Rol semantikoen etiketatze automatikoa

Haritz Salaberri, Olatz Arregi, Beñat Zapirain

Lengoaia eta Sistema Informatikoak Saila, Informatika Fakultatea (UPV/EHU)

haritz.salaverri@ehu.es

olatz.arregi@ehu.es

benat.zapirain@ehu.es

Jasoa: 2014-06-13 Onartua: 2014-07-21

Laburpena: Rol semantikoen etiketatze automatikoa (SRL), azaleko analisi semantikoa ere deitua, hizkuntzalaritza konputazionalaren ikerlerro garrantzitsua da eta bertan, zehatz finkatu nahi dira testu bateko gertakarietan, ekintza eta honetan parte hartzen dutenen arteko erlazio semantikoak edo rolak; berez, *nork*, *nori*, *zer* egin zion, *non* eta *noiz* gertatu den jakin nahi da. Rolek eskaintzen duten informazioak berebiziko garrantzia dauka testuak automatikoki prozesatu eta ulertzeko bidean. Ataza hau zeresan handia ematen ari da hizkuntzaren prozesamenduan ez ezik, besteak beste, Interneteko bilatzaileetan, itzulpen automatikoko eta galdera-erantzun sistemetan, sare sozialen azterketa automatikoan, eta dokumentuen informazio erauzketan.

Hitz gakoak: semantika konputazionala, rol semantikoen etiketatzea, lengoaia naturalaren prozesamendua.

Abstract: The main task of semantic role labeling (SRL), sometimes also called shallow semantic parsing, is to detect the semantic relations hold among the predicate of a sentence and its associated participants and properties and the classification into their specific roles. Performing sentence-level semantic analysis can help determine *who* did *what* to *whom*, *where*, *when*, and *how* within an event. The predicate of a clause (typically a verb) establishes *what* took place, and other sentence constituents express the participants in the event (such as *who* and *where*), as well as further event properties (such as *when* and *how*). The information provided by semantic roles is crucial in order to process texts automatically, and in addition to the applications in Natural Language Processing (NLP), semantic roles can help improve Internet search engines, question answering and translation systems. Nowadays, roles are on the edge regarding information extraction and social network research tasks.

Keywords: computational semantics, semantic role labeling, natural language processing.

1. SARRERA

Semantikan **gertaera** edo **gertakizun** esaten zaio errealitatean jazotzen den edozeri. *Mikelek sagarra jan zuen* esaldiak, esate baterako, *Mikelek sagarra* jan izanaren gertaeraren berri ematen du. Jakina da hizkuntzan gertakari bera esapide ezberdinak erabilita adieraz daitekeela, hau da, adierazpen desberdinek gertakari berari egin diezaioketela erreferentzia. Zenbaitetan gainera, adierazpen horiek esanahi bera izan dezakete. Hona ondoko bi esaldiak, esate baterako:

Mikelek liburu gorria irakurri du.
Liburu gorria Mikelek irakurri du.

1. Adibidea

Biek adierazten dute *Mikelek liburu gorria irakurri* duela. Esaldi biek gertakari berari egiten diote erreferentzia, eta, gainera, **esanahi** bera daukate. Kontuan hartu behar da, hala ere, erreferentzia-gertakari bera duten esaldiek ez dutela zertan esanahi bera izan. Hurrengo esaldien kasuan:

Mikelek Amaiari liburu gorria saldu zion.
Amaiak Mikeli liburu gorria erosi zion.

2. Adibidea

Bi esaldiek erreferentzia gertakizun berari egiten badiote ere, ikuspuntua ezberdina da, eta, beraz, esaldiek duten esanahia ere bai. Izan ere, lehenbizikoan burutu den ekintza *liburu gorria*ren *erosketa* izan da, eta bigarrenean, ordea, *salmenta*.

Esaldi mailako analisi semantikoa egiteak, beraz, gertakarietan *nork*, *nori*, *zer* egin zion ezartzen lagun dezake. Analisi semantikoa egiteko orduan, kontuan hartu behar da esaldiak **proposizio** izeneko unitateetan antolatzen direla, eta unitate horietako bakoitza gertaera bati dagokiola. Esaldi bateko proposizioen arteko harremana analisi sintaktikoak deskribatzen du, eta horrek berebiziko garrantzia dauka analisi semantiko egokia egin eta gertakari bakoitzaren parte-hartzaileak egoki identifikatzeko garaian.

Proposizio bateko **predikatua** gehienetan aditza izaten da, baina dena den, kontuan eduki behar da bi predikatu mota daudela: izenezkoak edo aditzezkoak. Lehen kasuan predikatu funtzioa izenak edo izenondoak betetzen du eta bigarrenean ordea aditzak. Esate baterako:

Zelulen $[apurketa]_{Pred}$ *nabarmena da.*
Mikelek leihoa $[apurtu]_{Pred}$ *du.*

3. Adibidea

Lehenbiziko esaldian *apurketa* izenak (nominalizazioak) betetzen du predikatu funtzioa, eta bigarrenean ordea *apurtu* aditzak.

Predikatuak ezartzen du *zer* izan den jazo dena, hau da, zein izan den gertakaria. Esaldiko gainerako osagaiek gertakarian parte hartu dutenak, (**argumentuak**) eta bestelako **propietateak** ezartzen dituzte [1].

1.1. **Rol semantikoak**

Rol semantikoen etiketatzearen (*Semantic Role Labeling*) eginbehar nagusia esaldi bateko predikatuaren eta parte-hartzaile eta propietateen arteko erlazio semantikoak detektatu eta dagokien "**paper**" edo "**rol**" semantikoa automatikoki esleitzea da. Predikatu baten parte-hartzaileei **argumentuak** esaten zaie, propietateei, ordea, **adjuntuak**.

[Atzo]$_{Adjuntua}$ *[Mikelek]* *[Amaiari]* *[liburu gorria]*$_{Argumentuak}$ *[saldu]*$_{PREDIKATUA}$ *zion*.

4. Adibidea

4. adibidean, *Mikel*, *Amaia* eta *liburu gorria* dira *saldu* ekintzan parte-hartzaileak, hau da, hauek dira saldu predikatuari lotutako argumentuak. *Atzo* berriz, modifikatzaile funtzioa duen adjuntua, denbora adjuntua, da.

Behin gertakariaren argumentuak eta adjuntuak identifikatuta, rol semantikoak esleitu behar zaizkie argumentuei, alegia, bakoitzak gertakarian betetzen duen funtzio semantikoa zein den ezarri behar da.

Dagoen rol semantiko ezberdinen zerrenda 'estandarra' zehazturik ez badago ere, hizkuntzalari guztiek onartzen dituzte *Egilea*, *Jasalea, Gaia* edo *Esperimentatzailea,* bezalako rolak. Adjuntuen kasuan ordea, denbora, kokapena edo modua bezalako propietateak adierazten dituzte. Aurreko adibidea horrela geratuko litzateke rolak esleituta:

[Atzo]$_{Denbora}$ *[Mikelek]*$_{Egilea}$ *[Amaiari]*$_{Jasalea}$ *[liburu gorria]*$_{Gaia}$ *[saldu]*$_{PREDIKATUA}$ *zion*.

5. Adibidea

Horrela, *salmentan*, *Mikelek egile/saltzaile* rola betetzen du, *Amaiak,* ordea, *jasale/erosle* rola, eta *liburu gorriak* berriz, *gaia/erosgaia* rola. Alegia, *liburu gorria* da saldu dena, *Mikel* da liburu gorria saldu duena eta Amaia da erosi duena. Gainera hau *atzo*-ko egunez eman zen.

Kontuan eduki beharra dago esaldi elkartuak rol semantikoekin etiketatzeko garaian egokiro adierazi beharko dela argumentu eta adjuntu bakoitza zer perpausi dagokion.

[*[Argentinara]*$_{Helburua(1)}$ *[joan]*$_{\mathbf{Pred1}}$ *zen* *[taldea]*$_{Gaia(1)}$]$_{Gaia(2)}$ *[egongo]*$_{\mathbf{Pred2}}$ *da* *[finalean]*$_{kokapena(2)}$

6. Adibidea

Erlazio semantikoak gramatika sortzailean [2] 1965. eta 1970. urte bitartean erabili ziren lehenengo aldiz hizkuntzako predikatuen argumentuak parte-hartzaile motaren (rola) arabera sailkatzeko. Rol semantikoei *rol tematikoak*, *kasu semantikoak*, *theta-rolak* edo *kasu sakonak* (kasuzko gramatikan) ere deitzen zaie.

2. **SRL-REN APLIKAZIOAK**

Hizkuntzaren prozesamenduaren alorrean kokatzen diren aplikazioen garapena aurrera eraman ahal izateko, askotan, beharrezkoa izango da informazio semantikoa, eta, hain zuzen ere, rol semantikoak etiketatuta dauzkaten corpusak eskura izatea. Makina bidezko itzulpenean eta testu-laburpen automatikoan, esaterako, lagungarria izan daiteke rolek eskaintzen duten informazioa emaitza hobeak lortzeko.

Gaur egun, rol semantikoen etiketatze automatikoa ikerkuntza bide nagusietakoa da hizkuntza teknologiaren barnean. Frogatu egin da azaleko analisi semantikoaren erabilerak hobekuntza nabarmenak ekar ditzakeela galdera-erantzun sistemetan [3], makina bidezko itzulpenean [4], testuen laburpen automatikoan [5], informazio erauzketan [6], edota testuetatik ondorioak erauzteko sistemetan [7].

Euskarazko rolen etiketatze automatikoa ikertzearen ondorio zuzen bezala, aurrikusten da euskarazko itzulpen automatikoa, galdera erantzun sistemak, eta, oro har, ukitu semantikoa duen aplikazio oro hobetzea.

2.1. **SRL-ren aplikazioak: Adibidea**

Jarraian, rol semantikoek galdera-erantzun sistemetan eskain dezaketen hobekuntza erakusten duen adibidea ageri da. Demagun sistemari ondorengo galdera egiten zaiola:

Zer urtetan saldu zuen Errusiak Alaska?

7. Adibidea

Sistemak galdera hori erantzuteko behar duen jakintza zuzenean aurkitu dezake entziklopedia edo iturri berezituetan, *Errusia* eta *Alaska* hitz gakoen bitartezko bilaketa eginda. Demagun hau dela sistemak bilaketaren ondorioz eskuratu duen informazioa:

...goi-bilera iritsi baino lehenago Errusiak Estatu Batuei Alaska saldu zien 1867. urtean...

Informazio honetan oinarrituta nahikoa izango da zenbakizko balioa bilatzea: 1867. urtean erosi zioten Estatu batuek Errusiari Alaska.

Hala ere, gerta daiteke hau izatea *Errusia* eta *Alaska* hitz gakoen bitartezko bilaketak itzultzen duen informazioa:

...1870. urteko goi-bilera iritsi baino lehenago Errusiak Estatu Batuei Alaska saldu zien 1867. urtean...

Informazio hori erabilita, ez da nahikoa zenbakizko balioa bilatzea, eta orduan, nola jakin dezake sistemak erantzun egokia zein den? Nola jakingo du erantzun beharrekoa 1867. urtea dela, eta ez 1870. urtea?

Demagun, galdera-erantzun sistemak erlazio sintaktikoak era automatikoan etiketatzen dituen sistema bat daukala atxikita. Nahikoa izango da sintaxi-analizagailu hau erabiltzea aurkitutako informazio atala horrela etiketatzeko:

I. Lehenik, informazio ataleko predikatuak identifikatuko dira, *iritsi* eta *saldu*.

...1870. urteko goi-bilera [iritsi]$_{PRED1}$ baino lehenago Errusiak Estatu Batuei Alaska [saldu]$_{PRED2}$ zien 1867. urtean...

II. Ondoren, sintaxi-analizagailuak bi predikatuei lotutako osagaiak bilatuko ditu eta sintaxi-harremanak ezarriko ditu. Horrela, *saldu* predikatuaren osagai bezala *Errusia*, *Estatu Batuak, Alaska* eta *1867. urtean* etiketatuko dira. Bestalde, *iritsi* predikaturako *goi-bilera* eta *1870. urteko* etiketatuko dira.

...[1870. urteko]$_{(1)}$ [goi-bilera]$_{(1)}$ [iritsi]$_{PRED1}$ [baino lehenago]$_{(2)}$ [Errusiak]$_{(2)}$ [Estatu Batuei]$_{(2)}$ [Alaska]$_{(2)}$ [saldu]$_{PRED2}$ zien [1867. urtean]$_{(2)}$...

Errusiak Alaska zein urtetan saldu zuen jakiteko, egoera honetan nahikoa izango da *saldu* predikatuari lotutako zenbakizko balioa itzultzea.

Hala ere, gerta daiteke beste hau izatea *Errusia* eta *Alaska* hitz gakoen bitartezko bilaketak itzultzen duen informazioa:

...1870. urteko goi-bilera iritsi baino lehenago Errusiak Estatu Batuei Alaska saldu zien 1867. urtean. Bost urte beranduago, 1872. urtean, Errusiak Alaskako urrea saldu zion Indiari...

Kasu honetan sintaxi-analizagailuak honela etiketatuko lituzke predikatuak eta beren menpeko osagaiak:

...[1870. urteko]$_{(1)}$ [goi-bilera]$_{(1)}$ [iritsi]$_{PRED1}$ [baino lehenago]$_{(2)}$ [Errusiak]$_{(2)}$ [Estatu Batuei]$_{(2)}$ [Alaska]$_{(2)}$ [saldu]$_{PRED2}$ zien [1867. urtean]$_{(2)}$. [Bost urte beranduago]$_{(3)}$, [1872. urtean]$_{(3)}$, [Errusiak]$_{(3)}$ [[Alaskako]$_{(3)}$ [urrea]]$_{(3)}$ [saldu]$_{PRED3}$ zion [Indiari]$_{(3)}$...

Kasu honetan bi *saldu* predikatu daude (*Pred2* eta *Pred3*) eta bien osagai bezala ageri da *Alaska (Alaska* eta *Alaskako)*. Gainera, *Alaska* eta *Alaskako* osagaiek, bakoitzari dagokion *saldu* predikatuarekin (Pred2 eta Pred3) erlazio sintaktiko berdina daukate, perpausaz kanpoko modifikatzailea. Hau da, kasu honetan ezin esan daiteke analisi sintaktiko soilarekin erantzun egokia 1867. urtea edo 1872. urtea den.

Egoera honetan lagungarria da rol semantikoekin etiketatzea testua. Galdera-erantzun sistemak SRL tresna bat baldin badauka atxikita honela etiketatuko da testua:

...[1870. urteko]$_{(1)Denbora}$ [goi-bilera]$_{(1)Ekitaldia}$ [iritsi]$_{PRED1}$ [baino lehenago]$_{(2)Denbora}$ [Errusiak]$_{(2)Saltzailea}$ [Estatu Batuei]$_{(2)Eroslea}$ [Alaska]$_{(2)Salgaia}$ [saldu]$_{PRED2}$ zien [1867. urtean]$_{(2)Denbora}$. [Bost urte beranduago]$_{(3)Denbora}$, [1872. urtean]$_{(3)Denbora}$, [Errusiak]$_{(3)Saltzailea}$ [[Alaskako]$_{(3)Kokapena}$ [urrea]]$_{(3)Salgaia}$ [saldu]$_{PRED3}$ zion [Indiari]$_{(3)Eroslea}$...

Rol semantikoen etiketetan oinarrituta, ikusten da erantzun egokia 1867. urtea izango dela eta ez 1872.a. Izan ere, lehen *saldu* predikaturako (*Pred2*) *salgaia Alaska* da, bigarrenarentzat (*Pred3*) ordea *Alaskako urrea* ez Alaska.

3. **BALIABIDE KONPUTAZIONALAK: PREDIKATU-LEXIKOIAK**

Rol semantikoen etiketatze automatikoa egiten duten sistemak garatzeko orduan, erabilgarriak dira predikatu-lexikoiak. Predikatu-lexikoiak hizkuntza bateko predikatu zerrendak dira, non predikatu bakoitzarentzat zehazten den zein diren berak izan ditzakeen adiera ezberdinak eta adieretako bakoitzean jasotzen dituen argumentuak.

Lehenengo atalean esan den moduan, predikatuak aditzezkoak edo izenezkoak izan daitezke. Horregatik, aditzezko predikatu-lexikoiak eta izenezko predikatu lexikoiak daude. Ingeleserako izenezko predikatuak zerrendatzen dituen lexikoirik ezagunena 2007. urtean eraikitzen hasitako *NomBank* [8] da. Esate baterako, *Hautatu* aditz-predikatuak euskaraz bi adiera izan ditzake:

Mikel lehengo urtean presidente hautatu zuten.

Mikelek kolore urdinekoa hautatu zuen.

8. Adibidea

Lehenengo esaldian, *hautatu* aditzak Mikel presidente aukeratua izan zela adierazten du, lexikoi hipotetiko batean *hautatu* predikatuaren lehe-

nengo adiera izenda daiteke (*hautatu_01*). Bigarren esaldian, ordea, *hautatu* aditza Mikelek zerbait aukeratu duela adierazteko erabiltzen da, lexikoian *hautatu* predikatuaren bigarren adiera izenda daiteke (*hautatu_02*). Lexikoiak, adierak zehazteaz gainera, adiera bakoitzean *hautatu* predikatuak hartzen dituen rolak zehazten ditu honela:

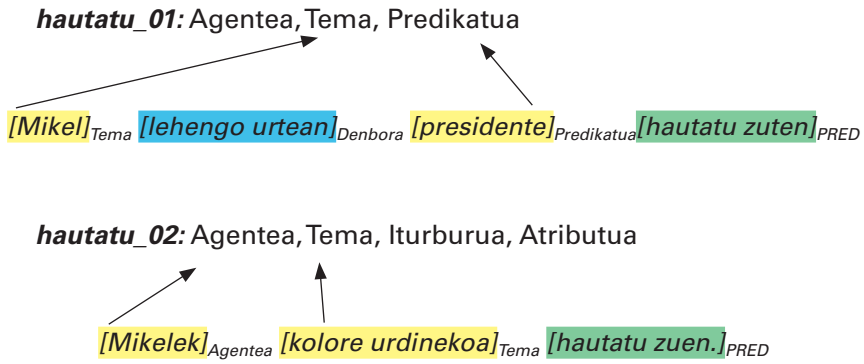
HAUTATU

hautatu_01: Agentea, Tema, Predikatua

hautatu_02: Agentea, Tema, Iturburua, Atributua

0. Irudia. Predikatu-lexikoi adibidea.

Informazio hori erabilita, rol semantikoak era automatikoan etiketatzen dituzten sistemek, esaldi bateko predikatua identifikatu eta gero, erabaki behar dute zein den predikatuak duen adiera. Ondoren, eta argumentu-identifikazioa eginda, erabakitzen dute zein den argumentu bakoitzak hartuko duen rola, predikatu-lexikoiak predikatu horren adiera horretarako ezartzen dituenen artetik.



Aditzezko predikatu-lexikoiei dagokienez, Ingeleserako ezagunenak, *VerbNet* [9] eta *FrameNet* [10] dira, lehenengoak Levin-en aditz klaseen sailkapena jarraitzen du [11], eta bigarrenak, ordea, *frame semantics* hurbilpena [12].

3.1. **VerbNet**

Levin klaseen aditz-sailkapenaren sorrera-lanean esaten denez, *Ezagutza lexikalaren* teorian oinarrituta, aditzen semantikaren eta sintaxiaren arteko harremanek aditz multzoak osatzeko bidea ematen dute. Honen arabera, izaera semantiko bereko aditzek egitura sintaktiko berak onartzen dituzte. Ondorioz, egitura sintaktiko horiek identifikatu eta beraietan oi-

narrituta, aditzak multzoka daitezke, multzokatze hauek semantika mailan gertatzen direlarik [11]. Aditz multzo horiei *Levinen aditz klaseak* deritze.

Aditz klaseak direlakoen idean, onartzen da badirela sintaktikoki eta semantikoki berdinak diren aditzak, hau da, klase berekoak diren aditzek txandaketa sintaktiko berak onartzen dituztela. Ondorioz, txandakaketa sintaktiko berak onartzen dituzten aditzek osagai semantiko berak izango dituzte.

Klasea: Objektu bat kokapen batean ezartzen duten aditzen klasea.

Aditzak: Dangle, Hang, Lay, Lean, Perch, Rest, Sit, Stand, Suspend.

1. irudia. *Levin*-en aditz klasea (*9.2. Verbs of putting in a Spatial Config.*).

Levin klaseak eta horiei dagozkien zehaztapenak erabilgarriak suertatu dira hizkuntzaren prozesamenduarekin lotutako hainbat eta hainbat atazatan, besteak beste rol semantikoen etiketatze automatikoan [13].

Levinen sailkapenean oinarritutako *VerbNet* izenekoa da gaur egun sarean aurki daitekeen eta domeinu jakin batera bideratu gabeko ingelesezko aditzen lexikoirik handiena. Modu hierarkikoan dago egituratuta, eta lexikoian bildutako informazioa publikoki eskuragarri dagoen beste hainbat baliabiderekin lotzen du. Esate baterako, loturak ezartzen ditu *WordNet* [14] eta *FrameNet* datu-base lexikalekin. Lexikoiko aditz-klase bakoitza deskribatzeko rol semantikoak, argumentuen gaineko hautapen murriztapenak, eta predikatu-argumentu egiturak erabiltzen dira besteak beste. Hauek dira *VerbNeteko* rol ohikoenak: *Actor*, *Agent*, *Attribute*, *Cause*, *Experiencer*, *Location*, *Patient*, *Predicate*, *Product*, *Recipient*, *Source*, *Stimulus*, *Theme*, *Destination*, *Instrument*, *Result eta Topic*. Bigarren irudian *Hit-18.1*. aditz-klaseari dagokion *VerbNeteko* sarrera ageri da.

Hit-18.1

Rolak: *Agent, Patient, Instrument, Result*

Klasekideak: Bang, Bash, Hit, Kick e.a

Frame-ak:

1. *Paula hit the Ball.* = Gertakaria = G

— Sintaxia: Agent_Aditza_Patient

— Semantika: ...

2. *Paula hit the Ball with a stick.*

— Sintaxia: Agent_Aditza_Patient_Preposizioa_Instrument

— Semantika: ...

...

2. irudia. VerbNeteko -eko Hit-18.1 aditz-klasea.

Irudian ikusten denez, h*it* predikatuaren lehen adierak a*gent*, *patient*, *instrument eta result* rolak har ditzake. *Klasekideak* atalean ezartzen da zein diren aditzaren egitura sintaktiko berak onartzen dituzten aditzak (Levin klase berekoak). Ondoren, *Frame* direlakoen atalean, egitura sintaktiko horiek banan banan aztertzen dira, kasu bakoitzean hartzen diren rolak eta rolen egituraketa sintaktiko zein semantikoak azalduz.

3.2. ***PropBank***

PropBank [15] *The Wall Street Journal Corpus* corpusetik hartutako milioi bat hitzez osatutako corpusa da. *PropBanken* predikatu-argumentu egiturak eta argumentuen rol semantikoak daude etiketatuta. Argumentuak anotatzeko *arg0, arg1, arg2,..., arg5* etiketak erabiltzen dira. *PropBank* sortu zenean, saiakera egin zen *arg0* etiketarekin *egile* papera betetzen zuten argumentuak etiketatzeko, eta *arg1* etiketarekin *gaia* papera jokatzen zuten argumentuak etiketatzeko [16]. Ondorioz, *PropBank*en *arg0* etiketa daramana, gehienetan, *egilea* izango da, eta *arg1* etiketa daramana, berriz, *hartzailea* edo *tema*. Adjuntuak ere bilduta daude *PropBanken,* honako etiketa hauekin: *TMP* (denbora), *LOC* (kokapena), *SRC* (iturburua), *ADV (adberbioa), CAU (kausa), DIS (dislokazioa), MNR (modua), MOD (modifikatzailea), NEG (ezeztapena)* (Oharra: *VerbNet* rolekin etiketatutako corporetan, adjuntuak ere etiketatzea erabaki denean, *PropBankeko* adjuntu-etiketak erabili izan dira). *PropBank* sortu zeneko saiakera honen bidez, saiakerarik gabe lortuko ziren baino emaitza hobeak lortu nahi ziren rol semantikoak etiketatzeko sistemek (*PropBank* erabilita sortutakoek) *arg0* eta *arg1* etiketatzeko orduan.

$$[\text{Mary}]_{arg0}\ [\text{drove}]_{PREDIKATUA}\ [\text{a red car}]_{arg1}\ [\text{around the block}]_{LOC}$$

9. Adibidea

3.3. ***FrameNet***

Frame semantics delakoan oinarritzen da *FrameNet*. Horren arabera esaldietan, zenbait hitzek gaitasuna daukate jakintza semantikoko *frame* edo egitura bat «aktibatzeko». Hitz hauei *Lexical Units* (LU) (unitate lexikalak) esaten zaie. Ikus dezagun ondorengo adibidea:

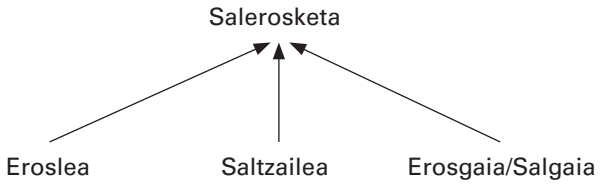
Mikelek Amaiari liburu gorria saldu zion.

Amaiak Mikeli liburu gorria erosi zion.

10. Adibidea

Bi esaldietan aktibatzen da *Salerosketa frame*a. Lehen esaldiaren kasuan, aktibazioa *saldu*k (*saldu* da kasu honetan *lexical unita*) eragiten du,

eta bigarrenean, ordea, *erosik* (*erosi* da kasu honetan unitate lexikala). Egitura semantiko honek ezartzen du *erosle (hartzailea)* rola izango duen argumentu bat egongo dela, *saltzaile (egilea)* rola betetzen duen beste argumentu bat izango dela, eta, azkenik *erosgaia/salgaia (gaia)* rola izango duen azken argumentu bat izango dela. Mikelek betetzen du *saltzaile* rola, Amaiak *erosle* rola, eta liburu gorriak *salgaia/erosgaia* rola.



3. irudia. *Frame semantics* adibidea.

FrameNet [10] markoen semantika (*frame semantics*) paradigma oinarri duen ingeleserako datu base lexikala da. 1.200 frame semantiko eta 13.000 unitate lexikal biltzen ditu. Honetaz gain, datu-baseak eskuz etiketatutako 170.000 esaldi baino gehiago biltzen ditu. Hurrengo atalean ikusiko den moduan, eskuzko etiketatzeak oso erabilgarriak dira rol semantikoak automatikoki etiketatzeko sistemak eraikitzeko orduan.

3.4. **Euskararako baliabideak**

Euskararako, dakigula behintzat, *Euskal Herriko Unibertsitateko IXA* taldeak sortutako *Basque Verb Index* (*BVI*) lexikoia da *SRL* sistemak gauzatzeko, eskuragarri dagoen aditzezko predikatu-lexikoi bakarra [17]. Ez dago izenezko predikatu-lexikoirik euskararako. Horretaz gainera, *PropBank-VerbNet* ereduari jarraituta etiketatutako *EPEC-RolSem* corpusa dago eskuragarri [18].

4. **EUSKARARAKO *SRL* SISTEMA BATEN ERAIKUNTZA**

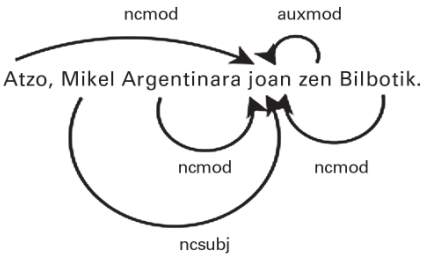
SRL sistemek arkitektura ezberdinak izan baditzakete ere, normalean, sekuentzialki antolatutako ondorengo bost osagaiez egoten dira osatuak: sintaxi-analizagailua, predikatu-identifikagailua, predikatu-desanbiguagailua, argumentu-identifikagailua eta argumentu-sailkagailua. Jarraian, euskararako sortutako *SRL* sistema erdiautomatiko baten gauzaketan emandako pausoak aurkezten dira [19]. Sistema horretan azkeneko hiru pausoak

bakarrik dira automatikoak. Eman dezagun sistema hori erabilita ondorengo esaldia etiketatu nahi dela rol semantikoekin:

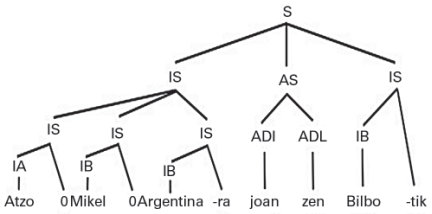
Atzo, Mikel Argentinara joan zen Bilbotik.

11. Adibidea

1. **Sintaxi-analizagailua:** Sarrera testuaren analisi sintaktikoa egiten duen osagaia da. Bi analisi mota egin daitezke, osagaietan oinarritutakoa (constituent-based syntax) [5. Irudia] edo dependentzietan oinarritutakoa (dependency-based syntax) [4. Irudia] [20].



4. irudia. Dependentzien analisia



5. irudia. Sakoneko analisia

Inplementatutako sistema erdiautomatikoan dependentzietan oinarritutako informazio sintaktikoa erabiltzen da. *SRL* sistema batean sintaxi-analizagailuaren garrantzia gainerako osagaiena baino handiagoa da; izan ere, sistema osoaren emaitza orokorra konputatzeko unean, sintaxi-analizagailuaren emaitzak % 50eko pisua dauka (CoNLL-2008 [21]). Horregatik, ezinbestekoa da informazio sintaktiko ahalik eta zuzenena edukitzea, esaldia rol semantikoekin ongi etiketatu nahi bada.

2. **Predikatu-identifikagailua:** Analisi sintaktikotik eskuratutako informazioan oinarrituta, osagai horrek erabakitzen du zein diren predikatuak, sarrerako testuan agertzen diren hitzen artetik.

Atzo, Mikel Argentinara $[joan]_{PRED}$ *zen Bilbotik.*

Automatikoki, bi hurbilpen erabil daitezke, lehen hurbilpena ikasketa automatikoa erabilita egin daiteke, eta bigarrena, berriz, heuristiko edo erregela linguistikoak erabilita. Esate baterako, hitzen kategorizazioari (*PoS* etiketei) erreparatuta aditz bezala etiketatuta daudenak predikatuak direla esaten duen heuristiko bat erabil liteke. Osagai horretan lortutako emaitzak ere guztiz zuzenak direla

onartzen dugu eskuzkoak izanagatik. Kontuan hartu beharra dago sortutako sistema erdiautomatikoak aditz-predikatuen argumentuak soilik etiketatuko dituela, ez baitago euskararako izen-predikatuen lexikoirik, hori dela eta, atal honetan aditz-predikatuak soilik identifikatu dira.

3. **Predikatu-desanbiguagailua:** Aurreko osagaiak identifikatutako predikatuei dagokien adiera ezartzen zaie predikatu-lexikoian oinarrituta.

Atzo, Mikel Argentinara [joan]$_{joan\_02}$ zen Bilbotik.

Desanbiguazio lan horretarako (aditzezko predikatuak dira identifikatutako guztiak) aurretik aipatutako *BVI* aditzezko predikatu-lexikoia erabiltzen da. Bertan 244 euskarazko aditzei dagozkien adierak, eta adiera bakoitzak jasotzen dituen argumentuak, rol semantikoak, postposizio-atzizkiak eta hautapen murriztapenak ageri dira. Kontuan izan beharra dago aipatutako *EPEC* corpusean 1.200 aditz baino gehiago ageri direla. Hori horrela, *SRL* sistemak gaitasuna izango du soilik lexikoian agertzen diren predikatuei adiera ezartzeko. Etorkizunean, *BVI* lexikoia osatu, eta, horrela *SRL* sistema ahaltsuagoa lortzeko asmoa dago.

4. **Argumentu (eta adjuntu) identifikagailua:** Osagai horrek sarrera testuko predikatu bakoitzari dagozkion argumentuak eta adjuntuak identifikatzen ditu.

[Atzo], [Mikel] [Argentinara] [joan]$_{joan\_02}$ zen [Bilbotik].

Euskararako *SRL* sistemak ataza hau automatikoki egiten du, %99 zehaztasuna lortzen duen heuristiko bat erabilita.

5. **Argumentu-sailkagailua:** Aurreko osagaiak identifikatutako argumentu eta adjuntu bakoitzari rol semantikoa esleituko dio osagai honek.

[Atzo]$_{TMP}$, [Mikel]$_{arg0}$ [Argentinara]$_{arg2}$ [joan]$_{joan\_02}$ zen [Bilbotik]$_{arg1}$

Garatutako sisteman pauso hori ere modu guztiz automatikoan egiten da. Ikasketa automatikoa erabiltzen da. Era berean, sistemak gaitasuna dauka argumentuei *VerbNet*eko rolak eta aldi berean adjuntuak ere esleitzeko. Aurreko adibidea *VerbNet*eko rolekin etiketatuta ageri da ondoko adibidean:

[Atzo]$_{TMP}$, [Mikel]$_{Agent}$ [Argentinara]$_{Destination}$ [joan]$_{joan\_02}$ zen [Bilbotik]$_{Source}$

4.1. **Ikasketa automatikoa**

Ikasketa automatikoa (*Machine learning*), adimen artifizialaren barnean kokatutako ikerkuntza arloa da. Bertan, ordenagailuak ataza jakin bat egiten ikasten du, horretarako aurretik programatua egon gabe. Lan atazak, sailkapen eta aurreikuspen lanak izaten dira.

Ikasketa automatikoa teknika oso erabilia da gaur egun hizkuntzaren prozesamenduan. Izan ere, hizkuntzaren prozesamenduaren barnean teknika hau erabiltzen dutenak dira besteak beste, testu idatzien sailkapena, korreferentzien erresoluzioa, entitateen desanbiguazioa eta nola ez, rol semantikoen etiketatzea.

Ezaugarriek ikasketa algoritmoari laguntzen diote sailkapena zuzen egiten. Esan bezala, rol semantikoen atazan, askotan erabili dira Adimen Artifizialeko (AA) teknikak. Ataza honen azken urratsa da AA-rekin gehien jorratu dena (5.osagaia). Argumentu-sailkatze prozesuan hainbat ezaugarri izan behar dira kontuan. Bibliografian aurkitu ohi diren ezaugarriak ondoren ageri dira zerrendatuta. Ulergarritasun arrazoiak direla eta, aurreko adibideko *Mikel*ek ezaugarri hauetan hartzen dituen balioak ere ageri dira:

— ***Predikatuaren lema:*** Proposizioaren predikatuaren lema (*joan*).
— ***Argumentuaren funtzio sintaktikoa:*** Argumentuak betetzen duen funtzio sintaktikoa *(Subjektua)*.
— ***Argumentuaren lema:*** *(Mikel)*.
— ***Argumentuaren kokapena predikatuarekiko:*** Argumentua proposizioko predikatuaren aurretik edo predikatuaren ondoan dagoen *(Aurretik)*.
— ***Argumentuaren kategoria:*** Argumentuaren PoS (*Part-of-Speech*) kategoria (*IZE-Izena*).
— ***Argumentuaren azpikategoria:*** Argumentuaren PoS azpikategoria *(IZE/B-Izen berezia)*.
— ***Argumentuaren postposizio-atzizkia:*** Argumentuaren deklinabide-kasua *(ABS-Absolutiboa)*.
— ***Distantzia hitz kopuruan:*** Predikatuaren eta argumentuaren arteko distantzia hitz kopurutan *(1)*.
— ***Distantzia argumentu kopuruan:*** Predikatuaren eta argumentuaren arteko distantzia argumentu kopurutan *(1)*.

5. **EUSKARARAKO *SRL* SISTEMA ERDIAUTOMATIKOAREN EMAITZAK**

SRL sistema garatzeko orduan, *Support Vector Machines* [22], *Erabaki Zuhaitzak* [23] *eta Erabaki Zuhaitzen konbinazioa* [24] algoritmoak erabili

dira. Ondorengo tauletan ageri dira algoritmo hauek erabilita lortzen diren emaitzak. F_1-Scorea lortzeko ondorengo formula aplikatzen da:

$$F_1 = 2 \times [(\text{Doitasuna} + \text{Estaldura}) / (\text{Doitasuna} \times \text{Estaldura})].$$

Formula hau erabili ahal izateko doitasuna eta estaldura kalkulatzen dira lehenik. Doitasunak adierazten du zer ehunekotan etiketatu diren zuzen argumentuak. Estaldurak, ordea, adierazten du dauden argumentuetatik zer ehuneko etiketatu den rol egokiarekin.

0. taula. Emaitza orokorrak *PropBank* rolekin.

PropBank	Doitasuna	Estaldura	F_1-Scorea
Support Vector Machines (SMO)	84,3	84,6	84,3
Erabaki Zuhaitzak (J48)	84,0	84,2	83,9
Erabaki Zuhaitzen konbinazioa (Random Forest)	77,4	78,3	77,7

0. Taulak erakusten dituen emaitzak, argumentuak *PropBank* rolekin etiketatzean lortzen direnak dira, hau da, argumentuak *arg0-arg5* etiketekin, eta adjuntuak *ADV, CAU, DIS, LOC, MNR, MOD, NEG, TMP e.a*. etiketekin etiketatzean lortzen direnak.

Argumentuak VerbNet rolekin etiketatzean lortzen diren emaitzak berriz 1. taulan ageri dira.

1. taula. Emaitzak *VerbNet* rolekin.

VerbNet	Doitasuna	Estaldura	F_1-Scorea
Support Vector Machines (SMO)	83,1	83,1	82,9
Erabaki Zuhaitzak (J48)	81,7	81,8	81,5
Erabaki Zuhaitzen konbinazioa (Random Forest)	72,2	72,9	72,1

Tauletan antzematen den moduan, emaitzarik onenak *Support Vector Machines* ikasketa algoritmoak ematen ditu, bai *PropBank* rolak etiketatzerakoan, eta bai *VerbNet* rolak etiketatzerakoan ere. 2. eta 3. tauletan erakusten da rol bakoitzerako lortutako emaitzak zein diren algoritmo hau erabiltzen denean.

2. Taulan ikusten den moduan, *arg0* eta *arg1* argumentuak dira *PropBank* roletatik emaitzarik onenak eskuratzen dituztenak. Izan ere, ikasketa-algoritmoa prestatzeko orduan beste rolei dagozkienak baino as-

2. taula. *PropBank* rolak.

PropBank	F_1-Scorea
arg0	95,0
arg1	93,7
arg2	81,6
arg3	57,9
arg4	15,4
ADV	50,8
CAU	80,5
DIS	41,6
LOC	73,9
MNR	67,8
MOD	54,3
NEG	99,2
TMP	78,9

3. taula. *VerbNet* rolak.

VerbNet	F_1-Scorea		F_1-Scorea
actor	89,7	*topic*	87,7
agent	96,2	*ADV*	50,5
attribute	92,4	*CAU*	78,3
cause	79,2	*DIS*	44,9
experiencer	66,9	*LOC*	73,5
location	80,1	*MNR*	68,7
patient	80,6	*MOD*	54,5
predicate	74,6	*NEG*	99,2
product	91,6	*TMP*	78,5
recipient	83,2		
source	74,4		
stimulus	87,3		
theme	88,0		

koz ere gehiago dira erabilitako *arg0* eta *arg1* adibideak, eta gainera, *arg0* eta *arg1* argumentuen izaera homogeneoa da. Berez, *arg2-arg4* argumentuak etiketatzea baino errazagoa da argumentu hauek etiketatzea, normalean *arg0* argumentuak *egile* rola betetzen duelako eta *arg1* rolak *jasale* edo *tema* rola.

VerbNeteko rolei dagokienez, batez beste, emaitzak ez dira *PropBank*eko rolenak bezain altuak; izan ere, aukeratzeko rol kopuru handiagoa izateak emaitzak hain onak ez izatea dakar, ikasketa algoritmoaren lana zailtzen baita.

Adjuntuen kasuan, emaitzak mota guztietakoak dira; bi kasuetan (NEG) ezeztapeneko adjuntuak lortzen ditu emaitzarik onenak (99,2). Emaitzarik okerrenak dislokazio adjuntua (DIS) etiketatzen lortzen dira, bai *PropBank* eta baita *VerbNet*erako ere (41,6 eta 44,9). Ezeztapeneko adjuntuetarako lortzen diren emaitzak oso onak dira, aurreikusten «oso errazak» direlako; izan ere, normalean, ia beti, *ez* hitza NEG moduan etiketatuko da. Orokorrean, adjuntuak etiketatuta lortutako emaitzek menpekotasun handia daukate ikasketa-algoritmoa prestatzeko orduan erabilitako mota bakoitzeko adibideen kopuruarekin.

6. **ONDORIOAK ETA ETORKIZUNEKO LANAK**

Garatutako lan honetatik, ondorioztatzen dugu rol semantikoek eskaintzen duten informazioak hobekuntza nabarmenak ekar ditzakeela egun mo-

dan dauden hainbat eremu eta aplikaziotan. Esate baterako, interneteko bilaketa-tresnetan. Horren ondorio zuzena izango da euskarak sarean ikusgarritasuna handiagoa lortzea posible izango baita bilaketei emaitza semantikoki esanguratsuagoekin erantzutea.

Etorkizuneko lanei dagokienez, lehenik eta behin, beharrezkoa izango da SRL sistema guztiz automatiko baten garapena. Sistema guztiz automatikoa inplementatuta, posible izango da testu bilduma handiak eskuzko etiketatzearen beharrik gabe rol semantikoekin etiketatzea. Kontuan eduki behar da eskuzko etiketazioak dauzkan denbora eta diru-kostuak.

Azkenik, aurrera eraman beharreko beste ataza bat euskararako izen-predikatu lexikoi baten garapena da, lexikoi hori izanda, aukera izango baita testuetan argumentu eta predikatu kopuru handiagoa etiketatu eta SRL-an emaitza hobeak lortzeko.

7. **BIBLIOGRAFIA**

[1] MÀRQUEZ Lluís, *et al*. 2008. «Semantic role labeling: an introduction to the special issue.» *Computational linguistics*, bol. **34**, no 2, o. 145-159.

[2] CHOMSKY Noam. 1965. *Aspects of the Theory of Syntax*. *MIT press, USA*.

[3] SHEN Dan eta LAPATA Mirella. 2007. «Using Semantic Roles to Improve Question Answering.» *EMNLP-CoNLL*. o. 12-21.

[4] BOAS Hans Christian. 2002. «Bilingual FrameNet Dictionaries for Machine Translation.» *LREC*.

[5] MELLI Gabor, *et al*. 2005. «Description of squash, the sfu question answering summary handler for the duc-2005 summarization task.» *safety*, bol. **1**, o. 14345754.

[6] SURDEANU Mihai, *et al*. 2003. «Using predicate-argument structures for information extraction.» *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume **1***. Association for Computational Linguistics, o. 8-15.

[7] HICKL Andrew, *et al*. 2006. «Recognizing textual entailment with LCC's GROUNDHOG system.» *Proceedings of the Second PASCAL Challenges Workshop*.

[8] MEYERS Adam, *et al*. 2004. «The NomBank project: An interim report.» *HLT-NAACL 2004 workshop: Frontiers in corpus annotation*. o. 24-31.

[9] SCHULER Karin. 2005. «VerbNet: A broad-coverage, comprehensive verb lexicon.»

[10] BAKER Collin F., *et al*. 1998. «The berkeley framenet project.» *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume **1***. Association for Computational Linguistics, o. 86-90.

[11] LEVIN Beth. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, USA.
[12] FILLMORE Charles J., 1976. Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, bol. 280, no 1, o. 20-32.
[13] SWIER Robert S eta STEVENSON Suzanne. 2004. «Unsupervised semantic role labelling.» *Proceedings of EMNLP*. o. 102.
[14] MILLER George A. 1995. «WordNet: a lexical database for English.» *Communications of the ACM*, bol. **38**, no 11, o. 39-41.
[15] KINGSBURY Paul eta PALMER Martha. 2003. «Propbank: the next level of treebank.» *Proceedings of Treebanks and lexical Theories*.
[16] LOPER Edward, *et al., 2007*. «Combining lexical resources: mapping between propbank and verbnet.» *Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands*.
[17] ALDEZABAL Izaskun., 2004. «*Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean: 100 aditzen azterketa, Levin-en (1993) lana oinarri hartuta eta metodo automatikoak baliatuz.*»
[18] ALDEZABAL Izaskun, *et al*. 2013. «A methodology for the semiautomatic annotation of EPEC-RolSem, a basque corpus labeled at predicative level following the PropBank-Verb Net model.» *UPV/EHU/LSI/TR*.
[19] SALABERRI Haritz, *et al.,* 2014. «First approach toward Semantic Role Labeling for Basque.» *The 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik, Iceland*
[20] SGALL Petr eta PANEVOVÁ Jarmila., 1989. Dependency syntax-a challenge. *Theoretical Linguistics*, bol. 15, no 1/2, o. 73-86.
[21] SURDEANU Mihai, *et al.,* 2008. «The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies.» *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, o. 159-177.
[22] PLATT John, *et al.,* 1998. «Sequential minimal optimization: A fast algorithm for training support vector machines.»
[23] QUINLAN J. Ross., 1996. «Bagging, boosting, and C4. 5.» *AAAI/IAAI, bol. 1*. o. 725-730.
[24] BREIMAN Leo., 2001. «Random forests.» *Machine learning*, bol. 45, no 1, o. 5-32.