eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

QA-RDF: Galdera-erantzuna Wikipediako infotauletatik erauzitako RDFaren gainean

Maddalen Lopez de Lacalle
Tutorea: Montse Maritxalar Anglada

hap

Hizkuntzaren Azterketa eta Prozesamendua Masterreko titulua lortzeko bukaerako proiektua

2012ko iraila

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

LABURPENA

Gero eta informazio gehiago dago webean erabilgarri, baita egituratutako informazioa ere, azken hauen artean, Web Semantikoaren etorrerarekin, RDF ezagutza-baseak aurkitzen direlarik. Euskararen kasurako ere Wikipediako artikuluen infotauletako informaziotik erauzitako RDF ezagutza-basea dago eskuragarri. Bestalde, erabiltzaile arruntek informazio hau atzitu ahal izateko beharrezkoak dira hizkuntza natural bidezko interfazeak. Lan honetan, zehazki, Wikipediako artikuluen infotauletako informaziotik erauzitako RDF sarea kontsultatzeko galdera-erantzun sistema bat garatu dugu, zeinek sarrerako hizkuntza naturalean idatzitako galdera SPARQL lengoaiara itzuli ondoren RDF sarean erantzuna bilatzen duen. Erabilitako hurbilpenak baliabide eta teknologia linguistiko gutxi erabiltzen ditu: RDFtik eta Wikipediatik erauzitako sinonimo hiztegiak eta galdera analizatzailea. Hala ere, landu ditugun galdera-patroientzat, burututako ebaluazioan, sarrerako galdera SPARQL lengoaiara itzultzean emaitza onak lortu ditu, baita galderarentzat erantzun zuzena itzultzean. Honez gain, guk egindako proben arabera, erantzunak testu-hutsaren gainean bilatzen dituen IHARDETSI galdera-erantzun sistemak baino asmatze-tasa altuagoa lortu du.

Hitz gakoak: RDF, SPARQL, Wikipedia, Infotaula, galdera-erantzun sistemak

ABSTRACT

There is a vast amount of data available in the web that is growing and creates more and more (semi)structured data. The Semantic Web has promote the proliferation of Resource Description Framework (RDF) knowledge-bases. In particular, Wikipedia's infoboxes has given the opportunity to extract structured information and store it RDF format. Not only for languages such as English and Spanish, it also has done for minority language such as the Basque. On the other hand, users require natural language interfaces to access easily to this huge amount of information. In this work, we present a question answering system which is based on a RDF knowledge-base generated from Basque Wikipedia's infoboxes. It translates the source natural language question to SPARQL language in order to obtain the right answer. The approach presented here deploys few resources and non-sophisticated language technology. However, the preliminary evaluation has shown that based on the synonym dictionary extracted from Wikipedia and RDF bases and the use of the query analyzer, the system is capable of obtaining good results for the addressed query patterns. It shows that the translation from the source query to SPARQL language is performed accurately as well as obtaining the right answer. In addition, the proposed approach outperforms IHARDETSI, a basque text-based question answering system, in terms of accuracy.

Key words: RDF, SPARQL, Wikipedia, Infobox, question-answering systems

Eskerrak

Bereziki eskertu nahi ditut Olatz Ansa eta Xabier Arregi, IHARDETSI sistemarekin emandako laguntza guztiarengatik. Baita ebaluazioaren diseinuan emandako gomendio eta laguntzagatik ere.

Garapenean eta ebaluazioan erabili diren galderak sortzen laguntzeagatik Oier Lopez de Lacalle eskertu nahi dut.

Igor Leturia ere eskertu nahi dut gure hurbilpenaren diseinua eta ideiekin laguntzeagatik.

Azkenik, lan honen zuzendariari, lan osoan zehar eskainitako denbora eta laguntzagatik eskertu nahi diot.

Aurkibidea

1 Sarrera

Interneten gero eta informazio gehiago dago erabiltzaileen eskura. Informazio hau modu egoki batean atzitu ahal izateko beharrezkoak dira informazioa berreskuratzeko sistemak, beraien artean, bilatzaileak bezalako tresnak. Bilatzaileek nabarmenak diren gaitasun asko dituzte eta beraien errendimendua hobetzeko etengabeko aurrerapenak egiten ari dira. Hala ere, bilatzaileek ez dituzte erantzun zehatzak itzultzen, hau da, askotan ez dugu gai baten inguruko informazioa bilatu nahi, baizik eta galdera batentzat erantzun zehatz, bakar eta egoki bat lortu nahi izaten dugu. Azken hau egiteaz, galdera-erantzun sistemak arduratzen dira. Galdera-erantzun sistemek dokumentu bilduma baten gainean (Web osoaren gainean ala dokumentu bilduma txikiago batean), hizkuntza naturalean idatzitako galderen erantzunak eskuratzeko gai izan behar dute.

Bai bilatzaileak, bai galdera-erantzun sistemak informazioaren berreskurapenaren (Information Retrieval) ikerketa alorrean kokatzen dira. Lehenengo sistemen egitekoa, erabiltzaile batek planteatutako kontsulta batentzat dokumentu esanguratsuena itzultzea da. Dokumentu hauek bilduma batean ala Internet osoaren gainean bilatu daitezke. Bigarrengo sistemek, galdera-erantzun sistemak, aldiz, hizkuntza naturalean idatzitako galdera batetik abiatuta, erantzuna den zati zehatza itzuli behar dute eta ez dokumentu esanguratsua (erantzuna duen dokumentua).

Gaur egun, badago galdera-erantzun sistemen eta Webaren gaineko bilaketa sistemen arteko fusioa aurkitzeko interesa. Google eta Microsoft bezalako enpresak galdera-erantzunaren gaitasun batzuk bere bilatzaileetan integratzen hasi dira, eta espero da integrazio hori etorkizunean garrantzitsuagoa izango dela.

Lan honetan, galderak erantzuteko sistema bat garatzea proposatzen da, hain zuzen, euskarazko Wikipediako infotauletatik erauzitako informazio egituratuaren gaineko galdera-erantzun sistema bat. Infotaulak Wikipediako orri batzuek albo batean dituzten taulak dira, non orriak lantzen duen kontzeptuari edota entitateari buruzko informazio erdiegituratua dagoen. Euskaraz, badago galderak erantzuteko sistema bat (Ansa et al., 2008), baina hau, testuen gainean saiatzen da galderarentzako erantzuna aurkitzen. Guk hemen proposatzen dugun sistema egituratutako informazioaren gaineko galdera-erantzun sistema bat da, zehazki, Wikipediako infotauletatik erauzitako RDF-Resource

Descriptor Framework[1] ezagutza-basearen gainean erantzunak bilatuko dituena. Sarrerako galdera hizkuntza naturalean egongo da eta hau SPARQLra[2] itzuliko da RDF ezagutza-basea kontsultatu ahal izateko.

Wikipedia proiektuak eskura jartzen duen domeinu irekiko informazio entziklopedikoa ezagutza-base gisa erabiltzen dituzten galdera-erantzun sistemen ikerketa lanak ugariak dira azken aldian. Izan ere, Wikipedia, Interneteko entziklopedia handiena, ezagunena eta erabiliena izanik, informazio-iturri garrantzitsua bilakatu da gaur egun, Interneteko web bisitatuenetakoa izanik. Askotan jotzen dugu Wikipediara pertsona ala leku bati buruzko, besteak beste, informazioren bat bilatu nahi dugunean, beraz, Wikipediako informazioaren gaineko galdera-erantzun sistema bat oso interesgarria eta erabilgarria izan daiteke.

Testuaren gainean hizkuntza naturalean idatzitako galdera batentzat erantzun zehatza eta zuzena aurkitzea zailtasun handiko ataza da, baina guk hemen aurkezten dugun galdera-erantzun sistemak ez du erantzuna aurkitzeko euskarazko Wikipediako artikuluen testuen gainean bilatuko, baizik eta artikulu gehienek albo batean duten infotauletako informazioaren gainean. Infotaula hauetan, artikuluak tratatzen duen entitateari buruzko informazio esanguratsua dago, eta, gainera, egituratua. Dbpedia proiektuak[3], hainbat hizkuntzatarako, hauen artean euskara, Wikipediako artikuluetako infotauletako edukia RDF (Resource Descriptor Framework) moduan jarri dute eskuragarri.

RDF moduan dagoen ezagutza-base hau galdera-erantzun sistema bat eraikitzeko baliabide interesgarri bat iruditzen zaigu, alde batetik, Wikipedian lantzen diren entitateei buruzko informazio esanguratsua aurkitzen delako, eta, beste alde batetik, datuak modu egituratuan egoteak erantzuna testu hutsaren gainean baino modu errazagoan aurkitzen lagundu dezakeelako. Gainera, erantzuna artikulu batean baino gehiagotan banatuta duten galderentzat oso zaila izango da testuan erantzun zuzena aurkitzea, baina RDFaren egitura kontuan izanda eta SPARQL lengoaiak ematen dituen aukerak baliatuz, horrelako galderak ere erantzuten saiatuko gara hemen aurkezten den sistemarekin.

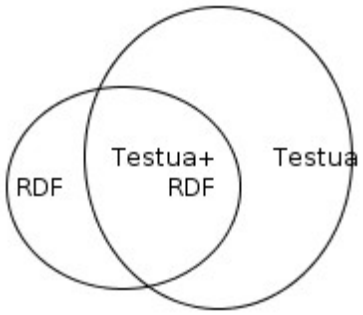
1http://www.w3.org/RDF/

2http://www.w3.org/TR/rdf-sparql-query/

3http://dbpedia.org/

Esan bezala, erabiliko dugun ezagutza-basea RDF moduan dago egituratua, beraz, galderen erantzunak lortzeko RDFak kontsultatzeko lengoaia erabili beharko dugu: SPARQL lengoaia. Horretarako, hizkuntza naturalean idatzita dagoen galdera SPARQL lengoaiara itzuli beharko dugu. Beste hizkuntza batzuetarako badauden arren (Unger et al., 2012), euskararentzako ez dugu hizkuntza naturala SPARQL lengoaiara itzultzen duen hizkuntza naturalerako interfazerik ezagutzen, beraz garrantzitsua iruditzen zaigu alor honetan euskararentzako bideak zabaltzen hastea.

SPARQL lengoaian oso galdera konplexuak adieraz daitezke, hala ere, lan honetan gure sistema galdera-mota eta konplexutasun maila zehatz batzuetakoetara mugatuko dugu. Guk dakigunera arte euskararako ez da aurretik honelako interfaze naturalekin lan egin, beraz, egokia ikusten dugu konplexutasun txikiagoa duten galderekin lanean hastea. Gainera, galdera sinpleagoak edota konplexutasun txikiagokoak izan arren, estaldura onargarria lor dezakeen sistema bat gara daiteke. Izan ere, orokorrean Wikipediako artikuluetako infotauletan testuan bertan baino informazio gutxiago egongo dela pentsatzea normala den arren, hau ez da beti gertatzen eta beste kasu batzuetan infotauletan testuan azaltzen ez den informazioa ere azaltzen da. Askotan, Wikipediako artikulu bat sortzean, errazagoa da datu zehatz batzuez infotaula sortzea, artikulu osoa idaztea baino.



1. irudia: Wikipediako edukien proportzioa

1. irudian azaldu nahi izan den moduan, Wikipediako artikuluetan informazio gehiago dago infotauletan baino, eta, honez gain artikuluetan eta infotaulek informazio asko konpartitzen dute, hau da, informazio berdina bai artikuluetako testuan eta bai infotauletan ere aurkitu daiteke. Hala ere, esan bezala, badago Wikipedia osoko informazioaren zati bat infotauletan soilik azaltzen dena, beraz, uste dugu, hemen aurkezten den sistema, infotauletatik erauzitako RDFaren gaineko galderak erantzuteko sistema, erabilgarria izan daitekeela informazio hori ere atzitzeko.

Honez gain, testuan eta RDFan konpartitzen den informazioa, erantzunak testuan bilatzen dituen galdera-erantzun sistema batekin ere bilatu daitezke, baina testuan dagoen informazio kopuru handia dela eta zarata gehiago ere sartzen da eta honek erantzun zuzena aurkitzeko lan handiago eskatzen du. Kasu hauetan ere, egituratutako informaziotik erantzuna bilatzeak galdera zuzen erantzuten lagundu dezake.

Azkenik, testuan soilik aurkitzen den informazio asko ere aurkitzen da Wikipedian, beraz, Wikipedia osoko informazioaren gaineko galdera-erantzun sistema batek bi hurbilpenak konbinatu beharko lituzke, informazio guztia atzitu ahal izateko. Adibidez, galdera batentzat RDFan erantzuna bilatzen ez denean testuan bilatzen saiatzen den sistema bat erabili (Jitkrittum, Haruechaiyasak eta Theeramunkong, 2009).

1.1 Lanaren ekarpenak

Hauek izan dira lan honen ekarpen nagusiak:

- Euskararentzako informazio egituratuaren gaineko galderak erantzuteko sistema bat aurkezten da. Zehazki, euskarazko wikipediako infotauletatik erauzitako RDFan bilatzen ditu erantzunak hizkuntza naturalean idatzitako galderentzat.
- Hizkuntza naturalean idatzitako galderak SPARQL lengoaiara itzultzen duen modulua garatu da euskararentzako.
- Esperimentazio ingurune bat sortu da, euskarazko galdera-zerrenda bat sortuz.
- Erantzunak testu hutsaren gainean bilatzen dituen galdera-erantzun sistemarekin osagarria den informazio-egituratuaren gaineko galdera-erantzun sistema bat aurkezten da.

2 Aurrekariak

Lan honetan aurkezten den sistema hizkuntza naturalean idatzitako galdera batetik abiatzen da euskarazko Wikipediako infotauletatik erauzitako RDF (Resource Description Framework) ezagutza-basean galdera horrentzat erantzun zuzena aurkitzeko. Bestalde, RDFak kontsultatzeko SPARQL lengoaia erabiltzen da, beraz, ezinbestekoa izango da hizkuntza naturalean idatzita dagoen hasierako galdera hori SPARQL lengoaiara itzultzea. Honela, aurkezten dena, alde batetik, galderak erantzuteko sistema bat da, hizkuntza naturalean idatzitako galdera batentzat, dagokion erantzun zehatza eta zuzena itzultzen saiatuko delako, baina, beste alde batetik, erabiltzaileak RDFak kontsultatzeko SPARQL lengoaiaren logika eta sintaxia ezagutu gabe, RDFa bera kontsultatzeko aukera emango du. Horregatik, hizkuntza natural bidezko interfaze bat (*Natural Language user Interface*) ere izango da. Ondoren azalduko den bezala, RDF eta SPARQL web semantikoko tresnak dira, beraz, lan honek web semantikoaren munduan ere koka dezakegu.

Hurrengo puntuetan, hiru kontzeptu hauek tratatuko dira, hala nola, hizkuntza natural bidezko interfazeak, galdera-erantzun sistemak eta web semantikoa. Honez gain, RDF eta SPARQL lengoaiaren gaineko azalpenak ere emango dira.

2.1 Hizkuntza natural bidezko interfazeak

Hizkuntza natural bidezko interfazeak konputagailu eta erabiltzaileen arteko komunikazioa errazten duten erabiltzaileentzako interfaze mota bat dira.

Hizkuntza natural bidezko interfazeek hizkuntza naturalaren prozesamenduaren eta hizkuntzalaritza konputazionalaren alorreko ikerketa eremu aktiboa osatzen dute. Web semantikoaren gaur egungo helburu garrantzitsuenetako bat da, oro har, hizkuntza natural bidezko interfaze intuitiboak sortzea.

Gero eta informazio gehiago aurki daiteke ontologietan oinarritutako ezagutza-baseetan biltegiratuta, web semantikoan, esaterako, edukiak RDF lengoaia estandarrean adierazten dira, beraz, eduki hauek erabiltzaile arruntentzat erabilgarria izateko ezinbestekoak dira hizkuntza natural bidezko interfazeak. Izan ere, hizkuntza natural bidezko interfazeei esker, eduki hauek kontsultatu ahal izango dituzte RDF, SPARQL

edota OWL bezalako lengoaiak ikasi beharra izan gabe. Beste modu batera esanda, hizkuntza natural bidezko interfazeek aipatutako lengoaia horien barne logika guztia ezkutatuz, erabiltzaileari kontsultak modu intuitibo eta erraz batean egiteko aukera eskaintzen die, hizkuntza naturala erabiliz.

Hizkuntza natural bidezko interfazeena ibilbide handiko ikergaia da eta RDF edota ontologiak bezalako ezagutza-baseen gaineko hizkuntza natural bidezko interfazeen aurretik datu-baseen gainekoak ere eraiki dira (Androutsopoulos, Ritchie eta Thanisch, 1995), (Chu eta Meng, 1999), (Popescu, Etzioni eta Kautz, 2003). Azken honetan, hizkuntza naturalean idatzitako galdera, SQL lengoaiara itzultzen du oinarritzat datu-base bat erabilita.

Azkenaldian, hizkuntza natural bidezko interfazeen beharra handitu egin dela esan daiteke, atzigarri dagoen informazioa ere izugarri eta etengabe handitzen ari delako.

Hizkuntza natural bidezko interfazeek "naturaltasun" gradu desberdinak izan dezakete, adibidez, gako-hitzetan oinarritzen den ohiko bilatzailea "azaleko" hizkuntza naturaleko erabiltzaileen interfaze gisa deskriba daiteke. Hizkuntza natural osoaren tratamenduak dituen zailtasunak kontuan hartuta, ulergarria dirudi mugatutako hizkuntza naturala edota menu bidez gidatutako hizkuntza natural bidezko interfazeak (Bernstein, Kaufmann eta Kaiser, 2005) proposatzen dituzten hurbilpenak egotea.

NLP-Reduce (Kaufmann, Bernstein eta Fischer, 2007) web semantikoko ezagutza-baseak kontsultatzeko hizkuntza natural bidezko interfazeak erabiltzaileari gako-hitzak ("*Chinese restaurant San Francisco"*), esaldi zatiak (*"Chinese restaurants that are in San Francisco"*) ala galdera esaldi osoak (*"Which Chinese restaurants are in San Francisco?"*) erabiltzeko aukera ematen dio. Sarrerako kontsultak SPARQLra itzultzeko, hizkuntza prozesatzeko teknika sinpleak erabiltzen ditu: lematizazioa eta sinonimoen hedapena. Sarrerako kontsultak gako hitz multzoak bezala prozesatzen ditu eta beraien artean triple egiturak identifikatzen saiatzen da sinonimoekin hedatutako ezagutza-basea erabilita. Ondoren, identifikatutako triple hauekin SPARQL galdera osatzen saiatzen da. Erabiltzen den ezagutza-basea sinonimoekin hedatzeko WordNet erabiltzen da, honela, erabiltzaileek galderak egiteko erabili dezaketen hiztegia zabalagoa bihurtzen da. *Querix* hizkuntza natural bidezko interfazeak (Kaufmann, Bernstein eta Zumstein, 2006) ingelesezko galdera osoak onartzen ditu soilik. Honek analizatzaile sintaktikoa erabiltzen du galderako hitzen artean, ezagutza-basean ematen

diren triple egiturak identifikatzeko eta hauekin SPARQL galdera sortzeko. Sarrerako galderarentzat semantika desberdina duten SPARQL galdera bat baino gehiago sortu ahal baldin badira, argibideetarako elkarrizketa kuadroak erabiltzen ditu erabiltzaileak berak anbiguotasuna argitu dezan. (Bernstein, Kaufmann eta Kaiser, 2005) artikuluan azaltzen den *Ginseng* sistemak galderak hizkuntza natural mugatu batean egitea eskaintzen dio erabiltzaileari. Hala ere, kontsultarako lengoaia kontrolatu hau ingeles osotik oso gertu dago. Galdera osatzen laguntzeko, modu dinamiko batean eta galdera posibleen gramatika batean oinarrituta, aukera anitzeko *pop-up* leihoetan galdera osatzeko aukerak ematen zaizkio erabiltzaileari. Modu honetan *Ginseng* hizkuntza natural bidezko interfazeak erabiltzaileari sistemak onartzen duen galdera bat sortzen gidatzen dio. Gramatika hau jarraituz, sortutako galderak SPARQLra itzultzen ditu galderak.

Kaufmann eta Bernsteinen (2007) lanean hizkuntza natural bidezko interfazeak azken erabiltzaile arruntentzat erabilgarriak direla ondorioztatzeaz gain, erabiltzaileek kontsulta moduan esaldi osoak erabiltzeko aukera ematen duten interfazeak nahiago dituztela erakutsi dute, gako-hitzak soilik edota kontsultak idazten gidatzen dituzten menuak erabiltzea baino.

Chong et. al (2007) galderaren izen sintagmetan oinarritzen da hizkuntza naturalean idatzitako galdera SPARQLra itzultzeko. Analizatzaile sintaktikoak itzulitako galderaren zuhaitz sintaktikoa tripletan oinarritutako datu-eredu batera bihurtzen da.

Jitkrittum, Haruechaiyasak eta Theeramunkong-en (2009) lanean thailanderako wikipediaren gaineko galdera-erantzun sistema bat aurkezten da, zeinek bi motatako informazioa konbinatzen den: informazio egituratua, wikipediako artikuluetako infotauletatik erauzitako RDFa, eta egiturarik gabeko wikipediako artikuluetako testua. RDFan erantzunak bilatzeko hizkuntza naturalean dagoen galdera SPARQL lengoaiara itzultzen dute, galdera aurretik definitutako patroi batzuekin konparatuz eta sailkatuz eta hauen araberako SPARQLa sortuz.

Hizkuntza natural bidezko interfazeek potentzial handia eskaintzen dute logikan oinarritutako web semantikoaren eta erabiltzaile errealen arteko tartea gutxitzeko, izan ere erabiltzaileek web semantikoa kontsultatzeko aukera ematen die lengoaia formal ezezagun bat ikasteko beharra izan gabe (Ding et. Al, 2004).

Euskararentzako, guk dakigunerarte behintzat, ez dago RDF sareen gainean hizkuntza naturalean idatzitako galderak erantzuten dituen sistemarik. Beraz, lan honetan hori lortzeko lehendabiziko pausuak emango direla esan dezakegu.

2.2 Galdera-erantzun sistemak

Webaren etengabeko hazkundea dela eta, erabilgarri dagoen informazio kopurua gero eta handiagoa da. Askotan informazio bila dabilen erabiltzaileak aurrean duen arazoa ez da dokumentu edo informazio eza, baizik eta modu azkar batean eskuragarri dagoen informazio guzti horren artean galdera batentzat erantzun egokia aurkitzea. Hau da, hain zuzen, galdera-erantzun sistemen helburu nagusia.

Galdera-erantzun sistemena ere ibilbide handiko ikerkuntza arloa da, izan ere, lehen galdera-erantzun sistemak 60ko hamarkadan garatu ziren. Hauek, funtsean, domeinu zehatz baterako sistema adituen hizkuntza natural bidezko interfazeak ziren. Adituek eskuz idatzitako datu-baseak ziren erabiltzen zituzten ezagutza-baseak. BASEBALL (Green et al. 1961) eta LUNAR (Woods et al., 1970) dira horrelako sistema famatuenetarikoak. BASEBALL-ek Estatu Batuetako beisbol ligari buruzko galderak erantzuten zituen eta LUNAR-ek Apollo misioak ekarritako harrien analisi geologikoari buruzko galderak erantzuten zituen.

1999. urtetik aurrera, galdera-erantzun sistemen ikerkuntza Text Retrieval Conference (TREC-8) barruan sartzen da. Lehiaketa honetan parte hartzen duten sistemek edozein gaietako galderak erantzun behar dituzte erantzuna testu-corpus baten gainean bilatuz. Cross-Language Evaluation Forum (CLEF) lehiaketak ere mota honetako sistemekiko interesa erakusten du, 2003tik aurrera TREC lehiaketaren hizkuntza arteko bertsioa antolatuz.

Galdera-erantzun sistemen aplikazio desberdinak sailkatu daitezke erabiltzaile motaren arabera (erabiltzaile arruntak ala adituak), zein motatako informazioaren gainean egiten duen bilaketa (egituratutako datuak, testu librea, bien konbinaketa...), bilduma motaren arabera (Web-a, dokumentu-bilduma...), domeinuaren arabera (librea, zehatza) etab. (Magnini, 2005)

Erantzuna bilatzeko erabiltzen den datu iturriaren motaren arabera bi talde desberdinetan sailkatu daitezke galdera-erantzun sistemak: alde batetik, erantzunak

egituratutako datuen gainean bilatzen dituztenak (Lopez, Motta eta Uren, 2006) (Unger et. Al, 2012), eta, beste alde batetik, testu hutsaren gainean bilatzen dituztenak (Ansa et. al, 2008), START[4], WikiQA[5] (Waltinger, Breuing eta Wachsmuth, 2011). Lehendabiziko galdera-erantzun sistemek egituratutako ezagutza-baseetatik ondorioztatzen dituzte erantzunak. Aldiz, bigarrengo galdera-erantzun sistemek makinek irakurtzeko moduan prozesatutako testu hutsekin sortutako bildumen edota webeko artikuluen gainean bilatzen dituzte erantzunak. Bi hurbilpenak konbinatzen dituzten sistemak ere badaude (Jitkrittum, Haruechaiyasak eta Theeramunkong, 2009).

Testu hutsaren gaineko gaur egungo galdera-erantzun sistemen arkitektura, askotan informazioaren berreskurapenerako sistema batean oinarritzen da. Galderako hitzak kontsultako gako-hitz bezala erabiltzen dira dokumentu esanguratsuenak bilatzeko. Hala ere, begi-bistakoa da erantzun zuzena bilatzeko hori baino gehiago falta dela, hau da, galderarentzako esanguratsua den dokumentua bilatzea ez da nahikoa, bertatik erantzun zuzena erauzi behar da. Horrela, galderako hitz guztiak ez dira erantzuna bilatzeko lagungarriak izango eta beste hitz batzuek, berriz, zein motatako erantzuna espero den jakiteko oso lagungarriak izango dira, besteak beste.

Honela, galdera-erantzun sistema gehienak oinarrizko hiru osagai hauetan oinarritzen dira: galdera-analizatzailea (galdera eta erantzun mota identifikatu, galderako gako hitzak erauzi etab.), informazioaren berreskurapenerako modulua (dokumentu edota paragrafo esanguratsuen identifikazioa) eta erantzun zuzenaren erauzlea. Informazioaren berreskurapenerako prozesua oso desberdina izango da erantzuna testu-hutsean ala egituratutako informazioan bilatu behar bada. Azkeneko kasu honetan, informazioaren berreskurapena egin ordez, sarrerako galdera datu egituratuak atzitzeko lengoaiara (SPARQL, SQL...) itzultzeko pausua burutu beharko da.

Erantzuna testu hutsaren gainean bilatzen dituzten galderak erantzuteko sistemen artean, START da webean oinarritzen den lehendabiziko sistema. Euskaraz egindako galdera naturalak erantzuteko sistema bakarra, guk dakigunera arte behintzat,

4http://start.csail.mit.edu/

5http://www.ulliwaltinger.de/wikiqa/

IHARDETSI[6] sistema da. Azkenik, erantzunak Wikipediaren gainean bilatzen dituen sistemen artean WikiQA aurkitu dezakegu.

Datu-egituratuen gaineko galdera-erantzun sistemak ere gero eta ugariagoak dira, izan ere, web semantikoaren etorrerarekin gero eta informazio gehiago aurkitu daiteke RDF ezagutza-baseetan, besteak beste. PANTO (Wang et al., 2007) sistemak hizkuntza naturalean idatzitako galderak SPARQL lengoaiara itzultzen ditu tripleetan oinarritzen den galderaren analisi sintaktikoan oinarrituta. Galderako hitzen eta ezagutza-basearen arteko *mapping*-a egiteko WordNet eta *string*-en antzekotasun metrikak erabiltzen dituzte. FREyA (Damljanovic, Agatonovic eta Cunningham, 2010) eta Querix (Kaufmann et al., 2006) sistemek ere ingeleserako analizatzaile sintaktikoa eta Wordnet bidezko sinonimo-hedapena erabiltzen dute hizkuntza naturaletik SPARQLrako itzulpena burutzeko, baina, honez gain, anbiguotasunak erabiltzaileekin argitzen dituzte, elkarrizketa-koadroak erabilita. QACID (Fernandez et al., 2009) domeinu zehatz batetako galdera-bildumetan oinarritzen da. Galdera hauek analizatu eta multzokatzen ditu *cluster*-etan eta eskuz sortzen dira galdera hauei dagozkien SPARQLak. AguaLog (Lopez et al., 2007) eta SMART (Battista et al., 2007) sistemek arrazonamendu automatikoa erabiltzen dute.

Azkenik, aipatzea, QALD-Question Answering over Linked data[7] lehiaketa antzeko bat dela, non RDF moduan dauden Dbpedia eta MusicBrainz bildumen gaineko galdera-erantzun sistemak konparatzen diren.

2.3 Web semantikoa

Webak hasieratik izan duen diseinua errepresentazio bisualera zuzenduta egon da. Hizkuntza naturalean eta HTML formatuan dagoen testuan oinarritutako diseinua da eta gizakiak interpreta dezan diseinatu da. Makinek momentuz, hizkuntza naturala ezin dute ulertu eta HTML etiketek ez dute horretan laguntzen.

W3C (World Wide Web Consortium, webaren jarraibideak, estandarrak eta eboluzioa gidatzen dituen erakundea) web semantikoaren sorkuntza bultzatzen ari da. Hau web paralelo bat izango litzateke informazioa RDF formatuan edukiko lukeena,

6http://ixa2.si.ehu.es/IhardetsiWebDemo/IhardetsiBezeroa.jsp

7http://www.sc.cit-ec.uni-bielefeld.de/qald-1

informazio egituratuarentzako formatu bat non etiketatzea semantikoa den, hots, testuaren esanahiarekin lotua. Honela, web semantikoan ondo egituratuta izango genuke objektuen, pertsonen eta beste entitate batzuen gaineko informazioa, beraien atributuekin eta elkarren arteko erlazioekin.

Web semantikoan, objektuak, pertsonak... eta haien arteko erlazioak deskribatzen dira etiketa bidez. Etiketetan, orriaren itxura eta egitura azaldu beharrean, orriko elementuen esanahia jasotzen da. Horri esker, HTML sarearekin batera existituko litzatekeen sare paralelo bat sor daiteke, makinek ulertzeko moduko ezagutza-base bat, semantika adierazteko formatuetan kodetua. Behin makinek ulertuta, modu eraginkorrean tratatu ahal izango lukete informazioa, eta aplikazio askotara arako bidea ireki.

W3Ck web semantikoari probetxua atera ahal izateko teknologiak definitu ditu. Adibidez, RDF-n dauden informazio-baseetan kontsulta nahi bezain konplikatuak egitea ahalbidetzen duen SPARQL kontsulta-lengoaia. Hala ere, web semantikoaren ideia oso ona izan arren, kontzeptu eta gauza ezberdin guztiak etiketatzeko estandarrak definitzeak lan ikaragarria suposatzen du. Gainera, web orri guztien RDF bertsio paralelo bat sortzeak ere lan asko suposatzen du, batez ere webguneak sortzeko tresnak ez badaude horretarako prestatuta eta eskuz egin behar bada. Zailtasun hauek direla eta, egi bihurtzea kostatzen ari zaio, baina, hizkuntzaren prozesamendua erabiltzen dituzten teknikak lagungarriak izan daiteke.

Euskararen aldetik ere gaiari heltzea ezinbestekoa da. Euskararentzako eskuragarri dauden RDF ezagutza-baseak, gure kasuan Wikipediako artikuluen infotauletako informaziotik erauzitakoa, euskara kontutan hartzen dute web semantikoarekin batera aterako diren erreminta eta zerbitzuak probatzeko proba banku egokiak dira. Honela euskara web semantikorako prestatzen hasi daiteke.

Esan bezala, Web semantikoa iristen joan ahala, horren ahalmena ustiatuko duten zerbitzuak eraikitzen joango dira. Galdera-erantzuna da zalantzarik gabe horietako aplikazio eremu garrantzitsu bat, eta proiektu horrelako zerbitzuak euskaraz eraiki ahal izateko oinarriak jarriko ditu.

Lan honetan gai honi heldu nahi diogu euskararen ikuspegitik. Galderak erantzuteko sistema bat garatzen saiatu nahi da, non galdera hizkuntza naturalean

legokeen eta berau SPARQLra bihurtuko litzatekeen RDF basea kontsultatu ahal izateko.

2.3.1 RDF- Resource Description Framework

RDF (Resource Description Framework) edozein motatako ezagutza zati txikietan deskonposatzeko *metodo* orokor bat bezala uler daiteke. Zati horiek ezaugarri batzuk edukiko dituzte. Oso modu sinplean datu edota gertaera asko adieraz daitezke eta era berean hain da egituratua aplikazio informatikoek beraiekin gauza baliagarriak egin ditzaketela. RDFen abantaila nagusia, sinplea eta helburu orokorrekoa izateko diseinatua dagoela da.

2. irudian ikus daiteke adibide sinple bat RDFak adierazteko N3 izeneko formatuan:

```
@prefix : <http://www.adibidea.org/> .
:Jon    :Da    :Pertsona .
:Jon    :Ama   :Susana .
:Jon    :Aita  :Pedro .
:Pedro      :Anaia     :Lukas .
```

2. irudia: N3 formatua

RDFa ez da soilik formatu bat, ezagutza errepresentatzeko modu bat da. RDFa XMLz ere idatzi daiteke. 3. irudian ikusi daiteke dago aurreko adibide berdina XMLz adierazita:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:ns="http://www.adibidea.org/#">
  <ns:Pertsona rdf:about="http://www.adibidea.org/#Jon">
    <ns:Ama rdf:resource="http://www.adibidea.org/#susana" />
    <ns:Aita>
      <rdf:Description rdf:about="http://www.adibidea.org/#Pedro">
        <ns:Anaia rdf:resource="http://www.adibidea.org/#Lukas" />
      </rdf:Description>
    </ns:Aita>
  </ns:Pertsona>
</rdf:RDF>
```

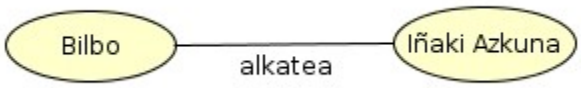
3. irudia: RDFa XML formatuan

RDFa esanahiarekin lotuta dago gehienbat, RDFan agertzen den guztiak esanahi bat dauka. Guk lan honetan erabiliko dugun formatua goian aipatu dugun N3 formatua izango da.

Web semantikoan datuak RDF lengoaia estandarrean adierazten dira. RDFetan informazioa egitura jakin batean adierazten da: RDF triplea (4. irudia). Hiru elementuk osatzen dute RDF triplea: subjektua, predikatua eta objektua. Predikatuak subjektua eta objektua lotzen ditu. Adibidez, "<Bilbo> <alkatea> <Iñaki_Azkuna>" RDF triplean *Bilbo* subjektua *alkatea* predikatuaren bidez *Iñaki Azkuna* objektuarekin erlazionatzen da. Horrelako RDF tripleak lotuz informazio konplexua adierazten duten RDF sareak eratzen dira.

Beste modu batera esanda, RDF tripleak, baliabidea, baliabidea deskribatzeko propietatea, eta propietate horrek hartzen duen balioarekin osatzen dira. RDF triplearen osagai bakoitza (subjektua edota baliabidea, predikatua edota propietatea eta objektua edota balioa) URL baten bidez identifikatua egon behar du. Baliabidea deskribatuko den hori da, propietatea baliabidearen ezaugarri bat da, deskribatu nahi dena, eta balioak, deskribatu nahi diren ezaugarrien balio konkretuak dira. Azken hau, balio literal bat izan daiteke, *string* bat, ala beste baliabide bat, URL bat.

Kasu honetan, entitate baten (Bilbo), propietatea (zein alkate duen adierazten duen *alkatea* propietatea) beste entitate bat da (Iñaki Azkuna). Horrelako hirukoteak lotuz informazio konplexua adierazten duten RDF sareak eratzen dira.



4. irudia: RDF triple baten adibidea (Subjektua: Bilbo ; Predikatua: alkatea ; objektua:Iñaki Azkuna).

```
<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/alkatea>
<http://dbpedia.org/resource/Inaki_Azkuna> .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/azalera>
"4126"^^<http://www.w3.org/2001/XMLSchema#int> .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/biztanleria>
"354860"^^<http://www.w3.org/2001/XMLSchema#int> .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/eskualdea>
<http://dbpedia.org/resource/Bilbo_Handia> .
```

5. irudia: RDF tripleen adibidea

RDF tripleak lotuz RDF sare bat eratzen da. RDF sare batean, hirukote baten subjektua beste hirukote baten objektua izan daiteke. 6. irudian agertzen den adibidean esaterako, *Bilbo* entitatea hirukote baten objektua eta beste hirukote baten subjektua da. Hau da, *Bizkaia-hiriburua-Bilbo* hirukotean objektua da *Bilbo*, aldiz, *Bilbo-alkatea-Iñaki Azkuna* RDF triplearen subjektua da.

6. irudia: Bilbo entitatea hirukote baten objektua eta beste hirukote baten subjektua da

RDF erabiliz Interneteko informazioa modu egituratuan (RDF-triple egitura jarraituz) adierazi daiteke. Honek hainbat informazio-iturri elkartzea ahalbidetzen du: RDF sare baten konektaturiko elementuak (Subjektuak, predikatuak edo objektuak) Interneteko hainbat baliabidetik etor daitezke eta.

Gainera, predikatuak semantikoak dira. Hau da, *izena* erlazioak edota predikatuak dagokion subjektuaren izenera eramango gaitu, *kokapena* erlazioak dagokion subjektuaren kokapenera eta abar. Beraz, zerbaiten kokapena jakin nahi baldin badugu, kokapena predikatua jarraitu besterik ez dugu.

Lan honetan, euskarazko Wikipediako infotauletatik erauzitako RDF sarea erabiliko dugu eta hau 3. atalean azalduko dugu.

2.3.2 SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) RDFen gainean galdetzeko lengoaia da. Lengoaia honekin RDF formatuan biltegiratutako datuak aldatu eta berreskuratu ahal izango ditugu. W3C-eko (World Wide Web Consortium, webaren jarraibideak, estandarrak eta eboluzioa gidatzen dituen erakundea) Data Access Working Group (DAWG) taldeak RDFak kontsultatzeko lengoaia estandarra izendatu du eta web semantikoko teknologia funtsezkoa kontsideratzen da. W3C berak ere gomendatutako lengoaia da.

SPARQL lengoaia erabiliz, RDF sareari edozein galdera egin diezaiokegu eta erantzuna jaso. SPARQL galderak hiruko patroiez (triple patterns), juntagailu, disjuntzio eta aukerazko beste patroi batzuekin osatzen da.

Ondoren azaltzen den SPARQL galderak esaterako, Bilboko alkatea zein den (*zein da Bilboko alkatea?*) bilatuko du. Galdera 7. irudian azaltzen den moduanadieraz daiteke SPARQL lengoaian:

```
PREFIX rs:<http://dbpedia.org/resource/>
PREFIX dc:<http://dbpedia.org/property/>
SELECT ?pertsona
WHERE
{ rs:?herria dc:alkatea ?pertsona .
FILTER regex(STR(?herria), "Bilbo")}
```

7. irudia: SPARQL galdera baten adibidea

Aldagaiak “?” karakterearekin adierazten dira. Adibidean subjektu gisa “Bilbo” duten eta predikatu gisa “alkatea” duten hirukote guztiak bilatuko ditu eta hauen objektua (adibidean ?pertsona) itzuliko du.

SPARQL galdera prozesatzaileak galderan adierazten diren patroiak betetzen dituzten RDF tripleak bilatuko ditu, galderako aldagai bakoitza dagokion RDF tripleko zatiarekin lotuz.

3 Erabilitako baliabideak eta tresnak

Jarraian, lan honetan erabili ditugun tresnak eta baliabideak deskribatuko ditugu. Alde batetik, erabilitako RDF ezagutza-basearen ezaugarriak azalduko ditugu, eta, bestetik, RDF honen gainean SPARQL lengoaian idatzitako galderak prozesatu eta erantzunak lortzeko eta hizkuntza naturalean idatzitako galderen analisia lortzeko erabili ditugun tresnak zeintzuk izan diren aipatuko ditugu.

3.1 Wikipediako infotauletatik erauzitako RDFa

DBpedia[8] komunitateak Wikipediatik egituratutako informazioa atera eta hau webean eskuragarri jartzeko ahalegina burutzen du. Honez gain, Dbpediak aukera ematen du ingelesezko Wikipediaren kontra SPARQL lengoaian hainbat kontzeptu erlazionatzen dituzten kontsulta sofistikatuak egiteko eta webeko beste datu bilduma batzuk Wikipediarekin lotzeko. Honela, web semantikoaren ahalmena erakusten saiatzen dira. Modu honetan, Wikipediako informazioa erabiltzeko modu berri eta interesgarrietan erabiltzeko aukerak sortzen dira.

DBpedia proiektua 2007an jarri zen abian eta esan bezala, Wikipediatik datuak erauzten ditu Wikipediaren bertsio semantikoa eskaintzeko (web semantikoan erabili daitekeen bertsioa). DBpedian informazioa RDF lengoaia erabiliz adierazten da eta euskararen kasurako Wikipediako infotauletatik, hau da, Wikipediako artikulu batzuk albo batean izaten duten informazio egituratuzko kutxetatik RDF sare bat erauzi dute (http://downloads.dbpedia.org/3.7/eu/). 8. irudian agertzen den kode zatia euskarazko wikipediako infotauletatik erauzitako RDF sarearen zati bat da, euskarazko Wikipediako *Bilbo* artikuluaren infotaulari dagokiona hain zuzen:

8http://dbpedia.org

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/alkatea> <http://dbpedia.org/resources/Iñaki_Azkuna>.

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/armarria> "Escudo heraldico de Bilbao.svg"@eu .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/azalera>"4126"^^<http://www.w3.org/2001/XMLSchema#int> .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/biztanleria>"354860"^^<http://www.w3.org/2001/XMLSchema#int> .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/biztUrtea>"2009"^^<http://www.w3.org/2001/XMLSchema#int> .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/eskualdea> <http://dbpedia.org/resource/Bilbo_Handia> .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/euskaldunenEhunekoa>"24.2"^^<http://www.w3.org/2001/XMLSchema#double> .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/garaiera> "16"^^<http://www.w3.org/2001/XMLSchema#int> .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/herritarra> "bilbotar"@eu .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/izena> "Bilbo"@eu .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/nongo> "Bilboko"@eu .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/ofiziala> "Bilbao"@eu .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/postakodea> "48001-48015"@eu .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/sorrera> "1300eko ekainaren 15a"@eu .

<http://dbpedia.org/resources/Bilbo> <http://dbpedia.org/property/webgunea> "http://www.bilbao.net/WebBilbaonet/home_e.jsp?idioma=e www.bilbao.net"@eu .

<http://dbpedia.org/resources/Bilbo><http://dbpedia.org/property/wikiPageUsesTemplate><http://dbpedia.org/resource/Template:Bizkaiko_udalerri_infotaula> .

8. irudia: Euskarazko RDFaren zati bat

RDF sare hau izango da hain zuzen garatu dugun sistemak erabiliko duen ezagutza-basea. Euskal Wikipediako infotauletatik erauzitako RDF ezagutza-sareak lantzen dituen baliabide edota entitateak eta hauen propietateak zeintzuk diren aztertu ditugu, erabiliko dugun RDF honek zer nolako galderak erantzun ahal izango dituen aztertzeko asmoarekin.

Hasteko, RDFak lantzen dituen entitate-motak aztertu ditugu. Entitateak deitzen diegu infotaula duten Wikipediako artikuluen tituluei edota RDF sareko tripleetako subjektuei (8. irudian *Bilbo* da lantzen den entitatea eta tripleetako subjektua). Izan ere, infotaula batetatik RDF tripleak sortzeko artikuluen tituluak subjektu bezala hartzen dira, predikatu bezala berriz infotaulako ezkerreko zutabeko eremuetako hitzak eta

azkenik, objektu bezala eskuin zutabeko balioak. Hau da, infotaulako lerro bakoitzeko RDF triple bat sortuko da (ikusi 9.irudia).

9. irudia: Wikipediako artikulu baten infotaula

9. irudian irudian Wikipediako *Bilbo* artikuluaren infotaularen zati bat azaltzen da. Kasu honetan, infotaula honetatik erauzitako RDF tripleetan *Bilbo* subjektua edota deskribatzen den baliabidea izango da, urdinez dagoen zutabeko hitzak *Bilbo* deskribatzeko erabiltzen diren propietateak (*Izen ofiziala, Estatua, Erkidegoa...*) eta azkenik, propietate hauen balioak edota objektuak zuriz dagoen zutabean azaltzen dena (Adibidez, <Bilbo> <Eskualdea> <Bilbo_Handia> RDF triplea sortuko da).

Artikuluetako infotaulak sortzeko txantiloi batzuk aplikatzen dituzte. Goiko adibidearekin jarraituz, *Bilbo* entitatearentzat *Bizkaiko udalerri infotaula* txantiloia erabiltzen da. 10. irudian ikusi daiteke mota hauetako entitateen infotaulentzat aurreikusitako propietateak.

```
{{Bizkaiko udalerri infotaula
| izena =
| bandera =
| armarria =
| nongo =
| ofiziala =
| eskualdea =
| postakodea =
| kokapena =
| koordenatuak =
| garaiera =
| azalera =
| biztanleria =
| bizt_urtea =
| dentsitatea =
| distantzia =
| sorrera =
| euskaldunen ehunekoa =
| herritarra =
| webgunea =
| oharrak =
}}
```

10. irudia: Infotaula baten txantiloila

Hala ere, aipatu behar da, kasu guztietan ez direla txantiloiko propietate guztien balioak betetzen, hutsak ere egon daitezke, alegia.

- **Entitateak**

RDFko entitateak sailkatzeko, entitate horren Wikipediako artikuluan erabiltzen den infotaularen txantiloiaren izena erabili dugu entitate-mota bezala. Adibide berdinarekin jarraituz, *Bilbo Bizkaiko udalerri* motako entitate bezala sailkatuko dugu (ikusi 1. taula).

Entitatea	Entitate-mota
Bilbo	Bizkaiko udalerri
Iñaki Azkuna	Agintari
Bizkaia	lurralde

1. taula: entitate--entitate-mota adibideak

Modu honetan, RDF sareko entitate guztientzat "*entitate—entitate-mota*" bikotea lortuko dugu. Hau, sarrerako galdera hizkuntza naturaletik SPARQLra itzultzean erabilgarria izango da galderan lantzen den entitatea identifikatzeko batetik (zein entitateri buruzko galdera den jakiteko) eta baita entitateari buruz zer galdetzen den jakiteko (entitateari buruzko zein propietateren balioa eskatzen den galderan jakiteko). Hau da, galderako entitatea identifikatu ondoren, entitate-mota horrentzat RDFan zein propietate dauden jakin ahal izango dugu.

Euskarazko Wikipediako infotauletatik erauzitako RDFan 61000 entitate desberdin inguru daude eta hauen artean 205 entitate-mota desberdin. Entitate-mota bakoitzeko zenbat entitate dauden ere neurtu dugu eta, aipatu behar da, entitate-mota batzuetarako oso entitate gutxi daudela: 100 entitate baino gehiago dituzten 38 entitate-mota daude eta 50 entitate baino gehiago dituzten 69 entitate-mota daude.

2. taulan, RDFan gehien errepikatzen diren entitate-mota batzuk azaltzen dira. Lehenengo zutabean, bigarren zutabean azaltzen den entitate-motako zenbat entitate dauden RDFan azaltzen da.

Entitate kopurua	Entitate-mota
35040	Frantziako udalerri
6820	Espainiako udalerri
1921	Taxotaula
1134	Hiri orokor
965	Biografia
595	Aktore biografia
594	Probintzia orokor
443	Agintari

2. taula: entitate-mota eta hauen kopuruak

Entitate-mota gehienek lekuei (*Frantziako_udalerri, Espainiako_udalerri, Hiri_orokor, Probintzia_orokor, Nafarroako_udalerri, Ibai_infotaula* ...) eta pertsona edota pertsonaiei (*Biografia, Aktore_biografia, Agintari, Futbolari, Musika_talde* ...) egiten die erreferentzia. Beraz, esan daiteke, RDFan gehienbat kontzeptu geografikoak eta biografikoak daudela. Hau kontuan izan dugu landu ditugun galdera-motak erabakitzerako orduan.

- **Propietateak**

Entitate-motez gain, RDF sarean erabiltzen diren propietateak edota predikatuak ere aztertu ditugu. "*Entitate-mota--propietate*" bikoteen zerrenda atera eta entitate-mota bakoitzarentzat zeintzuk diren RDFan gehien erabiltzen diren propietateak ikusi dugu.

10. irudian ikusi dugu nola entitate-mota bakoitzerako txantiloietan bertan propietateak definitzen diren, baina aipatu den bezala, batzuetan entitate guztientzat ez da informazio hori guztia betetzen (RDFan ez daude txantiloietan azaltzen diren propietate guztietarako tripleak, hauek entitatearen arabera aldatzen dira.)

Gehien errepikatzen diren propietateen azterketa hau erabili dugu landuko ditugun galderen galdetzaileak erabakitzeko.

3. taulan guk erabili dugun RDFan gehien errepikatzen diren propietate batzuk azaltzen dira.

Entitate-mota	Propietatea	Kopurua
Frantziako udalerria	alkatea	35031
Frantziako udalerria	postakodea	35029
Frantziako udalerria	kokapena	34936
Espainiako udalerria	azalera	6752
Espainiako udalerria	biztanleria	6659
Biografia	jaiotza herrialdea	960
Biografia	jaiotza data	958
Agintari	kargua	205
Agintari	jaiolekua	198

3. taula: entitate-mota propietate bikoteen kopuru batzuk

Gehien errepikatzen diren propietateen zerrenda hau aztertuta, ondorengo galdetzaileak inplementatuko ditugula erabaki dugu: *Noiz, Non, Nor, Nork eta Zein*. Galdetzaile hauekin RDFan biltegiratzen den informazio gehienari buruzko galderak sor daitezkeela uste dugu. Badakigu galdetzaile batzuk kanpoan ustean edota ez inplementatzean sistema galdera batzuetara mugatzen ari garela. Hala ere, aukeratutako galdetzaile horiekin galdera kopuru handia barneratzen dituztela uste dugu.

3.2 RDFa atzitzeko tresna (JENA/ARQ)

Webean web semantikoarekin lotutako hainbat tresna aurki daitezke, besteak beste, RDF sareen kudeaketarako sistemak. Hauek beharrezkoak dira RDF ezagutza-baseekin lan egiteko, izan ere RDFetako tripleak biltegiratu eta hauen gainean SPARQL galderak prozesatzeko aukera ematen dute.

Gure sistemak SPARQL galdera sortu ondoren, honelako tresna bat beharko dugu RDFtik erantzuna jasotzeko. Merkatuan RDFak kudeatzeko tresna asko dagoen arren (RDFStore, OpenLink Virtuoso, Sesame...), erabilienetako bat JENA izenekoa da. JENAk ARQ[9] izeneko kontsulta motore bat dauka zeinek SPARQL lengoaia onartzen duen. Hau izango da guk erabiliko dugun SPARQL prozesatzailea.

9http://jena.apache.org/documentation/query/index.html

3.3 Ihardetsiren galdera analizatzailea

Sarrerako euskarazko galderaren analisia lortzeko IHARDETSI galderak erantzuteko sistemak erabiltzen duen prozesatzaile linguistiko berdina berrerabiltzen da (Ansa et al., 2008). Sistema honek bi hizkuntza prozesadore erabiltzen ditu, oro har: MORFEUS izeneko lematizatzaile/etiketatzailea (Ezeiza et al., 1998) eta Eihera (Alegria et al., 2004) (Named Entity Recognition and Classification (NERC)). Galderako unitate lexiko bakoitzarentzat, bere lema eta kategoria itzuliko digu, bai hitz bakunentzat eta baita hitz anitzeko unitateentzat. Zenbakiak eta denborazko adierazpideak ere harrapatzen ditu lematizatzaile/etiketatzaileak. NERC prozesatzaileak, Ehierak, pertsona, erakunde eta leku entitateak harrapatzen ditu.

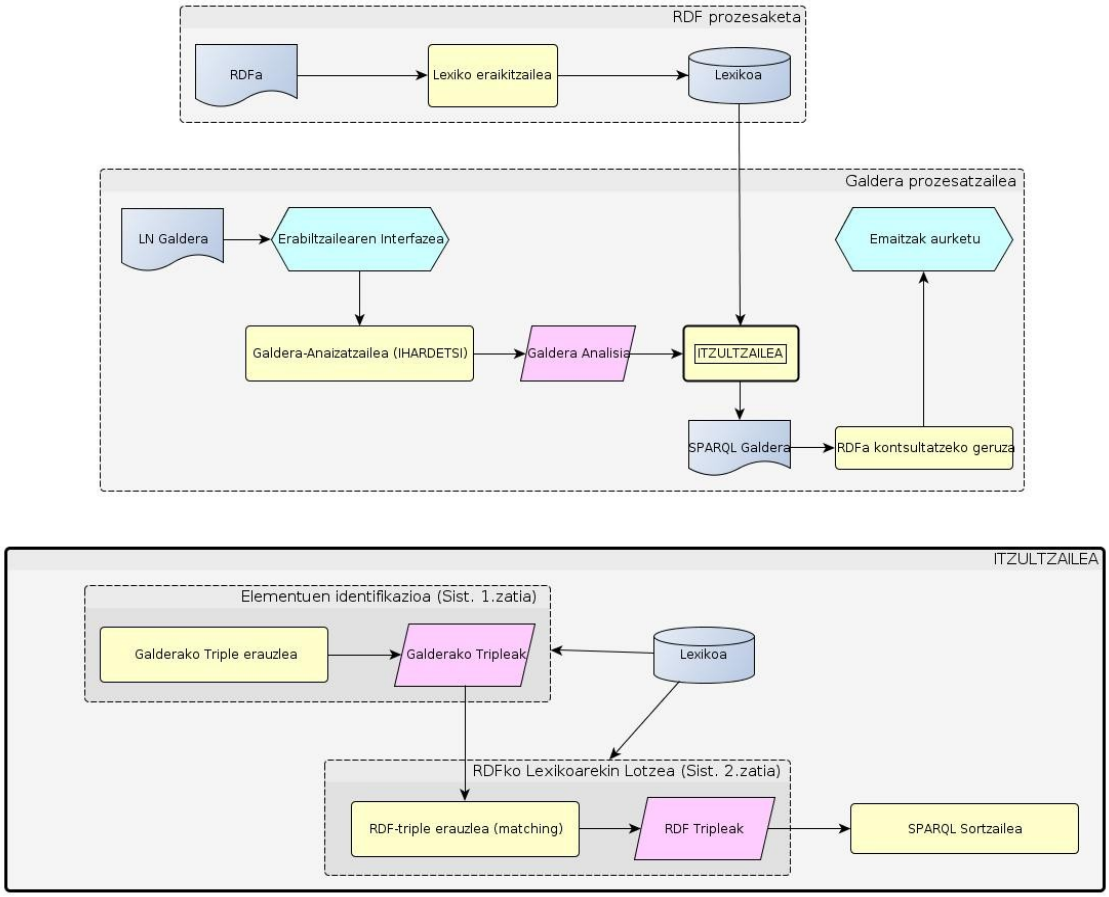
4 RDFen gaineko galdera-erantzuna

Memoriako atal honetan, galdera-erantzuna RDFaren gainean egiteko erabili dugun hurbilpenaren ezaugarriak azalduko ditugu. Lehendabiziko puntuan arkitektura deskribatuko dugu, ondoren, landu ditugun galderen ezaugarriak azalduko ditugu eta azkenik, aplikatu ditugun teknikak eta heuristikoak azalduko ditugu.

4.1 Arkitektura

Gure sistemak jarraitzen duen arkitektura azaltzen da memoriako atal honetan. 11. irudian azaltzen den arkitekturan bi modulu nagusi bereizten dira: alde batetik, erabiliko dugun RDF ezagutza-basetik ondoren erabiliko dugun lexikoaren prestakuntzara zuzendutako modulua eta beste alde batetik sarrerako hizkuntza naturalean idatzitako galdera SPARQL lengoaiara itzuli eta azken hau erabilita RDFtik galderarentzako itzulpena lortzeaz arduratzen den modulua.

Lehenengo pausuan RDFtik lexikoa erauzi eta sinonimoekin aberasten da, sistemaren estaldura handiagoa izan dadin. Gainera, galdera prozesatzen duen moduluan, sarrerako galdera analizatu, galdera-patroia identifikatu, galderaren osagai nagusiak identifikatu eta SPARQL galdera eraikitzen da. Azkenik, SPARQL galdera erabilita, RDFtik sarrerako galderarentzat erantzun egokia lortzen saiatzen da.



11. irudia: sistemaren arkitektura

Interfazea kenduta arkitekturako modulu guztiak inplementatuta daude. Erabiltzailearen interfazea garatu ez dugun arren, arkitekturaren irudian gehitu dugu, azkeneko sistemak eduki beharko lituzkeen osagai guztiak azaldu ahal izateko.

Jarraian, arkitekturako modulu guztiak azalduko dira: lexikoaren prestakuntza modulua, galdera SPARQL lengoaia itzultzen duen modulua, eta, azkenik, RDFtik erantzuna lortuko duen modulua.

4.1.1 Lexikoaren prestakuntza modulua

Modulu honetan, erabiliko dugun RDF ezagutza-basearen araberako lexikoa sortzen da. RDFtik bertatik erauziko dugu hurrengo moduluan, galderaren SPARQLaren sorkuntzan, hizkuntza naturalean idatzitako galderako subjektu edota entitatearen eta predikatuen edota entitateen propietateen identifikazioan lagunduko diguten hiztegiak. Erabiliko den RDFa aldatzen ez baldin bada, lexikoaren sorkuntza prozesu hau behin egitearekin nahikoa izango da.

Galderako entitatea modu egokian identifikatu ahal izateko erabilitako hiztegien artean hauek aurkitzen dira: RDFko entitate guztien zerrenda, Wikipediako artikuluen berbiderapenetatik sortutako hiztegia eta RDFko propietate jakin batzuetatik erauzitako hiztegia. Jarraian, hiztegi hauek nola sortu diren azalduko da.

- **RDFko entitate guztien zerrenda.** RDFko triple guztietatik subjektuetako entitate guztien zerrenda erauzi dugu. Modu honetan, sarrerako galderako hitz bakarreko ala hitz anitzeko entitatea identifikatu ahal izango dugu, galderan RDFan adierazten den forma berdinean adierazi baldin bada behintzat.
- **Wikipediako artikuluen berbiderapenetatik sortutako hiztegia.** RDFan lantzen diren entitateei dagozkien sinonimoak lortzeko wikipediako artikulueen berbiderapenen informazioa erabili dugu hiztegi hau lortzeko. Beheko irudian, wikipediako artikulu baten berbiderapen bat nola adierazten den ikus daiteke. 12. irudian azaltzen den berbiderapenari esker, *urtxintxa* eta *katagorri* terminoak sinonimoak direla ondorioztatu dezakegu:

```
<page>
  <title>Urtxintxa</title>
  <id>35923</id>
  <redirect />
  <revision>
    <id>242419</id>
    <timestamp>2006-10-27T17:04:20Z</timestamp>
    <contributor>
      <username>Barrie</username>
      <id>431</id>
    </contributor>
  <comment>[[Katagorri]] orrialdera berbideratzen</comment>
 <text xml:space="preserve">
  #REDIRECT [[Katagorri]]
 </text>
  </revision>
 </page>
```

12. irudia: Wikipediako artikulu baten berbiderapen-kodea

Wikipediako artikulu zehatz bat bilatzeko artikuluan azaltzen den izenburuaren edota entitatearen araberako bilaketa egiten dugu askotan. Bilaketa artikuluko izenburuan entitateak duen forma zehatzean egin ez arren, askotan artikulua aurkitzen da. Hau berbiderapenei esker gertatzen da. 12. irudian azaltzen den Wikipediako "*katagorri*" entitatean artikulura "*urtxintxa*" bezala bilatuz ere heldu gaitezkeela ikusten da.

13. irudia: Wikipediako artikulu baten berbiderapena

13. irudian azaltzen den moduko entitateen sinonimoez gain beste forma bateko sinonimoak ere lortzen dira berbiderapenei esker. Askotan, pertsonaia famatu bati buruzko informazioa galdetzeko erabiltzaileak ez du bere izen osoa idatziko, izan ere, pertsonaia asko bere abizenaz bakarrik ezagutzen dira edota abizenaz bakarrik adierazteko ohitura dago (ikusi 4. taula).

Berbiderapena	Entitatea
Galileo	Galileo Galilei
Copernico	Nikolas Koperniko
Koperniko	Nikolas Koperniko
Hitler	Adolf Hitler
Lenin	Vladimir Lenin
Kant	Immanuel Kant

4. taula: pertsona izenen berbiderapenak

Pertsona edota leku izen bat adierazteko modu desberdinak egon daitezke baita ere (ikusi 5. taula).

Berbiderapena	Entitatea
Enrike II.a Albretekoa	Henrike II.a Nafarroakoa

5. taula: entitate berdina idazteko modu desberdinak

Beste askotan, leku edo pertsona bat adierazteko forma bat zabaldua egon arren ofizialki beste modu batera idazten dira eta wikipediako artikuluen izenburuan eta ondorioz RDFan ere entitate hauek beste modu batera adierazita daude. Hauek ere berbiderapenen informazioa erabilita lortu ahal izango ditugu (ikusi 6. taula).

Berbiderapena	Entitatea
Madrid	Madril
Donosti	Donostia
Mozanbike	Mozambike
Joxemari Iturralde	Joxe Mari Iturralde
Penintsula Iberiko	Iberiar penintsula
Ertamerika	Erdialdeko Amerika
Bretainia Handia	Britainia Handia

6. taula: entitateen izen ofizial eta erabiliaren adibideak

Siglen erabilera ere kontrolatuko dugu berbiderapenei esker (ikusi 7. taula).

Berbiderapena	Entitatea
Alderdi popularra	PP
ETB	Euskal Telebista

7. taula: siglak eta forma osoak

Eta azkenik, askotan beste hizkuntza batean, normalean ingelesez edo latinez, adierazten diren termino batzuk ere, berbiderapenen hiztegiari esker identifikatu eta itzuli ahal izango ditugu (ikusi 8. taula).

Berbiderapena	Entitatea
London Tower	Londresko dorrea
Holanda	Herbehereak
Boletus aereus	Onddobeltz

8. taula: entitateak beste hizkuntza batzuetan

Berbiderapenen hiztegi honi esker, erabiltzaileek erabili ahal izango duten lexikoa zabalagoa izango da eta galdera gehiago modu egokian erantzutea ahalbideratuko digu. Izan ere, galderako entitatea ez baldin bada modu egokian identifikatzen, ezingo dugu hizkuntza naturalean idatzitako galderarentzako SPARQL lengoaian idatzitako galdera egokia sortu, eta, ondorioz, erantzun zuzena ere ez da aurkitu.

Modu honetan aurkitu den hiztegiak 41086 sarrera ditu, beraz, lagungarria izan daitekeela suposatzen dugu.

- **RDFko propietate jakin batzuetatik erauzitako hiztegia.** Entitate-mota batzuentzat erabiltzen diren infotaularen txantiloileko propietateak aztertuta, hainbat propietateek entitatea bera adierazteko dauden forma desberdinei buruzko informazioa ematen digula ohartu gara. Horrela, RDFko propietate hauek dituzten tripleak aukeratu ditugu eta hauetatik, subjektu eta objektuak lotuz hiztegi berri bat sortu dugu. Entitate biografikoetan, *"izena"* "*ezizena*" eta "*izenOsoa*" propietateen informazioa erauziko dugu. 9. taulan azaltzen dira, propietate hauen informazioari esker lortuko ditugun entitate batzuen sinonimoak.

Entitatea	Propietatea	Balioa
Alex Ferguson	izena	Sir Alex Ferguson
Alex Ferguson	ezizena	Fergie
Txema Vitoria	ezizena	Txiribiton
Alex Ferguson	izenOsoa	Alexander Chapman Ferguson
Bernardo Atxaga	izenOsoa	Jose Irazu Garmendia

9. taula: “izena”, “izenOsoa” eta “ezizena” propietateen balioen adibideak

Entitate geografikoetan, aldiz, “*berezkoIzena*” eta “*berezkoIzenaHizkuntzaOfizialean*” bezalako propietateen informazioa erauziko dugu. 10. taulan azaltzen dira, propietate hauen informazioari esker lortuko ditugun entitate batzuen sinonimoak.

Entitatea	Propietatea	Balioa
Tunisia	berezkoIzena	Tunisiako Errepublika
Andorra	berezkoIzenaHizkuntzaOfizialean	Principat d'Andorra

10. taula: “berezkoIzena” eta “berezkoIzenaHizkuntzaOfizialean” propietateen balioen adibideak

RDFko propietate jakin batzuetatik abiatuta sortutako hiztegi honi esker ere, erabiltzaileek erabili ahal izango duten lexikoa zabalagoa izango da.

Galderako propietateak modu egokian identifikatu ahal izateko, entitateekin egin dugun modu berdinean **RDFko propietate guztien zerrenda** erauzi dugu. Honetarako, RDFko triple guztietatik propietatea erauzi ditugu. Modu honetan, sarrerako galderako hitz bakarreko (*“hiriburua”*) edota hitz anitzeko propietateak (*“entrenatutakoTaldeak”*) identifikatu ahal izango ditugu galderan, RDFan adierazten den forma berdinean adierazi baldin badira behintzat. Guztira, 2500 propietate desberdin inguru erauzi dira.

Erabili ditugun hiztegi guztiak RDFtik bertatik sortuak izan dira, hau da, ez ditugu hitz arruntentzako bestelako sinonimoen hiztegirik edota galderako hitzen hiperonimo/hiponimoak identifikatzeko ontologiarik erabili.

4.1.2 Galdera SPARQL lengoaiara itzuli eta erantzuna lortzea

Hizkuntza naturalean idatzita dagoen sarrerako galdera SPARQLra itzultzeko burutzen diren pauso guztiak deskribatuko dira hemen. Jarraian, erabilitako galdera-analizatzailea, itzultzailea modulua (sarrerako galdera hizkuntza naturaletik SPARQL lengoaiara itzultzen duen modulua) eta RDFa kontsultatzeko erabili dugun tresna azalduko dira.

- **Galdera-analizatzailea:** sarrerako euskarazko galderaren analisia lortzeko Ihardetsi galderak erantzuteko sistemak erabiltzen duen prozesatzaile linguistiko berdina berrerabiltzen da (Ansa et al., 2008). Sistema honek bi hizkuntza prozesadore erabiltzen ditu, oro har: MORFEUS izeneko lematizatzaile/etiketatzailea (Ezeiza et al., 1998) eta Eihera (Alegria et al., 2004) (Named Entity Recognition and Classification (NERC)). Galderako unitate lexiko bakoitzarentzat, bere lema eta kategoria itzuliko digu, bai hitz bakunentzat eta baita hitz anitzeko unitateentzat. Zenbakiak eta denborazko adierazpideak ere harrapatzen ditu lematizatzaile/etiketatzaileak. NERC prozesatzaileak, Ehierak, pertsona, erakunde eta leku entitateak harrapatzen ditu. 14. irudian, galdera-analizatzaileak *"Non jaio zen Bilboko alkatea?"* galderarentzako itzultzen duen analisia ikus daiteke.

```
<?xml version="1.0" encoding="UTF-8"?>
<GALDERA_ANALISIA fasea="bilaketa terminoen erauzketa">
<JATORRIZKOGALDERA>Non jaio zen Bilboko alkatea?</JATORRIZKOGALDERA>
<GALDERA>Non jaio zen Bilboko alkatea?</GALDERA>
<TERMI FRM="Non">
<LEX LEM="non" KAT="ADB" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI=""/>
</TERMI>
<BILA_TERMI FRM="jaio">
<LEX LEM="jaio" KAT="ADI" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI=""/>
</BILA_TERMI>
<TERMI FRM="zen">
<LEX LEM="izan" KAT="ADL" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI=""/>
</TERMI>
<BILA_TERMI FRM="Bilboko">
<LEX LEM="Bilbo" KAT="IZE" KASUA="GEL" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="ENTI_LOC"/>
</BILA_TERMI>
<BILA_TERMI FRM="alkatea">
<LEX LEM="alkate" KAT="IZE" KASUA="ABS" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI=""/>
</BILA_TERMI>
<TERMI FRM="?">
<LEX LEM="" KAT="" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI=""/>
</TERMI>
<ESPEROERANTZUNA>LOCATION</ESPEROERANTZUNA>
<GALDERAMOTA>FACTOID</GALDERAMOTA>
<MINTZAGAIA>Bilbo alkate</MINTZAGAIA>
<GALDEGAIA>non</GALDEGAIA>
</GALDERA_ANALISIA>
```

14. irudia: Galderen analisiaren irteera formatua

Galderaren analisitik, unitate lexikal bakoitzerako lemak jasotzeaz gain (adibidez, 14. irudian, <LEX LEM=”alkate”), identifikatutako hitz anitzeko terminoen informazioa ere erabiliko dugu. Izan ere, identifikatutako hitz anitzeko terminoa,

normalean, bilatzen ari garen entitatea izango da, hala nola, sortu behar dugun SPARQL galderan subjektua izango dena.

Honez gain, analisian itzulitako < ESPEROERANTZUNA> eremuaren informazioa jasoko dugu, ondoren erantzun posible bat baino gehiago dagoen kasuetan erantzun zuzena lortzeko filtroak jartzeko.

IHARDETSIren galdera-analizatzailea erabiltzeak, ondoren burutuko den 2 sistemen arteko azterketarako, bi sistemak galdera batentzat itzulitako analisi berdinetatik abiatu direla ziurtatuko digu.

Galderaren analisia ITZULTZAILEA deitu diogun moduluari pasako diogu. Modulu hau izango da sistemaren muina.

- **Itzultzailea**: hizkuntza naturalean idatzitako galderak SPARQL lengoaiara itzultzeko prozesua burutzen da modulu honetan. Prozesu honetan bi pauso nagusi daude: SPARQL galderako elementuen identifikazio zuzena burutzea eta SPARQL galderako osagai hauen lexikoa erabiltzen ari garen RDF ezagutza-basean erabiltzen denarekin lotzea eta egokitzea. Hau da, lehendabiziko pausuan bukaerako SPARQL galderan subjektu, predikatu eta objektuak izango diren zatiak identifikatu behar dira sarrerako galderan eta bigarren pausuan aldiz, identifikatutako subjektu, predikatu eta objektuak RDFko hiztegiarekin lotu behar dira, erabiltzen ari garen ezagutza-basearekiko egokia den SPARQL galdera sortu ahal izateko. Bi pauso hauek hobeto nabarmentzeko, arkitekturaren irudian bi modulu desberdin marraztu ditugu eta honela azalduko dira. Hala ere, aipatu behar da, bi pausu hauek askotan nahastu egiten direla, izan ere, sparql galderako osagaien identifikazioan, RDFtik erauzitako lexikoa erabiltzen da.
    - **Galderako triple erauzlea:** honela deitu diogu SPARQL galderako elementuak identifikatzeaz arduratzen den zatiari. Honek hizkuntza naturalean dagoen galderan, ondoren SPARQL galderan subjektu ala predikatu izango direnak identifikatuko ditu. Pauso honetan, galderen egiturak du eragina, beraz, lan honetan landu ditugun galderen patroietan eta aplikatuko ditugu heuristikoetan oinarritu gara modulu hau inplementatzeko.
    - **RDF triple erauzlea:** identifikatutako osagaiekin erabiliko den ezagutza-basearekiko egokia den SPARQL galdera sortzea izango da helburua. Hemen, galderan erabiltzen den hiztegiak du eragina. Pauso hau, erabiliko

dugun ezagutza-basearekin lotuta dago, izan ere RDFan bertan erabiltzen den hiztegiarekin lotu behar dira galderako hitzak. Adibidez: *Zein da Bilboko agintaria?* galderan, lehenengo pausuan *Bilbo* subjektu bezala identifikatu dugu eta *agintaria* propietate bezala. Beraz, 15. irudian azaltzen den SPARQL galdera sor dezakegu. SPARQL galdera egokia izan arren, guk erabiliko dugun RDFan Bilbo entitateak edota subjektuak ez du *agintari* propietaterik, honen ordez, *alkate* propietatea dauka. Beraz, SPARQL galdera ondo sortua egon arren, ez luke erantzuna aurkituko. Kasu honetan RDFan erabiltzen den lexikora egokitu beharko genuke galdera.

```
SELECT ?erantzuna
WHERE
{ Bilbo agintari ?erantzuna}
```

15. irudia: SPARQL galdera

- **SPARQL sortzailea:** azkenik, aurreko pausuetan identifikatutako osagai guztiekin SPARQL formatuan sortuko dugu galdera eta RDFa kontsultatzeko erabiliko dugun moduluari pasako diogu.

- **RDFa kontsultatzeko geruza**: aurreko moduluan sortutako SPARQL galdera jasoko da hemen eta RDFak kudeatu eta kontsultatzeko ARQ SPARQL prozesatzaileari esker, RDFtik galdera horrentzako erantzuna jasotzen saiatuko gara hemen. Lehenengo saiakeran erantzunik jasotzen ez baldin bada, SPARQL aukera desberdinak sortuz, erantzuna bilatzen saiatuko da.

4.2 Landutako galderak

Gure sistema erantzuten saiatuko den galderak mugatu ditugu. Sistemaren lehenengo prototipoa garatzeko patroi jakin batzuk jarraitzen dituzten galderetan oinarritu gara (hauek 4.3.2 puntuan azaltzen dira).

Sistemaren garapenerako eta azken ebaluaziorako erabili ditugun galderak sortu ditugu. Alde batetik, garapen faserako 120 galdera *sinple* sortu ditugu eta beste alde batetik azken ebaluaziorako 73 galdera *sinple* eta 25 galdera *konplexu.*

Galdera *sinplea* esaten dugunean, sintaxia aldetik konplexutasun txikia duen galdera bat esan nahi dugu. Galdera hauen erantzuna galderako entitatearen artikuluan bertan

edota entitate horrentzako RDFko tripleetan aurkitu ahal izango da. Hau da, ez da entitate bat baino gehiago erlazionatu beharko erantzuna bilatu ahal izateko. Galdera hauek triple bakarreko SPARQL galderara itzuli beharko ditu gure sistemak. 16. irudian galdera sinple baten adibidea azaltzen da eta honi dagokion triple bakarreko SPARQL galderarena ere bai.

Galdera sinplea: *Non jaio zen Johannes Kepler?*
Triple bakarreko SPARQL galdera:

```
SELECT ?erantzuna
WHERE
{
?subjektua ?propietatea ?e
FILTER regex(str(?subjektua),"Johannes Kepler")
FILTER regex(str(?propietatea),"jaio")
}
```

16. irudia: Galdera sinple baten adibidea

Beste alde batetik, goian aipatu den bezala, sistemaren azken ebaluaziorako 25 galdera *konplexu* ere sortu ditugu. Galdera *konplexua* esaten dugunean, galdera hauen erantzuna aurkitzeko, entitate bat baino gehiago erlazionatu beharko ditugula esan nahi da. Hau da, galderan azaltzen den entitatean artikuluan bertan bakarrik begiratuta ez dugu erantzuna aurkituko, entitate bat baino gehiagoren artikuluak edota RDFko tripleak erlazionatuz bilatu beharko dugu erantzuna. Hala ere, galderen konplexutasuna ere mugatu da eta gehienez 2 tripletako SPARQL galdera eskatzen duten galderetara mugatu gara. 17. irudian galdera *konplexu* baten adibidea azaltzen da eta honi dagokion bi tripletako SPARQL galderarena ere bai.

```
Galdera: Noiz sortu zen Peruko hiriburua?
SPARQL galdera:
SELECT ?erantzuna
WHERE
{
?lotura ?propietateNagusia ?erantzuna
OPTIONAL { ?subjektuNagusia ?tartekoPropietatea ?lotura}
FILTER regex(str(?subjektuNagusia),"Peru")
FILTER regex(str(?tartekoPropietatea),"hiriburu","i")
FILTER regex(str(?propietateNagusia),"sortu","i")
}
```

17. irudia: Galdera konplexu baten adibbidea

Esan bezala, galdera hauen erantzuna bilatzeko artikulu bat baino gehiago konbinatu behar dira, beraz, testutik erantzuna ondorioztatzea lan zaila izan daiteke. Aldiz, RDFaren egitura dela eta, modu errazagoan lortu ahal izango dugu erantzuna, honetarako, SPARQL konplexuagoak sortu beharko ditugun arren. Adibidez, 17. irudian azaltzen den *"Noiz sortu zen Peruko hiriburua?"* galderaren erantzuna bilatzeko, lehendabizi "*Peruko hiriburua*" zein den lortu behar dugu.

Bai galdera sinpleen kasuan, bai galdera konplexuen kasuan, galderan entitatea azaldu beharko da beti eta galdetzaile hauek erabili ahal izango dira: *Non, Noiz, Nork, Nor* eta *Zein.*

Garapenerako eta azken ebaluaziorako galderen sorkuntza prozesua 5. puntuan azaltzen da xehetasun gehiagorekin.

4.3 Aplikatutako teknikak eta heuristikoak

Memoriako atal honetan, gure hurbilpena azalduko dugu. Lehendabizi SPARQL galderako osagaiak definituko dira, eta, ondoren, landutako galdera-patroiak zeintzuk diren azalduko dira. Azkenik, SPARQL galderako osagai bakoitzaren identifikazio prozesuan aplikatutako teknikak eta heuristikoetaz arituko gara.

4.3.1 SPARQL galdera osagaien definizioa

Hizkuntza naturalean idatzitako galderak SPARQL lengoaiara itzultzeko aplikatu ditugun teknikak eta oinarritu garen heuristikoak azaldu aurretik, hau guztia hobeto ulertzeko, kontzeptu batzuk definituko ditugu lehendabizi:

-**Subjektu nagusia**: galderan azaltzen den entitatea da. RDFko tripleetan subjektuaren aldean bilatu beharko dugu. *Noiz sortu zen Aconcagua mendia?* galderan "*Aconcagua*" da bilatzen ari garen *subjektu nagusia* eta *Zein da Frantziako hiriburuko alkatea?* galderan berriz "*Frantzia*" da bilatzen ari garen *subjektu nagusia*. *Subjektu nagusia* deitu diogu, SPARQL galderan sortuko ditugun tripleen artean ezaguna den subjektu bakarra delako. 18. eta 19. irudian, adibide gisa eman diren bi galderen SPARQL galderak azaltzen dira:

```
Galdera: Noiz sortu zen Aconcagua mendia?
SPARQL galdera:
SELECT ?e
WHERE
{
?subjektuNagusia ?propietateNagusia ?e
FILTER regex(str(?subjektuNagusia),"Aconcagua")
FILTER regex(str(?propietateNagusia),"sortu")
}
```

18. irudia: Galdera sinple batean subjektu nagusia

```
Galdera: Zein da Frantziako hiriburuko alkatea?
SPARQL galdera:
SELECT ?erantzuna
WHERE
{
?lotura ?propietateNagusia ?erantzuna
OPTIONAL { ?subjektuNagusia ?tartekoPropietatea ?lotura}
FILTER regex(str(?subjektuNagusia),"Frantzia")
FILTER regex(str(?tartekoPropietatea),"hiriburu","i")
FILTER regex(str(?propietateNagusia),"alkate","i")
}
```

19. irudia: Galdera konplexu batean subjektu nagusia

Bigarren adibidean, sparql galderan triple bat baino gehiago sortu ditugu, baina triple hauetako bakarrean dakigu aldez aurretik subjektuaren balioa zein den.

-**Propietate nagusia**: galderan aipatzen diren propietate guztien artetik, erantzunaren balioari erreferentziatzen dion propietatea da. *Noiz sortu zen Aconcagua mendia?* galderan "*sortu*" da bilatzen ari garen *propietate nagusia* eta *Zein da Frantziako hiriburuko alkatea?* galderan berriz "*alkate*" da bilatzen ari garen *propietate nagusia*.

-**Tarteko propietateak:** galderan aipatzen diren propietate guztien artetik, *propietate nagusia* ez den gainontzeko propietate guztiak dira. *Noiz sortu zen Aconcagua mendia?* galderan ez dago *tarteko propietaterik* eta *Zein da Frantziako hiriburuko alkatea?* galderan berriz, "*hiriburu*" da bilatzen ari garen *propietate nagusia*.

4.3.2 Landutako galdera-patroien zehaztapena

Hizkuntza naturalean idatzitako galderak SPARQL lengoaiara itzultzeko, goian azaldu diren osagaiak, hala nola *subjektu nagusia*, *propietate nagusia* eta *tarteko propietateak*, identifikatu behar ditugu lehendabizi. Osagai hauen identifikaziorako, galdera-patroi batzuetan oinarritzen gara. Galdera-patroi hauek 4.2 atalean aipatutako galdera-moten egituran oinarritzen dira. Aurrerago, 5.1 atalean, garapenerako eta ebaluaziorako sortutako galderen xehetasunak hobeto azalduko dira, hala ere, landu ditugun galderetako galdetzaileak *Non, Noiz, Zein, Nork* eta *Nor* izan dira eta entitateak galderan azaldu behar du. Galdetzaile hauek erabilita, euskarazko galderek, normalean behintzat, hiru patroi nagusi hauek jarraitzen dutela da gure hipotesi nagusia:

1. GALDETZAILEA [Propietate nagusia] ADL [Subjektu nagusia] [Tarteko propietateak]?

   GALDERATZAILEA: [Non|Noiz|Zein|Nork |Nor]

   ADL: aditz laguntzailea

Non [jaio]$_{PN}$ zen [Montpellierreko]$_{SN}$ [alkatea]$_{TP}$?
Noiz [jaio]$_{PN}$ zen [Bizkaiko]$_{SN}$ [hiriburuko]$_{TP}$ [alkatea]$_{TP}$?
Noiz [sortu]$_{PN}$ zen [Barbara Goenaga]$_{SN}$ [jaio]$_{TP}$ zen[hiria]$_{TP}$?
Nork [agintzen] $_{PN}$ du [Alessandro Voltaren] SN [jaiotza herrialdean]$_{TP}$?
Zein [lanbide]$_{PN}$ zuen [Orson Wellesek]$_{SN}$?
Zein [alderditako]$_{PN}$ da [Xabier Arzalluz]$_{SN}$?

20. irudia: 1. Patroia jarraitzen duten galderen adibideak

Patroi hau jarraitzen duten galderetan (ikusi 20.irudia), galdetzailea eta aditz laguntzailearen artean *propietate nagusia* aurkituko da, aditz laguntzailearen ondoan *subjektu nagusia* eta azkenik, *subjektu nagusiaren* ondoren, *tarteko propietateak* aurkituko dira.

2. GALDETZAILEA ADL [Subjektu nagusia] [Tarteko propietateak] [Propietate nagusia]?

Zein da [Peruko]$_{SN}$ [hiriburuaren]$_{TP}$ [probintzia]$_{PN}$?
Nor da [Potemkin korazatuaren]$_{SN}$ [gidoilaria]$_{PN}$?
Zein da [Urretxuko]$_{SN}$ [alkatea]$_{PN}$?
Zein da [Aimar Olaizolaren]$_{SN}$ [jaioterriko]$_{TP}$ [alkatea]$_{PN}$?
Zein da [Haile Selassieren]$_{SN}$ [heriotza herrialdeko]$_{TP}$ [hiriburua]$_{PN}$?
Non dago [Gipuzkoa] $_{SN}$ [kokatua]$_{PN}$?

21. irudia: 2. Patroia jarraitzen duten galderen adibideak

Patroi hau jarraitzen duten galderetan (ikusi 21.irudia), aditz laguntzailearen ondoan *subjektu nagusia* kokatuko da. *Propietate nagusia* berriz esaldiaren bukaerara pasatzen da eta azkenik, *subjektu nagusiaren* eta *propietate nagusiaren* artean *tarteko propietateak* aurkituko dira.

3. GALDETZAILEA ADL [Propietate nagusia] [Subjektu nagusia] [Tarteko propietateak]?

Gutxiagotan ematen bada ere, beste egitura hau ere jarraitu dezakete galderek.

Non zuen [egoitza] $_{PN}$ [Watson doktoreak]$_{SN}$?
Non dago [kokatua] $_{PN}$ [Aconcagua]$_{SN}$?
Nork du [agintzen] $_{PN}$ [Bizkaiko]$_{SN}$ [hiriburuan]$_{TP}$?

22. irudia: 3. Patroia jarraitzen duten galderen adibideak

Egitura hau jarraitzen duten galderetan (ikusi 22. irudia), lehendabizi entitatea identifikatu beharko da eta hau bera izango da *subjektu nagusia*. *Subjektu nagusiaren* eta aditz laguntzailearen artean aurkitzen dena *propietate nagusia* izango da eta azkenik, *subjektu nagusiaren* ondoren dagoena *tarteko propietateak* izan daitezke.

Laburbilduz, *subjektu nagusia* galderako esaldian aurkitzen den entitatea izango da beti eta normalean aditz laguntzailearen ondoan kokatzen da. Hala ere, hirugarren patroian ikusi dugun moduan, baliteke aditz laguntzailea eta *subjektu nagusiaren* artean *propietate nagusia* tartekatzea. Hau gutxitan gertatuko den arren, garrantzitsua izango da galderako entitatea modu egokian identifikatzea, bestela *propietate nagusia subjektu nagusitzat* hartu baitaiteke. Izan ere, galderan entitatea ezin izan bada identifikatu, aditz laguntzailearen ondoan dagoena hartuko dugu *subjektu nagusitzat*.

Propietate nagusiaren kokapenak hiru izan daitezke: galdetzailea eta aditz laguntzailearen artean, esaldiaren bukaeran ala aditz laguntzailea eta *subjektu nagusia* edota entitatearen artean. Galdetzailea eta aditz laguntzailearen artean zerbait dagoenean *propietate nagusia* bertan aurkituko da. Aldiz, galdetzailea eta aditz laguntzailearen artean ez badago ezer eta galderako entitatea eta galdetzailearen artean zerbait badago, hemen egongo da *propietate nagusia*. Azkenik, bi kasu hauek ez badira ematen, esaldiaren bukaerara joko dugu *propietate nagusiaren* bila.

Tarteko propietateak entitatea edota *subjektu nagusiaren* jarraian aurkituko dira beti, eta azkenengo honekin gertatzen den bezalaxe, beti aditz laguntzailearen ondoren kokatuko dira.

Propietate nagusia eta *tarteko propietateak* ez dira beraien artean inoiz nahastuko, hau da, entitatean ondoren propietate bat baino gehiago baldin badaude, azkena beti

propietate nagusia izango da eta tartean gelditzen diren propietateak, berriz *tarteko propietateak.*

11. taulan ikus daiteke, galdera berdina azaldutako hiru egituratan sor daitekeela.

1. patroia	2. patroia	3. patroia
Non [jaio]$_{PN}$ zen [Montpelliereko]$_{SN}$ [alkatea]$_{TP}$?	Zein da [Montpelliereko]$_{SN}$ [alkatearen]$_{TP}$ [jaiolekua]$_{PN}$?	Non zen [jaio]$_{PN}$ [Montpelliereko]$_{SN}$ [alkatea]$_{TP}$?
Nork [agitzen]$_{PN}$ du [Aimar Olaizolaren]$_{SN}$ [jaioterrian]$_{TP}$?	Zein da [Aimar Olaizolaren]$_{SN}$ [jaioterriko]$_{TP}$ [alkatea]$_{PN}$?	Nor da [alkatea]$_{PN}$ [Aimar Olaizolaren]$_{SN}$ [jaioterrian]$_{TP}$?
Noiz [jaio]$_{PN}$ zen [Bizkaiko]$_{SN}$ [hiriburuko]$_{TP}$ [alkatea]$_{TP}$?	Zein da [Bizkaiko]$_{SN}$ [hiriburuko]$_{TP}$ [alkatearen]$_{TP}$ [jaiotze data]$_{PN}$?	Noiz zen [jaio]$_{PN}$ [Bizkaiko]$_{SN}$ [hiriburuko]$_{TP}$ [alkatea]$_{TP}$?

11. taula: galdera berdina 3 patroiak erabilita

Noiz edo *Non* galdetzailea eta lehendabizi azaldu den galdera-patroia erabilita sortutako galdera bigarren patroia erabilita adierazi nahi baldin badugu, normalean *Zein* galdetzailearekin sortuko dugu.

Galderek hiru patroi hauetakoren bat jarraituko dutela da gure hipotesi nagusia, eta honen arabera hizkuntza naturalean idatzitako galderan SPARQL galdera sortzeko beharrezko osagaiak identifikatzen saiatuko gara. Honez gain, osagai hauen identifikazioa ziurragoa izan dadin, RDFtik erauzitako lexikoa erabiliko dugu. Honela, *Non dago kokatua Aconcagua?* bezalako galderetan, RDFtik erauzitako entitate-zerrendari esker, galderako entitatea identifikatu ahal izango dugu, "*kokatua*" *subjektu nagusitzat* hartzea ekidingo dugu.

Sistemaren garapenerako sortutako galderak erabilita, goian azaldutako galdera-patroi desberdinak kontuan hartzen dituzten heuristikoak erabiltzeaz gain, hizkuntza naturalean idatzitako galderan SPARQL galdera sortzeko osagai bakoitzaren (*subjektu nagusia, propietate nagusia eta tarteko propietateak)* identifikazioan laguntzeko, erabilitako RDF ezagutza-basetik erauzitako lexikoa eta hiztegiak nola erabili ditugun azalduko dugu jarraian.

4.3.3 Subjektu nagusiaren identifikazio prozesua

Galderako osagaien identifikazioan lehenengo burutuko den urratsa da *subjektu nagusiaren* identifikazioa. Hau, aurretik aipatu den bezala, galderako entitatea izango

da, beraz galderaren analisia eta RDFtik erauzitako lexikoa erabiliko ditugu eginkizun honetan.

Lehendabizi galdera-analizatzaileak itzulitako analisian, hitz anitzeko entitaterik markatu den begiratzen dugu. Adibidez *"Zein talde entrenatu ditu Alexander Chapman Fergusonek?"* galderarentzat galdera-analizatzaileak itzulitako analisian, 23. irudian azaltzen den moduan markatzen da galderako entitatea.

```
<BILA_TERMI FRM="Alexander_Chapman_Fergusonek">
        <LEX LEM="Alexander_Chapman_Ferguson" KAT="IZE" KASUA="ERG" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="ENTI_PER" />
        <BILA_TERMI_OSAG FRM="Alexander">
                <LEX LEM="Alexander" KAT="IZE" KASUA="" KM="@KM>" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="ENTI_HAS_PER" />
        </BILA_TERMI_OSAG>
        <BILA_TERMI_OSAG FRM="Chapman">
                <LEX LEM="Chapman" KAT="IZE" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="" />
        </BILA_TERMI_OSAG>
        <BILA_TERMI_OSAG FRM="Fergusonek">
                <LEX LEM="Ferguson" KAT="IZE" KASUA="ERG" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="ENTI_BUK_PER" />
        </BILA_TERMI_OSAG>
</BILA_TERMI>
```

23. irudia: Entitatearen analisiaren irteera kodea

Galderaren analisian markatutako hitz anitzeko entitaterik baldin badago, hau hartuko dugu SPARQL galderan *subjektu nagusia* izango den osagai gisa (goiko adibidean "Alexander Chapman Ferguson"). *Subjektu nagusia* identifikatua izango dugun arren, gure RDF ezagutza-basearen lexikora egokitu beharra daukagu SPARQL galdera egokia eraikitzeko, beraz, RDFtik erauzitako hiztegiak erabiliko ditugu (4.1.1 atalean aipatu diren berbiderapenetatik eta propietate jakin batzuetatik abiatuta erauzi ditugun hiztegiak). Adibide honetan, hiztegi hauek erabilita, *"Alexander Chapman Ferguson"* entitatea *"Alex Ferguson"* moduan itzuliko dugu, hain justu RDFan azaltzen den formara.

Galderaren analisian hitz anitzeko entitaterik markatu ez baldin bada, RDFtik erauzitako entitate edota subjektuen zerrenda (RDFko subjektu guztiak, berbiderapenak eta propietate jakin batzuetatik erauzitako subjektuen zerrenda) erabiliko dugu

galderako hitz bakarreko ala hitz anitzeko entitatea identifikatzeko. Kasu honetan ere, hiztegiak erabiliko ditugu entitatearen forma gure RDFan erabiltzen den lexikoarekin egokitzeko.

Analisian identifikatutako hitz bakarreko entitateak ez ditugu zuzenean SPARQL galderan *subjektu nagusia* izango den osagai gisa hartuko. Modu honetan 24. irudian azaltzen den adibidean "*Potemki*"-en ordez "Potemkin korazatua" identifikatu ahal izango dugu SPARQL galderan *subjektu nagusia* izango den osagai gisa.

```
<?xml version="1.0" encoding="UTF-8" ?>
<GALDERA_ANALISIA fasea="bilaketa terminoen erauzketa">
<JATORRIZKOGALDERA>Zein urtetakoa da Potemkin korazatua filma? </JATORRIZKOGALDERA>
        <GALDERA>Zein urtetakoa da Potemkin korazatua filma?</GALDERA>
        <TERMI FRM="Zein">
                <LEX LEM="zein" KAT="DET_NOLGAL" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="" />
        </TERMI>
        <BILA_TERMI FRM="urtetakoa">
                <LEX LEM="urte" KAT="IZE" KASUA="GEL_ABS" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="" />
        </BILA_TERMI>
        <TERMI FRM="da">
                <LEX LEM="izan" KAT="ADT" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="" />
        </TERMI>
        <BILA_TERMI FRM="Potemkin">
                <LEX LEM="Potemki" KAT="IZE" KASUA="INE" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="ENTI_LOC" />
        </BILA_TERMI>
        <BILA_TERMI FRM="korazatua">
                <LEX LEM="korazatu" KAT="ADI" KASUA="ABS" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="" />
        </BILA_TERMI>
        <BILA_TERMI FRM="filma">
                <LEX LEM="film" KAT="IZE" KASUA="ABS" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="" />
        </BILA_TERMI>
        <TERMI FRM="?">
                <LEX LEM="" KAT="" KASUA="" KM="" LEH="jatorrizkoa" IDF="kontsultatu_gabe" ENTI="" />
        </TERMI>
```

```
        <ESPEROERANTZUNA>TEMPORAL</ESPEROERANTZUNA
<ESPEROERANTZUNAERREGELAK>TEMPORAL</ESPEROERANTZUNAERREGELAK>
        <ESPEROERANTZUNAIKASKETA></ESPEROERANTZUNAIKASKETA>
        <GALDERAMOTA>FACTOID</GALDERAMOTA>
        <MINTZAGAIA>Potemki korazatu film</MINTZAGAIA>
        <GALDEGAIA>urte</GALDEGAIA>
</GALDERA_ANALISIA>
```

24. irudia: galdera analisi baten irteeraren adibidea

Azkenik, aipatutako moduan entitatea bilatu ez baldin bada, galderako galdetzailearen aditz laguntzailearen ondoko hitza hartuko dugu SPARQL galderako *subjektu nagusi* bezala (*Zein hizkuntza hitz egiten dute Ruandan?* galderatik "*Ruanda*" hartuko genuke).

4.3.4 Propietate nagusiaren identifikazio prozesua

Subjektu nagusia identifikatu ondoren, galderako *propietate nagusiaren*, hau da, SPARQL galderako azkeneko triplean kokatuko dugun propietatearen identifikazio urratsa burutuko dugu. Kasu honetan, goian azaldutako hiru galdera-patroiek eragina izango dute eta galdera-patroiaren arabera *propietate nagusia* bilatzeko galderako gunea aldatuko da. Jarraian, galdera-patroi bakoitzean *propietate nagusia* identifikatzeko emango ditugun pausuak azalduko dira.

1. **Lehenengo galdera patroia**

```
GALDETZAILEA [Propietate nagusia] ADL [Subjektu nagusia] [Tarteko propietateak]?
```

Galdetzailea eta honen aditz laguntzailearen artean hitzak baldin badaude (adibidez, *Noiz sortu zen Aconcagua mendia?* ala *Zein euskalki hitz egiten da Gipuzkoan?*), SPARQL galderako *propietate nagusia* bertan aurkitzen dela suposatuko dugu.

Galdetzailearen eta honen aditz laguntzailearen artean *hitz bakarra* baldin badago *(Noiz sortu zen Aconcagua mendia?),* hau izango da *propietate nagusi* izateko hautagai bakarra, beraz, zuzenean *propietate nagusi* moduan identifikatuko dugu. Normalean, hitz hau aurretik identifikatu dugun entitatearen entitate-motaren propietateen artean (entitate-mota horrentzat erabiltzen den infotaula txantiloilean aurkitzen diren propietateak) baldin badago, ziurrenik galdera *sinplea* izango da (Adibidez, *Zein*

eskualdetan dago Hernani kokatua? galderan *Hernani* "*Gipuzkoako udalerri*" motako entitatea da eta *eskualdea* propietatea dauka), hau da, triple bakarreko SPARQLa eraiki beharko dugu.

Galdetzailearen eta honen aditz laguntzailearen artean *hitz bat baino gehiago* baldin badaude *(Zein katetan emititzen da Barbara Goenagak lan egin zuen telesaila?* ala *Zein talde entrenatzen du Alex Fergusonek?), propietate nagusia* izateko hautagai bat baino gehiago edukiko dugu, beraz, ondorengo pausuak jarraituko ditugu:

Lehendabizi, propietateen-zerrenda erabilita hitz anitzeko propietatea den begiratuko dugu, honela, *Zein talde entrenatzen du Alex Fergusonek*? galderan adibidez, *"entrenatutakoTaldeak"* propietatea identifikatuko dugu *propietate nagusi* gisa, izan ere *"entrenatutakoTaldeak"* propietateak *"talde" eta "entrenatu"* lemak barneratzen baititu.

Hitz anitzeko propietaterik aurkitzen ez baldin bada, hautagaietakoren bat aurretik identifikatu dugun entitatearen entitate-motaren propietateen artean aurkitzen den begiratuko dugu, entitate mota horrentzat erabiltzen den infotaula txantiloilean. Hau normalean galdera *sinpleetan* gertatuko da, honela, *Zein hizkuntza mintzatzen dira Asturiasen?* galderan adibidez, *"hizkuntza"* hautatuko dugu *propietate nagusi* moduan. Azkenik, hautagaietako bat ere ez ezin izan bada entitatearen propietateen artean aurkitu, lehenengo hitzarekin saiatuko gara *propietate nagusi* moduan jartzen. Hau normalean galdera *konplexuetan* gertatuko da, *"Zein katetan emititzen da Barbara Goenagak lan egin zuen telesaila?"* galderan adibidez, ez *"kate"* eta ez *"emititu"* ez dira "Barbara Goenaga" entitatearen propietateen artean aurkituko, beraz, kasu honetan "kate" hartuko dugu *propietate nagusi* gisa.

2. Bigarren galdera patroia

GALDETZAILEA ADL [Subjektu nagusia] [Tarteko propietateak] **[Propietate nagusia]**?

Galdetzailea eta honen aditz laguntzailearen artean eta aditz laguntzailearen artean ezer ez dagoenean, *propietate nagusia* galdera bukaeran egongo dela uste dugu (*Zein da Asturiasko Printzerriko hiriburua? / Zein da Suediako hiriburuko webgunea?*)

Hala ere, kasu honetan ere hitz anitzeko propietateak egon daitezke eta hauek aurkitzeko, RDFko propietateen-zerrenda erabiliko dugu. Honela, *"Zein da Alex Fergusonen izen osoa?"* galderan adibidez, *"izenOsoa"* hitz anitzeko propietatea

aurkituko da eta *"Zein da Alessandro Voltaren jaio zen herrialdeko hiriburua?" galderan "jaiotzaHerrialdea"*.

3. Hirugarren galdera patroian

GALDETZAILEA ADL **[Propietate nagusia]** [Subjektu nagusia] [Tarteko propietateak]?

Batzuetan, *propietate nagusia* aditz laguntzailearen eta entitatearen artean kokatzen da (*Non du sorrera Hunte ibaiak?/ nork du agintzen Asturiasen?*) eta hau hartuko dugu *propietate nagusi* moduan, galdetzailea eta honen aditz laguntzailearen artean ez dagoenean ezer.

Kasu honetan, hitz bakarreko propietateak soilik hartu ditugu kontutan, beraz, etorkizuneko lanen artean, hitz anitzeko propietateak identifikatzea gehituko dugu.

Hala ere, uste dugu, forma naturalena *propietate nagusia* galdera bukaeran egotea dela, ala bestela, galdetzailea eta aditz laguntzailearen artean.

4.3.5 Tarteko propietateen identifikazio prozesua

Subjektu nagusia eta *propietate nagusia* ez den beste guztiaren artean *tarteko propietateak* aurki daitezke. Batzuetan ez dira hitz gehiago egongo, beraz, kasu honetan galdera *sinplea* izango da eta triple bakarreko SPARQL galdera sortu beharko dugu.

Tarteko propietateak beti entitatearen jarraian aurkituko dira, beraz, lehenengo eta hirugarren galdera-egituretan, *propietate nagusia* entitatearen jarraian ez dagoenez, kasu hauetan entitatearen ondoren dauden hitzak *tarteko propietate* gisa identifikatu ditzakegu ("*Noiz jaio zen Bilboko alkatea?*"). Bigarrengo galdera-patroia jarraitzen dutenetan aldiz, identifikatzen diren propietate guztien artetik azkenengoa izan ezik beste guztiak izango dira *tarteko propietateak* (*Zein da Hondarribi dagoen eskualdeko biztanleria?*)

Tarteko propietateen artean hitz anitzeko propietateak identifikatzen ere saiatuko gara entitatearen propietateen zerrendan begiratuta: *Nork agintzen du Alessandro Voltaren jaiotza herrialdean?* galderan adibidez, *"Alessandro Volta"* entitatea "zientzialarien biografia" motako entitatea da eta honen propietateen artean "*jaiotzaHerrialdea*" aurkitzen da.

Kasu batzuetan *subjektu nagusia* eta *propietate nagusia* ez den beste guztia *tarteko propietateak* izango dira, baina beste kasu batzuetan hautagai hauen artean, SPARQL galdera egokia sortzeko baztertu beharko ditugun hitzak ere aurkituko dira (*Zein da Bild aldizkariaren zuzendariaren jaioteguna? / zein da Hernani herriko alkatea?*).

Horrela, hasiera batean baztertuko ditugun hitzak ondorengoak izango dira:

-Entitatearen ondoko hitza entitate-mota baldin bada, baztertu egingo dugu *Zein da Yongzhou hiriaren webgunea? g*alderatik "hiri" hitza baztertuko dugu "Yongzhou" entitatea "Txinako hiri" motako entitatea baita eta *Noiz sortu zen Barbara Goenaga aktorea jaio zen hiria?* Galderatik *"aktorea"* hitza baztertuko dugu, "Barbara Goenaga" "aktore biografia" motako entitatea baita.

-Entitatea eta ondoko hitzaren tripletik sortutako SPARQL galderatik jasotako erantzunaren entitate-mota baldin bada ondorengo hitza, hau ere baztertu egingo dugu. *Nor da Barbara Goenagak lan egin zuen telesailko protagonista? g*alderan, *"Barbara Goenaga"* entitatearen *"lan"-en* artean *"Goenkale"* entitatea aurkitzen da, eta, hau *"telesail"* motakoa da, beraz, hitz hau baztertu ahal izango dugu SPARQL galderaren sorkuntzarako.

4.3.6 Erantzunaren berreskurapena

Subjektu nagusia, *propietate nagusia* eta *tarteko propietateak* identifikatu ondoren, SPARQL galdera osatuko dugu eta RDFtik erantzun zuzena lortzen saiatuko gara ARQ RDFa kontsultatzeko tresna erabilita.

Non eta *Noiz* galdetzailea duten galderen kasuan galdera-analizatzaileak itzulitako <ESPEROERANTZUNA> eremuaren balioa erabiliko dugu, "LOCATION" eta "TEMPORAL" hurrenez hurren, erantzun posible guztien artetik egokia filtratzeko. "LOCATION" motako erantzuna espero den galderen kasuan, sortutako SPARQL galderarentzat erantzun bat baino gehiago lortu baldin badira, *propietate nagusian* "herrialde", "hiri" edo "eskualde" bezalako hitza gehituta leku bati erreferentzia egiten dion erantzuna lortzen saiatuko gara. Horrela, "*Non jaio zen Alex Ferguson?*" galderarentzako "*jaio*" *propietate nagusiarekin* "*jaiotzaHerrialde*" propietatearen balioa itzultzeaz gain, "*jaiotzaData*" propietatearen balioa ere itzultzen da (ikusi 25. irudia), aldiz, *propietate nagusian* "*herrialde*", "*hiri*" edo "*eskualde*" duten propietateak soilik hartzeko jarritako filtroarekin emaitza zuzena bakarrik itzultzen da (ikusi 26. irudia).

```
SELECT ?e
WHERE
{
?subjektuNagusia ?propietateNagusia ?e
FILTER regex(str(?subjektuNagusia),"alex_ferguson","i")
FILTER regex(str(?propietateNagusia),"jaio","i")
}
```

25. irudia: SPARQL galdera

```
SELECT ?e
WHERE
{
?subjektuNagusia ?propietateNagusia ?e
FILTER regex(str(?subjektuNagusia),"alex_ferguson","i")
FILTER regex(str(?propietateNagusia),"jaio","i")
FILTER regex(str(?propietateNagusia),"herrialde|hiri|eskualde","i")
}
```

26. irudia: SPARQL galdera esperotako erantzuna aplikatuta

Aldiz, “TEMPORAL” motako erantzuna espero den galderen kasuan, sortutako SPARQL galderarentzat erantzun bat baino gehiago lortu baldin badira, *propietate nagusian* “data” hitza gehituta data bati erreferentzia egiten dion erantzuna lortzen saiatuko gara. Horrela, “*Noiz jaio zen Alex Ferguson?*” galderarentzako “*jaio*” *propietate nagusiarekin* “*jaiotzaData*” propietatearen balioa itzultzeaz gain, “*jaiotzaHerrialde*” propietatearen balioa ere itzultzen da (ikusi 25. irudia), aldiz, *propietate nagusian* “*data*” duten propietateak soilik hartzeko jarritako filtroarekin (ikusi 27. irudia) emaitza zuzena bakarrik itzultzen da.

```
SELECT ?e
WHERE
{
?subjektuNagusia ?propietateNagusia ?e
FILTER regex(str(?subjektuNagusia),"alex_ferguson","i")
FILTER regex(str(?propietateNagusia),"jaio","i")
FILTER regex(str(?propietateNagusia),"data","i")
}
```

27. irudia: SPARQL galdera esperotako erantzuna aplikatuta

Aurretik aipatu den bezala, identifikatutako osagaien lexikoaren egokitzapenerako ez dugu hiztegirik erabili lan honetan, estaldura hobetuko liratekeen arren. Hala, ere identifikatutako osagaiekin sortutako SPARQL galderarekin erantzunik aurkitzen ez baldin badugu, entitateen propietateen artetik identifikatutako propietatearen antzekoenarekin ordezkatuta SPARQL berri bat sortu eta erantzuna lortzen saiatzen gara. Adibidez, *Nork sortu zuen Watson doktorea pertsonaia?* galderarentzako hasiera batean ondorengo SPARQL galdera sortutako da (ikusi 28. irudia)

```
SELECT ?erantzuna
WHERE
{
?subjektuNagusia ?propietateNagusia ?erantzuna
FILTER regex(str(?subjektuNagusia),"Watson_doktore","i")
FILTER regex(str(?propietateNagusia),"sortu","i")
}
```

28. irudia: SPARQL galdera lexikoa egokitu gabe

SPARQL galdera honek ez du erantzunik aurkituko, "*Watson"* entitateak *"sortu"* propietaterik ez daukalako RDFan. Beraz, "Watson doktorea" entitate-motaren (pertsonaia) propietateen artetik, berdin hasten diren (lehenengo 3 hizkiak berdinak izatea) propietateen artetik antzekoenarekin ordezkatzen saiatuko gara (ikusi 29.irudia):

```
SELECT ?erantzuna
WHERE
{
?subjektuNagusia ?propietateNagusia ?erantzuna

FILTER regex(str(?subjektuNagusia),"Watson_doktore","i")
FILTER regex(str(?propietateNagusia),"sortzaile","i")
}
```

29. irudia: SPARQL galdera lexikoa egokituta

Azkenik, aurretik aipatutako guztiarekin emaitzarik jasotzen ez baldin bada, identifikatutako osagaiak modu desberdinean konbinatuz SPARQL galdera desberdinak sortzen saiatuko gara. Identifikatutako osagaien artean *propietate nagusia* soilik aurkitzen baldin bada, hau SPARQL galderatik kendu eta jasotako erantzunak esperotako erantzunarekin filtratzen saiatzen gara, adibidez *Non dago kokatua Comoko probintzia?* galderan "*Comoko probintzia*" entitateak ez dauka "*kokatu*" ez antzeko propietaterik "*kokatu*" *propietate nagusiarekin* ez dugu erantzunik jasotzen. Aldiz, *propietate nagusia* kendu eta esperotako erantzunarekin filtratuz (ikusi 30.irudia), "*eskualdea*" propietatearen balioa jasoko dugu, *Lonbardia* erantzun zuzena alegia.

```
SELECT ?erantzuna
WHERE
{
?subjektuNagusia ?propietateNagusia ?erantzuna
FILTER regex(str(subjektuNagusia),"Comoko_probintzia","i")
FILTERregex(str(?propietateNagusia),"herrialde|hiri|
eskualde","i")
}
```

30. irudia: SPARQL galdera propietate nagusia kenduta

Osagaien artean *propietate nagusia* eta *tarteko propietate* bakarra dauzkagunean, bietako bakarrarekin sortutako SPARQL galderarekin soilik probatzen ere saiatzen gara (*tarteko propietatea* ezabatu ala *tarteko propietatea propietate nagusi* moduan jartzen saiatuta). Eta, *propietate nagusia* eta *tarteko propietate* hautagai asko daudenean,

propietate nagusiarekin soilik sortutako SPARQL galderarekin erantzuna lortzen ere saiatuko gara.

5 Sistemaren ebaluazioa

Atal honetan, garatu dugun RDFaren gaineko galdera-erantzun sistemaren ebaluazioa nola burutu den azaltzen da. Alde batetik, garatutako sistemaren emaitzak hobetzeko gehitzen joan garen baliabide eta estrategia desberdinek ekarri dituzten hobekuntzak neurtu ditugu. Beste alde batetik, euskararentzako dagoen beste galdera-erantzun sistema batekin ere prozesatu ditugu galderak eta ebaluazioan erabili ditugun galdera hauek erantzuteko bi sistemek duten gaitasuna konparatu dugu.

Bigarren galdera-erantzun sistema hau IHARDETSI da (Ansa et al., 2008), euskararentzako testu-hutsaren gaineko galderak erantzuteko sistema. Honek, gureak ez bezala, testuan bilatzen ditu galderarentzako erantzunak, beraz, euskarazko wikipediako artikuluetan bilatu ditu ebaluaziorako erabilitako galdera-zerrendarako erantzunak, artikuluetako infotauletako informazioan bilatu ordez.

Galdera batentzat erantzuna testu hutsaren gainean bilatzea ala egituratutako ezagutza-base baten gainean bilatzea bi hurbilpen desberdin jarraitzen dituzten sistemak dira. Gainera, zailtasun maila desberdineko atazak dira. Aipatutako bi motatako sistemak konparatzeko, Wikipediako artikuluen egiturak aukera paregabea ematen digu. Izan ere, Wikipediako artikulu askotan bertan lantzen den entitateari buruzko testua (IHARDETSIk hemen burutuko du bilaketa) eta entitateari buruzko datu esanguratsuekin osatutako infotaula (gure sistemak hemen burutuko du bilaketa) izaten dituzte.

Hurrengo ataletan, lehendabizi garapen eta test faseetarako erabili ditugun galderak sortzeko prozesua azalduko da, eta, ondoren, sortutako galdera hauen ezaugarriak azalduko ditugu. Ebaluazioan erabilitako metrikak ere azalduko dira, eta azkenik, garapen eta test fasean burututako esperimentuak azalduko dira.

5.1 Garapen eta testerako galderak sortzeko gida-lerroa

Gure sistema garatu eta ondoren ebaluatu ahal izateko, galderak sortu behar izan ditugu. Galdera hauen sorkuntzarako gida-lerro bat idatzi dugu eta sistemaren garapenean parte hartu ez duen pertsona bati eskatu diogu galderak sortzeko.

Garapenerako eta azken ebaluaziorako galderak sortzeko pauso berdinak jarraitu dira eta pertsona berdinak sortu ditu galdera guztiak. Galderak sortu dituen pertsona honi, euskal Wikipedian lantzen diren 29 entitateren artikuluak eta infotaulak pasa dizkiogu garapenerako galderak sortzeko eta beste 18 entitateren artikulu eta infotaulak azken ebaluaziorako galderak sortzeko. Garapenerako entitateen artetik 14 entitate biografikoak edota pertsonen izenak dira eta beste 15 geografikoak edota leku izenak. Ebaluaziorako entitateen artean aldiz, 8 biografikoak eta 10 geografikoak dira.

Entitate hauetako bakoitzerako, dagokion artikulu eta infotaulatik abiatuta behean aipatzen diren murriztapenak betetzen dituzten galderak sortzeko eskatu zaio galdera sortzaileari:

- Galderak esaldi *sinpleak* izan behar dira, hau da, sintaxia aldetik konplexutasun txikia izan behar dute.
- Galdetzaile hauek erabili ahal izango dira: *Non, Noiz, Nork, Nor* eta *Zein*.
- Galderan entitateak azaldu behar du. Hala ere, entitateak ez du zertan artikuluko tituluan azaltzen den forma zehatzean azaldu behar. Adibidez:
    - Pertsonen izenetan izena ala abizena ez aipatzea posible izango da: *Non jaio zen Johannes Kepler?* galdetu beharrean *Non jaio zen Kepler?*
    - Entitatea beste modu batera ezagutzen baldin bada, galderan forma horrekin erabiltzea posible izango da*: Non dago Penintsula Iberikoa? ala Non dago Iberiar penintsula? galdetu ahal izango da.*

Galdera hauen erantzuna galderako entitatearen artikuluan bertan edota entitate horrentzako RDFko tripleetan aurkitu ahal izango da. Hau da, ez da entitate bat baino gehiago erlazionatu beharko erantzuna bilatu ahal izateko.

Aipatutako murriztapenez gain, galdera hauen sorkuntza prozesuan bete beharreko beste zenbait ezaugarri definitu ditugu:

- **Ordena:** galdera batzuk artikulutik sortuko dira eta beste batzuk aldiz, infotaulako informazioa erabilita. Lehendabizi, artikulutik sortuko ditu galderak. Hau da, infotaulako informazioa ikusi gabe, artikulua irakurri eta goian deskribatutako murriztapenak betetzen dituzten galderak sortuko ditu. Ondoren, infotaula ikusi eta bertan aurkitzen den informazioari buruzko galderak sortuko

ditu. Testutik galderak sortzean, infotaulan azaltzen den informazioak eraginik izan ez dezan jarraituko du ordena hau.

- **Proportzioa:** galderak sortuko dituen pertsona entitate bakoitzarentzat 4-5 galdera sortzen saiatuko da. Infotauletatik sortutako galderak gehiago izan beharko dira. Galderen %70a, gutxi gorabehera, infotauletatik sortutakoak izango dira eta %30a testuetatik sortutakoak (entitate bakoitzarentzat 3 infotauletatik sortutakoak eta 1-2 testuetatik sortutakoak). Lan honetan, RDFaren gaineako galdera-erantzun sistema garatu eta ebaluatzea da helburu nagusia, beraz RDFa erabilita erantzun ahal diren galdera kopurua handia izatea nahi dugu. Bestalde, Wikipediako artikuluen %60-70ak infotaula dute, beraz, galderen sorkuntzan ere proportzio hau jarraitzea erabaki dugu.

- **Roll-a:** galderak sortzeko prozesua errazteko, galderak sortuko duen pertsonari roll bat ematea erabaki genuen. Galdera sortzaileak bere burua lehen hezkuntzako irakasle baten paperean imajinatuko du. Ikasleek testuak irakurri ondoren testu horiei buruzko galderak proposatu nahi dizkie irakasle honek. Sortutako galderak irakasle honek sortuko lituzkeenak izango dira.

- **Formatua**: sortutako galdera bakoitzarentzat, galdera zein entitateri dagokion, nondik sortua izan den (testutik ala infotaulatik) eta erantzuna non aurki daitekeen (testuan, infotaulan ala bietan) markatuko da. Galdera bakoitzeko honelako erro bat sortu da:

Entitatea[Tabuladorea]Galdera[Tabuladorea]Nondik[Tabuladorea]Testuan[Tabuladorea]Infotaulan

-*Entitatea:* galdera zein entitateri buruzkoa den.

-*Galdera:* sortu den galdera.

-*Nondik:* galdera testutik ala infotaulatik sortua izan den. Jaso ditzakeen balioak bi dira: *testua* (galdera testutik sortua izan baldin bada) ala *infotaula* (galdera infotaulatik sortua izan baldin bada).

-*Testuan:* galderarentzako erantzuna testuan aurki daitekeen ala ez. Jaso ditzakeen balioak bi dira: 0 (erantzuna testuan ez bada aurkitzen) eta 1 (erantzuna testuan aurkitzen bada).

-*Infotaulan:* galderarentzako erantzuna infotaulan aurki daitekeen ala ez. Jaso ditzakeen balioak bi dira: 0 (erantzuna infotaulan ez bada aurkitzen) eta 1 (erantzuna infotaulan aurkitzen bada).

Adibidez, "*Johannes Kepler Non jaio zen Kepler? Infotaula 1 1*" galdera, Johannes Kepler entitatearentzat sortua izan da infotaulako informazioa erabilita eta erantzuna testuan eta infotaulan aurki daiteke.

Guztira, 120 galdera sortu dira sistemaren garapen faserako eta beste 73 galdera sistemaren azken ebaluaziorako. Garapenerako galderak erabilita, sistemari egingo dizkiogun hobekuntzak ebaluatuko ditugu eta IHARDETSIk lortzen dituen emaitzen hasierako azterketa bat egingo dugu. Ondoren, ebaluaziotako galderekin bi sistemek lortzen dituzten emaitzak konparatuko ditugu.

Galdera hauek sinpleak diren arren, gero ikusiko den moduan ez da hain prozesu sinplea izango dagokion SPARQL galdera egokia sortzea, beraz egokia ikusten dugu galdera hauekin hastea.

Honez gain, galdera konplexuagoekin, hau da entitate bat baino gehiago erlazionatzen dituzten galderekin ere hasierako ebaluaketa txiki bat ere burutu nahi izan dugu. Galdera hauen erantzunak lortzeko, RDFko triple bat baino gehiago erlazionatzen dituzten SPARQL galderak sortu beharko ditugu. Aldiz, testuan galdera hauentzako erantzuna aurkitzea ez da erraza izango, izan ere entitate desberdinen artikuluetako informazioa konbinatzea eskatzen baitituzte.

Galdera hauek sistemaren garatzaileok sortu ditugu 14 entitateren infotauletatik abiatuta. Sortu ditugun 25 galdera hauek guztietarako RDFan erantzuna dagoela ziurtatu dugu. Gehienez bi entitate edo RDFko subjektu erlazionatzen dituzten galderak sortu ditugu, beraz, bi triple dituzten SPARQL lengoaian idatzitako galderetara itzuli beharko ditugu galdera hauek.

Esan bezala, entitateen infotauletan edota RDFan dituzten tripleetako informazioan oinarritu gara galderak sortzeko, hau da, hautatutako entitate horren RDFko tripleen artetik, tripleko objektu moduan RDFko beste entitate baten subjektua duten tripleak aukeratu ditugu eta informazio horretatik abiatuta galdera sortu dugu. Adibidez, "*Peru*" entitateak "*hiriburua*" propietatearekin beste entitate bat, "*Lima*", lotzen du RDFan. Aldi berean, "*Lima*" entitatearen propietateen artean "*Lima"* noiz sortu zen adierazten

duen "sorrera" propietatea dago, beraz, "*Noiz sortu zen Peruko hiriburua?*" galdera sortu ahal izango dugu.

5.2 Esperimentazio ingurunea

Memoriako atal honetan, esperimentuetan erabili ditugun galderen ezaugarriak azalduko dira.

5.2.1 Entitate bakar baten gaineko galderak

- **Garapen faserako galderak**

Sistemaren garapenean parte hartu ez duen pertsona batek 5.1 puntuan azaldu den galderen sorkuntzarako gida-lerroa jarraituz, euskarazko Wikipedian lantzen diren 29 entitateren artikulu eta infotauletatik abiatuta sintaxia aldetik konplexutasun txikia duten 120 galdera sortu ditu.

12. taulan galdetzaile-mota bakoitzarentzat testutik eta RDFko informaziotik abiatuta sortutako galdera kopurua azaltzen da.

	NON	NOIZ	NOR	NORK	ZEIN	GUZTIRA
Testutik	7	13	0	0	23	43
RDFtik	13	10	3	2	49	77

12. taula: galdetzaile bakoitzeko garapenerako galdera kopurua

13. taulan sortutako galderen erantzunak non aurkitzen diren neurtu da. Gerta daiteke, galdera batentzat bai Wikipediako artikuluko testuan zein infotaulan aurkitzea erantzuna, baina beste galdera batzuek testuan ala infotaulan soilik izango dute erantzuna.

RDFtik abiatuta sortu diren 77 galderen artetik 49entzat testuan ere aurkitzen da erantzuna (%63'6) eta testutik abiatuta sortu diren 43 galderetatik 22 erantzun daitezke RDFa erabilita (%51'2).

	NON	NOIZ	NOR	NORK	ZEIN	GUZTIRA
Testuan	17	22	1	1	55	92
RDFan	16	15	3	2	63	99

13. taula: testuan eta RDFan erantzuna duten garapeneko galderen kopuruak

- **Test faserako galderak**

Garapenerako galderen kasuan bezalaxe, sistemaren garapenean parte hartu ez duen pertsona batek sortu ditu ebaluazioan erabili diren galderak (pertsona berdinak sortu ditu sistemaren garapen faserako galderak eta ebaluaziorako galderak) eta galderen sorkuntzarako gida-lerro berdinak jarraitu ditu. Euskarazko wikipedian lantzen diren 18 entitateren artikulu eta infotauletatik sintaxia aldetik konplexutasun txikia duten 73 galdera sortu ditu.

14. taulan galdetzaile-mota bakoitzarentzat testutik eta RDFko informaziotik abiatuta sortutako galdera kopurua azaltzen da. Ikus daitekeenez *"Zein"* galdetzailea erabilita sortu dira galdera gehienak, bai testuko informaziotik abiatuta eta baita RDFko informaziotik abiatutako galderen artean. Bestalde, *"nor"* eta *"nork"* galdetzailearekin ez dira ia galderarik sortu. Galderak sortzeko murriztapenen artean, galderan bertan entitateak azaldu behar duela izanik, "*nork*" galdetzailea asko mugatzen da, batez ere entitate biografikoekin (adibidez, "*Nork irabazi du nobel sari bat?*" bezalako galderak ez ditugu landu, galdera honetan, ez baita RDFan lantzen den entitaterik azaltzen). Bestalde, *"nor"* galdetzailearen lekuan, *"zein"* galdetzailea erabili da (adibidez, "*Nor da Toulonjac herriko alkatea?* galdera sortu beharrean *Zein da Toulonjac herriko alkatea?*" galdera sortu da), beraz, hau izan da galdetzaile honekin sortutako galderen kopurua hain txikia izateko arrazoia.

	NON	**NOIZ**	**NOR**	**NORK**	**ZEIN**	**GUZTIRA**
Testutik	4	6	0	0	18	28
RDFtik	5	6	3	1	30	45

14. taula: galdetzaile bakoitzeko testerako galdera kopurua

15. taulan sortutako galderen erantzunak non aurkitzen diren neurtu da. Gerta daiteke, galdera batentzat bai Wikipediako artikuluko testuan zein infotaulan aurkitzea erantzuna, baina beste galdera batzuek testuan ala infotaulan soilik izango dute erantzuna. Zehatzago esanda, testutik abiatuta sortu diren galdera batzuentzat RDFan ez da erantzuna aurkituko eta aldiz, RDFtik abiatuta sortu diren galdera batzuen erantzuna ez da testuan aurkituko. Azken kasu hau gutxiago gertatzen da, testuan infotauletan baino informazio gehiago ematen delako.

RDFtik abiatuta sortu diren 45 galderen artetik 30entzat testuan ere aurkitzen da erantzuna (%66'6) eta testutik abiatuta sortu diren 28 galderetatik 11 erantzun daitezke RDFa erabilita (%39'2).

	NON	NOIZ	NOR	NORK	ZEIN	GUZTIRA
Testuan	9	12	0	0	37	58
RDFan	8	9	3	1	35	56

15. taula: testuan eta RDFan erantzuna duten testeko galderen kopuruak

Beste modu batera esanda, badakigu bai testu gaineko galderak erantzuteko sistemak eta baita RDFaren gaineko galdera-erantzun sistemak sortutako galderen 15-17 galdera hurrenez hurren ezingo dutela erantzun (%20'5 eta %23'3). Hau honela izanda, Wikipediaren gaineko galdera-erantzun sistema osoa sortzeko, bi sistemak konbinatu beharko genituzke.

Galdera hauek, esaldi mailan konplexutasun txikikoa izateaz gain hauen erantzuna galderan lantzen den entitatean artikuluan bertan aurkitzen da. Hau da, entitate zehatz baten propietate baten balioa bilatzea eskatzen dute, beraz, gure sistemaren kasuan, triple bakarreko SPARQL galderak sortu beharko dira, 4.2 atalean azaldu den moduan.

5.2.2 Entitate bat baino gehiago erlazionatzen dituzten galderak

Goian aipatutako galderak sortzeaz gain, bai sintaktikoki konplexuagoak diren eta baita erantzuna bilatzeko entitate desberdinen arteko erlazioa identifikatzea eskatzen dituzten galderak ere sortu ditugu test faserako.

Entitate bat baino gehiago erlazionatzen dituzten galderak sortzeko 14 entitate aukeratu ditugu ausaz, 6 geografikoak eta 8 biografikoak. Entitate hauetatik abiatuta 25 galdera sortu ditugu. Galdera hauek guztientzat RDFan erantzuna aurkitzen da. Testuan aldiz, galdera hauen artetik 15ek dute erantzuna. Sortutako galderen kopurua handia ez den arren, hasierako proba bat egiteko balioko digute.

16. taulan, galdetzaile bakoitzarentzat zenbat galdera sortu diren ikus daiteke. Galdera *sinpleen* kasuan bezala, *zein* galdetzailearentzat sortu dira galdera gehienak.

ZEIN	NON	NOIZ	NORK	NON
17	1	4	1	2

16. taula: galdera konplexuentzat mota bakoitzeko zenbat

Alde batetik gure sistemak mota hauetako galderak nola erantzuten dituen aztertuko dugu eta beste alde batetik, bai testuan eta baita RDFan erantzuna duten galderentzat bi sistemek lortutako emaitzak aztertuko ditugu. Aldez aurretik badakigu testuan galdera hauen erantzuna bilatzea oso zaila izango dela. Hala ere, hasierako ebaluazio txiki bat egitea egokia iruditu zaigu.

5.3 Erabilitako metrikak

Galdera-erantzun sistemen ebaluaziorako neurri desberdinak erabiltzen dira. Lan honetan, QA@CLEF lehiaketako edizio desberdinetan erabili diren ebaluazio metriketan oinarritu gara. Lehiaketako urtearen arabera neurri desberdinak aukeratu dituzte aurkeztutako sistemak sailkatzeko, hala ere, normalean neurri bat baino gehiago erabiltzen dira sistemen errendimenduari buruzko informazio gehiago lortzeko. Hauek dira ebaluazioan erabiliko ditugun neurriak:

- **Mean Reciprocal Rank (MRR):** neurri hau galdera-erantzun sistemak galdera batentzat erantzun bat baino gehiago itzultzen duenean aplikatzen da. Erantzun hauek egokitasunaren (*confidence* ingelesez) arabera ordenatuta egoten dira. MRR neurriaren arabera, erantzunen rankingean erantzun zuzenak duen posizioaren arabera galdera horrek puntuazio bat jasotzen du. Adibidez, galdera bakoitzarentzat 3 emaitza posible itzultzen baldin badira, galdera horrek 1, 0.5, 0.333 ala 0 (3 emaitzetako bat bera ere ez denean zuzena) balioa jaso ahal izango du erantzun zuzenak rankingean duen posizioaren arabera. Azken ebaluazio-emaitza, galdera guztiek lortutako puntuazioaren batezbestekoa izango da. MRR neurria informazioaren berreskurapenean erabiltzen den MAP (Mean Average Precision) neurriarekin erlazionatuta dago. Guk garatutako sistemaren kasuan, galderek jasoko duten balioa 1 ala 0 izango da. Emaitza bat baino gehiago itzultzen diren kasuetan, ez dira egokitasun neurri baten arabera ordenatzen, beraz, erantzun zuzenak ausaz hartu duen posizioaren arabera

kalkulatu beharrean, galdera horrentzako balioa, erantzun zuzena lehenengo posizioan itzuli denean 1 balioa izango da eta bestela 0. Ihardetsik normalean emaitza bat baino gehiago itzultzen ditu, gainera, egokitasunaren arabera ordenatuta. Gehienez lehen 3 emaitzak hartuko ditugu kontuan, eta, erantzun zuzenaren posizioaren araberako balioa jasoko du galderak. Neurri hau IHARDETSIren emaitzak itzultzeko moduari begira erabili da gehienbat.

- **Zehaztasuna (asmatze-tasa, accuracy ingelesez)** : CLEF QAn gehien erabiltzen den neurria da. Honek modu egokian erantzun diren galderen proportzioa neurtzen du, erantzun diren galderetatik zenbatek jaso duten erantzun zuzena neurtuz. Erantzun bat baino gehiago itzuli diren kasuetan, lehendabiziko erantzuna soilik hartzen da kontutan. Gure sistemaren kasuan, erantzuna jaso ez duten galderak, gaizki erantzundako galderatzat hartu ditugu.

- **C@1**: 2009an CLEF QAn erabili zen neurri nagusia da. Neurri hau erabili ahal izateko, galdera guztiek gutxienez erantzun zuzen bat izan behar dute helburu bilduman (eta kasu honetan RDFan). Sistemek galdera erantzun gabe usteko aukera dute lortutako erantzunetan erantzun zuzena aurkitzen ez dela uste bada. Hain zuzen ere, c@1 neurriak zuzen erantzundako galderen kopurua mantentzeko gaitasuna saritzen du, honetarako galderak erantzun gabe usteko aukera emanez, gaizki erantzundako galderen kopurua txikituz. Galdera-erantzun sistema batzuetan (adibidez, diagnostiko mediko bat) galdera erantzun gabe uztea hobe da erantzun oker bat ematea baino; hau da neurri honen oinarrizko arrazoia. Hau sistema erreal bat erabiltzaileek onartzeko funtsezkoa izango da. 31. irudian ikus daiteke c@1 neurriaren formulazioa:

$$C@1=1/n(n_R + n_U \frac{n_R}{n})$$

non,

n_R: zuzen erantzundako galderen kopurua

n_U: erantzun ez diren galderen kopurua

n: galdera kopurua guztira

31. irudia: c@1 neurriaren formulazioa

5.4 Ebaluazioa

Atal honetan garapen fasean egindako esperimentuak azalduko dira lehendabizi. Ondoren azken ebaluazioan edota test fasean lortutako emaitzak erakutsiko dira. Azkenik, erantzunak testuan bilatzen dituen IHARDETSI galdera-erantzun sistemak lortutako emaitzak komentatuko dira.

5.4.1 Sistemaren garapen fasea

Aurretik aipatu den bezala, garapen fasean entitate bakarraren gaineko 120 galdera erabili dira guztira. RDFan 99 galderek dute erantzuna eta testuan aldiz, 92 galderek. Aldiz, bai RDFan, bai testuan erantzuna duten galderen kopurua 72 da. Galdera guztien artetik, 43 (%36) galdera testutik sortuak izan dira eta 77 (%64) RDFtik.

Baseline moduan erabili dugun hurbilpenak galderaren galdera-patroia bakarrik identifikatzen du eta ondoren, honen arabera galderako osagaien identifikazioa burutzen du. Galderako entitatearen identifikazioa ez denez burutzen, 2. eta 3. galdera patroiaren artean ezingo dugu bereizi, beraz, hurbilpen honek osagaiak modu honetan identifikatuko ditu: *propietate nagusitzat* galdetzailea eta aditz laguntzailearen artean aurkitzen den lehenengo hitza ala galderako azkeneko hitza hartuko da eta *subjektu nagusitzat,* aldiz, aditz laguntzailearen ondoko lehendabiziko hitza. Gainontzekoak *tarteko propietateak* izango dira. Adibidez, *Zein hizkuntza mintzatzen dira Asturiasen?* galderan *hizkuntza* hartuko da *propietate nagusi* moduan eta *Asturias subjektu nagusi* moduan.

Hasierako hurbilen honekin, ikusi nahi dugu, ea ahalik eta baliabide gutxien erabilita, hau da, galderako hitzen posizioak soilik kontutan hartuta, zenbat galdera zuzen erantzuten diren. Ondoren, SPARQL galdera zuzenen sorkuntza hobetzeko eta ondorioz lortutako erantzun zuzenen kopurua handitzeko, hainbat baliabide eta estrategia aplikatu ditugu.

17. taulan, aplikatutako hobekuntza bakoitzarekin lortutako emaitzak azaltzen dira. Jarraian erabili ditugun gure sistemaren hurbilpen desberdinak azaltzen dira:

-RDF-QA1: galdera-patroiak bakarrik aplikatuta.

-RDF-QA2: RDF-QA1 + Entitateen identifikazioa. Hau da, galdera-patroiak erabiltzeaz gain, RDFtik erauzitako subjektu edota entitateen zerrenda erabili dugu galderako entitatea identifikatzeko. Hitz bakarreko eta hitz anitzeko entitateak

identifikatzen dira. Honela, *Noiz hil zen Anna Magnani?* galderan adibidez, *subjektu nagusi* moduan *Anna Magnani* hitz anitzeko entitatea identifikatu ahal izango dugu.

-**RDF-QA3**: RDF-QA2 +Entitateen Hiztegiak. Entitateen zerrenda erabiltzeaz gain, Wikipediako artikuluen berbiderepenetatik ateratako entitateen sinonimoen hiztegia eta RDFko propietate jakin batzuk (*ezizena, izenOsoa, izenOfizia*l...) dituzten tripleetatik ateratako entitateen hiztegia erabili dugu galderako subjektua identifikatzeko. Honela, *Non jaio zen Fergie?* galderan adibidez, *subjektu nagusi* moduan identifikatzen den *Fergie*, ezizena dela identifikatu ahal izango dugu eta *Alex Ferguson* moduan itzulita *subjektu nagusiaren* forma zuzena lortu ahal izango dugu.

-**RDF-QA4**: RDF-QA3 + Hitz anitzeko propietateen identifikazioa. Entitateez gain propietateak modu egokian identifikatzen saiatzen gara. Honetarako RDFko propietate guztien zerrenda erabiltzen dugu galderan egon daitezkeen "hitz anitzeko propietateak" bilatzeko. Honela, *Zein talde entrenatzen du Alex Fergusonek?*galderan *Alex Ferguson* hitz anitzeko entitatea subjektu moduan identifikatzeaz gain, galdetzailea eta honen aditz laguntzailearen artean dagoen guztia hitz anitzeko propietate bat dela identifikatu ahal izango dugu, *entrenatutakoTalderak* propietatea hain zuzen.

-**RDF-QA5**: RDF-QA4 + esperotako erantzuna aplikatuta. Aurretik aipatutakoaz gain, galderaren analisian lortutako esperotako erantzunaren informazioa erabiltzen dugu, non eta noiz galdetzailea duten galderen kasuan, erantzun egokia aukeratzeko.

-**RDF-QA6**: RDF-QA5 + baztertzeko hitzak+ SPARQL konbinaketa desberdinak. Aurreko guztiaz gain, galderako hitz batzuk baztertzen ditugu (entitate-entitate mota segida ematen denean, entitate-mota ezabatzen dugu). Azkenik, hau guztiarekin erantzunik jasotzen ez baldin bada, SPARQL galdera desberdinak sortzen dira hautatutako osagaiekin.

-Ihardetsi: garapenerako galdera hauek IHARDETSI sistemarekin ere probatu ditugu, ematen dituen emaitzen hasierako azterketa bat egin ahal izateko.

Hurbilpen bakoitzerako, zenbat erantzun zuzen aurkitu diren kalkulatzeaz gain, sarrerako galdera SPARQL galderara itzultzeko prozesuan noraino iritsi garen aztertu dugu baita ere. Hau da, lehenengo pausua ondo burutu den, SPARQL galdera egokia sortu ahal izateko osagaiak, *subjektu nagusia* eta propietateak, ondo identifikatu diren, eta bigarren pausua, osagai bakoitza RDFko lexikora egokitzea ondo burutu den aztertu dugu.

Gure sistemaren kasuan, erantzun zuzena itzuli dela erabakiko da, galderak erantzun bakarra eta zuzena itzuli duenean. IHARDETSIren kasuan berriz, itzulitako lehenengo erantzuna zuzena denean. Gure sistemaren kasuan, erantzunik jaso ez dituzten galderak, erantzun gabeko galdera gisa hartu ditugu. Aldiz, IHARDETSIren kasuan, galderak beti jasotzen dute erantzuna, beraz, ez da erantzun gabeko galderarik egongo. Honek, c@1 balioa kalkulatzerakoan izango du eragina.

Alde batetik, c@1 balioa eta beste alde batetik, lortutako zehaztasuna kalkulatu dira. c@1_99 gure sistemarentzat RDFan erantzuna duten galderak kontutan hartuta kalkulatu da. c@1_72 bai testuan eta bai RDFan erantzuna duten galderak kontutan hartuta kalkulatu da eta azkenik, c@1_92 IHARDETSI sistemarentzat kalkulatu da, testuan erantzuna duten galderak kontutan hartuta.

Sistema	Erantz. zuzen kop.	Erantz. oker kop.	Zehaztasuna _120	Zehaztasuna _72	c@1_99	c@1_72	c@1_92	Osagai Egokiak	Lexiko egokia
RDF-QA1	14	106	0,12	0,11	0,26	0,20	-	31	14
RDF-QA2	30	90	0,25	0,24	0,45	0,36	-	67	27
RDF-QA3	30	90	0,25	0,24	0,45	0,36	-	67	29
RDF-QA4	40	80	0,33	0,39	0,58	0,55	-	75	42
RDF-QAQ5	49	71	0,41	0,47	0,70	0,68	-	76	52
RDF-QA_Q6	65	56	0,54	0,61	0,84	0,81	-	96	67
Ihardetsi	17	103	0,14	0,24	-	0,24	0,18	-	-

17. taula: garapenerako galderekin lortutako emaitzak

Galdera patroiak soilik aplikatuta, 14 galdera erantzuten dira modu egokian eta 31 galderarentzat SPARQL galdera sortzeko osagaiak ondo aukeratzen dira. Galdera patroiak identifikatu eta honen arabera hitzek duten posizioaren arabera galderako osagaiak identifikatzeak esfortzu txikia eskatzen du eta, hala ere, ia testutik modu egokian erantzundako galdera kopuruaren berdina erantzun dira modu egokian.

Entitateen identifikazioa burututa eta entitateentzat hiztegi desberdinak erabilita, heuristikoak soilik erabilita baino galderen bikoitza erantzuten da modu egokian. Entitateen hiztegiak erabilita emaitzak hobetzen ez diren arren, zenbait entitate modu egokian identifikatzea lortzen da (*Zein talde entrenatu ditu Alexander Chapman Fergusonek?* galderan "*Alexander Chapman Ferguson"* RDFn azaltzen den Alex Ferguson formara itzuli da), eta honi esker hurrengo hurbilpenetan aplikatzen diren hobekuntzekin batera hainbat galderarentzat SPARQL egokia sortzea lortzen da eta ondorioz emaitza egokia aurkitzea. Gainera, entitateen sinonimoen hiztegiak erabiltzea beharrezkoa izango da sistema erreal batean, hau da, erabiltzaileei entitate bat modu desberdinetan adierazteko aukera emango die eta.

Propietateak modu egokian identifikatzeko hitz anitzeko propietateak identifikatzen saiatuta ere asko hobetzen dira emaitzak.

Esperotako erantzunarekin erantzun posibleak filtratzea ere garrantzitsua da, emaitzak nabarmen hobetzen direlako.

Azkenik, gure sistemaren azken emaitzetan ikusi daiteke, nola ia galdera guztientzat SPARQL galderako osagaiak ondo aukeratu diren, baina, lexikoaren egokitzapen faltagatik galdera hauek guztiak ez dira modu egokian erantzun. Adibidez, *Zein izan ziren Enerst Fourneauren ekarpen zientifikoak?* galderan *"ekarpen zientifikoak"* ezin izan da *"lanak"* propietatearekin parekatu edota *Noiz hil zen Anna Magnani?* galderan *"hil"* ezin izan da *"heriotzaData"* propietatearekin parekatu.

Aipatu behar da, gure sistemak erantzuna ematen duen ia kasu guztietan erantzun zuzena dela, erantzuna jaso duten galderen artetik 5 bakarrik erantzun dira gaizki. Hau dela eta, c@1 neurrian balio altua jasotzen du gure hurbilpenak.

Ihardetsik aldiz, galdera guztietarako itzultzen ditu erantzun posibleak, hauek zuzenak izan ala ez. IHARDETSIk lortu dituen emaitzak ikusita, garapenean erabili ditugun mota honetako galderen erantzuna bilatzeko, datu-egituratuen gainean erantzunak bilatzen dituen hurbilpena lagungarria edota osagarria izan daitekeela pentsa genezake erantzunak testu-hutsean bilatzen dituzten sistementzat.

5.4.2 Sistemaren test fasea

Ondorengo ebaluazioan, testerako sortutako galderak gure sistemak eta IHARDETSIk nola erantzun dituen aztertu ditugu. Alde batetik, gure sistemaren

garapenean parte hartu ez duen pertsonak sortutako 73 galderen emaitzak konparatu ditugu eta beste alde batetik, guk sortutako beste 25 galdera konplexuagoekin burutu dugu azken ebaluazioa.

IHARDETSI galdera-erantzun sistemarekin galdera hauentzako erantzunak lortu ahal izateko, lehendabizi, guk erabili dugun RDFan azaltzen diren entitate edota subjektu guztien artikuluak, edo beste modu batera esanda, infotaula duten euskal wikipediako artikulu guztiak aukeratu eta indexatu ditugu. IHARDETSI Wikipediako artikuluen gainean saiatuko da erantzunak aurkitzen, infotaulei dagokien testuen gainean, alegia, eta, bestetik, gure sistemak RDFan bilatuko ditu erantzunak. IHARDETSIk galdera batentzat erantzun zuzena bilatzeko, galderan hitz-gakoak hautatu ondoren, hitz-gako hauek dituzten paragrafoak bilatzen ditu eta ondoren paragrafo hauetatik erantzun zehatza eta zuzena erauzten saiatzen da. Bildumako hitzak sinonimoekin hedatzeko aukera ere erabiltzen du, baina kasu honetan, hau egiteak eskatzen duen denbora dela eta, IHARDETSIk sinonimoekin hedatu gabeko Wikipediako artikuluetan bilatuko ditu erantzunak.

Aipatu behar da, ebaluazio honetarako, gure sistemari goian aipatutako hobekuntza guztiak aplikatu dizkiogula, hau da RDF-QA_Q6 hurbilpena erabiliko dugu.

- **Galdera sinpleen ebaluazioa**

4.2 puntuan azaldu diren entitate baten gaineko galderak erabili dira. Aipatu den bezala, galdera guztietarako ez dago testuan erantzuna, ezta RDFan ere. Beraz, bi tauletan banatu ditugu emaitzak: 18. taulan, bi lekuetan, testuan eta RDFan, erantzuna duten galderentzat atera dira emaitzak eta 19. taulan berriz, galdera guztiak kontutan hartuta atera ditugu emaitzak. Guztira 42 galdera dira bai testuan eta baita RDFan erantzuna duten galderak.

Sistema	Erantzun zuzen kop.	Erantzun oker kop.	c@1	MRR	Zehaztasuna
RDF-QA	28	14	0,76	0,66	0,66
IHARDETSI	8	38	0,19	0,21	0,19

18. taula: testuan eta RDFan erantzuna duten galderentzat emaitzak

IHARDETSIk lehenengoan 8 erantzun zuzen soilik aurkitu ditu (%19). Hauetatik 4 testutik abiatuta sortuak izan dira eta beste 4 infotauletatik. IHADETSIren kasuan c@1 neurriak jasotzen duen balioa zehaztasunaren berdina da, galdera guztiak erantzuten

dituelako beti. Lehendabiziko 3 posizioetako erantzunak kontutan hartuta, IHARDETSIk beste bi galdera erantzuteko gai da. Arrazoi honegatik, jasotzen duen MRR balioa zehaztasuna baino zerbait handiagoa da.

Gure sistema 28 galdera modu egokian erantzuteko gai izan da (%66).

19. taulan, galdera guztiak kontutan hartuta, 73 galdera, lortu diren emaitzak azaltzen dira:

Sistema	Erantzun zuzen kop.	Erantzun oker kop.	MRR	Zehaztasuna
RDF-QA	38	35	0,52	0,52
IHARDETSI	8	65	0,12	0,11

19. taula: testeko galdera guztientzat lortutako emaitzak

IHARDETSI sistemak Wikipediako testuan erantzuna duten galderak soilik kontutan hartuta, 58 galdera, lortu dituen emaitzak 20. taulan azaltzen dira.

Sistema	Erantzun zuzen kop.	Erantzun oker kop.	c@1	MRR	Zehaztasuna
IHARDETSI	8	58	0,13	0,15	0,13

20. taula: IHARDETSIrentzan erantzuna testuan duten galderentzat emaitzak

RDFan erantzuna duten galderak soilik kontutan hartuta, 56 galdera, gure hurbilpenak lortu dituen emaitzak 21. taulan azaltzen dira.

Sistema	Erantzun zuzen kop.	Erantzun oker kop.	c@1	MRR	Zehaztasuna
RDF-QA	38	18	0,81	0,67	0,67

21. taula: RDF-QArentzat erantzuna infotaulan duten galderentzat emaitzak

22. taulan, gure sistemarekin ala IHARDETSI sistemarekin bakarrik modu egokian erantzun diren galderen kopuruak azaltzen dira

Erantzun zuzena Ihardetsik bakarrik	Erantzun zuzena RDF-QAk bakarrik	Biek erantzun zuzena
4	33	4

22. taula: RDF-QAak ala IHARDETSIk bakarrik zuzen erantzundako galdera kopuruak

Gure sistemarekin erantzun zuzena 38 galderentzat aurkitu den arren, beste 12 galderarentzat modu egokian hautatu dira SPARQL egokia sortzeko osagaiak. Hala ere,

lexikoa gure RDFan erabiltzen denarekin bat ez datorrenez, ezin izan da galdera hauentzat emaitza egokia aurkitu (ikusi 23. taula).

Osagai egokiak	Lexiko egokia	Emaitza zuzena
50	32	38

23. taula: osagai edota lexiko egokiarekin sortu diren SPARQL galdera kopuruak

Erantzunak RDFan bilatzen dituen hurbilpenarekin garapenerako eta testerako erabilitako galderekin emaitza antzerakoak lortu dira, garapen fasean 0,54ko zehaztasuna edota asmatze-tasa lortu du eta test-ean berriz 0,52koa. Erantzunak testuan bilatzen dituen hurbilpenarekin ere antzekoa gertatu da, garapen fasean 0,14ko zehaztasuna edota asmatze-tasa lortu du lortu du eta test-ean berriz 0,11koa.

RDFarekin erantzun daitezkeen galderak 56 dira eta galdera hauetatik 50 galderentzat SPARQL galdera sortzeko osagaiak (*subjetu-nagusia* eta *propietate nagusia*) modu egokian identifikatu dira hizkuntza naturalean idatzita dagoen sarrerako galderan. Hala ere, emaitza zuzena 38 galderentzat lortu da, izan ere, osagai hauen identifikazioa zuzena izan arren hauen lexikoa, galderan erabili dena, gure RDFaren lexikora egokitzea falta delako. Beraz, etorkizunean emaitzak hobetu nahi baldin badira, ezinbestekoa izango da lexikoaren egokitzapen fasea lantzea. Adibidez "*Zein ogibide du Jeff Albertsonek*?" galderan, "*Jeff Albertson" subjektu nagusi* moduan eta "*ogibide" propietate nagusi* moduan identifikatzen dira, baina RDFan "*Jeff Albertson"* entitateak "*ogibide"* propietatearen ordez "*lan"* propietatea dauka.

Bestalde, aipatu behar da, kasu gutxi batzuetan, sortu den SPARQL galderaren osagaien lexikoa RDFan erabiltzen denarekin guztiz bat etortzen ez den arren, erantzun zuzena itzuli dela. Adibidez, "*Noiz igo zen Aconcagua lehen aldiz?"* galderan, "*Aconcagua" subjektu nagusi* moduan eta "*igo" propietate nagusi* moduan jarrita, erantzun zuzena aurkitu da, RDFan, "*Aconcagua"* entitateak "*igo"* duen propietate bakarra, "*lehenIgoera",* duelako.

- **Galdera konplexuen ebaluazioa**

24. eta 25. tauletan galdera konplexuekin lortutako emaitzak azaltzen dira. Erabilitako galderen kopurua handia ez den arren, gure hurbilpena mota honetako galderekin probatzeko eta hasierako ondorio batzuk ateratzeko balio izan digute. 4.2 atalean azaldu den bezala, bi entitate erlazionatzen dituzten galderak dira, eta galdera

hauen erantzuna bilatzeko artikulu bat baino gehiago konbinatu behar dira. Beraz, testutik erantzuna ondorioztatzea lan zaila izan daiteke. Aldiz, RDFaren egitura dela eta, modu errazagoan lortu ahal izango dugu erantzuna, honetarako, SPARQL konplexuagoak sortu beharko ditugun arren (2 triple dituzten SPARQL galderak). Guztira 25 galdera erabili dira ebaluazio honetan (ikusi 24. taula), baina galdera hauetatik 15entzat ziurtatu ahal izan dugu Wikipediako artikuluetako testuan erantzuna aurkitzen dela, beraz, IHARDETSI sistema 15 galdera hauekin soilik ebaluatu dugu (ikusi 25. taula).

#Osagai egokiak	#Lexiko egokia	#Emaitza zuzena	c@1	Zehaztasuna
23	17	17	0,81	0,68

24. taula: galdera konplexuentzat emaitzak

Sistema	Erantzun zuzen kop.	Erantzun oker kop.	MRR	Zehaztasuna	c@1
RDF-QA	11	4	0,73	0,73	0,83
IHARDETSI	0	15	0	0	0

25. taula: erantzuna testuan eta RDFan duten galdera konplexuentzat emaitzak

Erantzunak RDFan bilatzen dituen hurbilpenak zehaztasun altua lortu du erantzuna lortzeko bi entitate erlazionatu behar diren galderekin. Aipatu den bezala, ebaluazio honetarako galderak guk sortu ditugu, beraz, honek sortu ditzakeen desbiderapenak ekiditeko, etorkizunean mota hauetako galderen ebaluazioa, beste pertsona batek sortutako galderekin errepikatzea egokia litzatekeela uste dugu. Hala ere, artikuluetako testuetatik mota hauetako galderak erantzutea zailtasun handiko ataza izanik, galdera-patroi jakin batzuk jarraitzen dituzten galdera konplexu hauek RDFa erabilita modu errazagoan erantzun ahal direla ikusi da.

Kasu honetan ere, gaizki erantzun diren galdera gehienen kasuan, SPARQL galderako osagaiak RDFaren lexikora egokitu ez direlako gertatu da.

5.4.3 Emaitzen gaineko hausnarketa

Gure sistemak lortu dituen emaitzen eta IHARDETSIk lortu dituen emaitzen arteko aldea dela eta, emaitzen gaineko hausnarketa hau egitera bultzatu gaitu.

IHARDETSIk esperotako emaitzak baino txarragoak lortu ditu. Honen arrazoi nagusietako bat izan daiteke IHARDETSI sistemak ez duela garapen faserik izan gure

esperimentuen arabera. Kontuan hartu behar dugu, lan honetan, patroi jakin batzuk jarraitzen dituzten galderak erabili ditugula ebaluazioan eta gure sistema galdera-patroi horiek jarraitzen dituzten galderak erantzuteko diseinatu eta garatu dugula.

Gainera, IHARDETSIk aurretik egindako beste esperimentu eta lanetan helburu bildumako hitzak sinonimo eta erlazionatutako hitzekin hedatuta emaitza hobeak lortu ditu (Agirre et. al, 2009) (Agirre et. al, 2010), baina kasu honetan ez dugu hurbilpen hau erabiltzeko aukerarik izan. Izan ere, bildumaren hedapen semantiko hau burutzeak denbora asko eskatzen du eta ez gara iritsi hau egitera.

Memoriako 5.1 atalean azaldu den moduan, alderen %70a, gutxi gorabehera, infotauletatik sortuak izan dira eta %30a testuetatik sortutakoak. Lan honetan, RDFaren gaineako galdera-erantzun sistema garatu eta ebaluatzea da helburu nagusia, beraz RDFa erabilita erantzun ahal diren galdera kopurua handia izatea nahi izan dugu. Hala ere, ohartzen gara honek gure sistemari nolabaiteko abantaila ematen diola.

Azkenik, badakigu, IHARDETSIrentzat bildumaren indexazioan errore batzuk gertatu direla, eta honen ondorioz, Wikipediako artikulu batzuk kanpoan gelditu zirela. Garapen eta azken ebaluazioan erabilitako galderen erantzuna duten artikulu gehienak indexatu direla ziurtatu ahal izan den arren, ohartzen gara IHARDETSIrekin egindako esperimentu hauek berriz errepikatu beharko direla bilduma osoa modu egokian indexatuta, baina honek eskatzen duen denbora kontuan hartuta (hile bat pasa genuen Wikipediako artikulu guztiak indexatzen), ez dugu lan honetan azaltzen diren emaitzak berrikusteko aukerarik izan.

Etorkizunean sakonago aztertu beharko dira IHARDETSIk lortu dituen emaitzen arrazoiak zeintzuk izan diren.

6 Ondorioak eta etorkizuneko lanak

Lan honetan euskarazko Wikipediako infotauletatik erauzitako RDF ezagutza-basearen gaineko galdera-erantzun sistema bat aurkeztu da. Honek hizkuntza naturalean idatzitako galderak RDFak kontsultatzeko lengoaiara, hau da, SPARQL lengoaiara itzultzen ditu lehendabizi, eta, ondoren, RDFak kontsultatzeko JENAren ARQ[10] sistema erabilita emaitza zuzena lortzen saiatzen da.

Hizkuntza naturalean idatzitako galderak SPARQL lengoaiara itzultzeko prozesuan jarri dugu arreta. Honetarako erabili dugun hurbilpenak baliabide eta teknologia linguistiko gutxi erabiltzen ditu, eta, hala ere, doitasunari eta zehaztasunari begira nahiko emaitza onak lortu ditu hasierako ebaluazio honetan.

Erabili diren baliabide linguistikoen artean, zein hizkuntza naturala prozesatzeko baliabideen artean, alde batetik, galderaren analizatzailea dago eta bestetik, Wikipedia eta RDFtik bertatik erauzitako entitateen sinonimo hiztegiak.

Lan honetan aztertutako galdera-motak mugatu ditugu. Galderetan leku edota pertsona entitateen propietateei buruz galdetzen da. Honez gain, entitatea beti azaltzen da galderan eta galderen esaldiek konplexutasun maila jakin bat dute. Gainera, galdetzaile jakin batzuk erabilita sortutako galderak dira (*noiz, non, nor, nork eta zein*). Landu ditugun galdera-mota hauentzako, hizkuntza naturalean idatzitako galderan SPARQL galderako osagaien identifikazioan oso emaitza onak lortu dira. Bestalde, erantzun zuzena galderen %52an itzuli da.

Garapen fasean gure sistemaren hurbilpen desberdinentzat emaitzak aztertu ditugu eta gehitu dizkiogun hobekuntzak ebaluatu ditugu. Garapen fasean ikusi dugu, SPARQL galderako osagaien identifikazioan laguntzeko gure sistemari gehitu dizkiogun hobekuntza guztiek, bai entitateen identifikaziorako gehitutako sinonimoen hiztegiek baita hitz anitzeko propietateen identifikazioak, erantzun zuzena lortzen laguntzen dutela.

Gure sistema eta erantzunak testuan bilatzen dituen galdera-erantzun sistema baten arteko emaitzak konparatu ditugu. Azken sistema hau IHARDETSI galdera-erantzun sistema izan da. Hemen proposatzen dugun sistemak testu hutsean bilatzen duen

10http://jena.apache.org/documentation/query/index.html

galdera-erantzun sistema baten *estaldura* baino txikiagoa izan arren, *doitasun* altuagoa izateko aukerak ditu, bereziki patroi jakin batzuk jarraitzen dituzten galderen gainean.

Erantzun zuzena aurkitu ez den kasu gehienetan, SPARQL galderako osagaien lexikoaren egokitzapen faltarengatik gertatu da, hau da, SPARQL galdera egokia sortu den arren, hizkuntza naturalean idatzitako galderan osagaiak ondo identifikatuz, osagai hauen lexikoa, normalean propietateena, ez dator bat gure RDFan erabiltzen denarekin. Etorkizunean hau lantzea ezinbestekoa izango da emaitzak hobetu nahi baldin badira. Honetarako, sinonimoen hiztegiak edota Euskal Wordnet bezalako tresnak erabili ahal izango dira, SPARQL galderetan agertzen diren hitz hauen sinonimo, hiperonimo eta hiponimoekin aberasteko.

Konplexutasun maila desberdinetako galderak ebaluatu ditugu: alde batetik, emaitza Wikipediako artikulu bakarrean edota entitate bakarraren RDFko tripleen artean bilatzea eskatzen duten galderak, eta, beste alde batetik, erantzuna lortzeko entitate bat baino gehiago erlazionatu behar izatea eskatzen duten galderak. Lehendabiziko galderentzat triple bakarreko SPARQL galderak sortu behar izan dira eta bigarrengo galderentzat, aldiz, bi triplez osatutako SPARQL konplexuagoak sortu behar izan dira.

Hasiera batean triple bakarreko SPARQLa sortzea eskatzen duten galderak SPARQLra itzultzeak prozesu sinplea ematen badu ere, galderako osagaien identifikazio zuzena eta hauen lexikoaren egokitzapena beharrezkoak dira. Burututako ebaluazioan, galdera hauentzat gure sistemak emaitza hobeagoak eman ditu IHARDETSIk baino zehaztasuna eta doitasunari dagokionez. Garapeneko galderak erabilita gure hurbilpenak lortu duen zehaztasuna edota asmatze-tasa 0,54 da eta azken ebaluazioko galderak erabilita 0,52. IHARDETSIk zehaztasuna edota asmatze-tasa txikiagoa lortu du, garapenean 0,14, eta azkeneko ebaluazioan 0,11. IHARDETSIk esperotako emaitza baino txarragoak lortu ditu. Horregatik, IHARDETSIk ebaluatutako galdera hauentzat eman dituen emaitzak sakonago aztertu beharko ditugu zer gertatu den ondorioztatzeko.

Artikulu bat baino gehiago erlazionatzen dituzten galdera gehienak modu egokian itzuli dira SPARQLra, eta hemen ere, zailtasun nagusia lexikoaren egokitzapena izan da. Gainera, erantzunak testuan bilatzen dituen galdera-erantzun sistema batentzat oso zaila da galdera hauentzat erantzun zuzena bilatzea, beraz, RDF ezagutza-base

egituratuaren gainean galdera hauentzat erantzunak bilatzeak ekarpen berezia egiten duela uste dugu.

Lan honen ekarpen nagusiak modu honetan zerrendatu daitezke:

- Euskararentzako informazio egituratuaren gaineko galderak erantzuteko sistema bat aurkezten da. Zehazki, euskarazko wikipediako infotauletatik erauzitako RDFan bilatzen ditu hizkuntza naturalean idatzitako galderentzat erantzunak.
- Hizkuntza naturalean idatzitako galderak SPARQL lengoaiara itzultzen duen modulua garatu da euskararentzako.
- Esperimentazio ingurune bat sortu da, euskarazko galdera-zerrenda bat sortuz.
- Erantzunak testu hutsaren gainean bilatzen dituen galdera-erantzun sistemarekin osagarria den informazio-egituratuaren gaineko galdera-erantzun sistema bat aurkezten da.

Etorkizuneko lanak

- *Propietateen estaldura handitzeko sinonimoen hiztegia erabili*

Ondorioetan azaldu den bezala, gure sistemak erantzun zuzena aurkitu ez duen kasu gehienetan, SPARQL galderako osagaien lexikoaren egokitzapen faltarengatik gertatu da, hau da, SPARQL galdera egokia sortu den arren, hizkuntza naturalean idatzitako galderan osagaiak ondo identifikatuz, osagai hauen lexikoa, normalean propietateena, ez dator bat gure RDFan erabiltzen denarekin. Beraz, etorkizunean hau lantzea ezinbestekoa izango da emaitzak hobetu nahi baldin badira. Honetarako, sinonimoen hiztegiak edota ontologiak ere erabili ahal izango dira, SPARQL galdera hitz hauen sinonimo, hiperonimo eta hiponimoekin aberasteko.

- *Galdera patroi gehiago landu*

Lan honetan aztertu ditugun galderek murriztapen batzuk betetzen dituzte: leku eta pertsona entitateei buruzko propietateei buruz galdetzen dituzten galderak dira, entitatea beti azaltzen da galderan, galderen esaldiek konplexutasun maila jakin bat dute eta galdetzaile jakin batzuk erabilita sortutako galderak dira (noiz, non, nor, nork eta zein). Murriztapen hauek zabaldu eta galdera mota gehiago landu beharko genituzke

etorkizunean. Adibidez, galderan azaltzen den entitatea ez da beti SPARQL galderaren subjektua izango, objektua ere izan baitaiteke, beraz, hau kontuan hartu beharko dugu etorkizunean. Adibidez, "*Non da Iñaki Azkuna alkate?"* galderarentzako "*SELECT ? herria WHERE Iñaki_Azkuna alkate ?erantzuna"* SPARQL galdera lortzen da. SPARQL galdera zuzena den arren, gure RDFan "*Iñaki Azkuna"* entitateak ez dauka "*alkate"* propietaterik, horren ordez, "*Bilbo"* entitatearen alkate propietatearen balioa izan beharko litzateke, beraz, "*SELECT ?herria WHERE ?erantzuna alkate Iñaki_Azkuna"* SPARQL galdera sortu beharko genuke.

- *Galdetzaile gehiago aztertu*

RDFan aurki daitekeen informazio motak eta SPARQL lengoaiak ematen dituen aukerek beste galdetzaile batzuekin sortutako galderak lantzeko aukera ematen dute. Adibidez, "*Zenbat*" galdetzailea duten erantzun kuantitatibo bat espero duten galderak, edota bai/ez erantzuna espero duten galderak lantzeko aukera egongo litzateke. Baita ere, *handiena*, *txikiena*, *garaiena* bezalako propietateak eskatzen dituzten galderak (Adibidez, "*Zein da populazio txikiena duen estatua?"* galdera).

- *RDFa osatuz galdera-erantzun sistemaren estaldura handitu*

Aipatu den bezala, testuan RDFan baino informazio askoz gehiago dago, beraz, erantzunak RDFaren gainean bilatuko dituen sistemaren estaldura handitu nahi baldin badugu, RDFa informazio berriarekin osatu edota aberastu beharko dugu. Beste hizkuntza batzuetarako RDFak (ingelesa, gaztelera...) handiagoak dira eta artikuluetako infotaulak osatuagoak daude. Beraz, informazio hau erabili dezakegu modu erdi automatikoan euskarazko RDFa eta infotaulak osatzeko. Honez gain, testu hutsetatik triple erlazioak identifikatzea izan daiteke beste modu bat, baina erabiliko den propietateen lexikoa modu egokian kudeatu eta erabili beharko da. RDFa eguneratzean edota handitzean gure sisteman erabiltzen diren hiztegiak ere eguneratu beharko ditugu. Hala ere, sistemaren gainontzeko moduluetan ez du aldaketarik suposatuko.

7 Bibliografia

Agirre, E., O. Ansa, X. Arregi, M. Lopez de Lacalle, A. Otegi eta X. Saralegi. 2010. Document Expansion for Cross-Lingual Passage Retrieval.

Agirre, E., O. Ansa, X. Arregi, M. Lopez de Lacalle, A. Otegi ,X. Saralegi eta H. Zaragoza. 2009. Elhuyar-IXA: Semantic Relatedness and Cross-Lingual Passage Retrieval. CLEF, pp. 273-280.

Alegria, I., O. Arregi, I. Balza, N. Ezeiza, I. Fernandez eta R. Urizar. 2004. Development of a Named Entity Recognizer for an Agglutinative Language. In IJCNLP.

Androutsopoulos, I., G.D. Ritchie eta P. Thanisch. 1995. Natural language interfaces to databases - an introduction. Natural Language Engineering, pp. 29–81

Ansa, O., X. Arregi, A. Otegi eta A. Soraluze. 2008. Ihardetsi: A Basque Question Answering System at QA@CLEF 2008. Evaluating Systems for Multilingual and Multimodal Information Access, 9th Workshop of the Cross-Language Evaluation Forum, Lecture Notes in Computer Science, pp. 369-376.

Battista, A.D.L., N. Villanueva-Rosales, M. Palenychka and M. Dumontier, 2007. SMART: A web-based, ontology-driven, semantic web query answering application.

Bernstein, A., E. Kaufmann eta C. Kaiser. 2005. Querying the semantic web with ginseng: A guided input natural language search engine. In: 15th Workshop on Information Technologies and Systems, Las Vegas, NV, pp. 112–126 .

Chu, W., and F. Meng. 1999. Database Query Formation from Natural Language using Semantic Modeling and Statistical Keyword Meaning Disambiguation. Technical Report 990003, UCLA CS Dept., 16.

Damljanovic, D., M. Agatonovic eta H. Cunningham, 2010. Natural language interfaces to ontologies: Combining syntactic analysis and ontology-based lookup through the user interaction. Semantic Web: Res. Appli., 6088: 106-120. DOI: 10.1007/978-3-642-13486-9_8

Ding, L., T. Finin, A. Joshi, R. Pan, R. Scott Cost, Y. Peng, P.Reddivari, V. Doshi, eta J. Sachs. 2004. Swoogle: a search and metadata engine for the semantic web.Proceedings of the thirteenth ACM international conference on Information and knowledge management.652-659.ACM

Ezeiza, N., I. Aduriz, I. Alegria, J.M. Arriola eta R. Urizar. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. COLING-ACL, pp.380–384.

Fellbaum, Christiane (2005). WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 665-670

Fernandez, O., R. Izquierdo, S. Ferrandez and J.L. Vicedo, 2009. Addressing ontology-based question answering with collections of user queries.

Green, W., C. Chomsky eta K. Laugherty. 1961. BASEBALL: An automatic question answerer.

Proceedings of the Western Joint Computer Conference, pp. 219-224.

Jitkrittum, W., C. Haruechaiyasak and T. Theeramunkong. 2009. QAST: Question Answering System for ThaiWikipedia.Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions (KRAQ 2009.Association for Computational Linguistics.Suntec, Singapore,pp. 11—14.

Kaufmann, E., A. Bernstein. 2007. How Useful Are Natural Language Interfaces to the Semantic Web for Casual End-Users? ISWC/ASWC, pp. 281-294

Kaufmann, E., A. Bernstein eta L. Fischer. 2007. Nlp-reduce: A ”naïve” but domain independent natural language interface for querying ontologies. In: 4th ESWC, Innsbruck.

Kaufmann, E., A. Bernstein eta R. Zumstein. 2006. Querix: A natural language interface to query ontologies based on clarification dialogs. In: 5th ISWC, Athens, GA, pp. 980–981

Lopez, V., E. Motta eta V.S. Uren. 2006. AquaLog: An ontology-driven Question Answering System to interface the Semantic Web. HLT-NAACL.

Magnini, B. (2005) Open Domain Question Answering: Techniques, Systems and Evaluation. Tutorial of the Conference on Recent Advances in Natural Language Processing - RANLP. Borovetz, Bulgaria, September 2005.

Popescu, A. M., O. Etzioni eta H.A. Kautz. 2003. Towards a theory of natural language interfaces to databases. IUI 2003, pp. 149-157

Unger, C., L. Bühmann, J. Lehmann, A.C. Ngonga Ngomo, D. Gerber eta P. Cimiano. 2012. Template-based question answering over RDF data. WWW 2012, pp. 639-648

Waltinger, U., A. Breuing, eta I. Wachsmuth. 2011. Interfacing Virtual Agents With Collaborative Knowledge: Open Domain Question Answering Using Wikipedia-based Topic Models. Proc. of the International Joint Conference on Artificial Intelligence (IJCAI).

Wang, C., M. Xiong, Q. Zhou eta Y. Yu. 2007. PANTO - a portable natural language interface to ontologie . ESWC-2007.

Wang, C., M. Xiong, Q. Zhou. and Y. Yu. 2007. PANTO: A Portable Natural Language Interface to Ontologies. In: 4th ESWC, Innsbruck, A, pp. 473–487

Woods, J.A., Dickey Jr., J.S., Marvin, U., Powell, B.N. 1970. Lunar Anorthosites. Science 30: Vol. 167.no. 3918, pp. 602 – 604

8 Eranskinak

Garapenerako galderak:

Entitatea	Galdera	Testuan	Infotaulan
Aconcagua	Noiz sortu zen Aconcagua mendia?	1	0
Aconcagua	Non dago kokatua Aconcagua?	1	1
Aconcagua	Zein mendilerrotan dago Aconcagua	1	1
Aconcagua	Noiz igo zen Aconcagua lehen aldiz?	0	1
Alex Ferguson	Zein talde entrenatzen du Alex Fergusonek?	1	1
Alex Ferguson	Non jaio zen Fergie?	1	1
Alex Ferguson	Zein da Alex Fergusonen izen osoa?	1	1
Alex Ferguson	Zein talde entrenatu ditu Alexander Chapman Fergusonek?	1	1
Alex Ferguson	Noiz jaio zen Alex Ferguson?	1	1
Anna Magnani	Zein hiritan jaio zen Anna Magnani aktore italiarra?	1	1
Anna Magnani	Zein egunetan jaio zen Anna Magnani?	1	1
Anna Magnani	Noiz hil zen Anna Magnani?	1	1
Anna Magnani	Zein sari jaso zuen Anna Magnanik?	0	1
Anna Magnani	Zein lan egin ditu Anna Magnanik?	1	1
Asturiasko Printzerria	Zein hizkuntza mintzatzen dira Asturiasen?	1	1
Asturiasko Printzerria	Zein da Asturiasko mendirik altuena	1	0
Asturiasko Printzerria	Zein da Asturiasko Printzerriko hiriburua?	1	1
Asturiasko Printzerria	Zein da Asturiasko lehendakaria?	0	1
Asturiasko Printzerria	Noiz sortu zen Asturiasko estatutua?	0	1
Boada de Campos	Non dago Boada de Campos?	1	1
Boada de Campos	Zein da Boada de Camposko alkatea?	0	1
Boada de Campos	Zein eskualdetan dago Boada de Campos kokatua?	1	1
Boada de Campos	Zein probintziatan dago Boada de Campos?	1	1
Charles Bronson	Zein urtetan bidali zuten Charles Bronson Munduko Bigarren Gerrara?	1	0
Charles Bronson	Non lan egin behar izan zuen Charles Bronsonek?	1	0
Charles Bronson	Zein da Charles Bronsonen izen osoa?	1	1
Charles Bronson	Noiz jaio zen Charles Buchinski?	1	1
Charles Bronson	Non hil zen Charles Bronson?	1	1
Comoko probintzia	Zein da Comoko hiri nagusia?	1	1
Comoko probintzia	Non dago kokatua Comoko probintzia?	1	1
Comoko probintzia	Zein da Como probintziako hiriburua?	1	1
Ernest Fourneau	Zein izan ziren Enerst Fourneauren ekarpen zientifikoak?	1	1
Ernest Fourneau	Noiz aukeratu zuten Fourneau Medikuntzaren Akademiako kide?	1	0
Ernest Fourneau	Zein arlotan lan egin zuen Ernest Fourneauk?	1	1
Ernest Fourneau	Non hil zen Fourneau?	1	1
Ernest Fourneau	Zein da Ernest Fourneauren heriotza data?	1	1
Gipuzkoa	Zein euskalki hitz egiten da Gipuzkoan?	1	0
Gipuzkoa	zein ziren Gipuzkoan bizi ziren tribuak?	1	0

Gipuzkoa	Non dago kokatua Gipuzkoa?	1	1
Gipuzkoa	Zein da Gipuzkoago herritarraren izena?	0	1
Gipuzkoa	Zein antolaketa motetan dago Gipuzkoa?	1	1
Hernani	Zein da Hernaniko alkatea?	1	0
Hernani	Noiz dira Hernani herriko jaiak?	1	0
Hernani	Noiz sortu zen Hernani?	1	1
Hernani	Zein euskualdetan dago kokatua Hernani?	1	1
Hernani	Zein alturetan dago Hernani?	0	1
Hunte	Non du sorrera Hunte ibaiak?	1	1
Hunte	Zein da Hunte ibaiak egiten duen ibilbidea?	1	0
Hunte	Zein da Hunte ibaiak duen luzera?	1	1
Hunte	Non du lekua Hunte ibaiak?	1	1
Hunte	Non bukatzen da Hunte ibaia?	1	1
Irlandako Errepublika	Noiz iritsi zen San Patrizio artzapezpikua Irlandara?	1	0
Irlandako Errepublika	Noiz bilakatu zen Irlanda independente?	1	1
Irlandako Errepublika	Zein gobernu mota du Irlandak?	1	1
Irlandako Errepublika	Zein da Irlandako hiriburua?	1	1
Irlandako Errepublika	Zein da Irlandako ereserki ofiziala?	0	1
Izaba	Zein dira Izabako hizkuntza ofizialak?	1	0
Izaba	Zein da Izabako biztanleri kopurua?	1	1
Izaba	Zein garaieratan dago Izaba?	0	1
Izaba	Zein izen du Izabako alkateak?	0	1
Izaba	Zein eukaldetan kokatzen da Izaba?	1	1
Jeff Albertson	Zein ogibide du Jeff Albertsonek?	1	1
Jeff Albertson	Non bizi da Jeff Albertson?	0	1
Jeff Albertson	Zein telesailetan agertzen da Jeff Albertson?	0	1
Jeff Albertson	Zein da Jeff Albertsonen sortzailea?	0	1
Joan III a	Noiz bihurtu zen Joanes IIIa Aita Santu?	1	1
Joan III a	Zein kargu izan zuen Joan IIIak?	1	1
Joan III a	Noiz hil zen Joan IIIa?	1	1
Joan III a	Non hil zen Joanes IIIa?	1	1
Julia Gillard	Noiz bilakatu zen lehen ministro Julia Gillard?	1	1
Julia Gillard	Zein kargu du Julia Gillardek?	1	1
Julia Gillard	Non bizi da Julia Gillard?	0	1
Julia Gillard	Noiz jaio zen Julia Gilard?	1	1
Kaxmir	Non dago Kaxmir?	1	0
Kaxmir	Zein moneta erabiltzen da Kaxmirren?	0	1
Kaxmir	Zein hizkuntza hitz egiten da Kaxmirren?	0	1
Luis Mariano	Non dago Luis Marianoren hilobia?	1	0
Luis Mariano	Zein da Luis Marianoren izena osoa?	1	1
Luis Mariano	Non jaio zen Luis Mariano?	1	1
Luis Mariano	Zein hiritan hil zen Luis Mariano?	1	1
Luis Mariano	Zein herrialdeetan hil zen Luis Mariano?	1	1

Maleville	Zein zen 2007an Malevillen zegoen etxebizitza kopurua?	1	0
Maleville	Zein departamendu administratibotan dago Maleville herria?	1	0
Maleville	Zein naziotan dago Maleville?	1	1
Maleville	Zein da Malevilleko agintaria?	0	1
Maleville	Zein da Malevillengo kantoia?	0	1
Marin Konderria	Noiz sortu zen Marin konderria?	1	1
Marin Konderria	Zein da Marin konderriko hiriburua?	1	1
Marin Konderria	Zein da Marin konderriko dentsitate maila?	1	1
Marin Konderria	Zein da Maringo biztanleria?	1	1
Neil Jordan	Zein lanbide du Neil Jordanek?	1	1
Neil Jordan	Zein filma zuzendu zuen Neil Jordanek?	1	1
Neil Jordan	Zein saru jaso zituen Neil Jordanek?	1	1
Neil Jordan	Noiz jaio zen Neil Patric Jordan zinema zuzendaria?	1	1
Neil Jordan	Zein hiritan jaio zen Neil Jordan?	1	1
Orson Welles	Zein film zuzendu zituen Orson Wellesek?	1	1
Orson Welles	Zein lanbide zuen Orson Wellesek?	1	0
Orson Welles	Zein sari jaso zuen Orson Wellesek?	0	1
Orson Welles	Noiz jaio zen Orson Welles?	1	1
Orson Welles	Zein da Orson Wellesen izen osoa?	1	1
Ruanda	Noiz irabazi zuen independentzia Ruandak?	1	1
Ruanda	Zein da Ruandako hiriburua	1	1
Ruanda	Zein txanpon erabiltzen dute Ruandan?	0	1
Ruanda	Zein hizkuntza hitz egiten dute Ruandan?	1	1
Ruanda	Zein da Ruandako agintari nagusia?	0	1
Urretxu	Noiz pairatu zuen Urretxuk sute handi bat?	1	1
Urretxu	Noiz sortu zen Urretxuko herria?	1	0
Urretxu	Non kokatzen da Urretxu?	1	1
Urretxu	Zein da Urretxuko alkatea	0	1
Urretxu	zein azalera du Urretxuk?	1	1
Xabier Arzalluz	Noiz hartu zuen Arzalluzek EAJren buruzagitza?	1	1
Xabier Arzalluz	Noiz sartu zen Xabier Arzalluz Jesusen Konpainian?	1	0
Xabier Arzalluz	Zein alderdietan aritu zen Arzalluz?	1	1
Xabier Arzalluz	Non jaio zen Xabier Arzalluz?	1	1
Sadam Hussein	Nork agindu zuen Sadam Husseinen aurretik Iraken?	0	1
Watson Doktorea	Nork sortu zuen Watson pertsonaia?	1	1
Watson Doktorea	Nor zen Watsonen adiskide mina?	1	1
Watson Doktorea	Non zuen egoitza Watson doktoreak?	0	1
Caceres	Nor da alkate Caceresen?	0	1
Potemkin Korazatua	Nor da Potemkin korazatuaren gidoilaria?	0	1

Testerako galderak:

Entitatea	Galdera	Testuan	Infotaulan
Agatha Christie	Zein motatako eleberriak idazten zituen Agatha Christiek?	1	0
Agatha Christie	Non egin zuen lan Agatha Christiek?	1	0
Agatha Christie	Zein da Agatha Christieren izen osoa?	1	1
Agatha Christie	Zein du jaiolekua Agatha Christiek?	1	1
Agatha Christie	Noiz hil zen Agatha Christie?	1	1
Arratzu	Zein da Arratzuko euskaldunen ehunekoa?	1	1
Arratzu	Non dago Arratzu?	1	1
Arratzu	Zein da Arratzuko alkatea?	0	1
Arratzu	Zein da Arratzuko biztanle kopurua?	1	1
Bali	Zein da Baliko biztanle kopurua?	1	1
Bali	Non kokatzen da Bali?	1	1
Bali	Zein da Baliko hiriburua?	1	1
Indira Ghandi	Zein zen Indira Ghandiren familia?	1	1
Indira Ghandi	Noiz ezkondu zen Indira Ghandi?	1	1
Indira Ghandi	Zein izan zen Indira Ghandiren kargua?	1	1
Indira Ghandi	Zein alderdi politikokoa zen Indira Ghandi?	1	1
Indira Ghandi	Noiz hil zen Indira Ghandi?	1	1
Isaac Newton	Zein lege aurkitu zuen Newtonek?	1	1
Isaac Newton	Non jaio zen Isaac Newton?	1	1
Isaac Newton	Zein arlotan egin zuen lan Newtonek?	1	1
Isaac Newton	Noiz jaio zen Isaac Newton?	1	1
Islandia	Zein izan ziren Islandiara iritsitako lehen gizakiak?	1	0
Islandia	Noiz eratu zuen Islandia bere lehen batzar nazionala?	1	0
Islandia	Zein txanpon erabiltzen dute Islandian?	0	1
Islandia	Zein da Islandiako hiriburua?	1	1
islandia	Zein gobernu mota du Islandiak?	1	1
Jean Renoir	Nor izan ziren Jean Renoirren gurasoak?	1	0
Jean Renoir	Noiz jaso zuen Jean Renoirrek ohorezko oscar saria?	1	1
Jean Renoir	Noiz zuzendu zuen Jean Renoirrek “La Grande Ilusion” filma?	1	1
Jean Renoir	Non jaio zen Jean Renoir?	1	1
Jean Renoir	Zein herrialdeetan hil zen Jean Renoir?	1	1
Lantaron	Zein da Lantarongo ekonomiako arlo nagusia?	1	0
Lantaron	Noiz sortu zen Lantarongo udalerria?	1	1
Lantaron	Zein alderdi politikokoa da Lantorongo alkatea?	1	1
Lantaron	Zein da Lantarongo agintaria?	1	1
Lantaron	Non dago Lantaron?	1	1
Leslie Nielsen	Zein filmetan parte hartu Leslie Nielsenek?	1	1
Leslie Nielsen	Non jaio zen Leslie Nielsen	1	1
Leslie Nielsen	Zein egunetan jaio zen Leslie Nielsen?	1	1

Leslie Nielsen	Noren bikote izan zen Leslie Nielsen?	1	1
Leslie Nielsen	Non hil zen Leslie Nielsen?	1	1
Miguel de Cervantes	Noiz argitaratu zen Cervatesen Kixote eleberria?	1	0
Miguel de Cervantes	Noiz jarri zen Cervantes Madrillen bizitzen?	1	0
Miguel de Cervantes	Noiz jaio zen Miguel de Cervantes?	1	1
Miguel de Cervantes	Noiz hil zen Miguel de Cervantes?	1	1
Miguel de Cervantes	Zein da Cervantesen izen osoa?	1	1
Puerto Rico	Zein da Puerto Ricoko klima?	1	0
Puerto Rico	Zein da Puerto Ricoko estatuko uharte garrantzitsuena?	1	0
Puerto Rico	Zein da Puerto Ricoren gobernu mota?	1	1
Puerto Rico	Zein izen du Puerto Ricoko eserkiak?	0	1
Puerto Rico	Zein txanpon erabiltzen dute Puerto Ricon?	0	1
Tokio	Zein irletan kokatzen da Tokio?	1	1
Tokio	Zein da Tokioko metropoliaren biztanle kopurua?	1	0
Tokio	Zein azalera du Tokiok?	0	1
Tokio	Nola deitzen da Tokioko biztanlea?	0	1
Tokio	Zein dentsitate du Tokiok?	0	1
Toulonjac	Nongo udalerria de Toulonjac?	1	1
Toulonjac	Zein zen Toulonjac herriko diru-sarrera fiskala?	1	0
Toulonjac	Zein da Toulunjac herriko alkate?	0	1
Toulonjac	Zein da Toulunjaceko posta kodea?	0	1
Vadim Demidov	Zein taldetan jokatzen du Demidovek?	1	1
Vadim Demidov	Zein taldetan hasi zen Demidov?	1	0
Vadim Demidov	Non jaio zen Demidov?	1	1
Vadim Demidov	Zein posiziotan jokatzen du Demidovek?	0	1
Xuancheng	Zein mailako prefektura da Xuancheng?	1	1
Xuancheng	Zein da Xuanchengo agintaria?	0	1
Xuancheng	Zein azalerako hiria da Xuancheng?	1	1
Xuancheng	Zein da Xuanchen hiriko biztanle kopurua?	1	1
Ziordia	Zein da Ziordiako alkatea?	1	1
Ziordia	Zein da Ziordiako ondare aipagarria?	1	1
Ziordia	Zein eskualdetan dago Ziordia?	1	1
Ziordia	Zein da Ziordiako biztanleria?	1	1
Haile Selassie	Nor da Haile Selassieren bikotekidea?	0	1
Bizkaia	Nor da Bizkaiako agintaria?	0	1
Montpellier	Nork agintzen du Montpellierren?	0	1
Montpellier	Nor da Montpellierreko alkatea?	0	1

Testerako galdera konplexuak:

Entitatea	Galdera	Testuan	Infotaulan
Sadeko Markesa	Zein da Sadeko markesa jaio zen herrialdeko hiriburua?	1	1
Montpellier	Non jaio zen Montpellierreko alkatea?	?	1
Nacho Vigalondo	Zein da Nacho Vigalondo jaio zen hiriko alkatea?	?	1
Haile Selassie	Zein da Haile Selassieren heriotza herrialdeko hiriburua?	1	1
Bizkaia	Zein da Bizkaiako agintariaren jaioterria?	1	1
Berria	Zein da Berriako zuzendariaren izena?	1	1
Suedia	Zein da Suediako hiriburuko webgunea?	?	1
Barbara Goenaga	Nor da Barbara Goenagak lan egin zuen telesailko protagonista?	?	1
Barbara Goenaga	Zein katetan emititzen da Barbara Goenagak lan egin zuen telesaila?	1	1
Barbara Goenaga	Noiz sortu zen Barbara Goenaga jaio zen hiria?	?	1
Aimar Olaizola	Zein da Aimar Olaizolaren jaioterriko biztanleria?	1	1
Aimar Olaizola	Zein da Aimar Olaizolaren jaioterriko alkatea?	1	1
Alessandro Volta	Nork du alkatetza Alessandro Voltaren jaiotza herrialdean?	?	1
Alessandro Volta	Nork agintzen du Alessandro Voltaren jaiotza herrialdean?	?	1
Alessandro Volta	Zein da Alessandro Voltaren jaio zen herrialdeko hiriburua?	1	1
Frantzia	Noiz jaio zen Frantziako agintaria?	?	1
Frantzia	Zein da Frantziako hiriburuko alkatea?	?	1
Frantzia	Zein da Frantziako hiriburuko eskualdea?	1	1
Hondarribi	Zein da Hondarribi dagoen eskualdeko herririk handiena?	1	1
Hondarribi	Zein da Hondarribi dagoen eskualdeko biztanleria?	1	1
Hondarribi	Zein da Hondarribi dagoen eskualdeko azalera?	1	1
Peru	Noiz sortu zen Peruko hiriburua?	1	1
Peru	Zein da Peruko hiriburuaren sorrera data?	1	1
Peru	Zein da Peruko hiriburuaren probintzia?	?	1