiak kimatzeko.

Tesi-lan honetan hiztegien egituran oinarritutako metodoak [Tanaka and Umemura, 1994] eta corpus konparagarrietan oinarritakoak [Kaji et al., 2008, Gamallo and Pichel, 2010] konbinatzeko zenbait esperimentu eta horren gaineko ebaluazio sakonak egin ditugu, VI. kapituluan azalduko ditugunak.

II.6 IR eta CLIR euskararen gainean

Euskararen inguruan zenbait ikerketa burutu dira informazioaren berreskurapenaren alorrean. Jarraian esanguratsuenak aipatuko ditugu.

[Leturia et al., 2007] artikuluan EusBila aurkeztu zuten, lematizazioa eta hizkuntza-detekzioa inplementatzen zituen euskararako web-bilatzailea. EusBila bilatzaile komertzialen APIetan oinarrituta zegoen. Lematizazioa eta hizkuntza-detekzioa inplementatzeko kontsultaren hedapen morfologikoa eta hizkuntza-hitzak erabiltzen ziren hurrenez hurren. Bi teknika hauen ebaluazio sakona [Leturia et al., 2013] artikuluan azaltzen da.

Galdera-erantzun (ingelesez *Question Answering* edo QA) sistemak informazio soila berreskuratzeaz baino harago doaz. Informazio-beharra galdera baten bidez adierazten da, informazio adierazgarria galderari dagokion erantzuna delarik. Euskararako *Ihardetsi* izeneko galdera-erantzun sistema azaltzen da [Alegria et al., 2009] artikuluan. Sistemak hiru modulu ditu: galdera aztertzeko modulua, pasarteen berreskurapenerako modulua, eta erantzunen erauzlea. CLEF 2008ko[4] *Basque to Basque monolingual QA* izeneko atazan %13ko zehaztasuna lortu zuen sistemak [Ansa et al., 2008].

[4] http://clef.isti.cnr.it/2008.html

Ihardetsi sistemaren bertsio berriago bat prestatu zen CLEF 2009 eta 2010eko ResPubliQA[5] [Agirre et al., 2009, 2010] atazetarako. Atazaren ariketako bat hizkuntza arteko QA zen. Galdera aurreprozesatzeko Ihardetsiren modulua erabili zen. Galderaren itzulpena hiztegietan oinarritutako algoritmo bat erabiliz egin zen. Bildumako dokumentuak WordNet baliatuz hedatu ziren. 2010ko edizioan 0,30ko MRRa (*Mean Reciprocal Rank*) lortu zuen aurkeztutako sistemak (sistema elebakarraren eraginkortasunaren %50). [Otegi et al., 2015] lanean ezagutzan oinarritutako antzekotasun semantikorako teknikak baliatzen dituzte, kontsulta eta dokumentu-bildumako hiztegien arteko dibergentziei aurre egiteko. Euskararako zientzia eta teknologia domeinuko galdera-erantzun atazan oinarri-lerroko emaitzak gainditzen dituzte, %3,16an eta %3,72an erantzun zehatzak eta pasarteak berreskuratzeko.

Tesi honetan aurkezten dugun CLIR hurbilketaren baliabide giltzarria da hiztegi elebiduna. Izan ere, bere sorkuntzarako pibotaje bidezko metodo bat aurkezten dugu. Komeni da aipatzea lexiko elebidunen sorkuntzaren inguruan euskara aintzat hartzen duten zenbait ikerlan egin direla iraganean. [Gurrutxaga et al., 2006] lanean corpus paraleloetatik terminologia elebiduna erauzteko teknikak ikertu ziren. Erauzketarako proposatutako estrategiak hiru urrats nagusi ditu: 1) terminologia erauztea hizkuntza bakoitzean, 2) itzulpen-unitate bakoitzeko itzulpen-bikote hautagaiak sortzea, eta 3) itzulpen-unitate onenak kognatuen detekzioaz eta elkartze-neurrien bidez aukeratzea. *Itzulterm*[6] izeneko zerbitzu irekia da ikerlan horren emaitza. [Saralegi et al., 2008] lanean terminologia elebiduna corpus konparagarrietatik erauzteko teknikak landu ziren. Testuinguru-antzekotasuna eta kognatuen detekzioa baliatu ziren corpus konparagarrietatik ateratako itzulpen-bikoteak identifikatzeko.

II.7 Ebaluazioa

IR sistemen eraginkortasuna ebaluatzeko *Crandfield* [Cleverdon, 1967] izeneko metodologia edo ingurune estandarra erabiltzen da. Ingurune horrek honako osagaiak ditu: eduki-bilduma bat, gai-sorta bat, eta adierazgarritasun-epaiak. Atazan zenbat eta dokumentu adierazgarri gehiago itzuli rankingean, orduan eta arrakasta handiagoz beteko litzateke.

Literaturan hainbat neurri proposatu dira IR sistema baten eraginkortasuna estimatzeko. Doitasuna (*Precision*) eta estaldura (*Recall*)

[5] http://nlp.uned.es/clef-qa/repository/resPubliQA.php
[6] http://itzulterm.elhuyar.eus/

dira IR eszenatoki batean erabilitako neurri oinarrizkoenak. Doitasuna (P) IR sistema batek itzulitako dokumentu adierazgarrien kopurua zati itzulitako dokumentu guztien kopurua da (Ikusi II.3 formula). Estaldura (R) sistemak itzulitako dokumentu adierazgarrien kopurua zati bildumako dokumentu adierazgarri guztien kopurua da (Ikusi III.4 formula). F puntuazioa (*F score*) doitasuna eta estaldura bateratzen dituen neurria da (Ikusi II.5 formula). β parametroaren bidez doitasunari eta estaldurari ematen zaien pisua zehazten da. $\beta = 1$ pisu berdina ematen zaie biei, $\beta = 2$ pisu gehiago doitasunari, eta $\beta = 0.5$ pisu gehiago estaldurari.

$$P(N) = \frac{\sum_{i=1}^{N} rel(i)}{N} \tag{II.3}$$

$$R(N) = \frac{\sum_{i=1}^{N} rel(i)}{R} \tag{II.4}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{P(N) \cdot R(N)}{(\beta^2 \cdot P(N)) + R(R)} \tag{II.5}$$

Badaude bestelako neurri osatuagoak sistemak itzulitako rankingean dokumentu adierazgarriek rankingean dituzten posizioei erreparatzen dietenak. Horien artetik, *MAP* (*Mean Average Precision*) eta *DCG* (*Discounted Cumulative Gain*) dira IRko esperimentuetan erabilienak.

MAP *AP*an (*non-interpolated average precision*) dago oinarrituta:

$$AP = \frac{\sum_{i=1}^{N} P(i) \cdot rel(i)}{R} \tag{II.6}$$

Sistemak itzulitako dokumentu kopurua da N, $rel(i)$ rankingeko i. dokumentuak duen adierazgarritasun-maila (1 edo 0) da, $P(i)$ i. posizioan neurtutako doitasuna da, eta R kontsultarekiko bilduman dauden dokumentu adierazgarrien kopurua da. *AP*k ranking posizio guztietako doitasunak kalkulatu eta haien batezbestekoa hartzen du. *MAP* sistema ebaluatzeko erabili nahi diren kontsulta guztien *AP*en batezbestekoa da.

MAP IRko ikerketa gehienetan neurririk erabiliena izan da duela gutxi arte. Azkeneko urteotan ikerlari batzuek zalantzan jarri dituzte epaiketa boolearretan oinarritutako ebaluazioak eta neurriak. Ildo horretatik Jarvelin eta Kekalainen-ek [2000] *CG* (*Cumulative Gain*) eta haren aldaera batzuk

proposatu zituzten. Sistemak itzulitako lehenengo dokumentuei dagozkien adierazgarritasun-mailen ($rel(i)$) batura da CG, adierazgarritasun-maila bi baliotik gorako eskala batean egon daitekeelarik.

$$CG = \sum_{i=1}^{N} rel(i) \tag{II.7}$$

CGk ez ditu kontuan hartzen rankingeko dokumentuen posizioak. Hau problema bat da erabiltzailearen ikuspuntua ez delako ondo jasotzen. Sistema batek, dokumentu adierazgarriak lehenengo posizioetan itzultzen baditu, eraginkortasun hobea dauka erabiltzailearen ikuspuntutik. Hori dela eta, DCG aldaera proposatzen da, non dokumentuen adierazgarritasun-mailak zigortzen diren posizio atzeragoko dokumentuetakoak izanez gero:

$$DCG = \sum_{i=1}^{N} \frac{rel(i)}{log_b(i)} \tag{II.8}$$

Tesi-lan honetan MAP eta DCG neurriak erabili ditugu III. IV. eta V. kapituluetan azaldutako CLIR esperimentuetan. VI. kapituluan aurkeztutako pibotaje bidezko hiztegien eraikuntzarako metodoen ebaluazioan doitasuna, estaldura, eta F puntuazioa neurriak baliatu ditugu.

II.8 Erabilitako baliabideak

Esperimentuetan erabilitako baliabideak azalduko ditugu kapitulu honetan.

Indri tresna

Tesi-lan honetako III. IV. eta V. kapituluetan aurkeztutako esperimentuetan Indri tresnak eskaintzen duen Indri berreskurapen-algoritmoa [Zhai and Lafferty, 2001] erabili zen izen bereko tresnaren[7] bidez. Algoritmoa kontsulta itzulitakoan aplikatzen da dokumentu adierazgarrien rankinga

[7] https://www.lemurproject.org/indri.php

kalkulatzeko. Indri algoritmoa hizkuntza-ereduen eta inferentzia-sareen berreskurapen-frameworken konbinazio bat da.

V. kapituluko esperimentuan, non Clueweb bilduma erabiltzen den, Indri-ren *Batch Query service*[8] izeneko zerbitzua erabili zen. Web-zerbitzu honek Clueweb bilduma Indri algoritmoaren bidez kontsultatzeko aukera ematen du.

CLEF lehiaketako datu-multzoak

III. eta IV. kapituluetako esperimentuetan CLEF lehiaketaren barruan prestatutako datu-multzoak erabili dira. LA Times 1994 (113.005 dokumentu) eta Glasgow Herald 1995 (56.472 dokumentu) ingelesezko testu-bildumak erabili ziren. CLEF 2001eko gai-multzoa (41-90 tarteari dagokiona) IR sistemen parametroak doitzeko eta CLEF 2005ko eta 2006ko gai-multzoak (251-300 eta 301-350 tarteei dagozkionak) IR sistemak ebaluatzeko erabili ziren. Gai guztiak euskarara itzuli ziren euskara-ingelesa hizkuntza arteko berreskurapen-ataza ebaluatu ahal izateko. Gai batek honelako eremuak ditu:

- Izenburu labur bat (*title*), erabiltzaileak formulatzen dituen kontsulten modukoa.
- Informazio-beharraren deskribapen labur bat (*desc*), esaldi bakarrekoa.
- Informazio-beharraren deskribapen sakonago bat (*narr*), nahi diren dokumentuak zehatzago deskribatuz.

CLEF AdHoc-News Test Suites (2004-2008) - Evaluation Package izeneko ELRAko paketean[9] aurkitu ditzakegu arestian aipatutako bildumak eta gaiak (euskarazkoak izan ezik).

TREC Session Track 2010eko datu-multzoa

TREC 2010 Session track atazarako prestatutako kontsulta-saioak[10] ingelesetik euskarara itzuli ziren. Guztira bi kontsultaz osatutako 150 kontsulta-saio dira. Kontsultak Clueweb09[Callan et al., 2009] bildumari dagozkio. Kontsulta-saio hauek V. kapituluko esperimentuetan erabili dira.

[8] http://lemurproject.org/clueweb09/
[9] http://catalog.elra.info/retd/product_info.php?products_id=949
[10] http://trec.nist.gov/data/session10.html

<**title**> Impact of foreign textile imports on U.S. textile industry <**desc**> Document must report on how the importation of foreign textiles or textile products has influenced or impacted on the U.S. Textile industry. <**narr**> The impact can be positive or negative or qualitative. It may include the expansion or shrinkage of markets or manufacturing volume or an influence on the methods or strategies of the U.S. textile industry. "Textile industry"includes the production or purchase of raw materials; basic processing techniques such as dyeing, spinning, knitting, or weaving; the manufacture and marketing of finished goods; and also research in the textile field.

II.2 taula: CLEF datu-multzuko gai baten adibidea.

SemCor 1.6

Semantikoki etiketatutako testu-bilduma bat da SemCor[11]. Eskuzko anotazioa WordNet 1.6 baliabidearen arabera eginda dago SemCor bertsio honetan. IV. kapituluan erabilitako baliabidea da, web bilatzaileen bidez bildutako testuinguruen adierazgarritasun linguistikoa neurtzeko.

Corpus konparagarriak

Pibotaje bidezko hiztegien sorkuntzako esperimentuetarako (VI. kapituluan aurkezten direnak) euskara-gaztelania corpus konparagarri bat eraiki zen. Berria eta Diario Vasco egunkari digitalak iturburu gisa hartuta bildu genuen corpusa. Konparagarritasun-maila handia lortze aldera, denbora tarte bereko (2006-2010) albisteak soilik hartu ziren. Guztira, 149.892 eta 306.925 albiste bildu ziren Berria eta Diario Vascotik hurrenez hurren.

[11] http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor

III. KAPITULUA

Kontsultak itzultzeko hiztegietan oinarritutako metodoen azterketa eta konparaketa

[Saralegi and Lopez de Lacalle, 2009] artikuluan kontsultak itzultzeko bi teknikaren azterketa eta konparaketa azaltzen dira. Itzulpen-hautagaiak hiztegi elektroniko batetik eskuratzen dira. Itzulpen anbiguoak ebazteko bi metodo aztertu eta alderatu dira: Pirkolaren metodoa [Pirkola, 1998], eta agerkidetzan oinarritutako algoritmo iteratibo bat [Monz and Dorr, 2005]. Bien metodo hibrido bat ere aztertu da. [Saralegi and Lopez de Lacalle, 2010a] artikuluan hiztegien bidezko itzulpen-estrategian gertatzen diren erroreak aztertzen dira. Errore bakoitzak berreskurapen-prozesuaren eraginkortasunean duen eragina aztertu eta erroreak berak tratatzeko metodoak proposatu eta ebaluatzen dira.

- Xabier Saralegi and Maddalen Lopez de Lacalle. Comparing different approaches to treat translation ambiguity in CLIR: Structured queries vs. target co-occurrence based selection. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 398–404. IEEE, 2009

- Xabier Saralegi and Maddalen Lopez de Lacalle. Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010a

III.1 Comparing different approaches to treat translation ambiguity in CLIR: Structured queries vs. target co-occurrence based selection

Comparing different approaches to treat Translation Ambiguity in CLIR: Structured Queries vs. Target Co-occurrence Based Selection

X. Saralegi, M. Lopez de Lacalle
Elhuyar R&D
Zelai Haundi kalea, 3.
Osinalde Industrialdea, 20170 Usurbil. Basque Country
{xabiers, maddalen}@elhuyar.com

Two main problems in Cross-language Information Retrieval are translation selection and the treatment of out-of-vocabulary terms. In this paper, we will be focusing on the problem concerning the translation selection. Structured queries and target co-occurrence-based methods seem to be the most appropriate approaches when parallel corpora are not available. However, there is no comparative study. In this paper we compare the results obtained using each of the aforementioned methods, we specify the weaknesses of each method, and finally we propose a hybrid method to combine both. In terms of mean average precision, results for Basque-English cross-lingual retrieval show that structured queries are the best approach both with long queries and short queries.

III.1.1 Introduccion

The importance of Cross-language Information Retrieval (CLIR) nowadays is patent in multiple contexts. In fact, communication is more global, and the access to multilingual information is more and more widespread within this globalized society. However, unless some *lingua franca* is established in specific geographic areas and discourse communities, it is still necessary to facilitate access in the native speaker's language.

In our case, we are developing a CLIR system to allow Basque speakers to access texts in other languages. Since Basque has relatively few speakers (about 1,000,000) CLIR is an attractive technology for providing Basque speakers access to those global contexts. Even though lately most extended CLIR approaches are based on parallel corpora, Basque is a less resourced language, and that is why we have to turn our gaze to parallel corpora free approaches. The work presented in this paper compares the performance of

two methods for the translation selection problem which do not require the use of parallel corpora. In addition, we have also designed and evaluated a hybrid algorithm that combines both methods in a simple way.

The CLIR topic and its problematic are introduced in the next section. Section III.1.3 addresses the specific problem of the translation selection. The two approaches proposed for dealing with the translation ambiguity are presented in subsections III.1.3.1 and III.1.3.2 Following (subsection III.1.3.3), we propose a simple combination of both methods. Then, in section III.1.4 we evaluate and compare the different methods for the Basque-English pair, in terms of MAP (Mean Average Precision) and using CLEF (Cross Language Evaluation Forum) collections and topics. Finally, we present some conclusions and future works in section III.1.5.

III.1.2 The translation methods for CLIR

CLIR does not differ too much from Information Retrieval (IR) and only the language barrier requires specific techniques, which are mainly focused on the translation process. The different approaches differ essentially with respect to which available information is translated (queries, documents or both), and in the method used to carry out the translation.

There are three strategies for tackling a cross-language scenario for IR proposes: a) translating the query into the language of the target collection, b) translating the collection into the language of the source query, and c) translating both into an interlingua. The majority of the authors have focused on translating queries mainly due to the lower requirements of memory and processing resources [Hull and Grefenstette, 1996]. However, richer context information is useful for dealing with disambiguation problems, and it has been proved that the quality of the translation and retrieval performance improve when the collection is translated [Oard, 1998]. Translating both queries and documents into an interlingua provides even better results [McCarley, 1999, Chen and Gey, 2003].

As for the translation methods, they can be classified into three main groups: Machine Translation (MT)-based, parallel corpus-based, and bilingual Machine Readable Dictionary (MRD)-based. In general, authors point out that using MT systems is not adequate for several reasons: the quality of precision is often poor and the system requires syntactically well formed sentences, while in IR systems the queries are often sequences of words [Hull and Grefenstette, 1996].

The corpus-based approach implies the use of parallel (and also

comparable) corpora to train statistical translation models [Hiemstra, 2001]. The main problem is the need for large corpora. The available parallel corpora are usually scarce, especially for minority languages and restricted domains. The advantage of this approach is that the translation ambiguity can be solved by translating the queries by statistical translation models. Comparable corpora, which are easier to obtain, can be used in order to improve the term coverage [Talvensaari, 2008].

Lastly, MRD-based translation guarantees enough recall but does not solve the translation ambiguity problem. Thus, two main problems arise when using dictionaries to translate: ambiguities in the translation, and also the presence of some out-of-vocabulary terms. Many papers have been published about these two issues when queries are translated [Knight and Graehl, 1997, Ballesteros and Croft, 1998, Gao et al., 2001, Monz and Dorr, 2005].

Among the displayed alternatives, the MRD-based approach has been explored, because of the lack of sufficient parallel corpora for Basque, and because we assume that this situation will be similar for other minority languages. Specifically, we have concentrated on testing two methods to deal with translation ambiguity: structured queries and co-occurrence-based methods. Although the influence level of the errors derived from using dictionaries depends on the quality of the resources used and the tasks done, Qu et al. [2000] point out that the wrong translation selection is the most frequent error in an MT-Based translation process. So, we assume that this error distribution will be similar in MRD-based systems.

We have translated only the queries in our experiments. The reasons for this decision are, on the one hand, that the methods we want to analyze have been tested in such an experimental setup. On the other hand, the results of this research will be used for the development of a commercial web searcher, and so the processing and memory consumption are also important factors.

III.1.3 Selecting the correct translation from a dictionary

In order to deal with the translation selection problem affecting queries derived from bilingual dictionaries (MRD), there are several methods proposed in the literature. An extended approach to tackle the problem of ambiguity is by using structured queries, also called Pirkola's method [Pirkola, 1998]. All the translation candidates are treated as a unique token in the calculation of relevances estimating term frequency (TF) and document frequency (DF) statistics separately. Thus, the disambiguation takes place implicitly during the retrieval instead of during the query

formulation. A more advanced variant of this algorithm, known as probabilistic structured queries [Darwish and Oard, 2003], allows to weight the different translation candidates offering better performance.

Other approaches to tackle ambiguity in query translation are based on exploiting statistically monolingual corpora in the target language. Specifically, these methods try to select the most probable translation of the query, choosing the set of translation candidates that most often co-occur in the target collection. The algorithms differ in the way the global association is calculated and in the translation unit used (e.g., word, noun phrases...).

In [Ballesteros and Croft, 1998] a co-occurrence method and a technique using parallel corpora are compared, leading to the conclusion that the co-occurrence method is significantly better at disambiguating than the parallel corpus-based technique. In [Gao et al., 2002], the basic co-occurrence is extended by adding a decaying factor that takes into account the distance between the terms when calculating their Mutual Information. Hence, if the distance between the terms increases, the decaying factor does too. In the basic co-occurrence model, when calculating the coherence for a translation candidate, not only are the selected translations taken into account, but also those which are not selected. Liu et al. [2005] propose a statistical model called "maximum coherence model" that estimates all the translations of all query terms simultaneously and these translations maximize the overall coherence of the query. In this case, the coherence of a translation candidate is independent from the selection of other query terms translations. This new model is compared with a co-occurrence model similar to the one proposed by [Gao et al., 2001], which takes into account all the translations of the rest of words in the query. The model that they propose performs substantially better, but it is computationally very expensive. Jang et al. [1999] propose a co-occurrence method that only takes into account the consecutive terms when calculating the mutual information. Monz and Dorr [2005] introduce an iterative co-occurrence method which combines term association measures with an iterative machine learning approach based on expectation maximization.

This work compares two alternatives proposed in the literature which do not require parallel corpora. The unique resources used are a bilingual MRD and a corpus in the target language for the co-occurrence-based method, which makes them suitable for less resourced languages like Basque. We have chosen a specific method for each approach: Pirkola's method, and a co-occurrence-based method. Among all the co-occurrence-based algorithms we have chosen the Monz and Dorr's algorithm assuming that being iterative yields better estimations, although we do not have any references that

confirm this. In addition, we have designed an algorithm that combines both approaches. In this last case, we have used Darwish and Oard's probabilistic structured queries as a framework and Monz and Dorr's algorithm to estimate the weights of the translation candidates.

III.1.3.1 Dealing with ambiguous translations using Structured Queries

The #syn operator of structured queries is a suitable technique for dealing with ambiguous translations because among other things it is fast, offers good results and does not need external resources such as parallel corpora. The basic idea is to group together the translation candidates of a source word, thus making a set and treating them as if they were a single word in the target collection [Pirkola, 1998]. Hence, when estimating the term frequency (TF) and document frequency (DF) statistics for query terms, the occurrences of all the words in the set are counted as occurrences of the same word. If we assume that s_i is a query term, D_k is a document term, d is a document and $T(s_i)$ is the set of translation candidate terms of s_i given by the MRD.

$$TF_j(s_i) = \sum_{\{k|D_k \in T(s_i)\}} TF_j(D_k) \tag{III.1}$$

$$DF(s_i) = |\bigcup_{\{k|\in D_k \in T(s_i)\}} \{d|D_k \in d\}| \tag{III.2}$$

where $TF_j(s_i)$ is the term frequency of s_i in document j, and $DF(s_i)$ is the number of document that contain s_i.

If the translation candidates are correct or semantically related, the effect is an expansion of the query. The problem arises especially when wrong translations that are common words occur, because DF of the #syn set can take high scores and the correct translation loses weight in the retrieval process. TF statistics can also be altered when wrong translations appear in the retrieval documents. But the probability that many wrong translations occur in retrieved documents is low. That is what we call retrieval time translation selection.

In order to test this method in development experiments, we have prepared a list of Basque topics translated from the English ones belonging to the CLEF 2001 edition (41-90), and the LA Times 94 collection and the corresponding relevance judgments, which will be explained more fully in section III.1.4. First, we have calculated the MAP for different numbers of translation candidates from the MRD (figure III.1), because a high coverage

of translations and the precision level of the MRD affects the performance of this method [Larkey et al., 2002]. Moreover, the translation equivalents of source words are usually ordered by frequency use in a MRD. Therefore, we can exploit that order to prune the least probable translations in the interests of query translation precision.

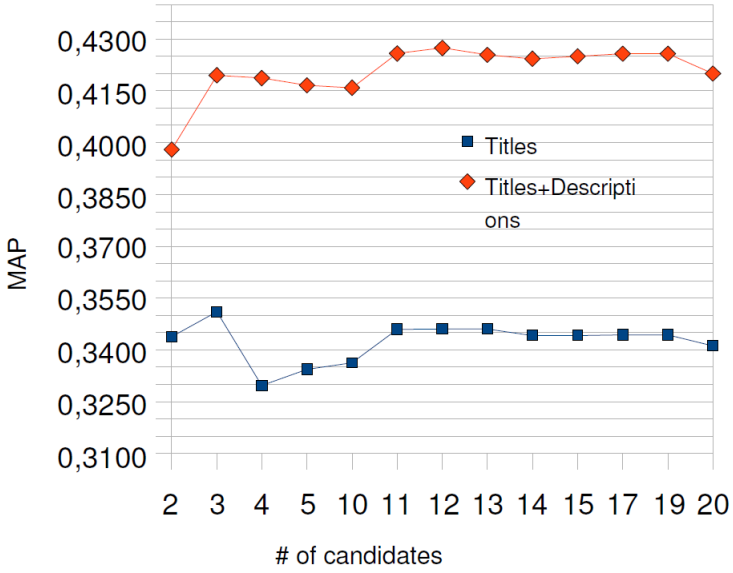


Figure III.1: MAP values for different numbers of translation-candidates.

In the graph (see figure III.1), we can see how the number of translation candidates from the MRD accepted for each source word affects the MAP. MAP curves are similar for both titles and titles+descriptions queries. They have local maximum in near points but the maximum global is reached by taking more candidates with the title+description set. The maximum MAP is achieved by taking the first three candidates for short queries, and the twelve first candidates for the long queries. This seems logical because there are more context words that can improve the retrieval-time disambiguation.

III.1.3.2 Target co-occurrence-based selection

As explained above, structured queries do not really do translation selection, and translations and statistics (TF and DF) can be wrong in some cases and decrease the retrieval performance. An alternative to executing the translation selection without using parallel corpora is to guide the selection

by using statistics of the co-occurrence of the translation candidates in the target collection. The basic idea is to choose the ones that co-occur more frequently, assuming that the correct translation equivalents of query terms are more likely to appear together in target document collection than incorrect translation equivalents. The main problem of this idea is to compute that global correlation in an efficient way, because the maximization problem is *NP-hard*.

The algorithm we have used for the translation selection is the one introduced by Monz and Dorr [2005]. Basically, it selects the translation candidates combination which maximizes the global coherence of the translated query by means of an *EM* (Expectation Maximization) type algorithm.

Initially, all the translation candidates are equally likely. Assuming that t is a translation candidate for a query term s_i given by the MRD, then:

Initialization step:

$$w_T^0(t|s_i) = \frac{1}{|tr(s_i)|} \tag{III.3}$$

In the iteration step, each translation candidate is iteratively updated using the weights of the rest of the candidates and the weight of the link connecting them. Iteration step:

$$w_T^n(t|s_i) = w_T^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} w_L(t,t') \cdot w_T(t'|s_i) \tag{III.4}$$

where *inlink(t)* is the set of translation candidates that are linked to t.

After re-computing each term weight they are normalized.

Normalization step:

$$w_L^n(t|s_i) = \frac{w_L^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w_L^n(t_{i,m}|s_i)} \tag{III.5}$$

The iteration stops when the variations of the term weights become smaller than a predefined threshold.

There are different association measures to compute the association strength between two terms ($w_L(t,t')$). We experimented with Mutual Information and Log Likelihood Ratio, and obtained the best results with the second one. That is the measure we use in the evaluation.

The question is whether by choosing the best translation of each query term we obtain a better MAP than grouping all the translation candidates by means of structured queries. As mentioned before, although in the structured

queries some weights and translations can be wrong, an expansion that can benefit the MAP is also produced. For example, for the Basque query *"gene gaitz"*, when we select the best English translation *"gene disease"* and run it, we obtain an AP of 0.5046. However, when all the translation candidates given by the MRD are put in sets with the #syn operator, *gene #syn(harm disease flaw ailment hurt malady defect difficult)*, even if we incorporate incorrect translations, we get a greater AP value, 0.5548. So, in this example it is clear that the noise expanded translation gives a higher AP score than the best translation. Nevertheless, for the Basque query *"gose greba"* we construct a translated query like *#syn(hunger yearning desire famine urge ravenous craving famished hungry) #syn(#1(work stoppage) strike walkout)* obtaining an AP of 0.0741. Whereas if we choose the best translation manually, we get the query *"hunger strike"* and obtain an AP of 0.6743. Looking at this example, it seems that our co-occurrence method could provide a margin for improving the MAP compared with structured queries when query terms have many incorrect translation candidates. In order to estimate whether this case is general, a lexicographer manually disambiguated some Basque queries (built from 41-90 CLEF queries) translated into English by an MRD. We preprocessed the queries by keeping only the lemmas of content words and then translated them using the MRD. The work by the lexicographer was to select the best translation candidate for each source term of the queries (example on table III.1).

Then, we calculated the MAP by processing Basque queries (table III.2) (titles and titles+description separately) for the different translation methods including the manual-based one. The MAP results show the MAP obtained by manual disambiguation does not reach that obtained using structured queries. So it seems that there is no margin for improvement for the co-occurrences-based method. However, the co-occurrences-based method outperforms structured queries when we are dealing with short queries. It even outperforms the theoretical threshold marked by the manual disambiguation. It could be due to a more statistical selection of short queries, more adequate for relevances in that collection.

III.1.3.3 Combining structured queries and co-occurrence-based algorithm

We think that we could take advantage of both techniques. Structured queries contribute to the translation less restrictiveness and query expansion in the retrieval phase, and the co-ocurrence-based method contributes translation selection and weighting capability. To do this, we propose that probabilistic structures queries [Darwish and Oard, 2003] be used, and

English query	Tainted-Blood Trial
Basque query	kutsatutako odolaren epaia
English query	Tainted-Blood Trial
Basque query (content words)	kutsatu odol epai
Structured translation into English	#syn(pollute impregnate infect) #syn(blood kinship) #syn(sentence crest judgment ridge notch scratch mark cut incision)
Best manual translation	#syn(pollute impregnate infect) #syn(blood kinship) #syn(sentence crest judgment ridge notch scratch mark cut incision)
Best manual translations	#syn(pollute infect) blood sentence #syn(pollute ~~impregnate~~ infect) #syn(blood ~~kinship~~) #syn(sentence ~~crest~~ ~~judgment~~ ~~ridge~~ ~~notch~~ ~~scratch~~ ~~mark~~ ~~cut~~ ~~incision~~)

Table III.1: Selecting the best translation of the structured query.

Translation method	MAP	
	Titles	Titles + description
English monolingual	0.4639	0.4912
Structured query (3 and 13 candidates)	0.3510	0.4274
Structured query (all candidates)	0.3352	0.4200
Best manual translation	0.3218	0.4127
Concurrences-based	0.3564	0.3908
Best manual translations	0.3471	0.4308
Probabilistic structured query	0.3568	0.4268
Probabilistic structured query+threshold (0.8)	0.3594	0.4249

Table III.2: MAP results for 41-90 topics.

the weights be estimated according to Monz and Dorr's algorithm. Thus, assuming $w_L(D_k|s_i)$ as the weight for the translation candidate D_k of a term s_i of a source query s we estimate TF and DF in this way:

$$TF_j(s_i) = \sum_{\{k|D_k \in T(s_i)\}} TF_j(D_k) \cdot w_L(D_k|s_i) \quad \text{(III.6)}$$

$$DF_j(s_i) = \sum_{\{k|D_k \in T(s_i)\}} DF_j(D_k) \cdot w_L(D_k|s_i) \quad \text{(III.7)}$$

As we did in subsection III.1.3.2, in order to estimate the possible improvement margin of this method, a lexicographer manually removed the wrong translations of the development queries, while maintaining only the correct ones (see table III.1). We maintained all the possible candidates since this method is capable of selecting more than one candidate. Thus, for the Basque query *"gene gaitz"* (*"gene disease"* in English) we obtained a query (*gene #syn(disease ailment malady)*) achieving an AP of 0.5946. A higher score than the one achieved taking all candidates. However, contrary to what we expected, the MAP for 41-90 topics is not much higher than that achieved without doing any kind of selection (although pruning some translations of the MRD can be considered to be a general disambiguation method) for long queries and for short queries it is even worse (see table III.2). Therefore, better quality in the translations does not seem to imply a big improvement in MAP. A further analysis will be conducted in the next section.

III.1.4 Evaluation and discussion

We evaluated the proposed translation methods using the collection from CLEF 2001 composed by LA Times 94 and Glasgow Herald 95. We translated from English to Basque two sets of topics: one for development (41-90) and the other one for test purposes (250-350). MAP values are calculated automatically with respect to existing human relevance judgments for queries and documents of the collections. The translation of the topics was carried out by professional translators and correctors of the Elhuyar foundation. The process was done in two steps: firstly, a translator translated the English topics into Basque, and then a corrector corrected the translations in order to minimize the possible bias -and the possible lack of naturalness- caused by the translation process.

We used the Indri as ranking model and the Porter Stemmer both for collections and translated topics. Before applying the proposed translation

methods we removed words like *"documentuak...(documents)"* and selected the content words manually. Specifically, nouns, adjectives, verbs and adverbs. Postpositions like *"artean (between), buruz (about)..."* were also removed. We used a Basque-English MRD which includes 34167 entries. For the treatment of OOV (Out-Of-Vocabulary) words we looked for their cognates in the target collection. Transliteration rules (see figure III.2) were applied and then LCSR (Longest Common Sequence Ratio) was computed. Those which reached a threshold (0.8) were taken as translation candidates in the translation phase.

ph- ⇨ f-, phase=fase
-tion ⇨ -zio, action=akzio

Figure III.2: Examples of transliteration rules.

The runs were done by taking the titles as queries (short queries), and also by taking the titles and descriptions as queries (long queries) and carrying out Basque to English translation:

1) Monolingual: Titles and titles+descriptions of CLEF 250-350 English topics.

2) First translation: First translation from dictionary

3) Structured query: Group translation candidates from the dictionary in a #syn set using Pirkola's method.

4) Structured query (Optimized dictionary): first translation candidates of the dictionary grouped in a #syn set (three for titles and twelve for the titles+descriptions maximize MAP on development experiments) using Pirkola's method.

5) Co-occurrence-based translation: Best translation selected by Monz and Dorr's co-occurrence-based algorithm.

6) Probabilistic structured query: all translation candidates of the dictionary grouped in a #wsyn set using Darwish and Oard's method, and weighted according Monz and Dorr's co-occurrence-based algorithm.

7) Probabilistic structured query+threshold: Best translations selected according to a threshold and weighted by Monz and Dorr's

co-occurrence-based algorithm and grouped by #wsyn set using Darwish and Oard's method.

The results are presented in table III.3 and figures III.3 and III.4.

Run	MAP		% of Mon.		Improvement Over First %	
	Short	Long	Short	Long	Short	Long
English monolingual	0.3176	0.3778				
First	0.2118	0.2500	67	66		
Structured query	0.2342	0.2959	74	78	9.56*	15.51*
Structured query (optimized dictionary)	0.2359	0.2960	74	78	10.22*	15.54*
Co-occurrences -based	0.2338	0.2725	74	72	9.41*	8.26*
Probabilistic structured queries + threshold	0.2404	0.2920	76	77	11.9*	14.38*
Probabilistic structured queries	0.2371	0.2941	75	78	10.67	14.99*

Table III.3: MAP values for 250-350.

The achieved MAP is higher with long queries than with short queries in both cases, monolingual and cross-lingual. In the cross-lingual retrieval the translation methods proposed also offer greater improvement with long queries. This is logical because more context words help in the translation selection. Unlike the results in the development experiments, the methods do not show a different performance depending on the length of the queries. We have examined the queries translated by Monz and Dorr's method and the quality is quite adequate except for a few cases due to false associations. For example, the Basque query *"kutsatu odol **epai**"* is translated as *"infect blood **cut**"* by Monz and Dorr's method instead of *"infect blood **sentence**"*. We can assume that it happens due to the stronger relation between -*epai* source word's translation candidate and *infect* and *blood* and *cut* -*epai* source

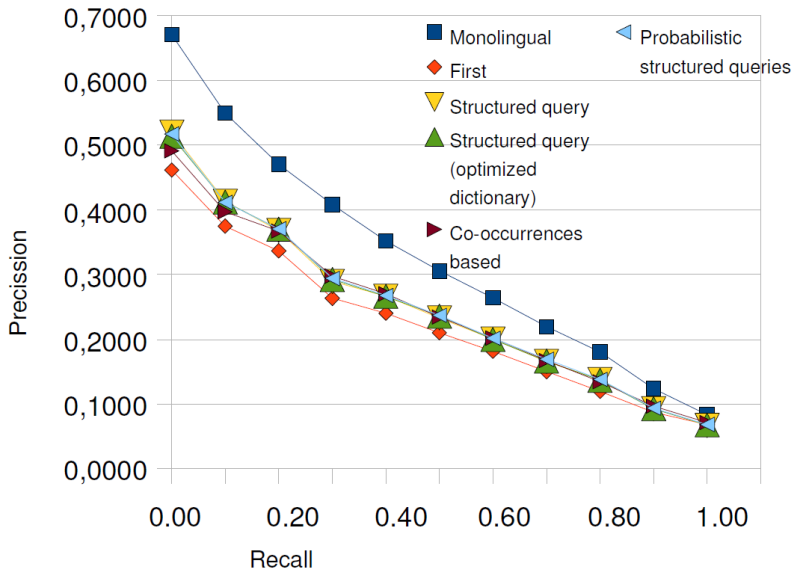


Figure III.3: PR curves (Titles).

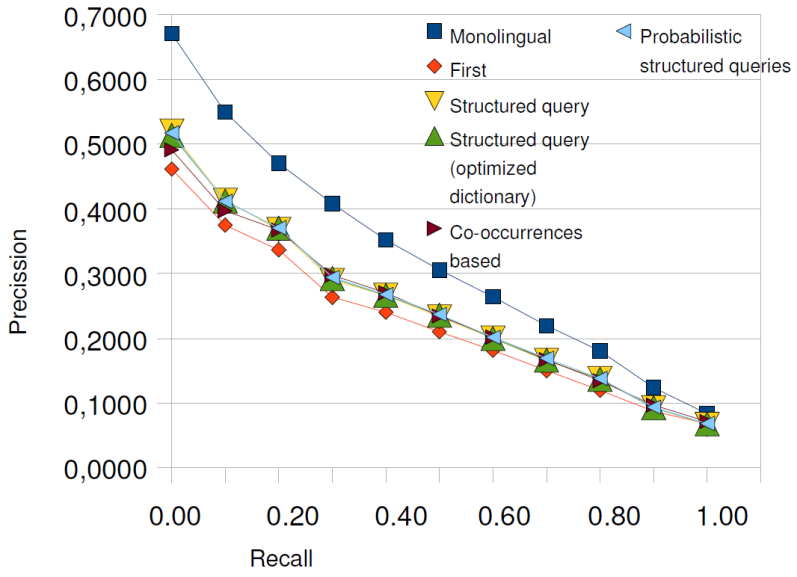


Figure III.4: PR curves (Titles + Description).

word's translation candidate- than between *infect* and *blood* and *sentence* another translation candidate for *epai*. It seems to be because of the the limited representativity of the target collection where some words rarely co-occur. So this could be mitigated by using a bigger corpus. For short queries, too, the hybrid method shows the best results, but statistically does not outperform Pirkola's method significantly. Pirkola's method achieves the best results when dealing with long queries. The optimized MRD improves the MAP but not significantly. All improvements that are statistically significant according to the Paired Randomization Test with α=0.05 are marked with an asterisk in table III.3.

It seems that selecting and weighting translation candidates by means of Monz and Dorr's method in order to include them in structured queries do not imply a significant improvement in MAP terms with respect to Pirkola's method. As in the earlier case, the queries translated by the hybrid method are adequate except for a few cases of false associations. In any case, as we have seen in subsection III.1.3.3, improving the quality of the translation does not always improve the MAP.

In our opinion, apart from the query expansion effect and retrieval time selection, another positive effect produced with structured queries is that the weight of some non-relevant terms are smoothed. It is a collateral effect that happens because non-relevant words tend to be common words which inflate the DF statistic. We have examined the differences between AP values corresponding to 41-90 queries (when titles and descriptions are taken) translated by taking all translations of the MRD and by pruning the wrong ones manually. In theory, all the AP values corresponding to each query will be better with the pruned ones. However, there are 6 queries where AP is significantly higher when all translation candidates are taken, despite many of them being wrong (see figure III.5).

If we analyze these queries more deeply, we can detect two factors that explain this effect:

1. Wrong translations can turn out to be relevant terms: In the example (46) of table III.4 among all the translation candidates of the Basque source word *bahitura* only *kidnapping* appears in the relevant documents of the collection for that query.

2. Wrong translations can reduce non relevant or noise producer source term weight: in the example (81) of table III.4. None of the translations of *erreserba* and *ehiza* appear in the relevant documents. Thus, taking all candidates decreases the weight of these irrelevant sets, leading to a better AP score.

Translation phase	query	AP
English query (46)	**Embargo on Iraq**	
Basque query (46)	*Irakeko bahitura*	
Basque (content words)	Irak bahitura	
Structured translation	Iraq #syn(seizure mortgage kidnapping confiscation)	0.2989
Best translations	Iraq #syn(seizure)	0.1302
English query (81)	**The reserve in the Antarctic in which hunting for whales is forbidden**	
Basque query (81)	*Baleak ehizatzea debekatuta dagoen Antarktikako erreserba*	
Basque (content words)	balea erreserba antarktika ehiza debekatu	
Structured translation whale	#syn(reservation reserve) Antarctica #syn(game hunting prey) prohibit	1.000
Best translations	whale #syn(~~reservation~~ reserve) Antarctica #syn(~~game~~ hunting prey) prohibit	0.3333

Table III.4: Selecting the best candidates from the structured query (Topics 46 and 81).

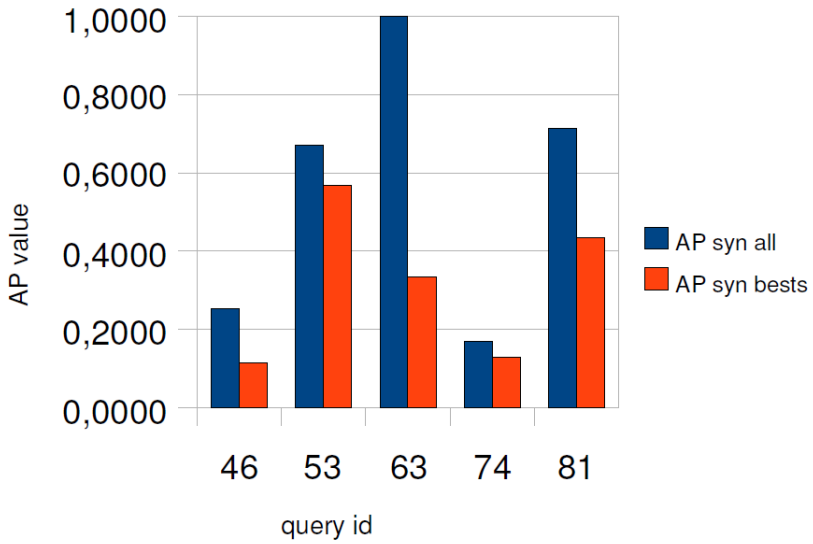


Figure III.5: AP values for queries with significantly improved AP when taking all translations candidates.

III.1.5 Conclusions

We have seen that query translation guided by MRD is useful for the Basque-English pair. Structured queries seem to be a useful method to deal with translation ambiguity. In fact, this method outperforms significantly both first translation method and selection method based on target collection co-occurrence in terms of MAP. Although the co-occurrences-based method significantly outperforms first translation election, the translation probabilities used in probabilistic structured queries do not improve the MAP achieved when using simple structured queries. Otherwise, the MAP is close to the MAP of monolingual retrieval (74% and 78% for short and long queries, respectively) applying only the synonymy expansion provided by the dictionary.

III.2 Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR

Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR

Xabier Saralegi, Maddalen Lopez de Lacalle
R&D Elhuyar Foundation
Usurbil, Spain
x.saralegi@elhuyar.com, m.lopezdelacalle@elhuyar.com

This paper deals with the main problems that arise in the query translation process in dictionary-based Cross-lingual Information Retrieval (CLIR): translation selection, presence of Out-Of-Vocabulary (OOV) terms and translation of Multi-Word Expressions (MWE). We analyse to what extent each problem affects the retrieval performance for the Basque-English pair of languages, and the improvement obtained when using parallel corpora free methods to address them. To tackle the translation selection problem we provide novel extensions of an already existing monolingual target co-occurrence-based method, the Out-Of-Vocabulary terms are dealt with by means of a cognate detection-based method and finally, for the Multi-Word Expression translation problem, a naive matching technique is applied. The error analysis shows significant differences in the deterioration of the performance depending on the problem, in terms of Mean Average Precision (MAP), the translation selection problem being the cause of most of the errors. Otherwise, the proposed combined strategy shows a good performance to tackle the three above-mentioned main problems.

III.2.1 Introduction

CLIR is becoming an increasingly relevant topic due to the growth in multilingual information and the fact that most inhabitants are polyglots. A typical CLIR system offers the user searching topics in his or her mother tongue and retrieves documents in other languages. Different strategies exist to tackle the crosslinguality depending on what information is translated: topics, documents or both. The best results are obtained by translating the collections into the language of the queries. However, this approach is computationally expensive and most of the works have focused on query translation methods. These methods can be based on MT systems, parallel corpora or dictionaries. MT systems and parallel corpora are scarce for the

majority of language pairs. That is why we think that the dictionary-based query translation approach must be explored, since bilingual dictionaries are more abundant and easier to obtain. That is the circumstance of the Basque-English language pair. In the dictionary-based query translation task well-known problems arise that need to be solved, some of the most relevant being translation selection, presence of OOV terms and MWE translations. We propose methods based on target co-occurrences to deal with translation selection, cognate detection to deal with OOV terms, and a naive matching process to detect MWEs. It is important to notice that all the methods presented in this paper are parallel corpora free. In addition to addressing these problems, we are also interested in measuring exactly how each problem affects retrieval performance in dictionary-based query translation and how good the proposed methods deal with them. We need a gold standard to do that evaluation. So we detect and fix the aforementioned three problems manually, and we consider this to be the reference theoretical optimum or topline performance of the system. The paper is organized as follows: first, we review some related works in which different methods to treat inherent problems in CLIR are presented. Next, the strategy we are proposing for translating the query is introduced, along with the methods it involves. That is followed by an appraisal of how each problem affects retrieval performance and how well the proposed methods tackle it. Finally, evaluation results and conclusions are presented.

III.2.2 Related Work

CLIR can be seen as IR with a language barrier placed between the query and the collection. Even though most authors choose to translate the queries into the language of the target collection, mainly due to the lower requirements of memory and processing resources [Hull and Grefenstette, 1996], documents have richer context information than queries, are useful in the translation selection process, and have more examples to reduce error rate of translations. Oard [1998] proved that under certain conditions the quality of the translation and retrieval performance improve when the collection is translated. Furthermore, translating both queries and documents and merging the obtained ranks provides even better results [McCarley, 1999, Chen and Gey, 2003]. The different techniques to carry out the translation can be grouped as follows, depending on the translation-knowledge source: MT-based, parallel corpus-based, and bilingual dictionary-based. For the last two groups different statistical frameworks are proposed; cross-lingual probabilistic relevance models and

cross-lingual language models. The first one offers useful operators to treat the ambiguous translations and is usually used along with dictionaries. The second one incorporates translation probabilities on a more formal and unified framework which are obtained from parallel corpora [Hiemstra, 2001]. The results depend on the quality of the resources but usually better results are achieved with cross-lingual language models [Xu et al., 2001].

However, parallel corpora are a scarce resource. Dictionaries are more accessible but the ambiguous translations must be dealt with. For the translation selection Pirkola [1998] proposed to use structured queries along with probabilistic relevance models. In this approach all translations of a source word are treated as the same token when TF and DF statistics are calculated for the translations of that source word. Darwish and Oard [2003] introduce a probabilistic structured query where weights are applied to translation candidates when TF and DF values are calculated. It offers improvement over non-probabilistic structured queries but only when parallel corpora are used to estimate the weights. As an alternative, Saralegi and Lopez de Lacalle [2010b] proposed that these weights be estimated by calculating the cross-lingual distributional similarity between contexts of the translation candidates obtained from the web, using the web as a comparable corpus.

Other authors propose using the target collection as a language model to solve the translation selection problem [Monz and Dorr, 2005, Ballesteros and Croft, 1998, Gao et al., 2001]. The proposed algorithms try to select the translation candidates which show the highest association degree in the target collection. The algorithms differ in the way the global association is calculated and in the translation unit used (i.e., word, noun phrases...) [Monz and Dorr, 2005, Gao et al., 2001, 2002, Liu et al., 2005].

Structured queries and co-occurrences-based methods were compared in [Saralegi and Lopez de Lacalle, 2009]. There was no significant difference in results when dealing with short queries. But when dealing with long queries, structured queries offer a significantly better MAP than the co-occurrences-based method. This is probably due to the synonym expansion effect produced and the implicit retrieval time selection, which is better when a long context is provided.

The other main problems which affect the translation process are the presence of OOV terms and the translations of MWEs. Cognate detection is the main strategy used for OOV terms treatment [Knight and Graehl, 1997].The translation of the MWE is also explored in some papers [Ballesteros and Croft, 1997].

III.2.3 Proposed Query Translation Method

In this work we have designed a global method that combines state-of-the-art and novel techniques to tackle the aforementioned problems in query translation. We propose a cognate detection-based method to find the translations of the OOV words in the target collection. To address the translation selection problem we propose a target co-occurrences-based method, based on the one proposed by Monz and Dorr [2005]. Although this method did not obtain better results compared with the ones obtained with structured queries in previous works [Saralegi and Lopez de Lacalle, 2009], the truth is that the syn operator of the structured queries is not provided by all retrieval models. Hence, we carried out our experiments with the co-occurrence-based approach. For the MWE treatment we used a simple matching and translation technique based on a bilingual MWE list to detect and translate them.

III.2.3.1 Experimental Setup

We prepared two sets of topics: a set of topics belonging to the CLEF 2001 edition (41-90) that was used as the development set, and another set of topics for test purposes (250-350). All topics were translated by hand from English to Basque. These topics were lemmatized in both languages. We also used the corresponding collections and human relevance judgements. It must be noted that only the LA Times 94 collection is related to the queries of the development set whereas both LA Times 94 and Glasgow Herald collections are linked to the test queries. We adopted a dictionary-based method to carry out the translation process. We used the *Morris* Basque/English dictionary including 77,864 entries and 28,874 unique Basque terms, and the *Euskalterm* terminology bank including 72,184 entries and 56,745 unique Basque terms. According to Demner-Fushman and Oard [2003] the growth in mean average precision is evident between about 3,000 and 20,000 unique terms. They conclude that beyond that range, little further improvement is observed. Hence, we can assume that the coverage of our dictionary is sufficient for the query translation task. We used the *Indri* retrieval algorithm for all the runs.

III.2.3.2 Treating Out-Of-Vocabulary words

The proposed cognate detection approach consists of applying some transliteration rules to the OOV word and then looking for its cognates in the target collection, by computing the Longest Common Subsequence

Ratio (LCSR) measure between the transliterated OOV word and words in the target collection.

In order to measure the damage caused by OOV words in the translation and retrieval processes, we first quantified out these kinds of words in the development set of topics. A total of 64 OOV terms were quantified out and they account for 15.46% of all query terms. This is a normal number taking into account the size of our dictionary. Afterwards, we determined the number of OOV words translated correctly by applying cognate detection. There were 89% in all, and almost all of them were named entities like in the study carried out by Demner-Fushman and Oard [2003]. Despite the fact that this was a good result, we realized that only a total of 7 (10.94%) OOV words needed transliteration and LCSR to detect their translation (Examples in table III.5). The rest of the resolved OOV words were named entities and words that are written equally in both languages. We classified the OOV words depending on their POS (see table III.6).

OOV word	Trans. rules	Transliteration	Max. LCSR
txetxenia	tx → ch	chechenia	(chechenia,chechenya) =0.89
korrupzio	zio→tion k→c	corruption	(corruption,corruption)=1

Table III.5: Example of an OOV word resolved using cognate detection.

Named Entities	Nouns	Adj.	Numbers
82.81%	12.5%	3.13%	1.56%

Table III.6: Distribution of OOV words depending on their POS.

We can see that even if the number of OOV words resolved with the cognate detection-based method are only a few, with respect to the MAP value, using the cognate detection-based method was effective (see table III.7). So, it seems that OOV words tend to be relevant terms in the query, named entities in their majority. We translated OOV words by hand and calculated the MAP value to estimate the topline. The fall produced when OOV words are not treated is 4-12% (First translation of the dictionary or the OOV word itself). But after the proposed method is applied, the fall is reduced to 0.58-3.4%.

Translation method	MAP			
	Title		Title + Descrip.	
		Impr. Over First. %		Impr. Over First. %
First translation	0.2703		0.3835	
First translation + OOV(by hand)	0.3085	12.38	0.3999	4.101
First translation + cognates	0.2969	8.96	0.3975	3.52

Table III.7: Retrieval performance for OOV words for 41-90 topics.

III.2.3.3 Translating Multi-Word Expressions

We identified the MWEs in the development set of topics by hand and analyzed whether they were compositional, in other words, whether they could be translated word by word or not. A total of 60 MWEs were quantified out and exactly 52 (86.67%) of them could be translated word by word (Example on table III.8).

Basque MWT	Words	Translations from Dictionary	Correct Candidate
bigarren mundu gerra	bigarren	second, secondary	second
	mundu	people, world	world
	gerra	war	war

Table III.8: Example of word-by-word MWT translation.

We compared retrieval performance by taking the first translation of each word in the MWE and taking the translation of the complete MWE from the dictionary when available. In addition, we translated all the MWEs by hand and calculated the MAP in order to estimate the topline resulting from the treatment of all of the MWEs. A total of 11 MWEs were directly translated from the dictionary. However, only the translations for the Basque MWE *"esku hartze"* and *"eguzki energia"* Basque MWE translations differ from the ones obtained with the word by word translation. Although they are very few, it seems they tend to be relevant, as a significant improvement

is achieved in terms of MAP (see table III.9). The proposed terminology list-based matching method does not offer a good result, maybe due to its dependence on the recall of the terminology bank. However, as the majority of MWEs are compositional, the co-occurrence-based translation selection method solves most of them.

Translation method	MAP			
	Title		Title + Descrip.	
		Impr. Over First. %		Impr. Over First. %
First translation	0.2703		0.3835	
First translation + MWE(by hand)	0.3371	19.81	0.4222	9.17
First translation + MWE	0.2860	5.49	0.3944	2.76

Table III.9: Retrieval performance for MWEs for 41-90 topics.

III.2.3.4 Translation selection based on target co-occurrences

Finally, we proposed an algorithm based on target collection co-occurrences to deal with the translation selection problem. We adopted the implementation proposed by Monz and Dorr [2005]:

Initially, all the translation candidates are equally likely. Assuming that t is a translation candidate of the set of all candidates $tr(s_i)$ for a query word s_i given by the dictionary, then:

Initialization step:

$$w_T^0(t|s_i) = \frac{1}{|tr(s_i)|} \tag{III.8}$$

In the iteration step, each translation candidate is iteratively updated using the weights of the rest of the candidates and the weight of the links connecting them.

Iteration step:

$$w_T^n(t|s_i) = w_T^{n-1}(t|s_i) + \sum_{t' \in inlink(t)} w_L(t,t') \cdot w_T(t'|s_i) \tag{III.9}$$

where $inlink(t)$ is the set of translation candidates that are linked to t, and $w_L(t,t')$ is the association degree between t and t' in the target collection, measured by the log-likelihood ratio.

After re-computing each translation candidate weight, they are normalized.

Normalization step:

$$w_L^n(t|s_i) = \frac{w_L^n(t|s_i)}{\sum_{m=1}^{|tr(s_i)|} w_L^n(t_{i,m}|s_i)} \tag{III.10}$$

The iteration stops when the variations of the term weights become smaller than a predefined threshold.

In order to measure how the translation selection problem affects retrieval performance we set up two toplines. One involved selecting the correct translation from among those candidates given by the dictionary by hand; in the other, a new translation was also provided if it was not in the dictionary. This new translation was taken from the corresponding source English query. We saw that the MAP results obtained in both experiments were notably better than those obtained with the baseline (First translation) (see table III.9). However, it is noteworthy that introducing new translations outperforms the method including the hand selected translations only. Hence, rather than a selection problem, it would depend on the translation recall of the dictionary used. So for this system the topline will be determined by "Translation selection by hand" results. The Monz and Dorr's selection algorithm (Target co-occurrence-based) achieves very similar results.

III.2.3.4.1 Adding a nearness factor to the degree of association

We introduced a variant into the Monz and Dorr's algorithm. We modified the iteration step by adding a factor $w_F(t,t')$ to increase the association degree $w_L(t,t')$ between translation candidates t and t' whose corresponding source words $so(t)$, $so(t')$ are near each other in the source query Q, and belong to the same MWE.

$$w'_L(t|t') = w_L(t|t') \cdot w_F(t|t') \tag{III.11}$$

$$w_F(t|t') = \frac{\max_{s_i,s_j \in Q} dis(s_i,s_j)}{dis(so(t),so(t'))} \cdot 2^{smw(so(t),so(t'))} \tag{III.12}$$

$$smw(s|s') = \begin{cases} 1 : \{s, s'\} \subseteq Z, Z \in MWU \\ 0 : otherwise \end{cases} \quad \text{(III.13)}$$

According to the MAP scores, the proposed variant (Target co-occurrences-based + nearness) does not achieve any improvement (see table III.10).

Translation method	MAP			
	Title		Title + Descrip.	
		Impr. Over First. %		Impr. Over First. %
First translation	0.2703		0.3835	
Translation selection by hand	0.3430	21.19	0.4266	10.10
Target co-occurrence based	0.3405	20.62	0.4123	6.99
Translation selection by hand + new translations	0.4004	32.49	0.4593	16.50
Target co-occurrence based+nearness	0.3399	20.48	0.4117	6.85

Table III.10: Retrieval performance for translation selection for 41-90 topics.

III.2.3.4.2 Calculating co-occurrences of senses instead of tokens

We also implemented another variant of the target collection co-occurrence-based algorithm, which instead of measuring the degree of association between the customary translation candidate words, it, measures the degree of association between the senses of the translations.

For example, for the source query word s_1 (e.g., metro) the senses of translations in the dictionary are C_1 and C_2; whose translation candidates are t_1 and t_2 (e.g., underground and subway) for the sense C_1, and t_3 and t_4 (e.g., metre and meter) for the sense C_2 $tr(s_1) = \{\{t_1, t_2\}, \{t_3, t_4\}\} = \{C_1, C_2\}$. In the same way, the translation candidate for the source query

word s_2 (e.g., geltoki) is t_5 (e.g., station) which belongs to the same and unique sense C_3, $tr(s_2) = \{\{t_5\}\} = \{C_3\}$. Thus, the frequency for a sense in the colection will be calculated as the amount of documents D_i where the translation candidate words i that belong to that sense appear.

Continuing with the example, the frequency of the sense C_1 will be calculated as the amount of documents including the words t_1 and t_2:

$$f(C_1) = |\bigcup_{i \in C_1} D_i| \tag{III.14}$$

and the frequency of the sense C_2 as the amount of documents including the words t_3 and t_4:

$$f(C_2) = |\bigcup_{i \in C_2} D_i| \tag{III.15}$$

lastly, the frequency of the sense C_3 as the amount of documents including the word t_5:

$$f(C_3) = |\bigcup_{i \in C_3} D_i| \tag{III.16}$$

Thus, the frequency with which the senses C_1 and C_3 appear together in the collection will be calculated as the size of the intersection of the documents D_i including translation candidate i belonging to each sense:

$$f(C_1, C_3) = |\bigcup_{i \in C_1} D_i \cap \bigcup_{j \in C_3} D_j| \tag{III.17}$$

In order to compute $f(C_1, C_3)$ faster, we built a new target collection which contained the senses of the words. The tokens of the collection will be formed by joining the corresponding source word and the sense taken from the dictionary (*source_word_id* + *sense_id*). So if a translation appears in more than one dictionary entry, all the senses will be taken for the new collection by introducing as many new tokens as senses where it appears.

The results show (see table III.11) that in the case of long queries the new method offers a significant MAP improvement over the Monz and Dorr's algorithm (target co-occurrence-based).

III.2.4 Evaluation

In order to carry out the evaluation we used a new set of topics belonging to the CLEF 2001 edition (250-350) and then used the corresponding collection (LA Times 94 and Glasgow Herald 95) and human relevance judgements.

Translation method	MAP			
	Title		Title + Descrip.	
		Impr. Over First. %		Impr. Over First. %
First translation	0.2703		0.3835	
Target token co-occurrence based	0.3405	23.29	0.4059	5.52
Target sense co-occurrence based	0.3323	18.05	0.4163	7.88

Table III.11: Retrieval performance for sense-based translation selection for 41-90 topics.

First, we evaluated each of the proposed methods to deal with the problems in the Basque to English query translation task. Then, we evaluated different combinations of all methods:

- English monolingual or topline.
- Baseline: Taking the first translation from the dictionary.
- OOV: First sense from the dictionary and cognate detection-based method to deal with OOV.
- MWE: MWE matching and first sense from the dictionary.
- Monz: Co-occurrence-based selection.
- Monz+Nearness: Co-occurrence-based selection including the nearness factor.
- Monz (senses): Sense co-occurrence-based selection.
- Monz (senses)+OOV: Sense co-occurrence-based selection and cognate detection-based method to deal with the OOV problem.

In the results obtained (see table III.12), we can see that the translation selection problem is the one which is better dealt with. The

Translation method	MAP					
	Title			Title + Descrip.		
		% over Mon	Impr. Over First. %		% over Mon	Impr. Over First. %
English monolin.	0.3176			0.3773		
Baseline	0.2195	67		0.2599	69	
OOV	0.2279	72	7.24	0.2670	71	2.66
MWE	0.2237	70	5.5	0.2601	69	0.08
Monz	0.2315	73	8.68	0.2642	70	1.63
Monz-Nearness	0.2318	73	8.8	0.2627	70	1.07
Monz (senses)	0.2362∗	74	10.5	0.2747	73	5.39
Monz (senses) +OOV	0.2424∗	76	12.79	0.2805	74	7.34

Table III.12: MAP values for 250-350 topics.

co-occurrence-based translation selection significantly outperforms the first translation approach when dealing with short queries. The improvement offered by the co-occurrence-based method for long queries is lower, probably because the first translation method achieves better results when queries provide many terms. In addition, this lower improvement may be caused by the greedy nature of the translation selection algorithm. Since it has to deal with more translation candidates, it is more likely to reach a maximum. On the other hand, the new proposed sense co-occurrence-based extension exceeds the MAP value obtained with Monz and Dorr's algorithm. Otherwise, as we have seen in the development experiments, the matching method proposed to deal with MWE translations offers a very poor performance. On the contrary, a cognate-based method for treating OOV words seems to be adequate. The best results are achieved by combining the sense co-occurrence-based translation selection method and the cognate-based OOV term translation method.

The improvements that are statistically significant according to the Paired Randomization Test with α=0.05 are marked with an asterisk in table III.11.

III.2.5 Conclusions

We have developed a query translation method which tackles three main problems in dictionary-based CLIR: presence of OOV words, translation of MWEs, and treatment of ambiguous translations. We have analyzed how each problem affects the retrieval performance in terms of MAP. Although results change depending on the length of the queries, the decrease produced by the translation selection (10-21% drop) and the one produced by MWEs (9-20% drop) seem to be the more determining ones. In the case of translation selection, we can distinguish two cases: wrong selection from the dictionary (10-21% drop), and incorrect translations in the dictionary (17-32% drop). OOV treatment (4-12% drop) seems to be the least influential factor, probably due to the similar orthography of both languages. Other pieces dealing with evaluation issues of errors, derived from the MT-based translation process have been carried out [Zhu and Wang, 2006, Qu et al., 2000]. Qu et al. [2000] point out that the wrong translation selection is the most frequent error in an MT-Based translation process. The same conclusion is obtained from our tests. In the development experiments, we have seen that the proposed methods for treating OOV words and ambiguous translations offer a good performance. The matching method proposed to treat MWEs offers a poor performance, but taking into

account that almost all of the MWEs are compositional, it is to be expected that they will be properly addressed by the co-occurrence-based translation selection method.

Otherwise, the improvements developed over the co-occurrence-based translation selection show a different performance behavior. Including the nearness factor provides a few better translations but this leads to no improvement in the overall retrieval performance. For example, the Basque query *"Antarktika balea ehiza debekatu"* is translated as *"Antarctic whale* ***hunting*** *forbidden"* adding the nearness factor while Monz and Dorr's algorithm provides a slightly worse translation: *"Antarctic whale* ***game*** *forbidden"*. Calculating co-occurrences between senses by means of the proposed method instead of between tokens provides better translation quality as well as better retrieval performance.

IV. KAPITULUA

Amarauna corpus konparagarri gisa erabiltzea kontsultaren itzulpena hobetzeko

[Saralegi and Lopez de Lacalle, 2010b] artikuluan corpus konparagarri elebidunetan modu inplizituan dauden itzulpen-probabilitateak CLIR prozesuan baliatzeko modua aztertu da. Itzulpen-probabilitateak antzekotasun distribuzionala konputatuz estimatu dira web-bilatzaileen bidez jasotako corpus konparagarrietatik. [Saralegi and Gamallo, 2013] artikuluan web-bilatzaileen bidez jasotako corpusen adierazgarritasun linguistikoa aztertu dugu. SemCor-en[1] (adierekin etiketatutako corpusa) eta bilatzaileen bidezko testuinguruen arteko aldea (adiera-banaketari dagokionez) neurtu dugu. Adiera-banaketaz gain adiera-aniztasuna eta koherentzia linguistikoa ere neurtu ditugu.

- Xabier Saralegi and Maddalen Lopez de Lacalle. Estimating translation probabilities from the web for structured queries on CLIR. In *European Conference on Information Retrieval*, pages 586–589. Springer, 2010b
- Xabier Saralegi and Pablo Gamallo. Analyzing the sense distribution of concordances obtained by web as corpus approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 355–367. Springer, 2013

[1] http://web.eecs.umich.edu/~mihalcea/downloads.html#semcor

IV.1 Estimating translation probabilities from the web for structured queries on CLIR

Estimating Translation Probabilities from the Web for Structured Queries on CLIR

Xabier Saralegi and Maddalen Lopez de Lacalle
Elhuyar Foundation, R & D
20170 Usurbil, Spain
xabiers,maddalen@elhuyar.com

We present two methods for estimating replacement probabilities without using parallel corpora. The first method proposed exploits the possible translation probabilities latent in Machine Readable Dictionaries (MRD). The second method is more robust, and exploits context similarity-based techniques in order to estimate word translation probabilities using the Internet as a bilingual comparable corpus. The experiments show a statistically significant improvement over non weighted structured queries in terms of MAP by using the replacement probabilities obtained with the proposed methods. The context similarity-based method is the one that yields the most significant improvement.

IV.1.1 Introduction

Several techniques have been proposed for dealing with translation ambiguity for the query translation task on CLIR, such as structured query-based translation (also known as Pirkola's method) [Pirkola, 1998], word co-occurrence statistics [Ballesteros and Croft, 1998] and statistical translation models [Hiemstra, 2001]. Structured queries are adequate for less resourced languages, rare pairs of languages or certain domains where parallel corpora are scarce or even nonexistent. The idea behind this method is to treat all the translation candidates of a source word as a single word (*syn* operator) when calculating TF and DF statistics. This produces an implicit translation selection during retrieval time. There are many works dealing with structured queries, and some variants are proposed. Darwish and Oard [2003] for example, propose that weights or replacement probabilities be included in the translation candidates (*wsyn* operator). One drawback with this approach is that it needs parallel corpora in order to estimate the replacement probabilities.

Following this line of work, we propose a simple method based on the implicit translation probabilities of a dictionary, and also a more robust one which uses translation knowledge mined from the web. We have analyzed

different ways of accessing web data: **Web As Corpus** tools, **News** search engines, and **Blog** search engines. Our aim is to examine how the characteristics of each access strategy influence the representation of the constructed contexts, and also, how far these strategies are adequate for estimating translation probabilities by means of the cross-lingual context similarity paradigm. All experiments have been carried out taking Spanish as source language and English as target.

IV.1.2 Obtaining Translation Probabilities from a Dictionary

The first method proposed for estimating translation probabilities relies on the hypothesis that, in a bilingual MRD (D), the position (pos) of the translation (w) among all the corresponding translation candidates f for a source word (v) is inversely proportional to its translation probability ($p(w|v)$). If we assume that it is an exponential decay relation, we can model the translation probability through this formula:

$$p(w|v) = \frac{1}{\sum_{(v,f)\in D} (\frac{1}{pos(D,v,f)}) \cdot pos(D,v,w)} \qquad \text{(IV.1)}$$

The principal problems of these assumptions are, firstly, that translations are not ordered in all MRD (partially or at all) by frequency of use, and secondly, that the proposed relation above does not fit all translation equivalents. So, we propose a method that is useful for ordering the translations of an MRD as well as for estimating more accurate translation probabilities, as presented in the following section.

IV.1.3 Translation Probabilities by Context Similarity

The idea is to obtain translation probabilities by using the web as a bilingual comparable corpus. This strategy is based on estimating the translation probability of the translation candidates taken from the MRD in accordance with the context similarity of the translation pairs [Fung and Yee, 1998]. The hypothesis is that the more similar the contexts are, the more probable the translation will be. The computation of the context similarity requires a large amount of data (contexts of words), which has to be representative and from comparable sources. The Internet is a good source of large amounts of texts, and that is why, we propose that different search-engines be analyzed to obtain these contexts. These search engines have different features, such as domain, coverage and ranking, which affect both the degree of comparability and the representativeness of the contexts, as follows:

WebCorp: This Web Concordancer is based on main search APIs. Therefore, navigational queries and popular ones are promoted. These criteria can reduce the representativeness of the contexts retrieved. Since we take a maximum number of snippets for each query, the selected contexts depend on the ranking algorithm. It guarantees good recall, but perhaps poor precision. Thus, the comparability degree between contexts in different languages can be affected negatively.

Google News Archive: The content is only journalistic. It seems appropriate if we want to deal with journalism documents but not with other registers or more specialized domains. In short, it offers good precision, enough recall and a good degree of comparability.

Google Blog search: The language used is more popular, and although the register is similar to that of journalism, the domain is more extensive. This could offer good recall but not very comparable contexts.

The method to estimate the translation probabilities between a source word (v) and its translations $f((v,f) \in D)$ starts by downloading, separately, the snippets of both words as returned by the search engines mentioned above. Then, we set up context vectors for the source $\vec{v}$ and the translation word $\vec{w}$ by taking keywordness (using log-likelihood scores) of the content words (nouns, adjectives, verbs and adverbs selected by using *Treetagger*) belonging to all their snippets. The next step is to translate the Spanish context vector $\vec{v}$ into English $tr(\vec{v})$. This is done by taking the first translation from a Spanish-English MRD (D) (34,167 entries). Cross-lingual context similarity is calculated according to cosine measure which is transformed into translations probabilities:

$$p(w|v) = \frac{cos(tr(\vec{v}), \vec{w})}{\sum_{(v,f)\in D} cos(tr(\vec{v}), \vec{f})} \quad \text{(IV.2)}$$

We analyze the differences between the translation rankings obtained with the different search engines and those in the original dictionary. We computed Pearson's correlation for the translation rankings obtained for the polysemous content words in all 300-350 Spanish CLEF topics. The correlation scores (cf. table IV.1) show that the different characteristics of each search engine produce translation rankings which are quite different from those in the dictionary (**Dic.**) and also from each other.

	WebCorp	News	Blog
Dic.	0.42	0.31	0.40
WebCorp		0.44	0.54
News			0.49

Table IV.1: Mean of Pearson's correlation coefficients for translation rankings compared to each other.

IV.1.4 Evaluation and Conclusions

We evaluated 50 queries (title+description) taken from 300-350 CLEF topics against collections from CLEF 2001 composed by LA Times 94 and Glasgow Herald 95 news. Previously, nouns, adjectives, verbs and adverbs were selected manually both in Spanish and English topics. *Indri* was used as the retrieval model and the queries were translated using several methods: taking the first translation of the MRD (**First**); taking all the translations and grouping them by the syn operator (**All** or **Pirkola**); and weighting the translations by using the *wsyn* operator and the methods described in sections IV.1.2 (**Dic.**) and 3 (**Webcorp**, **News** and **Blog**). The results are shown in table IV.2.

Method	MAP	% Monolingual	% Improv. over All
Monolingual (en)	0.3651		
First	0.2462	67.43	
All	0.2892	79.21	
Dic.	0.2951	80.83	2.04
WebCorp	0.2943	80.55	1.76
News	0.2993	82.63	3.49
Blog	0.2960	81.07	2.35

Table IV.2: MAP for 300-350 topics.

In the first column we show the MAP results obtained with each method, with the English monolingual results first. In the second column we show the percentage of the cross lingual MAP with respect to the monolingual result. We can see that using all translations with their replacement probability estimated according to the dictionary order produces better

results than using only the first translation or using all translations, with a significant improvement (according to the Paired Randomization Test with α=0.05) over the **All** method. So, exploiting the translation knowledge latent in the position of the translations improves the MAP when provided by the dictionary. Otherwise, the web-based estimation techniques also improve significantly over the **First** and **All** strategies (α=0.05). However, there is no significant improvement over the **Dic.** method. It seems that context similarity calculated from **Blog** or **News** sources is more suited to estimating translation probabilities since they significantly outperform **WebCorp** in terms of MAP. Therefore, comparability between sources of both languages, domain precision and informational snippets seem to be important factors in order to obtain useful context for context-similarity, although deeper analyses must be carried out to determine the importance of each more precisely. Finally, we conclude that translation knowledge obtained from the Internet, offers an adequate means, and by means of cross-lingual context similarity, it is useful for estimating replacement probabilities. Moreover, it could be an alternative when parallel corpora or MRDs with translations sorted according frequency of use are not available.

IV.2 Analyzing the sense distribution of concordances obtained by web as corpus approach

Analyzing the Sense Distribution of Concordances Obtained by Web As Corpus Approach

Xabier Saralegi[1] and Pablo Gamallo[2]
[1]Elhuyar Foundation,
Osinalde Industrialdea 3, 20160 Usurbil, Spain
x.saralegi@elhuyar.com
[2]Centro de Investigação em Tecnologias da Informação (CITIUS)
Universidade de Santiago de Compostela, Galiza, Spain
pablo.gamallo@usc.es

In corpus-based lexicography and natural language processing fields some authors have proposed using the Internet as a source of corpora for obtaining concordances of words. Most techniques implemented with this method are based on information retrieval-oriented web searchers. However, rankings of concordances obtained by these search engines are not built according to linguistic criteria but to topic similarity or navigational oriented criteria, such as page-rank. It follows that examples or concordances could not be linguistically representative, and so, linguistic knowledge mined by these methods might not be very useful. This work analyzes the linguistic representativeness of concordances obtained by different relevance criteria based web search engines (web, blog and news search engines). The analysis consists of comparing web concordances and SemCor (the reference) with regard to the distribution of word senses. Results showed that sense distributions in concordances obtained by web search engines are, in general, quite different from those obtained from the reference corpus. Among the search engines, those that were found to be the most similar to the reference were the informational oriented engines (news and blog search engines).

IV.2.1 Introduction

Most statistical approaches to solving tasks related to Natural Language Processing (NLP) as well as lexicographic works use corpora as a resource of evidences. However, one of the biggest problems encountered by these approaches is to obtain an amount of data that could be large enough for statistical and linguistic analysis. Taking into account the rapid growth of the Internet and the quantity of texts included in it, some researchers have

proposed using the Web as a source for building corpora [Kilgarriff and Grefenstette, 2003]. Two strategies have been proposed for exploiting the web with that objective in mind:

- *Web As Corpus*: The web is accessed directly as a corpus. This access is usually performed by means of commercial web search engines (*Bing*, *Yahoo*, *Google*...), which are used to retrieve concordance lines showing the context in which the user's search term occurs. *WebCorp* [Morley, 2006] is a representative linguistic tool based on this strategy.

- *Web For Corpus*: This strategy consists of compiling a corpus from the web to be accessed later. This compilation process can be performed by crawling the web and also by using commercial web search engines. This latter approach consists of sending a set of queries including seed terms corresponding to a certain topic or domain, and then retrieving the pages returned by the engine.

The two strategies usually rely on web search engines (SEs) in order to retrieve pages with text and in this way build word concordances or text corpora. Using APIs provided by SEs offers several advantages. It makes the treatment of spam and other low-quality, undesired contents easier. Besides, these APIs provide a high coverage of the web.

The Web as Corpus approach is more suitable than Web for Corpus for those tasks requiring an acceptable quantity of examples of concordances for any word (e.g., Distributional Similarity, Information Extraction, Word Sense Disambiguation, etc...). However, some problems can arise from using SEs for concordance compilation. For example, Agirre et al. [2000] found that the great number of erotic web pages strongly influenced their experiments on WSD. The set of pages retrieved by the web SE is dependent on ranking criteria[2], which are not specified according to linguistic features such as frequency of use of each sense. The users of commercial SEs have other needs than those focused on obtaining specific pieces of text information. Broder [2002] states that the target of search queries is often non-informational (more than 50%), since it might be navigational when the queries are looking for websites, or transactional when the objective is to shop, download a file, or find a map. Thus, criteria related to these needs are also reflected in the above mentioned ranking factors. The main page-ranking factors mainly rely on popularity, anchor text analysis and trusted domains, but not on content.

Our work is based on two main assumptions:

[2]An example of factors used for search engine rankings are listed here: http://www.seomoz.org/article/search-ranking-factors

- *Assumption 1*: *SemCor* is a sense-tagged corpus [Mihalcea, 1998], which can be regarded as a gold-standard reference of the "real" distribution of word senses in an open domain. According to Agirre and Martinez [2004], corpora with similar distribution to that of *SemCor* get the best results in the task of WSD for an open domain. Word Sense Disambiguation (WSD) systems with the best performance in Senseval-2 were trained with it. We are aware that the concept of reference corpora is arguable as some author point out [Kilgarriff, 2012].

- *Assumption 2*: The Web is, in terms of variety of genres, topics and lexicons, close to the most traditional open domain "balanced" corpora, such as the Brown or BNC [Baroni and Bernardini, 2006].

On the basis of these two assumptions, we aim to validate the following hypothesis: the ranked results obtained by SEs are not a representative sample in terms of sense distribution, since they follow ranking criteria focused on non-linguistic relevance and only first results are usually compiled. In other words, the linguistic examples or concordances extracted by web SEs are biased by non-linguistic criteria. A high bias would indicate that web-based concordances compilation is not useful at all in some NLP and lexicography tasks. For example, linguistic information obtained by an SE used for knowledge extraction such as cross-lingual distributional similarity [Nakov et al., 2007], semantic-class learning [Kozareva et al., 2008], or consultation (e.g., lexicographic, translators, writers or language learners) might not be reliable. In the remaining sections, we will attempt to confirm or reject such a hypothesis.

IV.2.2 Related Work

There are few works which deal with linguistic adequacy of concordances obtained by SEs. Chen et al. [2010] describe an experiment to manually evaluate concordances included in web documents retrieved from an SE for 10 test words. They annotated by hand about 1,000 to 3,000 instances for each test word. In particular, the authors evaluate two pieces of information: the quality of the web documents returned by the SE, and sense distributions. Concerning sense distribution, they concluded that, on the one hand, the most frequent senses from web-based corpora are similar to *SemCor* and, on the other, web-based corpora may provide more diverse senses than *SemCor*. However they do not perform any correlation analysis to draw that conclusion. As we will show later in section IV.2.5, a correlation

analysis performed over their results shows a low correlation in terms of sense distribution between web-based corpus and *SemCor*.

Other works analyze some aspects related to linguistic adequacy but for different purposes. Diversity in web-rankings is a topic closely related to word-sense distribution analysis. Santamaría et al. [2010] propose a method to promote diversity of web search results for small (one-word) queries. Their experiments showed that, on average, 63% of the pages in search results belong to the most frequent sense of the query word. This suggests that "diversity does not play a major role in the current *Google* ranking algorithm". As we will show in section IV.2.5, the degree of diversity of the concordances we have retrieved is still lower: in our experiments 72% of the concordances belong to the most frequent sense.

IV.2.3 Methodology for analyzing the adequacy of *Web As Corpus*

Our objective is to verify whether the distribution of senses in the rankings obtained from SEs are representative in linguistic terms for an open domain. Our analysis relies on *SemCor* as evidence, since it is a well-known, manually annotated open domain reference corpus for word senses according to *WordNet*. In addition, we also measure two further properties of the concordances retrieved by SEs, namely sense diversity and linguistic coherence (i.e., typos, spelling errors, etc...).

IV.2.3.1 How to obtain concordances of a word from the web?

Several SEs have been used in the literature in order to collect examples of concordances from the web. Most authors use SEs directly by collecting the retrieved snippets. Web As Corpus tools such as *WebCorp* are more linguistically-motivated tools. In that sense they offer parameters to post-process SE rankings (case sensitive searches to avoid named entities, no searches over links to avoid a somehow navigational bias...). Anyway they still depend on SE rankings. So they raise the same problems mentioned in section IV.2.1. Other emergent SEs are those focused on specific domains and genres such as news or blogs. These SEs are interesting for linguistic purposes because bias produced by factors related to navigational and transactional is avoided. In addition, their text sources are not domain restricted. In fact, newswire-based corpora are often built for open domain knowledge extraction purposes. However, some authors [Baroni and Bernardini, 2006] point out that, in terms of variety of genres and topics, Web is closer to traditional "balanced" corpora such as the BNC. Blog is a new genre

not present in traditional written sources but similar to them in terms of distribution of topics. In order to analyze and compare the influence of these characteristics, the following engines have been evaluated: *WebCorp*, *Google News Archive* (*GNews*), and *Google Blog Search* (*GBlog*). See table IV.3 for more details.

SE	Domain	Genre	Query	Ranking
WebCorp	Open	Open	inform. navig. transac.	topic popularity ...(see first note)
GBlog	Open	Blogs	inform.	topic
GNews	Open	News	inform.	topic

Table IV.3: Characteristics of SEs.

In order to guarantee a minimum linguistic cohesion of the concordances, the following parametrizations were used for each SE. English is selected as the target language in all of them. In *WebCorp*[3], Bing has been selected as the API because provides the best coverage. Case-sensitive searches were performed. The search over links option was disabled in order to mitigate navigational bias. Span of ± 5 words for concordances was established. *GNews* and *GBlog* do not offer choice for casesensitive searches. So, case-sensitive treatment was done after retrieving the snippets. In the cases of *GNews* and *GBlog*, searches were performed only on the body of documents and not on the titles (*allintext* operator was used).

IV.2.3.2 Selecting test words

10 test words (see table IV.4) are randomly selected from the *SemCor* 1.6 corpus, a corpus where all words are tagged with their corresponding sense according to *WordNet* 1.6. Due to the small size of the sample several conditions were established in order to guarantee the representativeness of the test words and the corresponding contexts:

- Nouns are selected because they are the largest grammatical group.
- More than 1 sense in *SemCor* because we want to focus on ambiguous words.

[3] *WebCorp* has included recently *GNews* and *GBlog* APIs

- Minimum frequency of 50 on *SemCor* corpus. As McCarthy et al. [2004] pointed out, *SemCor* comprises a relatively small sample of words. Consequently, there are words where the first sense in *WordNet* is counter-intuitive. For example, the first sense of *"tiger"* according to *SemCor* is an audacious person, whereas one might expect carnivorous animal to be a more common usage.

Word	Sense distribution
church	1=0.47, 2=0.45, 3=0.08
particle	1=0.63, 2=0.35, 3=0.02
procedure	1=0.73, 2=0.27
relationship	1=0.60, 2=0.21, 3=0.19
element	1=0.71, 2=0.21, 3=0.06, 4=0.02
function	1=0.58, 2=0.32, 3=0.09
trial	1=0.45, 2=0.03, 4=0.52
production	1=0.64, 2=0.21, 3=0.11, 4=0.04
newspaper	1=0.66, 3=0.02, 2=0.29, 2;1=0.02
energy	1=0.74, 2=0.10, 3=0.12, 4=0.05

Table IV.4: Selected test words from *SemCor* and their sense distribution.

IV.2.3.3 Annotation of web based concordances

Each test word is submitted to the three different SEs. The number of retrieved snippets, i.e., word concordances, may change depending on both the query word and the SE. So, in order to obtain more comparable samples, the first 250 concordances are retrieved for each case. As the number of test examples is still too much to analyze by hand, to save work without missing the rank information, only an interpolated sample of 50 concordances was analyzed. The hand analysis involves manually tagging the sense of the test words according to *WordNet* 1.6.

IV.2.3.4 Measuring the adequacy of concordances

The main objective here is to measure differences in terms of sense distribution between *SemCor* and the concordances retrieved by the SE. However, besides sense distribution, our aim is also to measure both sense

diversity and linguistic coherence. Let us describe first how we measure sense diversity, then linguistic coherence, and finally sense distribution.

IV.2.3.4.1 Sense diversity

We associate the term *"sense diversity"* with text corpora whose word occurrences cover a great variety of senses. It is possible to know to a certain extent the degree of diversity of a corpus by observing the senses of a sample of words. In particular, diversity can be measured by comparing the number of possible senses of the test words (e.g., their *WordNet* senses) with the number of different senses that are actually found in the corpus, i.e., in our collections of concordances. The higher the number of senses covered by the concordances, the greater their degree of diversity. Concordances with much sense diversity tend to be open to many domains.

IV.2.3.4.2 Linguistic coherence

The quality and linguistic coherence of the retrieved concordances can vary from totally nonsensical expressions to high quality texts. So, coherence, or more precisely "level of coherence", is also taken into account in our evaluation protocol. To do this, the annotators can assign four possible coherence values to each retrieved concordance:

- Score 0. The concordance has no problems.
- Score 1. The concordance has serious typographical errors or morphosyntactic problems, but it can be understood.
- Score 2. The query word is part of a Named Entity, e.g., *"town"* in *"Luton Town FC Club"*.
- Score 3. The concordance is totally nonsensical and can not be understood at all.

The range of values is from 0 (coherent) to 3 (totally incoherent or nonsensical). It should be borne in mind that values 1 and 2 could be unified since named entities written in lower-case seem to be typographical errors. However, we preferred to keep the two coherence levels because value 1 still allows us to assign a *WordNet* sense to the key word, but it is not the case when the coherence level is 2. On the basis of the notion of level of coherence, we define "degree of incoherence", which is associated with a concordance

collection. The degree of incoherence of a concordance collection, noted ϕ, is computed as follows:

$$\phi = \frac{\sum_{i}^{n} L(c_i)}{3 \cdot n} \tag{IV.3}$$

where $L(c_i)$ stands for the level of coherence of concordance c_i, and n is the number of concordances in the collection. Let us suppose that we have a collection of 4 concordances, with the following levels of coherence for each concordance: 0, 1, 0, 3. The degree of incoherence of the total collection is then $4/12 = 0.33$. The values of this function are ranged from 0 (fully coherent) to 1 (totally incoherent).

IV.2.3.4.3 Sense distribution

The distributions of senses found in the three SE concordance collections are compared with those extracted from *SemCor* by analyzing the Pearson correlation between them. The correlation between two sets of concordances is computed by considering the relative frequencies of those senses of the word that have been found in, at least, one of the two sets.

Besides the strict correlation, we are also interested in verifying properties concerning sense dominance. For instance, two collections may share (or not share) the same dominant sense. When their dominant senses are not the same, and one of them is domain-specific (scientific, technical, etc.), then the two collections should be considered very different in terms of sense distribution, regardless of their specific Pearson correlation. On the other hand, when they share the same dominant sense, it is also important to observe whether there are differences in terms of degree of dominance. If the main sense is very dominant in one collection and not so dominant in the other one, we may infer that there are significant differences in sense distribution. This is true even if the Pearson correlation is actually very high. Let us see an example. Word *"production"* has 4 senses with the following two sense distributions, in *SemCor* and *GNews*, respectively:

- *SemCor*: 0.64 0.21 0.11 0.04
- *GNews*: 0.98 0.02 0.0 0.0

The Pearson correlation between *SemCor* and *GNews* is very high: > 0.97. However, from a linguistic perspective, the distributions are very

different. While the sense distribution in *SemCor* may be considered as an evidence for content heterogeneity (there are three senses with more than 10% occurrences), sense distribution in *GBlog* shows that concordances are content homogeneous. As in *GBlog* only one sense covers more than 98% of the word occurrences, it means that concordances are retrieved from a domain-specific source. By contrast, concordances of *SemCor* seem to represent more open and balanced text domains. Here, we also should take into account the conclusions to which Kilgarriff [2004] came in, where he argued that the expected dominance of the commonest sense rises with the number of corpus instances. It follows that the dominance of one sense is higher in *GNews* because of the corpus size.

The rank of retrieved concordances for each SE is also analyzed. We are interested in observing the order of appearance of the senses among the web ranking. An adequate order will be that one in which concordances are ordered according to the probability of senses of the search word. Thus, concordances including the most probable ones should be on the top of the rank. For linguistic consultations, for instance, it is better to show concordances including the most common senses of the search word at the top of the ranking. In addition, those strategies that only retrieve the first concordances of the SE could also perform better if top concordances corresponded to the most common senses. Once again, we use *SemCor* to prepare a reference rank according to sense probabilities calculated from *SemCor*. In order to measure the adequacy of the rank of web-concordances, we compute the Spearman correlation between the web concordances rank and the *SemCor* based reference rank.

IV.2.4 Results

The results concerning sense diversity, linguistic coherence, and sense distribution are shown and analyzed as follows:

IV.2.4.1 Sense Diversity

In total, the 10 test words have 49 different *WordNet* senses. Among these 49 senses, the collection of concordances from *WebCorp* contains instances of 34 senses, two more (32) than the senses found in *SemCor* for the same 10 words. The concordances of *GBlog* contain a different 31 senses while those of *GNews* only 27. In table IV.5, we show the percentage of different senses we found in each corpus with regard to the total number of *WordNet* senses attributed to the 10 test words (first column), as well as to the number of

senses these words have in *SemCor* (second column). In the last column, we show the number of senses appearing in each collection of concordances that do not appear in *SemCor*. It follows that the *WebCorp* corpus may provide more diverse senses and, therefore, more domain diversity, than the two corpora built from the Google engines. In addition, we may also infer that the journalism articles seem to be more restricted in terms of domain diversity than the posts of blogs. However, we have to take into account that high diversity does not imply balanced sense distribution.

	%senses of test words	%senses in *SemCor*	#new-senses
WebCorp	69%	81%	8
GBlog	63%	78%	6
GNews	55%	72%	5

Table IV.5: Sense diversity.

IV.2.4.2 Linguistic coherence

Table IV.6 shows information on levels of incoherence associated with the three web-based concordance collections. The three first columns show the total values of 3 levels of coherence (level 0 is not shown but can be inferred). The forth column measures the degree of incoherence for each collection, according to formula IV.3. In the last column, we show the percentage of concordances having some positive incoherence value (i.e., 1, 2, or 3) for each collection.

We can observe that *WebCorp* is the SE that provides more incoherent concordances at the 3 levels. In addition, in *WebCorp* almost 1 context out of 4 has some problems of coherence. This is probably due to the fact that *WebCorp* covers the whole Web, containing many not very confident text sources. The degree of incoherence in *GBlog* is also relatively high (0.12), against only 0.04 of *GNews*, which is then the most reliable source of textual data in our experiments So, the linguistic quality of the corpora built with Google engines is clearly better than that of *WebCorp*.

IV.2.4.3 Sense distribution

IV.2.4.3.1 Pearson Correlation

	level 1	level 2	level 3	ϕ	incoherence
WebCorp	68	6	48	0.15	24%
GBlog	34	1	25	0.07	12%
GNews	32	0	9	0.04	8%

Table IV.6: Linguistic coherence.

The senses found in the web-based concordances are compared with those extracted from *SemCor* by analyzing the Pearson correlation between them (see table IV.7). This table is organized as follows. The test word is in the first column. The following columns show the Pearson correlation between the SemCor and sense distributions corresponding to the different web-based concordances (*WebCorp*, *GNews*, or *GBlog*).

	WebCorp	GBlog	GNews
church	0.78	0.85	0.70
particle	0.10	0.53	0.20
procedure	0.62	0.88	0.89
relationship	-0.28	-0.41	0.50
element	0.28	0.29	0.95
function	0.50	-0.03	0.03
trial	0.60	0.63	0.7
production	0.61	0.89	0.97
newspaper	0.93	0.99	0.66
energy	0.97	0.97	0.98
average	0.51	0.56	0.66

Table IV.7: Pearson correlation of sense distribution regarding to *SemCor*.

As far as the Pearson coefficient is concerned, the average correlation of *WebCorp* and *SemCor* is 0.51, which is the lowest correlation. The average correlation between *GBlog* and *SemCor* is 0.56, and the one between *GNews* and *SemCor* is the highest: 0.66. As the correlation values between 0.51 and 0.79 are interpreted as being "low", we may consider that there is always a low correlation between *SemCor* and our three web-based concordance collections. If we conduct a more detailed analysis word by word and compute the correlations of each test word, we can observe

that there are three words (*"newspaper"*, *"production"*, and *"procedure"*) with moderate correlations (between 0.80 and 0.86), four words (*"energy"*, *"church"*, *"trial"*, and *"element"*) with low correlations (between 0.51 and 0.79), and three (*"particle"*, *"function"*, and *"relationship"*) without any correlation at all, since their values are lower than the significance level (0.35, p<.01) established for tests with 50 pairs.

	Chen et al. [2010] concordances
author	1
back	0.94
cart	-1
case	0.85
center	0.15
core	0.35
mind	1
sequence	1
toast	0.99
average	0.59

Table IV.8: Pearson correlation between concordances reported in [Chen et al., 2010] and *SemCor*.

IV.2.4.3.2 Analysis on Dominant Senses

Besides the statistical test, it is also important to verify further qualitative aspects, in particular whether concordances are comparable in terms of sense dominance. More precisely, given two collections of concordances, we checked both whether they share the same dominant senses and whether the same dominant senses have a similar degree of dominance.

We observed that *SemCor* and each web-based concordances do not share the same dominant sense in several cases. In addition, for most of these words, their dominant senses in the web-based concordances are domain-specific senses, e.g., physics for *"particle"*, computer science for *"function"*, show business for *"production"*, or gossip news for *"relationship"*. It should be noticed that these specific senses are not dominant in *SemCor* and they do not correspond to the first sense in *WordNet*. The high relevance given by non-linguistic ranking criteria to webs dealing with scientific,

business or gossip topics could explain the large number of domain-specific senses in the concordances.

On the other hand, for many cases where the test word shares the same dominant sense in both *SemCor* and the web-based concordances, we observe that there are significant differences in terms of degree of dominance. In general, the sense distribution of *SemCor* seems to be more balanced than that of web-based concordances. Six words had same dominant sense in both *SemCor* and *GNews*, but in all cases the shared sense is clearly more dominant in *GNews*. The average degree of dominance in *SemCor* is 62% against 72% in the web-based concordances. As we showed above in 4.3.1, these differences may not be very significant for the Pearson correlation, but from a linguistic point of view, they are very significant since they denote that the web-based concordances are more homogeneous in terms of linguistic content. Once again, the ranking criteria of the SE could give more relevance to a very restricted subset of topics among all of those we can find in the open domain web.

In addition, we can find further qualitative differences between *SemCor* and web-based concordances. On the one hand, it should be noted that web-based concordances introduce new technical or domain-specific senses that are not in *SemCor*. On the other hand, we can find cases where the transactional function of some webs may influence sense distribution. For instance, the second sense of *"trial"* (very marginal in *SemCor*) is very important in *WebCorp* and *GBlog* because of the high number of commercial pages with *"free trial"* software.

IV.2.4.3.3 Spearman Correlation

Finally, the order of the senses appearing in the web concordances ranking is also analyzed. We check whether web-concordances of test words are sorted by the frequency of use of the included sense. For this purpose, the reference ranking for each test word and SE is prepared by sorting all the collected concordances according to sense probabilities mined from *SemCor*.

For example, suppose we collect the following concordances ranking from the web for a test word including 4 senses: $\{(context_1, sen_1), (context_2, sen_2), (context_3, sen_1), (context_4, sen_1)\}$. The *SemCor*-based ranking (the reference) is built by sorting all concordances according to sense probability estimated from *SemCor* ($sen_2 = 0.8, sen_1 = 0.2$): $\{(context_2, sen_2), (context_1, sen_1), (context_3, sen_1), (context_4, sen_1)\}$. Contexts with the same sense keep the original order. Then, the Spearman

correlation between original concordances ranking and *SemCor*-based reference ranking is calculated (see table IV.9).

Notice that if all concordances of the ranking are analyzed, we observe that only *GBlog* concordances are correlated. Other SEs provide some correlation only if the first 50 concordances are selected. So, it seems that top of rankings are more adequate in terms of sense probability.

	WebCorp		GBlog		GNews	
	All	top50	All	top50	All	top50
church	0.41	0.98	0.58	0.68	0.43	-0.25
particle	0.63	1	0.57	1	0.58	0.65
procedure	-0.59	-0.76	0.46	1	0.11	1
relationship	0.61	0.53	0.49	-0.25	0.70	0.37
element	0.53	1	0.53	0.99	0.38	0.95
function	-0.42	0.94	-0.08	-0.59	-0.75	-0.75
trial	0.13	0.58	0.80	0.99	0.55	0.82
production	-0.11	-0.01	1	1	0.19	0.82
newspaper	0.80	0.79	0.30	0.2	0.17	0.31
energy	0.42	0.66	0.68	1	0.51	0.4
average	0.24	0.57	0.53	0.6	0.30	0.43

Table IV.9: Spearman correlation of concordance ranking with respect to *SemCor*.

IV.2.5 Conclusions

We have proposed an experimental method to verify whether the distribution of senses in the rankings obtained from SEs are balanced, representative, and coherent in linguistic terms. Taking *SemCor* as a balanced reference, we observed that the concordances retrieved by different SEs have low correlation with *SemCor* with regard to sense distribution. If we consider that the diversity of topics and domains in the web is close to that of most traditional open domain balanced corpora, we may infer from our experiments that the sense distribution bias is due to the fact that web engines rank their pages using non-linguistic criteria. It should be noted that the best correlation was achieved with SEs that only cover a part of the web (news and blogs), whose text sources are thus rather far, in terms of topic and genre distribution, from a traditional balanced corpus. By contrast, the

worse correlation was achieved by the search engine (*WebCorp*) using the entire web as text source. We can surmise that ranking factors related to popularity or navigational queries introduce some non-linguistic bias in the concordances retrieved by general-purpose SEs. Furthermore, some SEs may retrieve concordances with serious problems concerning linguistic coherence (24% of concordances of *WebCorp* display problems of linguistic coherence). All these observations lead us to conclude that word sense information obtained by SEs used for knowledge extraction, word sense disambiguation, or lexicographic consultation might not be totally reliable. However, this final conclusion should be corroborated by carrying out further experiments over larger samples.

V. KAPITULUA

Saioen informazioa kontsultaren itzulpen-prozesua hobetzeko

[Saralegi et al., 2016] artikuluan kontsultaz gaindiko testuinguruak baliatu nahi izan ditugu hiztegian oinarritutako CLIR prozesuaren itzulpen-hautapena egiteko. Birformulatutako kontsulta itzultzeko saio bereko aurreko kontsultak adierazitako testuinguruak zenbateraino laguntzen duen aztertu dugu.

- Xabier Saralegi, Eneko Agirre, and Iñaki Alegria. Evaluating Translation Quality and CLIR Performance of Query Sessions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016

V.1 Evaluating Translation Quality and CLIR Performance of Query Sessions

Evaluating translation quality and CLIR performance of Query Sessions

Xabier Saralegi[1], Eneko Agirre[2], Iñaki Alegria[2]
[1]Elhuyar Fundazioa, Usurbil
[2]IXA NLP Group, The University of the Basque Country, San Sebastian
x.saralegi@elhuyar.com

This paper presents the evaluation of the translation quality and Cross-Lingual Information Retrieval (CLIR) performance when using session information as the context of queries. The hypothesis is that previous queries provide context that helps to solve ambiguous translations in the current query. We tested several strategies on the TREC 2010 Session track dataset, which includes query reformulations grouped by generalization, specification, and drifting types. We study the Basque to English direction, evaluating both the translation quality and CLIR performance, with positive results in both cases. The results show that the quality of translation improved, reducing error rate by 12% (HTER) when using session information, which improved CLIR results 5% (nDCG). We also provide an analysis of the improvements across the three kinds of sessions: generalization, specification, and drifting. Translation quality improved in all three types (generalization, specification, and drifting), and CLIR improved for generalization and specification sessions, preserving the performance in drifting sessions.

V.1.1 Introduction

The successful strategies for query translation in CLIR depend on resources such as Machine Translation systems or parallel corpora, but many languages can not rely on that kind of resources. Thus, they need other strategies based on less or more easily available resources, such as bilingual dictionaries. The translation is usually done from the language of the query to language of the target collection, mainly due to scalability reasons.

Lately, some authors have underlined the importance of using session information for obtaining better rankings [Carterette et al., 2011]. They claim that users often try to solve their information need by submitting more than one query, reformulating the initial query. Thus, a process regarding to an information need is often composed by several related queries. After entering an initial query, users tend to reformulate the

query in different ways such as specification (e.g., q_i=*"scyhe"* q_r=*"scythe mythology"*), generalization (e.g., q_i=*"computer worms"* q_r=*"malware"*) or drifting (e.g., q_i=*"sun spot activity"* q_r=*"sun spot earthquake"*). Studies on web search query logs showed that half of all Web users reformulated their initial query: 52% of the users in the 1997 Excite data set and 45% of the users in the 2001 Excite data-set [Wolfram et al., 2001]. This paper studies the use of this session context in order to improve the translation quality of the queries and the corresponding retrieval process.

We propose to use the previous queries of the same session in order to improve the query translation step in a CLIR system. Our hypothesis is that queries corresponding to the same session can be used as adequate additional context for improving the translation selection process. For instance, let us assume a session $s = \{q_i, q_r\}$ involving two queries in Basque: the initial query q_i="Neil Young diska", and its reformulation q_r=*"Neil Young bira data"* (*"Neil Young album"* and *"Neil Young tour date"*, respectively). Translation of q_r without using any context would be wrong, $tr(q_r)$=*"Neil Young turn date"*, but using q_i as context we are able to produce the correct translation, $tr(q_r|q_i)$=*"Neil Young tour date"*, because *"diska"* helps to disambiguate and select the correct translation for *"bira"*.

V.1.2 Related work

Several methods are proposed in the literature to deal with the query translation problem. The various techniques can be grouped depending on the translation-knowledge source as follows: MT-based, parallel corpus-based, and bilingual dictionary-based. For the last two groups different statistical frameworks are proposed: cross-lingual probabilistic relevance models and cross-lingual language models. The first framework offers operators to treat the ambiguous translations and it is usually used along with dictionaries. The second one incorporates translation probabilities on a more formal and unified framework which is obtained from parallel corpora [Hiemstra, 2001]. Although the results depend on the quality of the resources, usually better results are achieved with cross-lingual language models [Xu et al., 2001].

MT-systems and parallel corpora are a scarce for many languages (e.g., Basque). Dictionaries are more accessible for this kind of languages, but they are not free of problems: ambiguous translations must be dealt with. For the translation selection, Pirkola [1998] proposed to use structured queries along with probabilistic relevance models. In this approach all translations of a source word are treated as the same token when *TF* and *DF* statistics are

calculated for the translations of that source word. Other authors propose to use the target collection as a kind of language model to solve more precisely the translation selection problem [Monz and Dorr, 2005, Ballesteros and Croft, 1998]. Both kind of approaches were studied for the case of English-Basque pair on [Saralegi and Lopez de Lacalle, 2010a].

Research in Information Retrieval has traditionally focused on serving the best results for a single query. In practice however users often enter queries in sessions of reformulations. The different editions of Sessions Tracks at TREC, implement experiments to evaluate the effectiveness of retrieval systems over query reformulations. In the TREC 2010 Session track [Kanoulas et al., 2010] sessions were made up of two queries and three types of reformulations were considered:

1. *Generalization*: The user reformulates a more general query when the results are too narrow for him.

2. *Specification*: The user reformulates a more specific query when the results are too broad for him.

3. *Drifting*: The user reformulates another query with the same level of specification but moved to a different aspect or facet.

Overall, systems appeared to perform better over the generalization and drifting sessions than the specification ones [Kanoulas et al., 2010]. However only one team achieved a significant statistical improvement. The topics for next editions were collected from real user sessions with a search engine. Sessions were longer and reformulation types were not annotated. In 2011 and 2012 tracks about half of the submitted runs improved the baseline (no information about the session) by using the information about prior queries or using information about prior queries and retrieved results. In 2013 and 2014 editions most of the submitted runs were able to improve the baseline.

There are no papers dealing with query translation by using session information. The most similar works to this topic are those which exploit query logs and web clickthrough data for generation of cross-lingual query suggestions [Gao et al., 2007] and mining translation of web queries [Hu et al., 2008]. Gao et al. [2007] introduced a method of calculating the similarity between source language query and the target language query by exploiting, in addition to the translation information, a wide spectrum of bilingual and monolingual information, such as term co-occurrences and query logs with click-through. They used a discriminative model to learn the cross-lingual query similarity from a set of manually translated queries.

Hu et al. [2008] proposed a methodology for mining query translation pairs from the knowledge hidden in the click-through data. In a first step they identified bilingual URL pair patterns in the click-through data. In a second step they matched query translation pairs based on user click behaviour. Finally, query pairs are generated based on co-occurrence analysis of the click-through data.

V.1.3 Experimental setup

As mentioned before, our objective is to improve the performance of the query translation by using previous queries of the same session as context. In order to carry out the experiments we need a test collection including query sessions. We used the query session set built for TREC 2010 Session track. This set includes 150 pairs of initial and reformulated queries (q_i, q_r), grouped by their reformulation type (48 generalization, 52 specification, and 50 drifting). The query pairs were constructed from TREC 2009 and 2010 Web Track diversity topics by using the aspect and main theme of them in a variety of combinations to simulate a session composed of an initial and a second query. These topics correspond to the Clueweb09 collection.

Topic	Initial query q_i	Reformulation q_r	Reformulation type
2	*"hoboken"*	*"hoboken nightlife"*	specification
5	*"low carb high fat diet"*	*"list of diets"*	generalization
9	*"wooden fence"*	*"chain link fence"*	drifting

Table V.1: Examples of reformulation types from TREC 2010 Session track's dataset.

All the 150 query pairs were translated to Basque manually because the objective of the experiment was to evaluate the Basque to English retrieval process. Query pairs extracted from a Basque to English CLIR system's log would be more realistic. However, such log data was not available. In addition, by using 2010 Session track's test data we obtained a more standarized test-data. Due to effort limitations we choose the 2010 Session track's dataset over the next Session tracks' datasets because it provided reformulation types which offered us more information for the analysis stage.

V.1.4 Query translation algorithm

We adopted a bilingual dictionary based strategy for deal with the CLIR process. At first, Basque queries were translated into English. Translation candidates for query terms were obtained from Elhuyar Basque-English dictionary which includes 77,864 entries and 28,874 Basque headwords. Then, an iterative algorithm based on target language co-occurrences was applied for selecting the correct translation when multiple candidates were available. This iterative algorithm is based on [Monz and Dorr, 2005] and its application on the Basque to English CLIR task was evaluated by Saralegi and Lopez de Lacalle [2010a]. The algorithm selects the translation candidates that maximize the association degree between them according to a target collection.

Initially, all the translation candidates t given by the dictionary for each query term s_i are equally likely:

$$w_0(t|s_i) = \frac{1}{|tr(s_i)|} \tag{V.1}$$

In the iteration step, the translation weight $w_n(t|s_i)$ of each translation candidate is updated according to the translation weights $w_{n-1}(t'|s_i)$ of the rest of the candidates ($inlink(t)$) and the association degree between them $L(t,t')$:

$$w_n(t|s_i) = w_{n-1}(t|s_i) + \sum_{t' \in inlink(t)} L(t,t') \cdot w_{n-1}(t'|s_i) \tag{V.2}$$

Then, each translation's weight is re-computed and normalized. The iteration stops when the variations of the term weights become smaller than a predefined threshold. The weight of the association between candidates $L(t,t')$ is computed by calculating the Log-likelihood ratio association measure. We investigated extracting the frequencies (marginal and joint frequencies) required by the Log-likelihood ratio from three collections: ClueWeb09, Wikipedia, and the web (by using Bing search engine). There was not any difference between them, in terms of HTER (Translation Edit Rate). Finally, Wikipedia was used as source collection because of efficiency reasons.

V.1.5 Query translation using session as context

V.1.5.1 Evaluation of query translation

The quality of translations was evaluated by means of the HTER measure [Snover et al., 2006]. This measure computes the average amount of editing that a human would have to perform to correct the output of a system. A lower HTER value means a better translation. We do not penalize wrong word order in the translation because it does not have any negative effect on the retrieval process. In addition to this, according to [Snover et al., 2006], HTER achieves higher correlations than BLEU with human judgements. A fluent speaker in Basque and English performed the minimum number of edits over the English translations provided by the strategies which will be introduced in this section, in order to compute HTER. We analysed different strategies for combining the initial query and its reformulation in order to improve the translation of the reformulated query:

1. $tr(q_r)$: Translation of the reformulated query without previous query information. This would be the baseline.

2. $tr(q_r|q_i)$: Translation of the reformulated query using the previous query as additional context.

3. $tr(q_r|tr(q_i))$: translation of the reformulated query using the translation of the previous query as additional context.

The hypothesis behind the second strategy is that the words of the previous query contribute positively in the iterative algorithm when finding the correct translations for q_r. For performing this strategy the association degrees between the words in the initial and reformulated queries are taken into account when the translation algorithm is applied over the reformulated query. Thus, the translations of the words in the initial query are added to the $inlink(t)$ set. We tested a coefficient c for giving more or less importance to the words of the initial query when $L(t, t')$ was computed. $L(t, t')$ is modified by the coefficient c when t or t' belongs to q_i. Best results were achieved when more weight ($c = 2$) was given to the initial query words, with a one point absolute error reduction with respect to the baseline or first strategy, which does not use previous queries.

The third strategy consists on using the translation of the initial query q_i as additional context when translating q_r. The translation of q_i is performed by using the same iterative algorithm. In this case the idea is to to include less and more precise context words. The hypothesis behind this strategy

is that, in spite of including some wrong translations, these contexts are more helpful for the iterative translation selection algorithm. For example, the reformulated query *"PS 2 joko berriak"* (*"new PS 2 games"*) was wrongly translated to *"PS 2* ***set*** *new"* with the second strategy. The third strategy, which uses the translation of the initial query as context *tr("ps 2 games")="ps 2 joku"*, provides a correct translation *"PS 2 game new"*. Table V.2 presents the results, showing that better translations are obtained with this third strategy.

Strategy	Clueweb09
$tr(q_r)$	0.160
$tr(q_r\|q_i)$	0.157
$tr(q_r\|tr(q_i))$	**0.140**

Table V.2: Average HTER depending on translation strategy (smaller is better).

We also analyzed whether any reformulation type could benefit more than the others from using the session information on the translation selection process. The evaluation of translated queries with respect to the different reformulation types shows that the drifting and specification types are more susceptible to be improved (see table V.3). In the case of drifting reformulation words of initial query contribute with new information useful to disambiguate some words of reformulation. This is the case of q_i*="neil young* ***diska****"* (*"neil young album"*) q_r*="neil young* ***bira*** *data"* (*"neil young tour date"*) pair, where *"album"* helps to correctly disambiguate *"bira"*. The specification reformulations are improved because initial queries are more general and involve frequent phrases, which the iteration algorithm tends to translate correctly. Using these translations as context helps translation selection process. For example. the reformulated query q_r*="txakur adopzio erakunde"* is translated correctly to $tr(q_r)$*="dog adoption organization"* when including the translation of the initial query (q_i*="txakur adopzio"* translated as *"dog adoption"*), which is performed correctly because it is a frequent phrase, and thus providing useful context words.

We analysed manually some translations of reformulations that theoretically could be improved using the previous query as additional context in the translation selection stage. An analysis of the co-occurrence and Log-likelihood ratio values obtained from the collection was carried out. In some cases, we realized that the semantic relatedness we detected

manually between some initial and reformulated queries was not strong enough to affect the translation selection process of the reformulation. In other cases, the manually identified semantic relatedness was not reflected adequately in the collection used for mining co-occurrences. And due to the variety of the topics in the test-set, it is difficult to build an unique collection which fits all of them.

Reformulation type	HTER for $tr(q_r)$	HTER for $tr(q_r\|tr(q_i))$
generalization	0.118	**0.107**
specification	0.200	**0.179**
drifting	0.152	**0.130**

Table V.3: Improvement on translation quality depending on reformulation type (smaller is better).

V.1.5.2 Evaluation of the retrieval process

Next, the retrieval process was evaluated. The translated queries were processed with the Batch Query service for Clueweb09[1] which is based on the Indri search engine. We evaluated the three strategies mentioned above: a) $tr(q_r)$ translation of q_r without previous query information (the baseline), b) $tr(q_r|q_i)$ translation of q_r using the previous query as additional context, and c) $tr(q_r|tr(q_i))$ translation of q_r using translation of the previous query as additional context.

The results in table V.4 show that the best result is achieved by using the translation of the initial query as context for translating the reformulated query, with up to 5.1% improvement. This improvements of $tr(q_r|tr(q_i))$ over $tr(q_r)$ on p@10, MAP, and nDCG@10 are significant according to the Paired Randomization Test with α=0.05. The results correspond very well to the improvement in translation quality reported in the previous section.

A comparison of the IR results by reformulation type with respect to translation quality (table V.5 vs. table V.3) shows that, unlike the translation quality, the IR improvements for drifting reformulations is the weakest. Hand inspection showed that, in some cases, producing better translation does not necessarily mean that the information need is expressed better. However, this fact should be contrasted with a more extended topic-set including this type of reformulations.

[1] http://boston.lti.cs.cmu.edu/Services/batchquery/

Strategy	p@10	Impr. over $tr(q_r)$	MAP	Impr. over $tr(q_r)$	nDCG@10	Impr. over $tr(q_r)$
$tr(q_r)$	0.148	-	0.060	-	0.157	-
$tr(q_r\|q_i)$	0.152	2.7%	0.060	0%	0.162	3.2%
$tr(q_r\|tr(q_i))$	**0.154**	4.1%	**0.063**	5%	**0.165**	5.1%

Table V.4: Retrieval performance depending on query translation and building strategy.

generalization			
Strategy	p@10	MAP	nDCG@10
$tr(q_r)$	0.145	0.053	0.157
$tr(q_r\|q_i)$	**0.148**	**0.055**	**0.162**
specification			
Strategy	p@10	MAP	nDCG@10
$tr(q_r)$	0.143	0.056	0.148
$tr(q_r\|tr(q_i))$	**0.160**	**0.061**	**0.167**
drifting			
Strategy	p@10	MAP	nDCG@10
$tr(q_r)$	**0.155**	0.070	**0.165**
$tr(q_r\|tr(q_i))$	**0.155**	**0.071**	**0.165**

Table V.5: Retrieval performance depending on reformulation type.

V.1.6 Conclusions

This work shows that: 1) the quality of query translation can be improved using previous queries as context, 2) the improvements in translation quality transfer to improvements in CLIR performance, 3) translation quality improved in all three types of sessions (generalization, specification, and drifting), and CLIR improved for generalization and specification sessions, preserving the performance in drifting sessions, 4) the best strategy to include the initial query as context is to translate it and then to use the translation in the iterative translation selection algorithm.

The main limitation to obtain higher improvements is due to the weak semantic relatedness scores between the words in the initial query and in the reformulated query. In some cases the related words are not well represented in the collection. Future works will be focused on performing further experiments with different datasets including longer query sessions. We expect that using a longer context, in terms of amount of previous queries, can mitigate the problems derived from the aforementioned limitations.

VI. KAPITULUA

Hiztegi elebidunen sorkuntza pibotaje-tekniken bidez

[Saralegi et al., 2011] artikuluan bi hiztegi elebidunen gurutzaketa hutsaren bidez lortutako hiztegitik itzulpen okerrak kimatzeko bi teknika aztertu eta zenbait eszenatokitan ebaluatu dira. Bata, alderantziko kontsulta (*Inverse Consultation*, IC ingelesez) [Tanaka and Umemura, 1994], hiztegien egiturez soilik baliatzen da. Bestea [Kaji et al., 2008, Gamallo and Pichel, 2010], corpus konparagarri elebidunetatik kalkulatutako antzekotasun distribuzionalean dago oinarrituta.

- Xabier Saralegi, Iker Manterola, and Inaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011

VI.1 Analyzing methods for improving precision of pivot based bilingual dictionaries

Analyzing Methods for Improving Precision of Pivot Based Bilingual Dictionaries

Xabier Saralegi, Iker Manterola, Iñaki San Vicente
R&D Elhuyar Foundation
Zelai haundi 3, Osinalde Industrialdea
20170 Usurbil, Basque Country
{x.saralegi, i.manterola, i.sanvicente}@elhuyar.com

An A-C bilingual dictionary can be inferred by merging A-B and B-C dictionaries using B as pivot. However, polysemous pivot words often produce wrong translation candidates. This paper analyzes two methods for pruning wrong candidates: one based on exploiting the structure of the source dictionaries, and the other based on distributional similarity computed from comparable corpora. As both methods depend exclusively on easily available resources, they are well suited to less resourced languages. We studied whether these two techniques complement each other given that they are based on different paradigms. We also researched combining them by looking for the best adequacy depending on various application scenarios.

VI.1.1 Introduction

Nobody doubts the usefulness and multiple applications of bilingual dictionaries: as the final product in lexicography, translation, language learning, etc. or as a basic resource in several fields such as Natural Language Processing (NLP) or Information Retrieval (IR), too. Unfortunately, only major languages have many bilingual dictionaries. Furthermore, construction by hand is a very tedious job. Therefore, less resourced languages (as well as less-common language pairs) could benefit from a method to reduce the costs of constructing bilingual dictionaries. With the growth of the web, resources like Wikipedia seem to be a good option to extract new bilingual lexicon [Erdmann et al., 2008], but the reality is that a dictionary is quite different from an encyclopedia. Wiktionary[1] is a promising asset more oriented towards lexicography. However, the presence of less resourced languages in these kinds of resources is still relative -in Wikipedia, too-.

[1] http://www.wiktionary.org/

Another way to create bilingual dictionaries is by using the most widespread languages (e.g., English, Spanish, French...) as a bridge between less resourced languages, since most languages have some bilingual dictionary to/from a major language. These pivot techniques allow new bilingual dictionaries to be built automatically. However, as the next section will show, it is no small task because translation between words is not a transitive relation at all. The presence of polysemous or ambiguous words in any of the dictionaries involved may produce wrong translation pairs. Several techniques have been proposed to deal with these ambiguity cases [Tanaka and Umemura, 1994, Shirai and Yamamoto, 2001, Bond et al., 2001, Paik et al., 2004, Kaji et al., 2008, Shezaf and Rappoport, 2010]. However, each technique has different performance and properties producing dictionaries of certain characteristics, such as different levels of coverage of entries and/or translations. The importance of these characteristics depends on the context of use of the dictionary. For example, a small dictionary containing the most basic vocabulary and the corresponding most frequent translations can be adequate for some IR and NLP tasks, tourism, or initial stages of language learning. Alternatively, a dictionary which maximizes the vocabulary coverage is more oriented towards advanced users or translation services.

This paper addresses the problem of pruning wrong translations when building bilingual dictionaries by means of pivot techniques. We aimed to come up with a method suitable for less resourced languages. We analyzed two of the approaches proposed in the literature which are not very demanding on resources: Inverse Consultation (IC) [Tanaka and Umemura, 1994] and Distributional Similarity (DS) [Kaji et al., 2008], their strong points and weaknesses, and proposed that these two paradigms be combined. For this purpose, we studied the effect the attributes of the source dictionaries have on the performance of IC and DS-based methods, as well as the characteristics of the dictionaries produced. This could allow us to predict the performance of each method just by looking at the characteristics of the source dictionaries. Finally, we tried to provide the best combination adapted to various application scenarios which can be extrapolated to other languages.

The basis of the pivot technique is dealt with in the next section, and the state of the art in pivot techniques is reviewed in the third section. After that, the analysis of the aforementioned approaches and experiments carried out for that purpose are presented, and a proposal for combining both paradigms is included. The paper ends by drawing some conclusions from the results.

VI.1.2 Pivot Technique

The basic pivot-oriented construction method is based on assuming the transitive relation of the translation of a word between two languages. Thus:

if p (pivot word) is a translation of s (source word) in the A-B dictionary and t (target word) is a translation of p in the B-C dictionary, we can say that t is therefore a translation of s, or $translation_{A,B}(s) = p$ and $translation_{B,C}(p) = t \rightarrow translation_{A,C}(s) = t$

This simplification is incorrect because it does not take into account word senses. Translations correspond to certain senses of the source words. If we look at figure VI.1, t (case of t_1 and t_2) can be the translation of p (p_2) for a sense c (c_3) different from the sense for which p (p_2) is the equivalent of s (c_1). This can happen when p pivot word is polysemous.

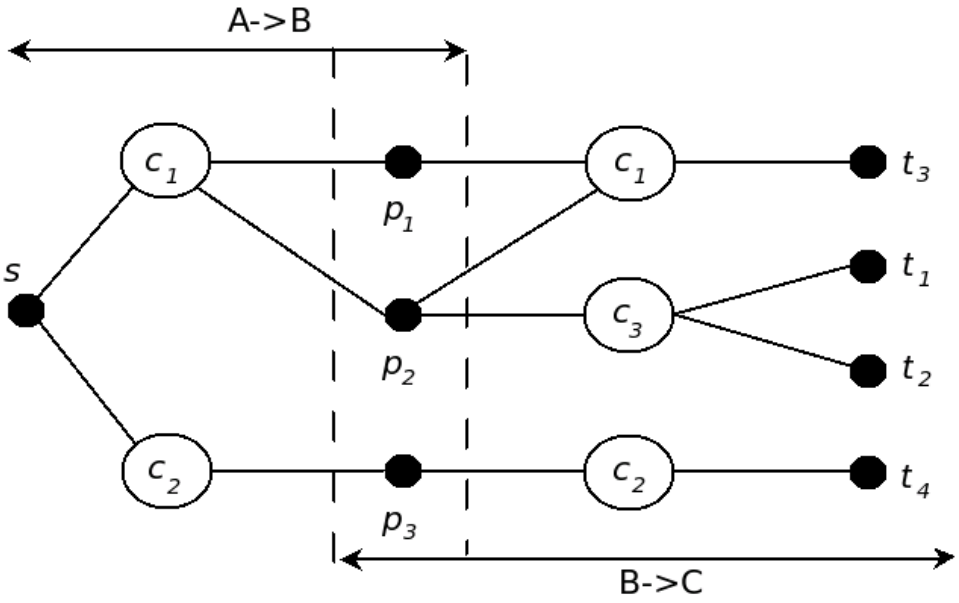


Figure VI.1: Ambiguity problem of the pivot technique.

It could be thought that these causalities are not frequent, and that the performance of this basic approach could be acceptable. Let us analyze a real case. We merged a Basque-English dictionary composed of 17,672 entries and 43,021 pairs with an English-Spanish one composed of 16,326 entries and 38,128 pairs, and obtained a noised Basque-Spanish dictionary comprising 14,000 entries and 104,165 pairs. 10,000 (99,844 pairs) among all the entries have more than one translation. An automatic evaluation shows that 80.32%

of these ambiguous entries contain incorrect translation equivalents (80,200 pairs out of 99,844). These results show that a basic pivot-oriented method is very sensitive to the ambiguity level of the source dictionaries. The conclusion is that the transitive relation between words across languages can not be assumed, because of the large number of ambiguous entries that dictionaries actually have. A more precise statement for the transitive property in the translation process would be:

if p (pivot word) is a translation of s with respect to a sense c and t is a translation of p with respect to the same sense c we can say that t is a translation of s, or $translation_{A,B}(s_{c_1}) = p$ and $translation_{B,C}(p_{c_2}) = t$ and $c_1 = c_2 \rightarrow translation_{A,C}(s) = t$

Unfortunately, most dictionaries lack comparable information about senses in their entries. So it is not possible to map entries and translation equivalents according to their corresponding senses. As an alternative, most papers try to guide this mapping according to semantic distances extracted from the dictionaries themselves or from external resources such as corpora.

Another problem inherent in pivot-based techniques consists of missing translations. This consists of pairs of equivalents not identified in the pivot process because there is no pivot word, or else one of the equivalents is not present. We will not be dealing with this issue in this work so that we can focus on the translation ambiguity problem.

VI.1.3 State of the Art

In order to reject wrong translation pairs, Tanaka and Umemura [1994] worked with the structure of the source dictionaries and introduced the IC method which measures the semantic distance between two words according to the number of pivot-words they share. This method was extended by using additional information from dictionaries, such as semantic classes and POS information in [Bond et al., 2001, Bond and Ogura, 2008]. Sjöbergh [2005] compared full definitions in order to detect words corresponding to the same sense. However, not all the dictionaries provide this kind of information. Therefore, external knowledge needs to be used in order to guide mapping according to sense. István and Shoichi [2009] proposed using WordNet, only for the pivot language (for English in their case), to take advantage of all the semantic information that WordNet can provide. Mausam et al. [2009] researched the use of multiple languages as pivots, on the hypothesis that the more languages used, the more evidences will be found to find translation equivalents. They used Wiktionary for building a multilingual lexicon. Tsunakawa et al. [2008] used parallel corpora to estimate translation

probabilities between possible translation pairs. Those reaching a minimum threshold are accepted as correct translations to be included in the target dictionary. However, even if this strategy achieves the best results in the terminology extraction field, it is not adequate when less resourced languages are involved because parallel corpora are very scarce.

As an alternative, [Kaji et al., 2008, Gamallo and Pichel, 2010] proposed methods to eliminate spurious translations using cross-lingual context or distributional similarity calculated from comparable corpora. In this line of work, Shezaf and Rappoport [2010] propose a variant of DS, and show how it outperforms the IC method. In comparison, our work focuses on analyzing the strong and weak points of each technique and aims to combine the benefits of each of them.

Other characteristics of the merged dictionaries like directionality [Paik et al., 2004] also influence the results.

VI.1.4 Experimental Setup

This work focuses on adequate approaches for less resourced languages. Thus, the assumption for the experimentation is that few resources are available for both source and target languages. The resources for building the new dictionary are two basic (no definitions, no senses) bilingual dictionaries (A-B, B-C) including source (A), target (C) and a pivot language (B), as well as a comparable corpus for the source-target (A-C) language pair. We explored the IC [Tanaka and Umemura, 1994] and DS [Kaji et al., 2008, Gamallo and Pichel, 2010] approaches. In our experiments, the source and target languages are Basque and Spanish, respectively, and English is used for pivot purposes. In any case, the experiments could be conducted with any other language set, so long the required resources are available.

It must be noted that the proposed task is not a real problem because there is a Basque-Spanish dictionary already available. Resources like parallel corpora for that language pair are also available. These dictionaries and pivot language were selected in order to be able to evaluate the results automatically. During the evaluation we also used frequency information extracted from a parallel corpus, but then again, this corpus was not used during the dictionary building process, and therefore, it would not be used in a real application environment.

VI.1.4.1 Resources

In order to carry out the experiments we used three dictionaries. The two dictionaries mentioned in the previous section (Basque-English $D_{eu \to en}$ and English-Spanish $D_{en \to es}$) were used to produce a new Basque-Spanish $D_{eu \to en \to es}$ dictionary. In addition, we used a Basque-Spanish $D_{eu \to es}$ dictionary for evaluation purposes. Its broad coverage is indicative of its suitability as a reference dictionary. Table VI.1 shows the main characteristics of the dictionaries. We can observe that the ambiguity level of the entries (average number of translations per source word) is significant. This produces more noise in the pivot process, but it also benefits IC due to the increase in pivot words. As for the directions of source dictionaries, English is taken as target. Like Paik et al. [2004] we obtained the best coverage of pairs in that way.

Dictionary	**#entries**	**#pairs**	**ambiguity level**
$D_{eu \to en}$	17,672	43,021	2.43
$D_{en \to es}$	16,326	38,128	2.33
$D_{eu \to es}$(reference)	57,334	138,579	2.42
$D_{eu \to en \to es}$(noisy)	14,601	104,172	7.13

Table VI.1: Characteristics of the dictionaries.

Since we were aiming to merge two general dictionaries, the most adequate strategy was to use open domain corpora to compute DS. The domain of journalism is considered to be close to the open domain, and so we constructed a Basque-Spanish comparable corpus composed of news articles (see table VI.2). The articles were gathered from the newspaper Diario Vasco (Hereinafter DV) for the Spanish part and from the Berria newspaper for the Basque part. Both publications focus on the Basque Country. In order to achieve a higher comparability degree, some constraints were applied:

- News in both languages corresponded to the same time span, 2006-2010.
- News corresponding to unrelated categories between newspapers were discarded.

In addition, as mentioned above, we extracted the frequencies of translation pairs from a Basque-Spanish parallel corpus. The corpus had

Corpus	#words	#docs
Berria(eu)	40Mw	149,892
DV(es)	77Mw	306,924

Table VI.2: Characteristics of the comparable corpora.

295,026 bilingual segments (4 Mw in Basque and 4.7 Mw in Spanish) from the domain of journalism.

VI.1.5 Pruning Methods

IC and DS a priori suffer different weak points. IC depends on the structure of the source dictionaries. On the other hand, DS depends on a good comparable corpus and translation process. DS is measured more precisely between frequent words because context representation is richer.

The conditions for good performance of both IC and DS are analyzed below. These conditions will then be linked to the required characteristics for the initial dictionaries. In addition, we will measure how divergent the entries solved for each method are.

VI.1.5.1 Inverse consultation

IC uses the structure of the D_{a-b} and D_{b-c} source dictionaries to measure the similarity of the meanings between source word and translation candidate. The description provided by Tanaka and Umemura [1994] is summarized as follows. To find suitable equivalents for a given entry, all target language translations of each pivot translation are looked up (e.g., $D_{b\to c}(D_{a\to b}(s))$). This way, all the "equivalence candidates" (ECs) are obtained. Then, each one is looked up in the inverse direction (following the previous example, $D_{c\to b}(t)$) to create a set of words called "selection area" (SA). The number of common elements of the same language between SA and the translations or equivalences (E) obtained in the original direction ($D_{a\to b}(s)$) is used to measure the semantic distance between entries and corresponding translations. The more matches there are, the better the candidate is. If only one inverse dictionary is consulted, the method is called "one time inverse consultation" or IC1. If n inverse dictionaries are consulted, the method is called "n time inverse consultation". As there is no significant difference in performance, we simply implemented IC1. Assuming that each element (x) of these two sets (SA, E) has a weight that is determined by the number

of times it appears in the set that belongs (X), this weight is denoted as $\delta(X, x)$. In the same way, the number of common elements between SA and E is denoted as follows:

$$\delta(E, SA) = \sum_{x \in SA} \delta(E, x) \tag{VI.1}$$

IC asks for more than one pivot word between source word s and translation candidate t. In our example:

$$\delta(D_{a \to b}(s), D_{c \to b}(t)) > 1 \tag{VI.2}$$

In general, this condition guarantees that pivot words belong to the same sense of the source word (e.g. *iturri→tap→grifo, iturri→faucet→grifo*). Consequently, source word and target word also belong to the same sense.

Conceptually, the IC method is based on the confluence of two evidences. Let us take our dictionaries as examples. If two or more pivot words share a translation t in the $D_{es \to en}$ dictionary ($|tr(t_c, D_{es \to en}| > 1$) (e.g. *grifo→tap, grifo→faucet*) we could hypothesize that they are lexical variants belonging to a unique sense c. If an entry s includes those translations ($|tr(s_c, D_{eu \to en})| > 1$) (e.g. *iturri→tap, iturri→faucet*)) in the $D_{eu \to en}$ dictionary, we could also hypothesize the same. We can conclude that entry s and candidate t are mutual translations because the hypothesis that *"faucet"* and *"tap"* are lexical variants of the same sense c is contrasted against two evidences. This makes IC highly dependant on the number of lexical variants. Specifically, IC needs several lexical variants in the pivot language per each entry sense in both dictionaries. Assuming that wrong pairs cannot fulfill this requirement (see Formula VI.2) we can estimate the probabilities of the conditions for solving an ambiguous pair (s, t) where s and $t \in c$, as follows:

(a) $p(|tr(s_c, D_{a \to b})| > 1)$: Estimated by computing the average coverage of lexical variants in the pivot language for each entry in $D_{a \to b}$.

(b) $p(|tr(t_c, D_{c \to b})| > 1)$: Estimated by computing the average coverage of lexical variants in the pivot language for each entry in $D_{c \to b}$.

(c) $p(|tr(s_c, D_{a \to b}) \bigcap tr(t_c, D_{c \to b})| > 1)$: Convergence degree between translations of s and t in $D_{a \to b}$ and $D_{c \to b}$ corresponding to c.

So, in order to obtain a good performance with IC, the dictionaries used need to provide a high coverage of lexical variants per sense in the pivot language. If we assume that variants of a sense do not vary considerably

between dictionaries, performance of IC in terms of recall would be estimated as follows:

$$R = p(|tr(s_c, D_{a\to b})| > 1) * p(|tr(t_c, D_{c\to b})| > 1) \quad \text{(VI.3)}$$

We estimated the adequacy of the different dictionaries in the experimental setup according to estimations (a) and (b). Average coverage of lexical variants in the pivot language was calculated for both dictionaries. It was possible because lexical variants in the target language were grouped according to senses in both dictionaries. Only ambiguous entries were analyzed because they are the set of entries which IC must solve. In the $D_{eu\to en}$ dictionary more than 75% of senses have more than one lexical variant in the pivot language. So, $p(|tr(s_c, D_{eu\to en})| > 1) = 0.75$. In $D_{es\to en}$ this percentage (23%) is much lower. So, $p(|tr(t_c, D_{es\to en})| > 1) = 0.23$. Therefore, $D_{eu\to en}$ dictionary is more suited to the IC method than $D_{es\to en}$. As the conditions must be met in the maximum of both dictionaries, performance according to Formula VI.3 would be: $0.75 * 0.23 = 0.17$. This means that IC alone could solve about 17% of ambiguous entries.

VI.1.5.2 Distributional Similarity

DS has been used successfully for extracting bilingual terminology from comparable corpora. The underlying idea is to identify as translation equivalents those words which show similar distributions or contexts across two corpora of different languages, assuming that this similarity is proportional to the semantic distance. In other words, establishing an equivalence between cross lingual semantic distance and translation probability. This technique can be used for pruning wrong translations produced in a pivot-based dictionary building process [Kaji et al., 2008, Gamallo and Pichel, 2010].

We used the traditional approach to compute DS [Fung, 1995, Rapp, 1999]. Following the "bag-of-words" paradigm, the contexts of a word w are represented by weighted collections of words. Those words are delimited by a window (± 5 words around w) and punctuation marks. The context words are weighted with regard to w according to the Log-likelihood ratio measure, and the context vector of w is formed. After representing word contexts in both languages, the algorithm computes for each source word the similarity between its context vector and all the context vectors corresponding to words in the target language by means of the cosine measure. To be able to compute the cross-lingual similarity, the context vectors are put in the same space by translating the vectors of the source words into the target language. This

is done by using a seed bilingual dictionary. The problem is that we do not have that bilingual dictionary, since that is precisely the one we are trying to build. We propose that dictionaries extracted from our noisy dictionary ($D_{eu \to en \to es}$) be used:

- Including the unambiguous entries only
- Including unambiguous entries and selecting the most frequent candidates according to the target language corpus for ambiguous entries
- The dictionary produced by the IC1 method

The second method performed better in the tests we carried out. So, that is the method implemented for the experiments in the next section.

DS calls for several conditions in order to perform well. For solving an ambiguous translation t of a source word s, both context representations must be accurate. The higher their frequency in the comparable corpus, the richer their context representation will be. In addition to context representation, the translation quality of contexts is also a critical factor for the performance of DS. Factors can be formulated as follows if we assume big and highly comparable corpora:

(a) Precision of context representation: this can be estimated by computing the frequency of the words

(b) Precision of translation process: this can be estimated by computing the quality of the seed dictionary

VI.1.6 Results

In order to evaluate the performance of each pruning method, the quality of the translations was measured according to the average precision and recall of translations per entry with respect to the reference dictionary. As we were not interested in dealing with missing translations, the reference for calculating recall was drawn up with respect to the intersection between the merged dictionary ($D_{eu \to en \to es}$) and the reference dictionary ($D_{eu \to es}$). F-score is the metric that combines both precision and recall.

We also introduced the frequency of use of both entry and pair as an aspect to take into account in the analysis of the results. It is better to deal effectively with frequent words and frequent translations than rare ones. Frequency of use of Basque words and frequency of source-target

translation equivalent pairs were extracted respectively from the open domain monolingual corpus and the parallel corpus described in the previous section. Corpora were lemmatized and POS tagged in both cases in order to extract the frequency information of the lemmas.

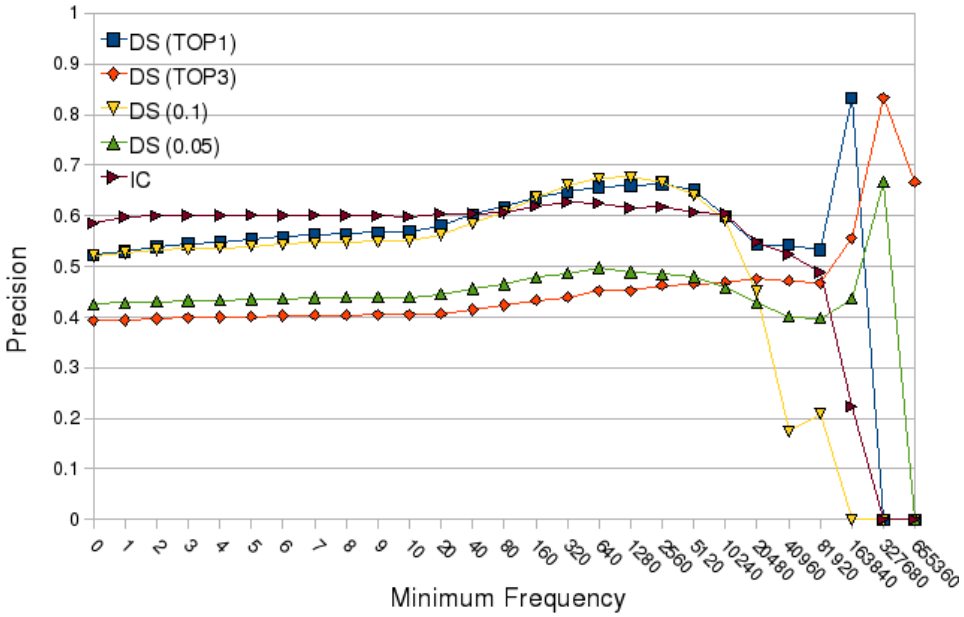


Figure VI.2: Precision results according to the minimum frequency of entries.

VI.1.6.1 Inverse Consultation

Results show that IC precision is about 0.6 (See figure VI.2). This means that many wrong pairs fulfill IC conditions. After analyzing the wrong pairs by hand, we observed that some of them corresponded to correct pairs not included in the reference dictionary. They are not included in the reference because not all synonyms -or lexical variants- are included in it, only the most common ones. This is an inherent problem in automatic evaluation, and affects all the experiments presented throughout section VI.1.6 equally. Other wrong pairs comprise translation equivalents which have the same stem but different gramatical categories (e.g., 'aldakuntza' (noun) *(change,*

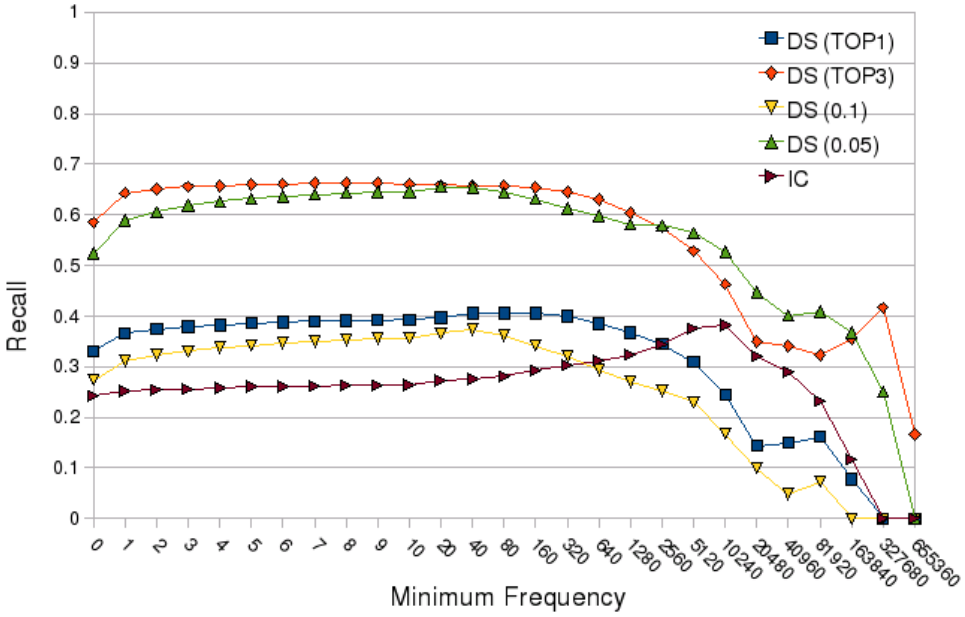


Figure VI.3: Recall results according to the minimum frequency of entries.

shift) → *'cambiar'* (verb) *(to change, to shift)*). These wrong cases could be filtered if POS information would be available in the source dictionaries.

Precision is slightly better when dealing with frequent words, a maximum of 0.62 is reached when minimum frequency is between 150 and 2,000. Precision starts to decline significantly when dealing with those entries over a minimum frequency of 10,000. However, only very few entries (234) reach that minimum frequency.

Recall is about 0.2 (See figure VI.3), close to the estimation computed in section VI.1.5.1. It presents a more marked variability according to the frequency of entries, improving the performance as the frecuency increases. This could be due to the fact that frequent entries tend to have more translation variants (See table VI.3). The fact that there are too many candidates to solve would explain why the recall starts to decline when dealing with very frequent entries.

Global performance according to F-score reflects the variability depending on frequency (See figure VI.4).

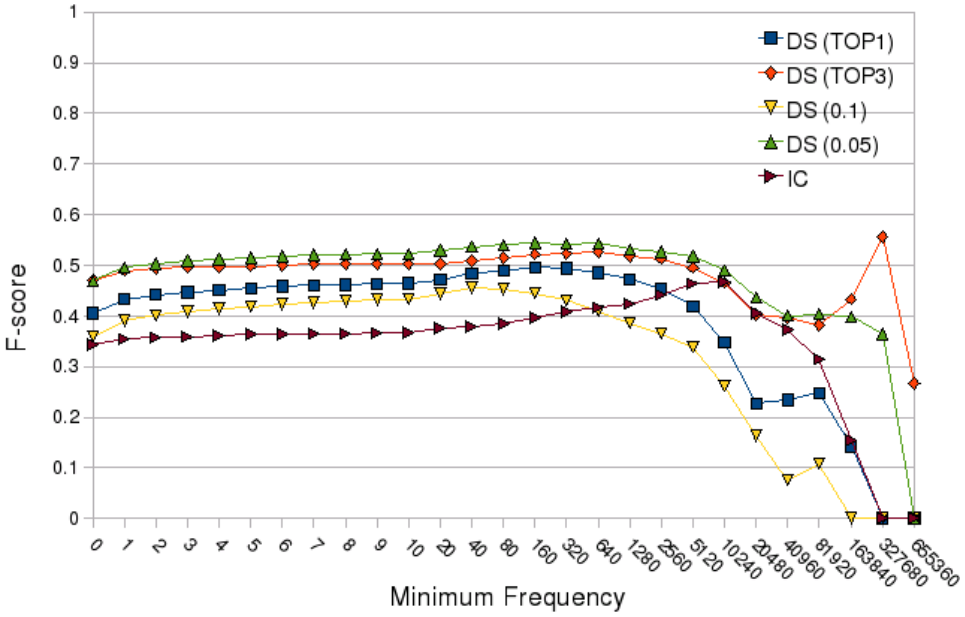


Figure VI.4: F-score results according to the minimum frequency of entries.

Recall according to frequency of pairs provides information about whether IC selects rare translations or the most probable ones (See figure VI.5). It must be noted that this recall is calculated with respect to the translation pairs of the merged dictionary $D_{eu \to en \to es}$ which appear in the parallel corpus (see section VI.1.4.1). Results (See figure VI.5) show that IC deals much better with frequent translation pairs. However, recall for pairs whose frequency is higher than 100 only reaches 0.5. Even if the maximum recall is achieved for pairs whose frequency is above 40,000, it is not significant because they suppose a minimum number (3 pairs). In short, we can conclude that IC often does not find the most probable translation (e.g. 'usain'→*'olor' (smell)*, 'zulo'→*'agujero' (hole)*,...).

VI.1.6.2 Distributional Similarity

DS provides an idea of semantic distance. However, in order to determine whether a candidate is a correct translation, a minimum threshold must be established. It is very difficult to establish a threshold manually because

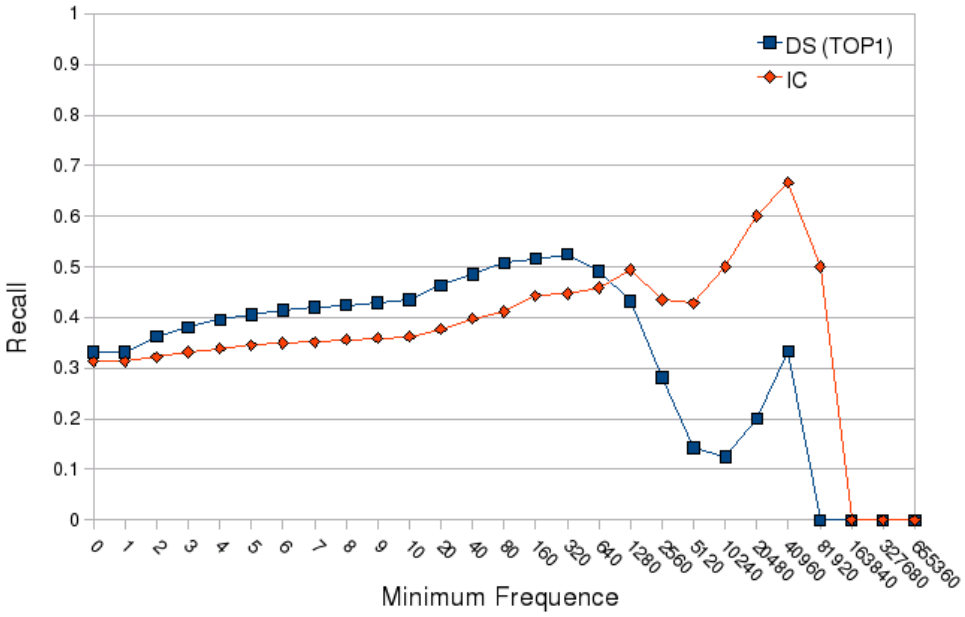


Figure VI.5: Recall results according to the minimum frequency of translation pairs.

its performance depends on the characteristics of the corpora and the seed dictionaries. The threshold can be applied at a global level, by establishing a numeric threshold for all candidates, or at local level by selecting certain top ranked candidates for each entry. The dictionary created by IC or unambiguous pairs can be used as a reference for tuning the threshold in a robust way with respect to the evaluation score such as F-score. In our experiments, thresholds estimated against the dictionary created by IC are very close to those calculated with respect to the whole reference dictionary (see figure VI.6).

There is not much variation in performance between local and global thresholds. Precision increases from 0.4 to 0.5 depending on the strictness level of the threshold (See figure VI.2), the stricter the better. In all cases, precision is slightly better when dealing with frequent words (frequency $>$ 20). This improvement is more marked with the strictest thresholds (TOP_1, 0.1). However, if global thresholds are used, performance starts to decline

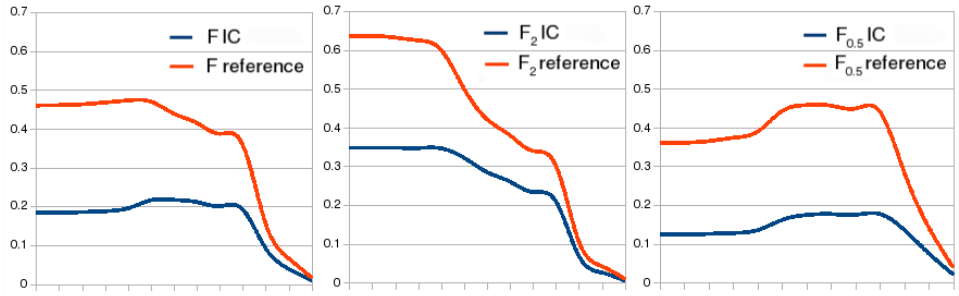


Figure VI.6: Threshold parameter tuning comparison for different F_n scores. Tuning against dictionary created by IC vs. Reference dictionary.

significantly when dealing with words whose frequency is above 1,000. So, it seems that local thresholds (TOP_3) perform more consistendly with respect to the high frequencies of entries.

Recall (See figure VI.3) goes from 0.5 to 0.7 depending on the strictness level of the threshold. It starts declining when frequency is above 50 depending on the type of threshold. In this case, global thresholds seem to perform better because the most frequent entries are handled better. These entries tend to have many translations. Therefore thresholds based on top ranks are too rigid.

There is no significant difference between global and local thresholds in terms of F-Score (See figure VI.4). Each threshold type is more stable in precision or recall. So the F-Score is similar for both. Variability of F-Score according to frequency is lower than in precision and recall. As performance peaks on both measures at different points of frequency, the variability is mitigated when measures are combined by F-Score.

We have plotted the recall according to the frequency of pairs calculated from a parallel corpus in order to analyze the performance of DS when dealing with frequent translation pairs (See figure VI.5). The performance decreases when dealing with pairs whose frequency is higher than 100. This means that DS's performance is worse when dealing with the most common translation pairs. So it is clear that it is very difficult to represent the contexts of very frequent words correctly.

The results show that DS rankings are worse when dealing with some words above a certain frequency threshold (e.g. 'on' *'good'*, 'berriz' *'again'*, 'buru' *'head'*, 'orain' *'now'*...). Although context representation of frequent words is based on many evidences, high polysemy level related to high frequency leads to a poorer representation. Alternatively we found that some

of those frequent words are not very polysemous. Those words do not have strong collocates, that is, they tend to appear freely in contexts, which also leads to poor representation. This low quality representation hampers an accurate computation of semantic distance.

VI.1.6.3 Comparison between IC and DS

As for average precision, IC provides better results than DS if all entries are taken into account. However, DS tips the scales in its favor if only entries with frequencies above 50 are considered and strict thresholds are used (TOP_1, 0.1).

DS clearly outperforms IC in terms of average recall of translations. Even if strict thresholds are used, DS outperforms IC for all entries whose frequency is lower than 640.

If average precision and recall are evaluated together by means of F-score, DS outperforms IC (figure VI.4). Only when dealing with very frequent entries (frequency $> 8,000$) is IC's performance close to DS's, but these entries make up a very small group (234 entries).

In order to compare the recall with respect to the frequency of translation pairs under the same conditions, we have to select a threshold that provides a similar precision to IC. TOP_1 is the most similar one (see figure VI.2). As figure VI.5 shows, again DS is better than IC. Even if IC's recall clearly surpasses DS's when dealing with frequent translation pairs (frequency $> 2,560$), it only represents a minimal number of pairs (39).

VI.1.6.4 Combining IC and DS according to different scenarios

In order to see how the methods can complement each other, we calculated the performance for solving ambiguous entries obtained by combining the results of both methods using various alternatives:

- Union: $IC \cup DS$: Pairs obtained by both methods are merged. Duplicated pairs are cleaned.
- Lineal combination (Lcomb): $IC * k + DS * (1 - k)$. Each method provides a value representing the translation score. For IC that value is the number of pivot words (see Formula VI.1), and the context similarity score in the case of DS. Those values are linearly combined and applied over the noised dictionary.

As mentioned in the first section, one of the goals of the paper was to analyze which method and which combination was best depending on the

use case. We have selected some measures which are a good indicator of good performance for different use cases:

- $AvgF$: Average F-score per entry.
- $wAvgF$: Average F-score per entry weighted by the frequency of the entry. Higher frequency increases the weight.
- $AvgF_2$: Average F-score per entry where recall is weighted higher.
- $AvgF_{0.5}$: Average F-score per entry where precision is weighted higher.

For the use cases presented in section VI.1.1, some measures will provide richer information than others. On the one hand, if we aim to build small, accurate dictionaries, $AvgF_{0.5}$ would be a better indicator since it attaches more importance to high precision. In addition, if we want the dictionaries to cover the most common entries (e.g., in a basic dictionary for language learners) it is also interesting to look at $wAvgF$ values because greater value is given to finding translations for the most frequent words. On the other hand, if our objective is to build big dictionaries with a high recall, it would be better to look at $AvgF_2$ measure which attaches importance to recall.

Method	$AvgF$	$wAvgF$	$AvgF_2$	$AvgF_{0.5}$
IC	0.34	0.27	0.27	0.46
DS	0.47	0.44	0.64	0.46
Union	**0.52**	**0.49**	0.65	0.49
Lcomb	**0.52**	**0.49**	**0.67**	**0.52**

Table VI.3: Performance results of methods for ambiguous entries according to different measures.

Table 3 shows the results for the different combinations. The parameters of all methods are optimized for each metric (as explained in section VI.1.6.2, see figure VI.6). In all cases, the combinations surpass the results of both methods separately. There is a reasonable improvement over DS (10.6% for $AvgF$), and an even more startling one over IC (52.9% for $AvgF$). IC only gets anywhere near the other methods when precision is given priority ($AvgF_{0.5}$). There is no significant difference in terms of performance between the two combinations, although Lcomb is slightly better. $wAvgF$ measure is stricter than the others since it takes frequency of entries into account. This is emphasised more in the case of IC where results decrease notably compared with $AvgF$.

VI.1.7 Conclusions

This paper has analyzed IC and DS, for the task of pruning wrong translations from bilingual dictionaries built by means of pivot techniques. After analyzing their strong and weak points we have showed that IC requires high ambiguity level dictionaries with several lexical variants per entry sense. With an average ambiguity close to 2 translation candidates DS obtains better results. IC is a high precision method, but contrary to our expectations, it seems that it is not much more precise than DS. In addition, DS offers much better recall of translations and entries. As a result, DS performs the best if both precision and recall are taken into account by F-score.

Both methods prune most probable translations for a significant number of frequent entries. DS encounters a problem when dealing with very frequent words due to the difficulty in representing their context. The main reason behind this is the high polysemy level of those words.

Our initial beliefs were that the translations found by each method would diverge to a certain extent. The results obtained when combining the two methods show that although the performance does not increase as much as expected (10.6% improvement over DS), there is in fact some divergence. As for the different use cases proposed, combinations offer the best performance in all cases. IC is indeed the poorer method, although it presents competitive results when precision is given priority.

Future experiments include contrasting these results with other dictionaries and language pairs.

VII. KAPITULUA

Ondorioak, ekarpenak eta etorkizuneko lanak

Tesi-lan honetan, baliabide urriko hizkuntzetarako egokiak diren CLIR teknikak aztertu eta garatu ditugu.

Horretarako, CLIR prozesuan ebatzi beharreko problema nagusia den kontsultaren itzulpena ebazteko teknikak landu eta ebaluatu ditugu, III. IV. eta V. kapituluetan azaldutako esperimentuetan. Murriztapenetako bat corpus paralelo eta itzultzaile automatikoen urritasuna izanik, hiztegi elebidunetan oinarritutako metodoetan jarri dugu arreta. Literaturan eta gure azterketetan ikusi dugun bezala, itzulpen anbiguoen trataera da hiztegien bidezko estrategian berariaz landu beharreko alderdi nagusia. Hori dela eta, esperimetu gehienak fenomeno hori tratatzera bideratu ditugu, eta aztertu eta garatutako teknika guztiak corpus paraleloen beharrik gabekoak izan dira. Zehazki, helburu-bilduma (III. kapituluan), corpus konparagarriak (IV. kapituluan), eta kontsulta-saioak (V.kapituluan) aztertu dira itzulpen-anbiguoak tratatzeko baliabide gisa.

Halaber, kontsulta itzultzeko metodo guztietan giltzarria den hiztegi elebiduna sortzeko teknikak aztertu eta garatu ditugu VI. kapituluan. Zehazki, pibotaje-ideian oinarritutako teknikak landu ditugu. Mota honetako tekniketan ere itzulpen-anbiguoen trataera ezinbestekoa da sortutako hiztegien doitasuna bermatze aldera. Problema hau ebazteko aztertu ditugun teknikek baliabide urriko hizkuntzen eskakizunak bete behar dituztenez gero, pibotaje-prozesuan erabilitako hiztegien egituratik eta corpus konparagarrietatik inferitutako informazioa baino ez dute baliatzen.

Egindako esperimentuetatik ondorioztatu dezakegu itzulpen automatikorako sistemak edota corpus paraleloak baliatu gabe posible dela CLIR sistema bat garatzea. Hau frogatu ahal izateko honako

hipotesiak egiaztatu ditugu:

- Hiztegietan oinarritutako teknikak balekoak direla itzulpen-hautapenaren arazoari aurre egiteko ere.
- Amarauna corpus konparagarri gisa erabiltzeak kontsultaren itzulpen-prozesua hobetzen laguntzen duela.
- Kontsulta-saioak kontsulten itzulpen-prozesua hobetzeko baliagarriak direla.
- Pibotaje bidezko metodoak eraginkorrak direla hiztegi elebidunak eraikitzeko hizkuntza pare batentzat hiztegirik ez dagoenean.

Kapitulu honetan esperimentuen emaitzen analisitik ateratako ondorioak, egin ditugun ekarpenak, eta etorkizunerako uzten ditugun lanak azalduko ditugu.

VII.1 Ondorioak

Esan bezala, tesi honetan kontsultak itzultzeko zenbait metodo ikertu ditugu, erraz eskuratu daitezkeen baliabideak baino behar ez dituztenak. Ezaugarri hau dela eta, baliabide urriko hizkuntzak dauden eszenatokietarako metodo egokiak dira. Gure lanak nagusiki euskara erabilera-kasutzat hartu badu ere, emaitzak estrapolagarriak dira antzeko egoerako hizkuntzetara.

III. kapituluko [Saralegi and Lopez de Lacalle, 2010a] artikuluan egindako euskara-gaztelania CLIR atazaren ebaluazioan, itzulpen-anbiguotasuna eta hitz anitzeko terminoak, elkarrekin erlazionatuta dauden fenomenoak, CLIR prozesuan kalitate-galera handiena eragiten duten fenomenoak direla ikusi dugu (ikusi VII.1 taula). Egindako esperimentuetan oinarrituta, egiaztatu dugu jaitsiera handiena kontsulta laburren kasuan gertatzen dela. Itzulpen anbiguoen tratamendu ezak %21,19ko eta %10,10eko beherakadak eragiten ditu kontsulta laburren eta luzeen kasuan, VII.1 taulak erakusten duen bezala. Hiztegitik kanpoko hitzen (*Out-Of-Vocabulary* edo OOV) presentzia ere eraginkortasunean eragiten duen fenomenoa da, %12.38ko jaitsiera eraginez kontsulta laburren berreskurapenean, eta %4,10koa kontsulta luzeenean, VII.1 taulan erakusten den bezala. Hiztegietan oinarritutako CLIR sistema baten garapenean, itzulpen-hautapenaren prozesuaren inplementazioa oso garrantzitsua dela erakusten dute emaitzak hauek. Halaber, OOV hitzen (hiztegitik

kanpoko hitzak) trataerak eragin nabarmena duela CLIR sistema baten eraginkortasunean.

Itzulpen anbiguoak		Hitz anitzeko terminoak		Hiztegitik kanpoko hitzak	
Laburrak	Luzeak	Laburrak	Luzeak	Laburrak	Luzeak
-21,19%	-10,10%	-19,81%	9,17%	12,38%	4,1%

VII.1 taula: Errore-mota bakoitzak CLIR prozesuaren eraginkortasunean eragindako beherakada portzentuala (MAP) $eu \rightarrow es$ norabiderako.

III. kapituluan aurkeztutako [Saralegi and Lopez de Lacalle, 2009] artikuluko esperimentuetan arreta berezia jarri dugu itzulpen-hautapenaren probleman, OOV hitzen trataera ere landu bada ere. Hiztegien bidezko kontsultaren itzulpenean gertatutako itzulpen-anbiguotasuna ebaztea ez da batere prozesu tribiala. Are gutxiago corpus paraleloak bezalako itzulpen-ezagutza esplizitua duten baliabideak ez erabiltzea murriztapen gisa jartzen badugu.

Itzulpen-hautapena inplementatzerakoan corpus paraleloak ez erabiltzeko alternatibak daudela frogatu dugu. CLIR atazan itzulpen-hautapena egiteko helburu-bilduma ustiatu daitekeela ikusi dugu. Funtsezko ideia helburu-bilduman elkarrekin agertzeko joera duten itzulpen-hautagaien multzoa aukeratzea da. Estrategia hau inplementatzeko bi teknika aztertu ditugu. Lehenengo teknikak (Pirkolaren metodoa) ez du aukeraketa esplizitutik egiten, bildumatik jasotako agerkidetza-estatistikak ranking-funtzioko *tf* eta *idf* estatistikoen kalkuluen eragiten dute eta. Hautapena berreskurapen-garaian modu inplizituan egiten dela esan daiteke. Bigarren teknika bildumako agerkidetza-estatistikak modu zuzenean baliatzen dituen *greedy* algoritmo batean oinarritzen da. Itzulpen-hautagaien agerkidetza-estatistikak baliatuz *greedy* metodoak elkartze-maila handiena duten itzulpen-hautagaien multzoa aukeratzen du.

Bi teknikek hiztegiko lehen itzulpena hartzen duen oinarri-lerroko emaitzak nabarmen hobetzen dituzte, eta ingelesezko elebakarraren eraginkortasunaren %80ra hurbiltzen dira (goren lerroa), VII.2 taulan erakusten den bezala. Esperimentuetan bi metodoen artean ez dugu alde esanguratsurik aurkitu kontsulta laburren kasuan. Kontsulta luzeak itzultzean Pirkolaren metodoak emaitza hobeak lortzen ditu. Bi metodoak konbinatuz kontsulta laburren kasuan lortzen da hobekuntza, baina oso txikia da.

Bi metodoen artean eraginkortasunean alde handirik ez badago ere,

nabarmentzekoa da Pirkolaren metodoa kostu konputazional txikiagokoa dela. Hori dela eta, egokiagoa litzateke konputazio-kontsumoa lehentasun duten eszenatokietarako. Halere, hainbat IR tresnetan zaila da Pirkolaren metodoa inplementatzea, ranking-funtzioaren kodean aldaketak eskatzen dituelako. Kasu horietan agerkidetza-estatistiketan oinarritutako *greedy* metodoa errazago integratu liteke, ranking-funtziotik kanpokoa baita.

	MAP		Ingelesezko elebakarrarekiko %		Lehen itzulpenarekiko hobekuntza	
	Lab.	Luz.	Lab.	Luz.	Lab.	Luz.
Ingelesezko elebakarra	0,3176	0,3778	-	-	-	-
Lehen itzulpena	0,2118	0,2500	%67	%66	-	-
Pirkola	0,2342	0,2959	%74	%78	+%9,56	+%15,51
Agerkidetzak (*greedy*)	0,2338	0,2725	%74	%72	+%9,41	+%8,26
Hibridoa	0,2404	0,2920	%76	%77	+%11,9	+%14,38

VII.2 taula: Helburu-bilduman oinarritutako itzulpen-hautapenerako metodoen ebaluaketa *eu* $\rightarrow$ *en* norabiderako. Hibridoa Pirkola eta Agerkidetzen (*greedy*) konbinazioa da. Kontsula laburrak (Lab.) eta kontsulta luzeak (Luz.).

III. kapituluan landutako teknikek helburu-bildumako agerkidetzak ustiatzen dituzte, baina informazio hori mugatua da zenbait itzulpen-hautagai ebazteko. Hori dela eta, beste bi informazio-iturri aztertu ditugu IV. eta V. kapituluetan: corpus konparagarriak eta kontsulta-saioak, hurrenez hurren. Bi informazio-iturri hauek ustiatuz itzulpena hautatzeko prozesua eta ondorengo berreskurapena gehiago findu daitezkeela egiaztatu dugu.

IV. kapituluan azaldatuko [Saralegi and Lopez de Lacalle, 2010b] artikuluan corpus konparagarrietatik itzulpen-ezagutza erauzi daitekeela frogatu dugu itzulpen-hautapena hobeto egiteko. Helburu-bildumatik erauzitako informazioa ez bezala, jatorrizko hitzaren eta bere itzulpenaren arteko elkartze-mailari lotutako informazioa erauzi daiteke corpus konparagarrietatik. Elkartze-maila hauek estimatzeko hizkuntza-arteko antzekotasun distribuzionala baliagarria dela ikusi dugu. Modu honetan

erauzitako itzulpen-probabilitateak Pirkolaren metodoan (pisuak onartzen dituen aldaera [Darwish and Oard, 2003]) integratuz (VII.3 taulan Pirkola+corpus konparagarriak) hobekuntza estatistikoki esanguratsua lortzen dugu, %3,49koa. Hizkuntza-arteko antzekotasun distribuzionala kalkulatzeko web-bilatzaileak corpus konparagarri iturri gisa erabil daitezkeela ere egiaztatu dugu. Proposatzen dugun metodoak hitz guztietarako antzekotasun distribuzionala kalkulatzeko behar adinako testuinguruen bilketa bermatzen du.

	MAP	Ingelesezko elebakarrarekiko %	Pirkolarekiko hobekuntza
Ingelesezko elebakarra	0,37	-	-
Lehen itzulpena	0,25	%67,43	-
Pirkola	0,29	%79,21	-
Pirkola+corpus konparagarriak	0,30	%82,63	+%3,49

VII.3 taula: Web corpus konparagarrietan oinarritutako itzulpen-hautapenerako metodoaren ebaluaketa $es \rightarrow en$ norabiderako.

IV. kapituluko [Saralegi and Gamallo, 2013] artikuluan web-bilatzaileen bidez lortutako testuinguruen azterketa linguistiko sakonago bat egin dugu. Adierazgarritasun linguistikoa neurtzeko adieren banaketa eta koherentzia linguistikoa hartu dira irizpide gisa, eta hiru web-bilatzaile aztertu dira: *WebCorp*, *Google Blog Search* eta *Google News Archive*. Adiera-banaketari dagokionez, eta *SemCor* erreferentzia hartuta, *Google News Archive* eta *Google Blog Search* bilatzaileek harekiko korrelazio handiena dute (ikusi VII.4 taula), korrelazio hau moderatua bada ere. Koherentzia linguistikoari dagokionez *WebCorp* bilatzaileak erakusten du portaera okerrena (testuinguruen %24k koherentziari lotutako akatsak dituzte). Emaitza hauek direla eta, esan dezakegu amarauneko bilatzaileen bidez eskuratutako hitz-adierei buruzko informazioa ez dela behar adineko zehatza ezagutza-erauzketa, edota adiera-desanbiguazioko atazetan ustiatzeko.

V. kapituluan saio-mailako informazioaren ustiaketak kontsulten itzulpen-hautapena hobetzen laguntzen duela ikusi dugu. Saio-mailako testuinguruak, semantikoki lortutako kontsulta-multzo bat izanik,

	Adiera-banaketaren Pearson korrelazioa SemCor-ekiko	Inkoherentzia linguistikoa (%)
WebCorp	0,51	%24
Google Blog Search	0,56	%12
Google News Archive	0,66	%8

VII.4 taula: Amarauneko bilatzaileen bidez eskuratutako testuinguruen adierazgarritasun linguistikoaren azterketa.

birformulatutako kontsulten itzulpena hobeto desanbiguatzeko balio du. Gure esperimentuetan lortutako emaitzen arabera hobekuntza lor daiteke, bai itzulpen-prozesuan, bai ondorengo berreskurapen-prozesuan (ikusi VII.5 taula). Bi kontsulta dituzten saioekin egindako esperimentuen arabera hobekuntza itzulpenaren hobekuntza birformulazio mota guztietan lortzen da, orokortzean, zehaztean, eta dibagazioan. Berreskurapen-prozesuan lortutako hobekuntza, ordea, orokortze, eta zehazte motako birfomulazioekin baino ez da gertatzen esanguratsua.

	Itzulpena (HTER)		Berreskurapena (MAP)	
Birformulazio-mota	-saioa	+saioa	-saioa	+saioa
Orokortzea	0,118	0,107	0,053	0,055
Zehaztea	0,200	0,179	0,056	0,061
Dibagazioa	0,152	0,130	0,070	0,071

VII.5 taula: Itzulpen-hautapenerako metodoen ebaluaketa *eu* $\rightarrow$ *en* norabiderako, saioa erabilita (+saioa) eta erabili gabe (-saioa). HTER txikiagoa hobea.

III. IV. eta V. kapituluetako ondorioak laburbilduz honakoa esan dezakegu: hiztegien bidezko estrategia hizkuntza arteko berreskurapenerako baliagarria dela, baina hainbat fenomeno (itzulpen-anbiguotasuna, hitz anitzeko terminoak eta hiztegitik kanpoko hitzak) modu berezian tratatuz, batez ere, itzulpen anbiguoei dagozkienak. Fenomeno hau tratatzeko corpus paraleloetan oinarritutako tekniken alternatibak daude, baliabide urriko hizkuntzetarako egokiak direnak. Hiru informazio-iturri mota ezberdinetan oinarritutako teknikak aurkeztu ditugu, eta guztietan hobekuntza lortu

dugu itzulpen-hautapen prozesuaren eraginkortasunean. Hortaz, hiztegiek, helburu-bildumak, web-bilatzaileak eta saioek emandako informazioa ustiatuta eraginkortasun handiko CLIR sistema bat garatu daiteke.

Metodoa	Batazbesteko F_1-scorea	Batasbesteko $F_2 - puntuazioa$ (Estaldura lehenetsita)	Batasbesteko $F_{0.5} - puntuazioa$ (Doitasuna lehenetsita)
AK	0,34	0,27	0,46
AS	0,47	0,64	0,46
AK+AS	0,52	0,67	0,52

VII.6 taula: Pibotaje metodoen portaera ($Fpuntuazio$ren arabera) pibotaje-prozesuaren bidezko kasu anbiguoak ebazteko. AK alderantzizko kontsulta, AS antzekotasun distribuzionala.

Tesi-lan honetako azken esperimentuetan, **VI. kapituluan** azaldutakoetan, hiztegi elebidunaren faltan hiztegiak modu automatikoan eraiki ahal direla ikusi dugu. Pibotaje bidezko estrategiaz baliatuta, baliabide urriko hizkuntzak dituzten hiztegi elebidunak modu automatikoan sor daitezkeela frogatu dugu. D_{a-b} eta D_{b-c} hiztegiak gurutzatuz, eta b hizkuntza pibote gisa erabiliz, D_{a-c} hiztegia sortu ahal dira. B hizkuntzako hitzen anbiguotasunaren ondorioz eragindako itzulpen okerrak baztertzeko berezko teknikak aplikatu behar dira ordea. Bestela, zarata handiko hiztegiak eraikiko dira. Aztertu ditugun bi teknikek informazio-mota ezberdinak ustiatzen dituzte: batak (*Alderantzizko kontsulta* edo *AK*), gurutzatu beharreko hiztegien egitura, eta besteak (*Antzekotasun distribuzionala* edo *AS*), corpus konparagarrietatik kalkulatutako hizkuntza-arteko antzekotasun distribuzionala. Lehenengo teknika, aparteko baliabiderik behar ez duena, doitasun handiko hiztegientzat egokia da soilik. Corpus konparagarrietan oinarritutako teknikak, ordea, estaldura hobea bermatzen du doitasun berdina emanez. Biak konbinatuz lortuko genituzke doitasun eta estaldura handieneko hiztegiak (ikusi VII.6 taula). Hortaz, CLIR eszenatoki bati begira konbinazioa izango litzateke metodorik egokiena.

VII.2 Ekarpenak

Tesi honetan hiru motako ekarpenak egin ditugu: CLIR tekniken ebaluazioa, CLIR teknika berrien garapena, eta CLIR esperimentuak egiteko zenbait baliabideren sorkuntza. Jarraian azaltzen ditugu hiru multzo hauetako ekarpen nagusiak:

- Euskararako CLIR sistema oso bat garatzeko inplementatu behar diren urratsen azterketa osoa.
- Hiztegietan oinarritutako CLIR sistema bat garatzeko orduan ebatzi beharreko problemen deskribapena eta kuantifikazioa (III. kapituluan).
- Kontsultaren itzulpen-prozesuan gertatutako hiztegitik kanpoko hitzen presentziaren problemari aurre egiteko antzekotasun ortografikoan oinarrituko metodo baten egokitzapena eta azterketa berri bat (III. kapituluan).
- Kontsultaren itzulpen-prozesuan gertatutako itzulpen anbiguoen problema ebazteko zenbait metodoren egokitzapena eta metodo berriren garapena (III., IV. eta V. kapituluetan).
  - Helburu-bilduman oinarritutako bi metodoren azterketa eta egokitzapen berri bat (III. kapituluan). Bata (Pirkola) ranking-funtzioan integratutakoa, eta bestea kontsulta itzultzeko *greedy* algoritmo bat.
  - Amaraunetik estimatutako hizkuntza arteko antzekotasun distribuzionalean oinarritutako metodo berri baten garapena (IV. kapituluan).
  - Hiztegietan oinarritutako hizkuntza arteko bilaketa inplementatzen duen kode irekiko pakete bat, Bilakit[1] izenekoa, Solr/Lucene bilaketa-tresnaren gainean funtzionatzen duena.
  - Kontsulta-saioen informazioan oinarritutako metodo berri baten garapena (V. kapituluan).
- Hizkuntzen artean konparagarriak diren testuinguruak amaraunetik biltzen dituen metodo eta bertan oinarritutako tresna berri bat garatu dira IV. kapituluan. Tresnak *WebCorp*, *Google Blog Search* eta *Google News Archive* bilatzaileak erabiltzeko aukera ematen du.

[1] https://github.com/Elhuyar/Bilakit

- Amarauneko bilatzaileen bidez bildutako testuinguruen azterketa linguistiko berri bat (IV. kapituluan).

- CLIR esperimentutarako hainbat euskarazko gai-zerrenda (*topics*) prestatu dira. CLEF eta TREC datu-multzo estandarretako gaiak hartu eta itzuli egin dira euskarara (III. eta V. kapituluetan).

- Pibotaje bidez eraikitako hiztegi elebidunen doitasuna hobetzeko bi metodoren azterketa eta biak konbinatzen dituen metodo berri bat garatu dira, hizkuntza arteko antzekotasun distribuzionalean eta hiztegien egituran oinarritutakoa (VI. kapituluan).

- Hizkuntza arteko antzekotasun distribuzionala kalkulatzeko kazetaritzako corpus konparagarriak bildu dira euskara eta gaztelaniarako (VI. kapituluan).

VII.3 Etorkizuneko lanak

Etorkizuneko lanen artean aurreikusitako esperimentuetan egiteko geratu diren lanak honakoak dira:

- IV. kapituluko esperimentuak gaztelania-ingelesa bikoterako egin dira. Esperimentuetatik ateratako ondorioak teorikoki euskara-ingelesa bikotera estrapolagarriak badira ere, enpirikoki kontrastatu nahiko genuke esperimentua euskara-ingelesa CLIR ataza baten gainean.

- V. kapituluan aurkeztutako metodoa, saioaren informazioa ustiatzen duen itzulpen-hautapenerako metodoa, saio luzeagoak ustiatuz ebaluatu nahi genuke.

- III., IV. eta V. kapituluetan aurkeztutako itzulpen-hautapen metodo guztiak metodo bakarrean integratu nahiko genituzke. Horrela, ikusiko genuke itzulpen-hautapena hobetzeko baliatu ditugun informazio-iturburu mota guztiak modu bateratuan ustiatuz zenbateko hobekuntza metatua lor genezakeen.

- VI. kapituluan aurkeztutako pibotaje-metodoen eraginkortasuna eszenatoki teorikoetan ebaluatu da. CLIR ataza eszenatoki teoriko horietako bat bada ere, baliabideen ebaluazio estrinseko bat egin nahiko genuke CLIR ataza erreal baten barruan.

Bestalde, tesi-lan honek bidea irekitzen dio zenbait ikerketa-ildori.

Batetik, MT sistemen eta corpus paraleloen eskuragarritasuna handiagoa izaten ari da tesia hasi genuenean aurreikusi baino, euskararen kasuan behintzat. Hori dela eta, ondo legoke aztertzea CLIR atazatan zenbateraino diren osagarriak tesi-lan honetan proposatzen diren itzulpen-hautapenerako informazio-iturriak eta corpus paraleloak eta MT sistemak.

Bestetik, azkeneko urteetan *word embedding* teknikak hizkuntza naturaleko hainbat atazatan aplikatzen hasi dira. *Word embedding* kontzeptua erabat berria ez bada ere, hitzen bektoreak eraikitzeko eredu berriak proposatu dira literaturan. Eredu berri hauek kontaketetan oinarritutakoek baino emaitza hobeak ematen dituzte [Baroni et al., 2014]. Tesi-lan honetan, hizkuntza arteko antzekotasun distribuzionala neurtzeko kontaketa bidezko hitzen bektoreak erabili ditugu kontsultaren itzulpenean eta pibotaje bidezko hiztegien sorkuntzan. *Word embedding* teknika berri hauek hizkuntza arteko antzekotasun distribuzionala hobeto neurtzen lagun dezakete, antzeko atazetan frogatu den bezala [Artetxe et al., 2016].

Glosategia

- Antzekotasun distribuzionala (*distributional similarity*): Hitzen arteko antzekotasun semantikoa kalkulatzeko hurbilpena da. Banaketa-hipotesian oinarritzen da: antzeko esanahiko hitzak antzeko banaketak dituzte.

- Adierazgarritasun-epaia (*relevance judgement*): Kontsulta bakoitzeko dokumentu esanguratsuak zehazten ditu.

- Adierazgarritasun-maila (*relevance*): Dokumentu batek kontsultan adierazitako informazio-beharra asetzeko duen gaitasun-maila da. Adierazgarritasun-mailak gaiari ez ezik, autoritate, berritasun eta bestelako irizpideei erreparatu diezaieke.

- Agerkidetza (*co-occurrence*): Bi hitz testu-bilduma bateko testuinguru (dokumentua, esaldia, esalditik beherako hitz-leihoa) batean elkarrekin agertzea.

- Berreskurapen-algoritmoa (*retrieval algorithm*): Kontsulta baterako adierazgarriak diren dokumentuak bilduma batean aurkitu eta ranking-funtzioaren arabera ordenatzen dituen algoritmoa.

- Birformulazioa (*reformulation*): Kontsulta saio bateko kontsulta baten ondorengo kontsulta.

- Corpus konparagarria (*comparable corpora*): Adierazle baten arabera antzekoak diren corpusak. Askotan bi hizkuntzetako bi corpus, alor edo gai berekoak direnak.

- Corpus paraleloa (*parallel corpora*): Elkarren itzulpenak diren testuen bilduma, askotan esaldi mailan lerrokatuta dagoena.

- Dokumentu adierazgarrien rankinga (*relevant document ranking*): Kontsulta baterako adierazgarriak diren dokumentuak adierazgarritasun-mailaren arabera sailkatuta.
- Erabiltzailea (*user*): IR sistema bat erabiltzen duen pertsona.
- Helburu-bilduma (*target collection*): Kontsultarekiko adierazgarria den informazioa aurkitzeko IR sistemak erabiltzen duen testu-bilduma.
- Hizkuntza arteko informazioaren berreskurapena (*Cross Lingual Information Retrieval* edo CLIR): Kontsulta eta IR sistemaren atzean dagoen bilduma hizkuntza desberdinetan daudenean.
- Informazioaren berreskurapena (*Information Retrieval* edo IR): Informazio adierazgarria lortzeko ataza, informazio-behar batetik abiatzen dena, askotan gako-hitz multzo baten bidez adierazita.
- Itzulpen automatikoa (*Machine Translation* edo MT): Testuak modu automatikoan itzultzeko teknologia.
- Itzulpen-hautagaiak edo itzulpen anbiguoak (*translation candidates*): Itzuli nahi den hitzerako itzulpen-hautagaiak.
- Kontsulta (*query*): Erabiltzailearen informazio-beharra adierazten duena eta IR sistemari pasatzen diona.
- Kontsulta-saioa (*query session*): IR sistema batean, erabiltzaileak bere helburua topatu duen arte egindako kontsulta segida.

Bibliografia

Eneko Agirre and David Martinez. The effect of bias on an automatically-built word sense corpus. In *Proceedings of the 4rd International Conference on Languages Resources and Evaluations (LREC)*, 2004.

Eneko Agirre, Olatz Ansa, Eduard H. Hovy, and David Martínez. Enriching very large ontologies using the WWW. In *ECAI Workshop on Ontology Learning*, 2000.

Eneko Agirre, Olatz Ansa, Xabier Arregi, Maddalen Lopez De Lacalle, Arantxa Otegi, Xabier Saralegi, and Hugo Zaragoza. Elhuyar-ixa: Semantic relatedness and cross-lingual passage retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 273–280. Springer, 2009.

Eneko Agirre, Olatz Ansa, Xabier Arregi, Maddalen Lopez De Lacalle, Arantxa Otegi, and Xabier Saralegi. Document Expansion for Cross-Lingual Passage Retrieval. In *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

Iñaki Alegria, Antton Gurrutxaga, Pili Lizaso, Xabier Saralegi, Sahats Ugartetxea, and Ruben Urizar. A XML-Based Term Extraction Tool for Basque. In *LREC*, 2004.

Iñaki Alegria, Olatz Ansa, Xabier Arregi, Arantxa Otegi, and Ander Soraluze. Ihardetsi: A Question Answering system for Basque built on reused linguistic processors. In *Proc. SALTMIL 2009 workshop: Information Retrieval and Information Extraction for Less Resourced Languages*, pages 37–43, 2009.

Olatz Ansa, Xabier Arregi, Arantxa Otegi, and Ander Soraluze. Ihardetsi: a Basque question answering system at QA@ CLEF 2008. In *Workshop*

of the Cross-Language Evaluation Forum for European Languages, pages 369–376. Springer, 2008.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *EMNLP*, 2016.

Hosein Azarbonyad, Azadeh Shakery, and Heshaam Faili. Using Learning to Rank Approach for Parallel Corpora Based Cross Language Information Retrieval. In *ECAI*, volume 12, pages 79–84, 2012.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Lisa Ballesteros and W Bruce Croft. Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum*, volume 31, pages 84–91. ACM, 1997.

Lisa Ballesteros and W Bruce Croft. Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71. ACM, 1998.

Marco Baroni and Silvia Bernardini. *WaCky!: working papers on the web as corpus*, volume 6. Gedit, 2006.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247, 2014.

Francis Bond and Kentaro Ogura. Combining linguistic resources to create a machine-tractable Japanese-Malay dictionary. *Language Resources and Evaluation*, 42(2):127–136, 2008.

Francis Bond, Ruhaida Binti Sulong, Takefumi Yamazaki, and Kentaro Ogura. Design and construction of a machine-tractable Japanese-Malay dictionary. In *MT Summit VIII*, pages 53–58, 2001.

Andrei Broder. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM, 2002.

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. Clueweb09 data set, 2009.

Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. Adapting ranking SVM to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM, 2006.

Ben Carterette, Evangelos Kanoulas, and Emine Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 611–620. ACM, 2011.

Aitao Chen and Fredric C Gey. Combining query translation and document translation in cross-language retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 108–121. Springer, 2003.

Ping Chen, David Brown, Andrew Tran, Noble Ozoka, and Rafael Ortiz. Word sense distribution in a web corpus. In *Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on*, pages 449–453. IEEE, 2010.

Cyril Cleverdon. The Cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd, 1967.

Kareem Darwish and Douglas W Oard. Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 338–344. ACM, 2003.

Crystal David. English as a global language. *UK: Cambridge University Press. Print*, 1997.

Dina Demner-Fushman and Douglas W Oard. The effect of bilingual term list size on dictionary-based cross-language information retrieval. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*, pages 10–pp. IEEE, 2003.

Maike Erdmann, Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. An approach for extracting bilingual terminology from wikipedia. In *International Conference on Database Systems for Advanced Applications*, pages 380–392. Springer, 2008.

Flash Eurobarometer. User Language Preference Online. *Analytical Report.[online][cited 2015. 3. 30.]¡ http://ec. europa. eu/public_opinion/flash/fl_313_en. pdf*, 2011.

Pascale Fung. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 173–183, 1995.

Pascale Fung and Lo Yuen Yee. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics, 1998.

Pablo Gamallo and José Pichel. Automatic Generation of Bilingual Dictionaries Using Intermediary Languages and Comparable Corpora. *Computational Linguistics and Intelligent Text Processing*, pages 473–483, 2010.

Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, and Changning Huang. Improving query translation for cross-language information retrieval using statistical models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 96–104. ACM, 2001.

Jianfeng Gao, Ming Zhou, Jian-Yun Nie, Hongzhao He, and Weijun Chen. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 183–190. ACM, 2002.

Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. Cross-lingual query suggestion using query logs of different languages. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 463–470. ACM, 2007.

Julio Gonzalo. Scenarios for interactive cross-language information retrieval systems. In *Proceedings of SIGIR 2002 Workshop on Cross-Language IR*, 2002.

Antton Gurrutxaga, Xabier Saralegi, Sahats Ugartetxea, and Iñaki Alegria. Elexbi, a basic tool for bilingual term extraction from Spanish-Basque parallel corpora. In *Atti del XII Congresso Internazionale di Lessicografia: Torino, 6-9 settembre 2006*, pages 159–165, 2006.

Felix Hieber and Stefan Riezler. Bag-of-words forced decoding for cross-lingual information retrieval. In *Proceedings of the Conference*

of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies, Denver, Colorado, 2015.

Djoerd Hiemstra. *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata, 2001.

Rong Hu, Weizhu Chen, Jian Hu, Yansheng Lu, Zheng Chen, and Qiang Yang. Mining Translations of Web Queries from Web Click-through Data. In *AAAI*, pages 1144–1149, 2008.

David A Hull and Gregory Grefenstette. Querying across languages: a dictionary-based approach to multilingual information retrieval. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–57. ACM, 1996.

Varga István and Yokoyama Shoichi. Bilingual dictionary generation for low-resourced language pairs. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 862–870. Association for Computational Linguistics, 2009.

Myung-Gil Jang, Sung Hyon Myaeng, and Se Young Park. Using mutual information to resolve query translation ambiguities and query term weighting. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 223–229. Association for Computational Linguistics, 1999.

Kalervo Järvelin and Jaana Kekäläinen. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48. ACM, 2000.

Gareth Jones, Tetsuya Sakai, Nigel Collier, Akira Kumano, and Kazuo Sumita. Exploring the use of machine translation resources for englishjapanese cross-language information retrieval. In *In Proceedings of MT Summit VII Workshop on Machine Translation for Cross Language Information Retrieval*, pages 181–188. Citeseer, 1999.

Hiroyuki Kaji, Shin'ichi Tamamura, and Dashtseren Erdenebat. Automatic Construction of a Japanese-Chinese Dictionary via English. In *LREC*, volume 2008, pages 699–706, 2008.

Evangelos Kanoulas, Paul Clough, Ben Carterette, and Mark Sanderson. Session track at TREC 2010. *Simulation of Interaction*, page 13, 2010.

Adam Kilgarriff. How dominant is the commonest sense of a word? In *International Conference on Text, Speech and Dialogue*, pages 103–111. Springer, 2004.

Adam Kilgarriff. Getting to know your corpus. In *International Conference on Text, Speech and Dialogue*, pages 3–15. Springer, 2012.

Adam Kilgarriff and Gregory Grefenstette. Introduction to the special issue on the web as corpus. *Computational linguistics*, 29(3):333–347, 2003.

Alexandre Klementiev and Dan Roth. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 817–824. Association for Computational Linguistics, 2006.

Kevin Knight and Jonathan Graehl. Machine transliteration. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 128–135. Association for Computational Linguistics, 1997.

Zornitsa Kozareva, Ellen Riloff, and Eduard H Hovy. Semantic Class Learning from the Web with Hyponym Pattern Linkage Graphs. In *ACL*, volume 8, pages 1048–1056, 2008.

Frederick Wilfrid Lancaster and Emily Gallup. Information retrieval on-line. Technical report, 1973.

Leah S. Larkey, James Allan, Margaret E. Connell, Alvaro Bolivar, and Courtney Wade. UMass at TREC 2002: Cross Language and Novelty Tracks. In *TREC*, 2002.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria, and Aitzol Ezeiza. EusBila, a search service designed for the agglutinative nature of Basque. In *Proceedings of Improving non-English web searching (iNEWS07) workshop*, pages 47–54, 2007.

Igor Leturia, Antton Gurrutxaga, Nerea Areta, Iñaki Alegria, and Aitzol Ezeiza. Morphological query expansion and language-filtering words for improving Basque web retrieval. *Language resources and evaluation*, 47 (2):425–448, 2013.

Yi Liu, Rong Jin, and Joyce Y Chai. A maximum coherence model for dictionary-based cross-language information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 536–543. ACM, 2005.

Walid Magdy and Gareth JF Jones. Studying machine translation technologies for large-data CLIR tasks: a patent prior-art search case study. *Information Retrieval*, 17(5-6):492–519, 2014.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008. ISBN 0521865719, 9780521865715.

Jennifer Marlow, Paul Clough, Juan Cigarrán Recuero, and Javier Artiles. Exploring the effects of language skills on multilingual web search. In *European Conference on Information Retrieval*, pages 126–137. Springer, 2008.

Mausam, Stephen Soderland, Oren Etzioni, Daniel S. Weld, Michael Skinner, and Jeff A. Bilmes. Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In *ACL/IJCNLP*, 2009.

J Scott McCarley. Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 208–214. Association for Computational Linguistics, 1999.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279. Association for Computational Linguistics, 2004.

Rada Mihalcea. Semcor semantically tagged corpus. *http://www.cse.unt.edu/rada/downloads.html*, 1998.

George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

Christof Monz and Bonnie J Dorr. Iterative translation disambiguation for cross-language information retrieval. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 520–527. ACM, 2005.

Barry Morley. WebCorp: A tool for online linguistic information retrieval and analysis. *Language and Computers*, 55(1):283–296, 2006.

Preslav Nakov, Svetlin Nakov, and Elena Paskaleva. Improved word alignments using the Web as a corpus. *Proceedigs of Recent Advances in Natural Language Processing (RANLP'07)*, pages 400–405, 2007.

Douglas W Oard. A comparative study of query and document translation for cross-language information retrieval. In *Conference of the Association for Machine Translation in the Americas*, pages 472–483. Springer, 1998.

Arantxa Otegi, Xabier Arregi, Olatz Ansa, and Eneko Agirre. Using knowledge-based relatedness for information retrieval. *Knowledge and Information Systems*, 44(3):689–718, 2015.

Kyonghee Paik, Satoshi Shirai, and Hiromi Nakaiwa. Automatic construction of a transfer dictionary considering directionality. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, pages 31–38. Association for Computational Linguistics, 2004.

Ari Pirkola. The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 55–63. ACM, 1998.

Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

Yan Qu, Alla Eilerman, Hongming Jin, and David A Evans. The effect of pseudo relevance feedback on MT-based CLIR. In *Content-Based Multimedia Information Access-Volume 1*, pages 46–61, 2000.

Razieh Rahimi and Azadeh Shakery. A language modeling approach for extracting translation knowledge from comparable corpora. In *European Conference on Information Retrieval*, pages 606–617. Springer, 2013.

Reinhard Rapp. Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526. Association for Computational Linguistics, 1999.

Stephen E Robertson, Cornelis J van Rijsbergen, and Martin F Porter. Probabilistic models of indexing and searching. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 35–56. Butterworth & Co., 1980.

Fatiha Sadat, Masatoshi Yoshikawa, and Shunsuke Uemura. Bilingual terminology acquisition from comparable corpora and phrasal translation to cross-language information retrieval. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 2*, pages 141–144. Association for Computational Linguistics, 2003.

Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN 0070544840.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Celina Santamaría, Julio Gonzalo, and Javier Artiles. Wikipedia as sense inventory to improve diversity in web search results. In *Proceedings of the 48th annual meeting of the association for computational Linguistics*, pages 1357–1366. Association for Computational Linguistics, 2010.

Xabier Saralegi and Pablo Gamallo. Analyzing the sense distribution of concordances obtained by web as corpus approach. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 355–367. Springer, 2013.

Xabier Saralegi and Maddalen Lopez de Lacalle. Comparing different approaches to treat translation ambiguity in CLIR: Structured queries vs. target co-occurrence based selection. In *2009 20th International Workshop on Database and Expert Systems Application*, pages 398–404. IEEE, 2009.

Xabier Saralegi and Maddalen Lopez de Lacalle. Dictionary and Monolingual Corpus-based Query Translation for Basque-English CLIR. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010a.

Xabier Saralegi and Maddalen Lopez de Lacalle. Estimating translation probabilities from the web for structured queries on CLIR. In *European Conference on Information Retrieval*, pages 586–589. Springer, 2010b.

Xabier Saralegi, Iñaki San Vicente, and Antton Gurrutxaga. Automatic extraction of bilingual terms from comparable corpora in a popular science domain. In *Proceedings of Building and using Comparable Corpora workshop*, pages 27–32, 2008.

Xabier Saralegi, Iker Manterola, and Inaki San Vicente. Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 846–856. Association for Computational Linguistics, 2011.

Xabier Saralegi, Iñaki San Vicente, and Irati Ugarteburu. Cross-Lingual projections vs. corpora extracted subjectivity lexicons for less-resourced languages. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 96–108. Springer, 2013.

Xabier Saralegi, Eneko Agirre, and Iñaki Alegria. Evaluating Translation Quality and CLIR Performance of Query Sessions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

Daphna Shezaf and Ari Rappoport. Bilingual lexicon generation using non-aligned signatures. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 98–107. Association for Computational Linguistics, 2010.

Satoshi Shirai and Kazuhide Yamamoto. Linking English words in two bilingual dictionaries to generate another language pair dictionary. In *Proceedings of ICCPOL*, pages 174–179, 2001.

Jonas Sjöbergh. Creating a free digital Japanese-Swedish lexicon. In *Proceedings of PACLING*, pages 296–300. Citeseer, 2005.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200, 2006.

Stephen Soderland, Oren Etzioni, Daniel S Weld, Michael Skinner, Jeff Bilmes, et al. Compiling a massive, multilingual dictionary via probabilistic inference. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on*

Natural Language Processing of the AFNLP: Volume 1-Volume 1, pages 262–270. Association for Computational Linguistics, 2009.

Artem Sokolov, Felix Hieber, and Stefan Riezler. Learning to translate queries for clir. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pages 1179–1182. ACM, 2014.

Andrée Tabouret-Keller. Bilingualism in Europe. *The handbook of bilingualism*, pages 662–688, 2004.

Tuomas Talvensaari. *Comparable corpora in cross-language information retrieval*. Tampereen yliopisto, 2008.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Järvelin, Martti Juhola, and Heikki Keskustalo. Creating and exploiting a comparable corpus in cross-language information retrieval. *ACM Transactions on Information Systems (TOIS)*, 25(1):4, 2007.

Tuomas Talvensaari, Ari Pirkola, Kalervo Järvelin, Martti Juhola, and Jorma Laurikkala. Focused web crawling in the acquisition of comparable corpora. *Information Retrieval*, 11(5):427–445, 2008.

Kumiko Tanaka and Kyoji Umemura. Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, pages 297–303. Association for Computational Linguistics, 1994.

Takashi Tsunakawa, Naoaki Okazaki, and Jun'ichi Tsujii. Building Bilingual Lexicons using Lexical Translation Probabilities via Pivot Languages. In *LREC*, 2008.

Ferhan Ture, Jimmy Lin, and Douglas W Oard. Looking inside the box: Context-sensitive translation for cross-language information retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1105–1106. ACM, 2012.

Iñaki San Vicente and Xabier Saralegi. Polarity Lexicon Building: to what Extent Is the Manual Effort Worth? In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may 2016. European Language Resources Association (ELRA).

Srinivasarao Vundavalli. Mining the behaviour of users in a multilingual information access task. In *CLEF 2008 Workshop Notes, Aarhus, Denmark*, 2008.

Dietmar Wolfram, Amanda Spink, Bernard J Jansen, Tefko Saracevic, et al. Vox populi: The public searching of the web. *JASIST*, 52(12):1073–1074, 2001.

Hua Wu, Haifeng Wang, and Chengqing Zong. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 993–1000. Association for Computational Linguistics, 2008.

Mairidan Wushouer, Toru Ishida, and Donghui Lin. A heuristic framework for pivot-based bilingual dictionary induction. In *Culture and Computing (Culture Computing), 2013 International Conference on*, pages 111–116. IEEE, 2013.

Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 105–110. ACM, 2001.

Angel Zazo, Carlos G Figuerola, José Luis A Berrocal, and Viviana Fernández Marcial. Use of free on-line machine translation for interactive cross-language question answering. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 263–272. Springer, 2005.

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.

Jiang Zhu and Haifeng Wang. The effect of translation quality in MT-based cross-language information retrieval. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 593–600. Association for Computational Linguistics, 2006.