Euskal Herriko Unibertsitatea / University of the Basque Country

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Lengoaia eta Sistema Informatikoak Saila

Rol Semantikoen Etiketatze Automatikoa: Rol Multzoak eta Hautapen Murriztapenak

Beñat Zapirain Sierrak Eneko Agirre eta Lluís Màrquez-en zuzendaritzapean egindako tesiaren txostena, Euskal Herriko Unibertsitatean Informatikan Doktore titulua eskuratzeko aurkeztua.

Donostia, 2010eko urria.

Lan hau EHUko ikerketa beka baten laguntzaz egin da (2005-2009).

Oh, it's such a perfect day,
I'm glad I spent it with you.
Oh, such a perfect day,
You just keep me hangin' on.

Lou Reed

Esker onak

Tesi lan honen zuzendari izan diren Eneko eta Lluísi nire esker beroenak, azken urte hauetan aurrera eraman dugun kalitatezko lankidetzagatik.

Eskerrik asko Stanford unibertsitateko NLP taldeko Mihai Surdeanuri interesa eta laguntzagatik, eta baita Sussex unibertsitateko Diana McCarthyri ere, lankidetza eta begiruneagatik.

Eskerrik asko Lengoaia Naturalaren Prozesamendurako IXA talde osoari, ikerlari berriak prestatzen egiten duzuen esfortzuagatik.

Pantaila Gelako Akademikoei: `nunc est bibendum (!)`

Laburpena

Tesi honetan, Rolen Sailkatze Automatikoan (RSA) aski ezagunak diren bi arazo izan ditugu aztergai: (1) Rol multzo ezberdinen egokitasuna praktikan, eta (2) RSArako sistemek darabiltzaten ezaugarri lexikalen eragin mugatua eta pairatzen duten sakabanaketa. Lehen puntuari dagokionez, gaur egun gure arloan gehien erabiltzen diren PropBank eta VerbNeteko rol multzoen azterketa konparatibo sakona aurkeztuko dugu, rol multzo bakoitzarekin entrenatutako sailkatzaileen errendimendua, sendotasuna, eta orokortzeko gaitasuna, esperimentazio ingurune eta domeinu ezberdinetan neurtuz. Bigarren puntuari dagokionez, ezaugarri lexikoek planteatzen dituzten arazoak aztertuko ditugu eta, WordNet eta antzekotasun distribuzionaleko neurriekin sortutako hautapen murriztapenak erabiliz, arazo horien eragina modu esanguratsuan leunduko dugu. In-vitro egindako esperimentuekin, hautapen murriztapen horiek lexikotik eratorritako ezaugarriek baino sailkatze ahalmen handiagoa dutela ikusiko dugu. Azkenik, hautapen murriztapenetatik erauzitako ezaugarriak baliatuz, artearen egoeran dagoen RSA sistema baten errendimendua hobetuko dugu (domeinu barnean eta domeinuz kanpo).

Abstract

This thesis focuses on two well-known open issues in Semantic Role Classification (SRC) research: (1) the suitability of different role inventories in practice, and (2) the limited influence and sparseness of lexical features. About the former, we present an empirical comparative study on the use of PropBank vs. VerbNet roles, the two most widely used role inventories, testing the performance differences for unseen verbs and the robustness for new corpus domains. About the latter, we test the use of automatically learnt selectional preferences as a complement to lexical features, proposing both WordNet-based and distributional similarity based models. We show that all our selectional preference models improve over lexical features in in-vitro experiments, and that the models are complementary. Finally, we show that incorporating features based on selectional preferences, the overall performance of an state-of-the-art SRC system improves both in in-domain and out-of-domain corpora.

Gaien aurkibidea

Aurkibidea **i**

I Sarrera **1**
I.1 Rol semantikoak 4
I.2 Tesiaren motibazioa 6
I.3 Ekarpenak 10
I.4 Tesi lanaren egitura 12
I.5 Publikazioak 13

II Rolen etiketatze automatikoa: artearen egoera **15**
II.1 Rol semantikoak eta baliabideak 17
II.1.1 Proposizioen bankua: PropBank 17
II.1.2 FrameNet 23
II.1.3 VerbNet 24
II.1.4 Beste hizkuntzetarako errekurtsoak 31
II.2 Rol Semantikoen Etiketatzaile automatiko baten arkitektura 32
II.2.1 Rolen etiketatze automatikoa urratsez urrats . . . 32
II.3 Ezaugarrien ingeniaritza 38
II.4 Rol semantikoen etiketatze automatikoaren historia laburra 40
II.4.1 Gaur egungo erronkak 45

III PropBank eta VerbNeteko rolen azterketa konparatiboa **49**
III.1 Corpus semantikoak eta rol multzoak: ikuspegi kritikoa . 51
III.1.1 PropBank: teorikoki neutrala 51
III.1.2 VerbNet: rol tematiko orokorrak 53
III.2 Esperimentaziorako baliabideak 54
III.2.1 Datu multzoa 55
III.2.2 SRL sistema 56

III.3 Rol multzoen orokortzeko gaitasuna 63
III.3.1 Aditz ezezagunetara orokortzeko gaitasuna 67
III.3.2 Aditz ezaugarriekiko sentikortasuna 68
III.4 Mapaketa VNeko rol tematikoetara 70
III.5 Erlazionatutako lanak 72
III.6 Ondorioak 75

IV Ezaugarri lexikalak orokortzen: Hautapen Murriztapenak rolen sailkatze automatikorako 77
IV.1 Aurrekariak 80
IV.2 Hautapen Murriztapenen ereduak 83
IV.2.1 WordNeten oinarritutako Hautapen Murriztapen Ereduak 84
IV.2.2 Hautapen Murriztapen Distribuzionalak 88
IV.2.3 Preposizioentzako Hautapen Murriztapenak 90
IV.3 Argumentuen sailkapena HM ereduekin 92
IV.4 Esperimentazio ingurunea 93
IV.5 Emaitzak eta analisia 94
IV.6 Ondorioak 101

V Hautapen Murriztapenak Argumentuen Sailkapenerako 103
V.1 Hautapen Murriztapenak SRL sistema batean integratzen 104
V.1.1 *SwiRL* 104
V.1.2 Ezaugarriak 107
V.1.3 SwiRLen ikasketa eredua 107
V.2 HMen integrazioa *SwiRL*en 107
V.2.1 SwiRL-i ezaugarriak gehitzen 108
V.2.2 Sailkatzaileen arteko meta-sailkatzailea 112
V.3 Esperimentazio ingurunea 117
V.4 Emaitzak eta analisia 118
V.5 Ondorioak 123

VI Ondorioak eta etorkizuneko lanak 125
VI.1 Etorkizuneko lanak 130

Bibliografia 131

I. KAPITULUA

Sarrera

Testuen azaleko analisi semantikoak esaldiko predikatuek adierazten dituzten "gertaeren" atributu eta propietateak zehazteko xedea du, "nork", "nori", "zer" egin zion, "non" eta "noiz" adieraziz. Predikatuek normalean "zer" gertatu den zehazten dute, eta gertaera osatzen duten gainontzeko osagaiek parte hartzaileak ("nork" eta "nori", adibidez) eta beste zenbait propietate ("noiz" eta "non") adierazten dituzte. Rol semantikoen etiketatze automatikoaren helburua, beraz, testu bateko gertaeretan predikatua eta parte hartzaile eta propietateen arteko erlazio semantikoak zehaztea da. Horretarako predikatuarentzat aurretik definitutako *rol semantikoak* jarraian ikusten den adibidean[1] bezala erabiltzen dira. (iturria: Aldezabal *et al.*, 2010).

[$_{Agente}$ *Nemesiok*], [$_{Denbora}$ *joan baino lehen*], [$_{Paziente}$ *Alejandro adiskideari]* **eskatzen dio** [$_{Tema}$ *zaindu dezala bere "x" zakurra*]

Predikatuaren ekintzan parte hartzen duten argumentuek oinarrizko rolak jokatzen dituzte (*Agente*, *Paziente* ...); aldiz, ekintzaren propietateak zehazten dituzten hautazko argumentuek rol adjuntuak jokatzen dituzte (*Denbora*, *Norantza*...)

[1] Tesi lan honetan zehar erabiliko ditugun adibide gehienak ingeleserako PropBank corpusetik atera ditugu eta, beraz, ingelesez idatzita emango dira.

Mota honetako analisiak esaldiaren azaleko errepresentazio semantikoa ahalbidetzen du, esaldiko gertaeren propietateak eta entitate garrantzitsuenen arteko oinarrizko erlazioak adieraziz.

Hizkuntzalaritza konputazionala edo, zehazkiago, semantika konputazionala da testuetako predikatu-argumentu egiturak modu automatizatuan detektatzeko ardura bere gain hartu duen zientziaren adarra. Berrogeita hamargarren hamarkadan, hizkuntza prozesatzeko lehen saiakerak egin zirenetik, hizkuntzaren aspektu semantikoak konputazionalki lantzera bideratutako ikerketa lanek, batez ere, lexikoien, gramatiken eta bestelako errekurtso semantikoen eskuzko garapenena izan zuten jomuga (Hirst, 1987; Pustejovsky, 1995; Copestake eta Flickinger, 2000). Baliabide hauek hizkuntzaren analisi semantiko sakonaren oinarri izatea zuten helburu baina, sarritan, eskulan gogorra eskatzen zuten eta oso domeinu zehatzetara zeuden mugatuta.

Azken bi hamarkadetan errekurtso konputazionalen ugaritzeari eta ikasketa automatikoko metodoen garapen azkarrari esker, Lengoaia Naturalaren Prozesamenduko (LNP) arlo garrantzitsu askotan aurrerapen handiak egin dira. Metodo hauek egitura linguistiko konplexuak eskulan handiegirik gabe ikasteko bidea ireki zieten hizkuntza prozesatzeko sistemei. Hasieran, analisi katearen egiturari jarraituz, ikasketa automatikoa eta metodo estatistikoak morfologian eta sintaxian aplikatu ziren, eta alor hauetan emaitza positiboak lortu ahala, semantika lantzeko beharraz gero eta gehiago jabetzen hasi zen LNPko ikerketa komunitatea. Briscoe eta Carroll-ek (1997) azpikategorizazio egiturak automatikoki erauzteko egin zuten lanak adibidez, edota aditzak beren azpikategorizazioaren arabera sailkatzeko egin ziren beste zenbait lanek (Merlo eta Stevenson, 2001; Schulte im Walde, 2000) erakutsi zuten metodo konputazional hauek egokiak zirela predikatu-argumentu erlazioak ikasteko. Horrela, azken urteotan, rol semantikoen etiketatze automatikoari begira, rolekin etiketatutako lehen corpusak agertu ziren: FrameNet (Fillmore *et al.*, 2004) eta PropBank (Palmer *et al.*, 2005).

Corpus baliabide egokien eta ikasketa metodo eraginkorren etorrerak, beraz, rolen etiketatze automatikoaren garapena ekarri zuen eta, dagoeneko, LNParen barruan ongi definitutako ataza garrantzitsuan bilakatu da. Zeresan handiko ikerketa lan asko egin dira azken urte hauetan gai honen inguruan (Gildea eta Jurafsky, 2002; Surdeanu *et al.*, 2003; Xue eta Palmer, 2004; Pradhan *et al.*, 2005a) eta Senseval-3 (Litkowski, 2004), CoNLL-2004 (Carreras eta Màrquez, 2004), CoNLL-2005 (Carreras eta Màrquez, 2005) eta SemEval-2007 (Pradhan *et al.*, 2007a) bezalako ebaluazio saioek rolak etiketatzeko sistemen arteko ebaluazio konparatibo azpimarragarria eskaini

diote ikerketa komunitateari.

Gertaera eta predikatu-argumentuen arteko maila honetako analisiak nolabaiteko karga semantikoa duten LNPko hainbat aplikaziotan onurak ekar ditzake. Esaterako,

- Galdera-erantzun sistemek galdera egiten den predikatuaren (edo antzekoen) argumentuak bila ditzakete, adibidez, galdera erantzuten saiatzeko (Narayanan eta Harabagiu, 2004; Frank *et al.*, 2007). Modu horretan, *Who assassinated JFK?* galderaren erantzuna lortzeko aski izango litzateke, adibidez, *assassinate* predikatuak gobernatutako argumentu egitura batean *JFK* argumentu "Pazientea" bilatu eta argumentu "Agentea" itzultzea.

- Itzulpen automatikoan ere, argumentuak hizkuntza batetik bestera itzultzeko rol semantikoa kontuan har daiteke, rolaren arabera itzulpena ezberdina izan daitekeelako. (Boas, 2002; Fung *et al.*, 2007; Wu eta Fung, 2009a; Wu eta Fung, 2009b)

- Itzulpen automatikoaren makina bidezko ebaluaziorako erabiltzen diren zenbait neurri berritzaile kalkulatzeko beharrezkoak diren ezaugarri linguistikoak rolekin etiketatutako egituretatik erauz daitezke (Giménez eta Màrquez, 2007; Giménez eta Màrquez, 2008).

- Testuen laburpen automatikorako, adibidez, predikatuek eta rolen guneek testu baten edukia labur dezakete (Melli *et al.*, 2005)

- Informazio erauzketan ere, argumentuen rolen gaineko bilaketek patroi edota erregela ahaltsuagoak definitzen lagun dezakete (Surdeanu *et al.*, 2003).

- *Textual Entailment* edota testuetatik ondorioak erauzteko sistema batzuek rolak etiketatzeko sistemen analisiak kontuan hartzen dituzte beren inferentziak egiteko (Tatu eta Moldovan, 2005; Burchardt *et al.*, 2007).

- Hizkuntzaren barneratze goiztiarra aztertzen duten ikerketa lan batzuek, besteak beste, rolak etiketatzeko sistemak erabiltzen dituzte beren ereduak eraikitzeko (Connor *et al.*, 2008; Connor *et al.*, 2009;M. Connor eta Roth, 2010).

- Irudi bidezko komunikazio sistemek rolak automatikoki etiketatzeko baliabideak erabiltzen dituzte hizkuntza irudietara itzultzeko (Goldberg *et al.*, 2008).

Nahiz eta rolak etiketatzeko sistemek mundu errealeko aplikazioetan duten eragina oraindik nahiko mugatua izan, informazio semantikoaren beharra erakusten duten aplikazio askorentzako onuragarria izan daiteke etorkizunean.

I.1 Rol semantikoak

Rol semantikoek predikatuak zehazten duen gertaeran argumentuek zein paper jokatzen duten esaten digute. Baina zeintzuk dira rol semantiko posibleak? Zenbat dira? Zenbat motatako erlazio semantikoak adieraz ditzakete?

Fillmore-ek (Fillmore, 1968) bere "Case for case" ezaguna aurkeztu zuenetik asko ikertu da hizkuntzalaritzan rol semantikoen izaerari buruz. Hala ere, nahiz eta "Agente" edo "Paziente" bezalako rol semantiko garrantzitsuenetan nolabaiteko adostasunak egon daitezkeen, oraindik ere adituen artean ez dago erabateko akordiorik behin betiko rol multzo unibertsal bat zehazteko garaian (Dowty, 1991). Egileen arabera, beraz, egoera ezberdinetarako prestatutako rol multzo zerrenda espezifiko eta luzeak (`FOOD`, `COOK`, `HEATING-INSTRUMENT`, etab.) topa ditzakegu (Fillmore *et al.*, 2004), edo rol tematiko bezala ezagutzen ditugun rol multzo orokorrago eta murritzagoak (Gruber, 1965; Jackendoff, 1972), edota Dowty-k (1991) proposatzen duen *Proto-Agent* eta *Proto-Theme* bi roletako sistema. Hizkuntzalariek gai honen inguruan dituzten ikuspegi eta teoria ezberdinek bere eragina izan dute hizkuntzalaritza konputazionalean izaera ezberdineko hainbat rol inbentarioen sorrerarekin (II.1 sekzioan ematen da baliabide hauen berri).

Hizkuntzalaritzak rol semantikoen inguruan burututako lan garrantzitsu askok predikatu-argumentu egituren gauzatze sintaktikoak beren azaleko semantika edo rol semantikoen egiturarekin nola erlazionatzen diren deskribatu dute (besteren artean, Grimshaw, 1990; Levin, 1993; Levin eta Rappaport Hovav, 2005). Levinen lanak (1993) hain zuzen, antzeko gauzatze sintaktikoak (alternantziak) dituzten aditzen artean antzekotasun semantikoak daudela zehazten du, onartzen dituzten rol tematikoetan esaterako. Levinek ezaugarri komun horiek partekatzen dituzten aditzak "aditz klaseetan" multzokatu zituen. Klase hauek eta haien predikatu-argumentuen deskribapenak

LNPko hainbat ataza hobetzeko erabiliak izan dira (Habash *et al.*, 2003; Shi eta Mihalcea, 2005), haien artean rolen etiketatze automatikoa (Swier eta Stevenson, 2005). Aditz klaseek, gainera, VerbNet lexikoi konputazionalaren (Kipper *et al.*, 2000) oinarri teorikoak ezarri zituzten.

Levinek aditza eta bere gauzatze sintaktikoak lotu bazituen argumentuen semantikarekin, Fillmore-ek bere *frame semantics*-en (Fillmore, 1976) gertaera semantiko ezberdinekin lotutako hitzak biltzen ditu. Hitz hauek gertaera semantiko espezifikoekin daude lotuta (*apply-heat* frame semantikora adibidez) eta, ondorioz, baita frame elementu bezala ezagutzen diren rol semantiko zehatzagoekin ere (`COOK`, `FOOD`, `HEATING-INSTRUMENT` etab.). Levinen klaseko osagaiekin gertatzen ez zen bezala, frameetako osagaietan ez da homogeneotasun sintaktikorik eskatzen, baina, horren ordez, informazio semantiko aberatsago eta zehatzagoa eskaintzen dute.

Sintaxira orientatutako rol sistemen hurbilpenetan (ikusi PropBank eta VerbNet II.1 sekzioan) rolak bi kategoriatan banatu ohi dira: (1) Argumentuak, aditzaren ezinbesteko posizio sintaktikoak betetzen dituzten erlazioak eta (2) adjuntuak, hautazkoak edota predikatuaren ekintzan hain garrantzitsuak ez diren rolak. Frameen semantikan ere bi multzo bereizten dira: (1) core frame elementuak (`COOK`, `FOOD`, ...) eta (2) elementu periferikoak edota frame elementuak (`TIME`, `MANNER`, `PLACE` etab.). Hurbilpen bakoitzak proposatzen duen rol multzoak bere ezaugarri propioak ditu, bai rol kopuruan eta baita rol horietako bakoitzaren semantikan ere. Hizkuntzalari konputazionalen eginkizuna da multzo bakoitzetik eredu konputazionalak sortzea eta baita eredu horien arteko konparaketa eta analisia bideratzea. Tesi lan honetan hain zuzen, PropBank eta VerbNeteko rol multzoen azterketa konparatiboa egiten dugu III atalean.

Nahiz eta, normalean, predikatuek aditz forma izaten duten eta, ondorioz, errekurtso konputazional gehienak aditzentzat egin diren, zenbait izen eta adjektibok ere forma predikatiboa har dezakete. Jarraian datorren esaldiko predikatua, adibidez, *gift* (oparia) da eta, ikus daitekeenez, hiru argumentuz lagunduta agertzen zaigu: *giver* (emailea), *thing given* (emandakoa) eta *entity given to* (nori eman zaion).

$$[_{giver}\ Her]\ \textbf{gift}\ of\ [_{thing-given}\ a\ book]\ [_{entity-given-to}\ to\ John]\ \ldots$$

Frameen semantikan aditzak ez diren elementu lexikoak barneratzen dira, esan bezala, egoera semantikoak adierazteko gaitasuna dutelako. Hurbilpen

sintaktikoak, aldiz, aditzen klaseetan oinarritzen dira soilik beste kategorietako osagaiek ez dutelako, agian, aditzek erakusten duten jokaera sintaktiko garbia erakusten.

Rol semantikoen gaineko teoria edo ikuspegi ezberdinen gainean eraikitako baliabide konputazionalen zerrenda eta deskribapena (tesi lan honetan erabiliko direnena zein besteena) tesi lan honetako II.1 atalean egiten da:

- Ikuspegi sintaktikoagoa duten baliabideak: PropBank (Palmer *et al.*, 2005) eta VerbNet (Kipper *et al.*, 2000)
- Frameen semantikan oinarritutakoak: FrameNet (Fillmore *et al.*, 2004)
- Predikatu ez berbalen baliabideak: NomBank (Meyers *et al.*, 2004).

I.2 Tesiaren motibazioa

Tesi honen ikergai nagusia ingeles hizkuntzarako rolen etiketatze automatikoa da. Esan bezala, konputazionalki behintzat oso gai berria den arren, lan eskerga egin da ataza honen inguruan ingeleserako. Euskararako baliabideen eraketari dagokionez, esan beharra dago Euskal Herriko Unibertsitateko IXA ikerketa taldeak horretan diharduela azken urteotan. Zoritxarrez, tesi lan hau aurkeztu dugun unean, baliabide horiek garapen prozesuan zeuden oraindik eta ezin izan ditugu euskararako prototipo baten eraketarako erabili. Hori dela eta, geure ikerlan guztia ingeleserako eta ingelesezko baliabideak soilik erabiliz egin dugu.

Aurreko atalean aurkeztu ditugun eztabaida linguistikoak apur bat albo batera utziz, rolak automatikoki etiketatzeaz hitz egiten dugunean, esaldi bat eta predikatua emanda, bere argumentuak identifikatu eta, predikatuarekin jokatzen duten rol semantikoaren arabera, etiketa bat edo beste esleitzeaz ari gara. Mota honetako predikatu-argumentu egiturak identifikatu eta etiketatzeari azaleko analisi semantikoa esaten zaio. Tesi lan honetan aditz predikatuekin egingo dugu lan.

Zergatik rol semantikoak? Zein da rol semantikoak etiketatzeko beharra? Sintaxia eta semantikaren arteko azaleko konexioa bezala ikus ditzakegu rolak, izan ere, sintaxiak harrapa ezin ditzakeen zenbait orokorpen semantiko detektatzeko erabil daitezke. Adibide gisa azter dezagun jarraian datorren esaldiaren argumentu egitura (iturria: Yih eta Toutanova, 2006):

[$_{Temp.}$ *Yesterday*], [$_{Agent}$ *John*] **hit** [$_{Patient}$ *Kevin*] [$_{Instrument}$ *with a hammer*]

Adibideko *hit* predikatuak adierazten duen ekintza lau argumenturekin gauzatzen da. Aditz argumentu bakoitzak predikatuarekin jokatzen duen paperaren arabera, rol bat du esleituta:

- Denbora (*Temp.*): ekintza noiz gertatu zen.
- Agentea (*Agent*): ekintza nork behartu zuen.
- Pazientea (*Patient*): ekintzak nor kaltetu zuen.
- Tresna (*Instrument*): ekintza burutzeko zein tresna erabili zen.

Jatorrizko esaldiaren alternantzia sintaktiko posibleak aztertuz argumentu egitura berbera partekatzen dutela ikus daiteke:

Yesterday, **John** (hit) Kevin with a hammer
Kevin was (hit) by **John** *yesterday* with a hammer
Yesterday, Kevin was (hit) with a hammer by **John**
With a hammer, **John** (hit) Kevin *yesterday*
Yesterday Kevin was (hit) by **John** with a hammer
The hammer with which **John** (hit) Kevin *yesterday* was hard
John (hit) Kevin with a hammer *yesterday*

Goiko aldaerek jatorrizko esaldiaren esanahi berdina dutela esan dezakegu nahiz eta bakoitzak patroi sintaktiko ezberdina jokatzen duen. Bariazio sintaktiko ezberdin horien gainetik, ordea, esaldi batetik bestera "hit" predikatuak azaleko analisi semantiko berbera erakusten du eta, horregatik, bere argumentuek, nahiz eta sintaktikoki modu ezberdinean gauzatzen diren, rol semantiko berberak jokatzen dituzte. Rol semantikoek, beraz, sintaxiaren gainean kokatzen den geruza semantikoa osatzen dute.

Sintaxiak rolen etiketatze automatikoan berebiziko garrantzia du, predikatuen argumentuek, aurreko adibideetako esaldiekin ikusi ahal izan dugun moduan, oso maiz sintagma motako egiturekin bat egiten baitute. Horrela, rolak etiketatzea, atazaren ikuspegi konputazionaletik behintzat, sintagmei etiketa semantikoak esleitzea da.

1. [$_{Agent}$ *John*] **broke** [$_{Theme}$ *the window*]
2. [$_{Theme}$ *The window*] **broke**

3. [$_{Agent}$ *Sotheby's*] **offered** [$_{Recipient}$ *the Dorrance heirs*] [$_{Theme}$ *a money-back guarantee*]
4. [$_{Agent}$ *Sotheby's*] **offered** [$_{Theme}$ *a money-back guarantee*] *to* [$_{Recipient}$ *the Dorrance heirs*]

Sintaxiak berebiziko garrantzia du predikatuen argumentuak detektatzeko garaian, esan bezala, argumentuek sintagmekin edo analisi zuhaitzeko nodoekin maiz egiten dutelako bat. Dena dela, rola ez dago sintagmaren posizio sintaktikora hertsiki lotuta, kasuaren arabera rol ezberdinak ikus baititzakegu funtzio sintaktiko jakin bati lotuta. 1 eta 2 esaldiek erakusten duten alternantzian subjektuak rol bat ala beste har dezake gauzatze sintaktikoaren arabera (*Agent* eta *Theme*), eta berdin 3 eta 4 esaldietan objektu zuzenarekin(*Recipient* eta *Theme*). Orduan, rol semantikoen etiketatze automatikoa ataza sintaktikoa al da soilik? Bistan denez ez, aurreko adibideekin ikusi dugun bezalaxe, sintaxiak ezin baitu, zenbait kasutan, sintagma bati dagokion rola zein den zehazten lagundu; aspektu lexiko-semantikoak kontuan hartzea ere beharrezkoa da. Ikus dezagun adibide batekin:

1. [$_{Patient}$ *JFK*] *was assassinated* [$_{LOC}$ *in Dallas*]
2. [$_{Patient}$ *JFK*] *was assassinated* [$_{TMP}$ *in November*]

Goiko bi esaldiek, nahiz eta analisi zuhaitz (eta preposizio) berbera partekatzen duten, rol semantiko ezberdinak jokatzen dituzten argumentuak gordetzen dituzte: [*in Dallas*] eta [*in November*] preposizio sintagmak. Argumentu horiek, barnean gordetzen dituzten osagai lexikalen eraginez, rol ezberdina jokatzen dute aditzarekiko esaldi batean eta bestean. Lehen argumentuak "*Dallas*" hiria adierazten duen hitza gordetzen du bere baitan eta, beraz, predikatuaren ekintza "non" gertatu zen adierazten du lekuzko rol batekin. Bigarrenak, aldiz, ekintza "noiz", zein hilabetetan gertatu den zehazten du eta denborazko rol semantiko batekin etiketatu da. Modu horretan nabarmen geratzen da rol semantikoen etiketatze automatikoan ere, noski, semantikak bere garrantzia ere baduela eta beharrezkoa dela argumentuek gordetzen dituzten osagai lexikoen "esanahia" kontuan hartzea adibidean erakutsitako kasua eta beste zenbait kasu arazotsu ebazteko.

Ezagutza semantikoa rolen etiketatze automatikoan

Azken urteotan eraiki diren sistemek mota askotako arkitekturak diseinatu dituzte etiketatze arazoa gainditzeko. Hala ere, definizioz, sistema guztiek funtsezko bi urrats egiten dituzte etiketatzea gauzatzeko: *argumentuen identifikazioa* eta *argumentuen sailkapena*. Lehenak, izenak dioen moduan, esaldia eta predikatua emanda, argumentuaren mugak identifikatzean datza, eta bigarrenak, identifikatutako argumentu horiei, aditzarekiko jokatzen duten rol semantikoaren arabera, etiketa bat edo beste esleitzean. Pradhan eta Martinek (2008) adierazi zuten, rolak etiketatzeko sistemek erabiltzen dituzten ezaugarriek eragin ezberdina izan zezaketela ataza batean edo bestean. Zehazki, beren esperimentuekin erakutsi zuten analisi zuhaitz sintaktikotik sistemek erauzten dituzten ezaugarriek batez ere *argumentuen identifikazio* faserako direla egokiak, baina *argumentuen sailkapen* prozesurako ez direla hain erabakigarriak. Horrela, ikusi zuten corpus jakin baten gainean eraikitako (entrenatutako) sistemen errendimendua nabariki kaltetua gertatzen zela domeinuz kanpoko corpusak etiketatzen zirenean, eta galera hori ez zegoela sintaxiari -eta beraz *argumentuen identifikazioari*- lotuta, baizik eta *argumentuen sailkapenari*.bezala. Sistemek domeinuz kanpoko corpusekin *argumentuen sailkapen* prozesuan izaten zituzten galerak, besteak beste, argumentuek bere baitan gordetzen zituzten osagai lexikoetan (argumentuen guneetan) zuten jatorria, domeinuz kanpoko corpus horietan osagai lexiko berri edota ezezagunarekin topo egiten zutelako eta, hortaz, ezin zituztelako hitz horiek argumentuen rola iragartzeko erabili. Honekin eta JFKren adibidearekin jarraituz, zer gertatuko litzateke etiketatze sistemak *Dallas* eta *November* hitzak ezagutuko[2] ez balitu? Ba osagai lexiko horiek ezingo liratekeela argumentuei rol zuzena esleitzeko erabili eta, hortaz, sistemek informazio edota ezaugarri sintaktikoekin soilik erabaki beharko luketela argumentu arazotsu horien rola. Horrela, bi hitzak ezezagunak izanda eta analisi sintaktiko berdina partekatuta, sistemek bi argumentuei rol berbera (zuzena ala ez) esleituko liekete ziurrenik.

Tesi lan honen gune garrantzitsuenean *argumentuen sailkapen* prozesuan eragin negatiboa duten mota honetako arazo lexikoei irtenbidea bilatzen saiatuko gara. Horretarako aztertu eta aplikatu ditugun teknikak hautapen murriztapenak eta hitzen arteko antzekotasun neurriak izan dituzte oinarri.

[2] *Dallas* eta *November* hitzak ezezagunak izango dira entrenamenduko corpusean agertu ez badira.

Ideiak funtsean honetan datza:

Rol semantikoak etiketatzeko sistemak gauzatzeko erabiltzen diren teknika estatistikoek edota ikasketa automatikoko teknikek eskuz etiketatutako baliabideak behar dituzte argumentuak identifikatu eta argumentu horiei zein rol esleitu behar zaien ikasteko. Ikasketa corpus horietan markatzen diren argumentuak eta, zehazkiago, haien guneak hainbat rolekin geratzen dira lotuta. Adibidez, ikasketa corpusaren barnean "*in November*" argumentua topatuko bagenu TMP rolaz lagunduta, ondorioztatuko genuke "*November*" izen guneak nolabaiteko lotura izan dezakeela denborazko rolekin. Eta berdin "*Dallas*" eta lekuzko rolekin. Baina non sartzen dira jokoan hautapen murriztapenak eta antzekotasun neurriak? Demagun orain "*in March*" argumentua topatu dugula ikasketa corpusetik kanpo. Nola jakingo du sistemak argumentu horri lekuzko rola edota denborazko rola esleitu behar dion ez badu "*March*" lehenago posizio horretan ikusi? Hautapen murriztapenenek, aipatzen genituen corpus ebidentziak oinarri hartuta, lekuzko eta denborazko argumentuen guneen itxura zein den esango digute. Adibidearekin jarraituz, lekuzko rolak izango dituzten guneak, "*Dallas*" ("*Paris*", "*heaven*", "*garden*"...) itxura izango dute eta denborazko rolekin lotutakoak aldiz "*November*" ("*yesterday*", "*August*", "*century*"...) itxura. Horrela "*in March*" argumentu berria sailkatzerakoan, WordNet edo antzekotasun distribuzionaleko neurriak erabiliz, bere gunea ("*March*") rol bakoitzaren hautapen murriztapenekin alderatu ahalko genuke, zorte apur batekin denborazko rola duten guneetatik lekuzko roletatik baino gertuago dagoela ondorioztatzeko.

Esan bezala, gure iritziz, hau izan da ekarpen nagusia baina ez bakarra. Ikus ditzagun bada, ekarpen hori eta gainerako guztiak detaile gehiagorekin.

I.3 Ekarpenak

Domeinuz kanpoko corpusetan rolak etiketatzeko sistemek erakusten dituzten arazoak aztertzeko helburuarekin hasi genuen tesi lan hau. Helburu handinahia zalantzarik gabe, izan ere, rol semantikoen etiketatze automatikoak ez ezik, lengoaia naturalaren prozesamenduko beste ataza ia guztiek (guztiek ez esateagatik) erakusten dituzte ahuleziak domeinuz kanpo lan egitea tokatzen zainean. Gai horri helduta, beraz, ikerlan asko eraman dira aurrera nahiz eta zoritxarrez (ala zorionez) inork ez duen domeinu aldaketaren handicapa erabat gainditzea lortu. Guk ere, noski, ez dugu rolak etiketatzeko sistemen domeinuz kanpoko errendimendu eskasa guztiz ebaztea lortu, baina arazoa

neurri batean ahuldu eta planteamendu eta teknika garbiak erabiliz pausoak aurrera egiteko gai izan gara.

Ikus ditzagun jarraian gure ustez, tesi lan honekin egin ditugun ekarpen nagusiak:

- **PropBank eta VerbNeteko rol multzoen analisi konparatiboa (III. kapitulua)**: bi rol inbentario konputazional hauen eskuragarritasuna aprobetxatuz eta esperimentu egokien laguntzaz, inbentario bakoitza erabiltzearen abantantailez eta desabantailez eztabaidatu dugu. Planteatu ditugun esperimentuek erakutsi dute baldintza zailetan PropBankeko rol multzoarekin lan egiten duen sailkatzaile batek baldintza oso antzekotan entrenatutako VerbNet sailkatzaileak baino sendotasun eta orokortzeko gaitasun handiagoa erakutsiko duela, baita domeinuz kanpoko corpusekin ere.

- **Hautapen murriztapenak eta antzekotasun metodoak rolen etiketatze automatikoan (IV. kapitulua)**: WordNeteko eta antzekotasun distribuzionaleko metodotan oinarritutako hautapen murriztapenak erabili ditugu informazio lexiko-semantiko soila erabiltzen duten sailkatzaileak eratzeko. Sailkatzaile hauek informazio lexikoa bakarrik erabiltzen duten sailkatzaileek baino errendimendu hobea erakusten dute eta, lexikoa orokortzeko ahalmenei esker, domeinu berrietara hobeto egokitzen dira.

- **Preposizio-rol eta aditz-rol motako hautapen murriztapenak (IV. kapitulua)**: bi hautapen murriztapen mota hauek aurkeztu eta konbinatu ditugu, eta rolak automatikoki etiketatzeko ohikoagoak diren aditz-rol motako hautapen murriztapen soilak baino hobeak izan daitezkeela erakutsi dugu zenbait esperimenturekin. Hautapen murriztapen konbinatuak aditz-rol motako hautapen murriztapenak baino egokiagoak dira argumentuak sailkatzeko ezagutza iturri bezala, eta emaitzak, hain justu, nabariki hobetzen dituzte kapitulu honetan proposatuko dugun ebaluazio esperimentalean.

- **Artearen egoeran dagoen sailkatzaile baten hobekuntza hautapen murriztapenekin (V. kapitulua)**: Preposizio-rol eta aditz-rol motako hautapen murriztapenak rolak etiketatzeko errendimendu altuko sistema batean integratzeko urratsak deskribatzen ditugu kapitulu honetan, emaitzak modu esanguratsuan hobetuz. Lehen aldia da gure ikerketa arloan ataza hau modu arrakastatsuan gauzatzen dela.

I.4 Tesi lanaren egitura

- **Lehen kapitulua** - *Sarrera*: sarrera hau.

- **Bigarren kapitulua** - *Rolen etiketatze automatikoa: artearen egoera.* Kapitulu honetan rol semantikoen etiketatze automatikorako erabilgarriak diren baliabideen azterketa egingo dugu. Horrez gain, rolak etiketatzeko sistema tipikoen arkitektura aztertuko dugu, maila bakoitzak planteatzen dituen arazoak zeintzuk diren eta normalean nola ebazten diren adieraziz. Jarraian, sistemek beren lana egiteko erabiltzen dituzten ezaugarri garrantzitsuenen errepasoa egingo dugu eta, bukatzeko, rol semantikoen etiketatzearen historia laburra eta ataza honek etorkizunean planteatzen dituen erronkak zein diren azalduko dugu.

- **Hirugarren kapitulua** - *PropBank eta VerbNeteko rol multzoen arteko azterketa konparatiboa.* Rolen etiketatze automatikorako erabil daitezkeen bi rol multzo ezberdinen azterketa konparatiboa egingo dugu kapitulu honetan, bi rol inbentarioen arteko diferentziak zehaztuz, eta bakoitzaren abantaila eta desabantaila teorikoak azpimarratuz. Bukatzeko, rolak etiketatzeko sistema bana eratuko dugu (bat rol multzo bakoitzeko) eta bien errendimenduen azterketa bat egingo dugu corpus/etiketatze baldintza ezberdinetan, domeinuz kanpo ere zein jokaera erakusten duten adieraziz. Horrez gain, VerbNeteko rol tematikoen "filosofia" aplikazio errealetarako egokiagoa dela kontuan izanda, rol tematiko hauek eskuratzeko bide sendoak (edo sendoagoak) proposatuko ditugu.

- **Laugarren kapitulua** - *Ezaugarri lexikalak orokortzen: Hautapen Murriztapenak rolen sailkatze automatikorako.* Kapitulu honetan rolak etiketatzeko sistemek ezaugarri lexikalekiko duten dependentzia izango dugu aztergai. Dependentzia hauen eragin negatiboa domeinu barneko eta domeinuz kanpoko corpusetan aztertu eta hautapen murriztapenen laguntzaz, arazoa leuntzeko proposamenak egingo ditugu.

- **Bosgarren kapitulua** - *Hautapen Murriztapenak Argumentuen Sailkapenerako.* Aurreko kapitulutan lexikoaren orokortzeari buruz ikasitakoa, *SwiRL* sisteman modu arrakastatsuan integratzeko bideak ikusiko ditugu, hautapen murriztapen bakoitzak isla dezakeen informazio semantikoa sailkatzaile eta meta-sailkatzaileen laguntzaz modu eraginkor

batean harrapatuz. Horrela, gure jakintza arloko lehenak izango gara artearen egoeran dagoen argumentu sailkatzaile bat hautapen murriztapenen erabilerarekin hobetzen.

- **Seigarren kapitulua** - *Ondorioak eta etorkizuneko lanak.* Tesi honen ondorio nagusiak bilduko ditugu atal honetan eta baita etorkizunerako rolen etiketatze automatikoan irekita ikusten ditugun ikerlerroak ere.

I.5 Publikazioak

Euskal Herriko Unibertsitateak sustatzen duen "Tesi Europarraren" iniziatiba eta eskakizunekin bat eginez, tesi honen ingelesezko bertsio laburtu bat prestatu dugu (`http://ixa2.si.ehu.es/~bzapirain002/summary.pdf`).

Bestalde, tesi lan honetan egindako ikerketa eta aurrerapenen berri nazioarteko hainbat kongresutan eman dugu. Hona hemen argitaratutako artikuluen zerrenda dagokien tesi kapituluaz lagunduta:

PropBank eta VerbNeteko rol multzoen arteko azterketa konparatiboa (III. kapitulua):

- Beñat Zapirain, Eneko Agirre and Lluís Màrquez. **A Preliminary Study on the Robustness and Generalization of Role Sets for Semantic Role Labeling**. *Computational Linguistics and Intelligent Text Processing.* 9th International Conference, CICLing 2008, Haifa, Israel. LNCS 4919. Springer-Verlag, 2008.

  Laburpena: VerbNet eta PropBankeko rol multzoen azterketa enpirikoa rolak etiketatzeko sailkatzaileen laguntzaz. Rol multzoen sendotasuna eta orokortzeko gaitasunaren konparaketa esperimentazio ingurune ezberdinetan.

- Beñat Zapirain, Eneko Agirre and Lluís Màrquez. **Robustness and Generalization of Role Sets: PropBank vs. VerbNet** *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics*, ACL-08: HLT, Columbus, Ohio, 2008.

  Laburpena: VerbNet eta PropBankeko rol multzoen azterketa enpiriko zabala corpus handiekin. Rol multzoen sendotasun eta orokortzeko gaitasunaren konparaketa esperimentazio ingurune ezberdinetan.

VerbNeteko rol tematikoak PropBankeko argumentu zenbakituetatik abiatuta lortzeko teknikak.

Ezaugarri lexikalen orokortzen: hautapen murriztapenak rolen sailkatzen automatikorako (IV. kapitulua):

- Beñat Zapirain, Eneko Agirre and Lluís Màrquez. **Sequential SRL Using Selectional Preferences: An Approach with Maximum Entropy Markov Models**. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval 2007)*, pp.354-347, Prague, Czech Republic, Association for Computational Linguistics, 2007.

  Laburpena: WordNeten oinarritutako hautapen murriztapenak rolak etiketatzeko sistema sekuentzial batean integratzeko lehen saiakera.

- Beñat Zapirain, Eneko Agirre and Lluís Màrquez. **Generalizing Over Lexical Features: Selectional Preferences for Semantic Role Classification**. *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2009)*, Singapore, 2009.

  Laburpena: WordNeten eta antzekotasun distribuzionaleko tekniketan oinarritutako hautapen murriztapenen multzo zabal baten deskribapena. Hautapen murriztapenen konparaketa aditz argumentuak sailkatzeko ataza batean.

Hautapen murriztapenak argumentuen sailkapenerako (V. kapitulua):

- Beñat Zapirain, Eneko Agirre, Lluís Màrquez and Mihai Surdeanu. **Improving Semantic Role Classification with Selectional Preferences**. *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, Los Angeles, California, 2010.

  Laburpena: aditz-rol eta preposizio-rol moduko hautapen murriztapen multzo zabal baten deskribapena eta aplikazioa. Hautapen murriztapenen integrazio arrakastatsua artearen egoeran dagoen sailkatzaile batean.

II. KAPITULUA

Rolen etiketatze automatikoa: artearen egoera

Azken urteotan aurrerapauso handiak egin dira lengoaia naturalaren prozesamenduko arlo gehienetan, baina urrun gaude oraindik esaldien esanahi zehatza adierazteko gai izatetik. Puntako teknika estatistikoak eskuz etiketatutako corpus masa handien gainean aplikatzeari esker, baditugu dagoeneko esaldien analisi sintaktiko erabilgarriak egiten dituzten analizatzaile sintaktiko (*parser*) sendoak. Hala ere, *parser* hauek ez dira gai esaldi baten parte hartzaile semantikoak modu egoki batean identifikatu eta haien arteko erlazio semantikoak ezagutzeko. Ikus dezagun adibide sinple batekin:

Kevin broke the window[1]
The window broke[2]

Sintaxiak argi esango digu lehen esaldian *the window* osagaia aditzaren objektua dela, eta subjektua bigarrenean; baina ez da gai izango objektu eta subjektu horiek aditzarekiko jokatzen duten rol semantikoa bi adibideetan berdina dela adierazteko.

[1] Kevinek leihoa puskatu zuen.
[2] Leihoa puskatu zen.

Analizatzaile sintaktikoek azaleko semantikarekiko erakusten dituzten gabeziak gainditzeko berebiziko garrantzia izan zuten semantikoki etiketatutako baliabideen agerpenak. FrameNet (Fillmore *et al.*, 2004) eta PropBank (Palmer *et al.*, 2005) bezalako corpus semantikoen sorrerak rolak automatikoki etiketatzeko prototipoen ugaritzea ekarri du azken bost urteotan. Gildea eta Jurafsky-k (2002) beren "Automatic Labeling of Semantic Roles" lan aitzindarian FrameNeten oinarritutako SRL sistema automatiko baten deskribapena egin zutenetik, asko izan dira LNP munduko ikerlariak beren interesa eta esfortzua gai zirraragarri eta erlatiboki berri honetara bideratu dutenak. Pizkunde honetan berebiziko garrantzia izan zuten CoNLL-2004 (Carreras eta Màrquez, 2004) eta, batez ere, CoNLL-2005ean (Carreras eta Màrquez, 2005) egin ziren ebaluazio saioek. Hurrengo azpiataletan sakonago aztertuko ditugun saio haiek PropBanken entrenatutako hogeita hamar bat SRL sistema eta haien arteko analisi konparatiboa eskaini zioten LNP-ko ikerlari komunitateari eta, oraindik ere, ingeleserako SRL sistema berrien errendimendua ebaluatu eta konparatzeko erreferente izaten jarraitzen dute.

Kapitulu honen helburua rolen etiketatze automatikoa ikergaiaren errebisio global bat egitea izango da. Inolako zalantzarik ez dago ikerkuntzan ekarpenak egiteko eman beharreko lehen urratsa ikergai konkretu baten uneko egoera orokorra ezagutzea dela, eta horri ekingo diogu hurrengo ataletan zehar.

Kapitulu honen egitura ondokoa da: rolen etiketatze automatikoan erabiltzen diren baliabide semantikoak aztertuko ditugu hasteko II.1 atalean. Zehazki, FrameNet, VerbNet eta PropBank baliabideak ikusiko ditugu, baina azken honetan zentratuko dugu gure atentzioa batik bat, PropBank baita, askogatik, rolen etiketatze automatikoko sistemak entrenatzeko gehien erabili den baliabidea. SRL sistemei dagokienez, II.2 atalean oinarrizko sistema orokor baten deskribapena egingo dugu. Besteak beste, sistemek erabiltzen dituzten ikaste estrategia eta arkitekturen errepaso bat egingo dugu, eta baita ohikoak diren ikasketa ezaugarriena ere. Errendimendu altuko sailkatzaileen adibideak ere jarriko ditugu sailkatzaile arrakastatsuenek jarraitzen dituzten bideak/estrategiak azaltzeko asmoz; eta bukatzeko, II.4 atalean, CoNLL-2005eko ebaluazio saiora aurkeztu ziren sistemen analisi kualitatibo bat egingo dugu, erakutsi zituzten ahulezia nabarmenetan arreta ipiniz eta tesi lan honetan oinarri izan diren ikerlerroak identifikatuz eta justifikatuz.

II.1 Rol semantikoak eta baliabideak

Hiru dira ingeleserako rol semantikoekin etiketatuta aurki daitezkeen baliabide ezagunenak: PropBank (Palmer *et al.*, 2005), VerbNet(Kipper *et al.*, 2000) eta FrameNet(Fillmore *et al.*, 2004). Baliabide edo inbentario hauek, bakoitzak bere filosofiari jarraiki, rol semantikoen etiketatze automatikoan erabilgarriak diren rol eskemak definitzen dituzte, bakoitzak bere ezaugarriekin, bere ahulezia eta abantailekin. Eskema erabiliena PropBank da, bost roletako eskema sinple bat definitu eta ikasketa automatikoa eta teknika estatistikoak aplikatzeko oso egokia den corpus bat semantikoki etiketatzen duelako. VerbNetek Levinen (Levin, 1993) klaseetan oinarritutako aditzen hierarkia bat planteatu eta klase bakoitzeko aditzek joka ditzaketen 23 rol tematikotako eskema orokor eta independente bat definitzen du formalki, baina nahiz eta rolen erabilerari buruzko adibide erreal asko ematen diren, ez dira sistema automatikoak entrenatzeko nahikoak nahiz eta, aurrerago ikusiko dugun moduan, SemLink bezalako errekurtsoekin gabezia hau gaindi daitekeen. FrameNetek bizitza errealean gerta daitezkeen hainbat ekintza edo gertaera semantikoki deskonposatzen ditu *frame elements* izeneko rol semantiko multzo zabal batekin eta, VerbNetek bezala adibide erreal asko etiketatuta ematen dituen arren, ez du, berari dagokion azpiatalean azalduko dugun arrazoiengatik, rol semantikoak entrenatzeko ezagutza iturri gisa arrakasta handirik lortu.

II.1.1 Proposizioen bankua: PropBank

PropBank corpusa Penn Treebank IIko (Marcus *et al.*, 1994) egitura sintaktikoen gainean eraikitako geruza semantikoa da. Zehazkiago, treebankeko Wall Street Journal[3] sekzioko aditzak (milioi bat hitz guztira) predikatu-argumentu egiturekin eta rol semantikoekin hornitzen ditu. Horretarako erabiltzen duen rol multzoa "teorikoki neutrala" dela esaten da, definitzen diren rolek ez dutelako teoria linguistiko jakinik jarraitzen. Horrela, aditz guztien argumentuak etiketatzeko beti rol multzo berbera erabiltzen den arren, aditzaren arabera rol-label bakoitzaren interpretazio semantikoa ezberdina izan daiteke (beren esanahia aditzarekiko menpekoa da beraz).

[3]WSJko testuez gain, Brown corpusa ere etiketatzen ari dira.

ArgM-LOC: lekuzkoa	**ArgM-CAU**:kausa
ArgM-EXT: hedadura	**ArgM-TMP**: denbora
ArgM-DIS: diskurtso markatzailea	**ArgM-PNC**:helburua
ArgM-ADV: adberbiala	**ArgM-MNR**: moduzkoa
ArgM-NEG: ezeztapen marka	**ArgM-DIR**: norabidea
ArgM-MOD: modua	

Irudia II.1: PropBankeko rol adjuntuak.

Frameset *accept.01* "take willingly"
Arg0: Acceptor
Arg1: Thing accepted
Arg2: Accepted-from
Arg3: Attribute
Ex:[Arg0 He] [ArgM-MOD would][ArgM-NEG n't] *accept*
[Arg1 anything of value] [Arg2 from those he was writing about].

Irudia II.2: PropBankeko accept.01 *frameset*a eta gauzatze adibide bat (Ex:). *Roleseta* osatzen duten core argumentuak (adibidean Arg0-Arg3) gorriz agertzen dira goitik behera, lerroz lerro.

Diseinua

PropBankeko rol multzo teorikoki neutrala *core-argument* izeneko argumentu zenbakituek osatzen dute (Arg0, Arg1, ..., Arg5). Aditz baten erabilera bakoitzak onartzen duen *core* argumentu multzoari *roleset* esaten zaio. *Roleset* bakoitza hainbat *frame* sintaktikorekin lotuta egon daiteke baldin eta *frame* sintaktikoak rol multzo horrekin bateragarriak diren bariazio sintaktikoak badira. *Rolsetak* eta harekin lotutako *frame*ek *frameset* delako egitura osatzen dute. Aditz polisemiko batek *frameset* bat baino gehiago izan ditzake baldin eta adieren arteko aldea rol multzo ezberdinak behar izateko bezain handia bada. *Frameset*ean, hori guztiaz gain, *core* argumentu bakoitzaren deskribapena ere egiten da egituratu gabeko hizkuntza librean eta, normalean, rol multzoak baimendutako alternantzia sintaktikoen adibideak ere ematen dira.

*Roleset*etan deskribatzen diren *core* argumentuez gain, PropBanken argu-

Frameset *kick.01* "drive or impel with the foot"
Arg0: Kicker
Arg1: Thing kicked
Arg2: Instrument (defaults to foot)
Ex1: [ArgM-DIS But] [Arg0 two big New York banksi] seem [Arg0 *trace*i] to have kicked [Arg1 those chances] [ArgM-DIR away], [ArgM-TMP for the moment], [Arg2 with the embarrassing failure of Citicorp and Chase Manhattan Corp. to deliver $7.2 billion in bank financing for a leveraged buy-out of United Airlines parent UAL Corp].

Irudia II.3: PropBankeko kick.01 *frameset*a eta gauzatze adibide bat (Ex:). *Roleseta* osatzen duten core argumentuak (adibidean Arg0-Arg2) gorriz agertzen dira goitik behera, lerroz lerro.

Frameset *kick.03* "kick in"
Arg0: Contributor
Arg1: Contribution
Arg2: Given to
Ex1: [Arg0 Mary] [ArgM-MNR gamely] *kicked in* [Arg1 $5] to [Arg2 John's bail].

Irudia II.4: PropBankeko kick.03 *frameset*a eta gauzatze adibide bat (Ex:). *Roleseta* osatzen duten core argumentuak (adibidean Arg0-Arg2) gorriz agertzen dira goitik behera, lerroz lerro.

mentu adjuntu orokorrak ere etiketatzen dira bakoitza dagokion etiketarekin (ikusi rol adjuntuen etiketak II.1 irudian). Etiketa hauek, *core* rolak adierazteko erabiltzen direnak ez bezala, aditzarekiko independenteak dira eta, hortaz, haien esanahia ez da aditzarekin aldatzen.

II.2 eta II.3 irudietan PropBankeko bi *frameset*en adibideak ikus daitezke. Lehena *accept.01* (onartu) aditzaren erabilerari dagokiona da eta bigarrena, *kick.01* (ostikoz jo) aditzaren erabilerari dagokiona. *Frameset* bakoitzaren *roleset*a osatzen duten argumentuak gorriz agertzen dira lerroz lerro beren deskribatzaileaz lagunduta. *Core* rolek aditzarekiko erakusten duten dependentzia ongi ikus daiteke *roleset* bakoitzeko argumentuen deskribapenari erreparatuta. Arg0 eta Arg1 rolek, nahiz eta hurrenez hurren "*acceptor*"-"*kicker*" eta "*accepted thing*"-"*thing kicked*" bezala definitzen diren, agente eta tema rol prototipikoekin bat egiten dutela susma daiteke bai *frameset* batean eta bai bestean. Ez da antzekotasun semantikorik ikusten, ordea, Arg2 rolarekin, aditzaren arabera *accepted from* eta *instrument* interpretazioak har baititzake. Rol zenbakituen arteko diferentziak aditz lema

beraren adiera ezberdinen artean ere gerta daitezke. II.4 irudian adibidez, "*to kick*" aditzaren hirugarren PropBank adieraren roleseta ikus daiteke, eta bertan zehazten diren rolen esanahiak ez dute II.3 irudiko deskribatzaileekin bat egiten nahiz eta aditz lema berbera etiketatzeko erabiltzen diren.

PropBanken, nolabait esateko, frameset batek aditzaren "PropBank adiera" islatzen du. PropBank adiera hauek ez dira zertan bat etorri behar aditz jakin baten balizko adiera xeheagoekin, frameseten arteko bereizketa ez baita aspektu semantikoetan soilik oinarritzen eta sintaxia ere kontuan hartzen delako. Horrela, gaur egun, 3,300 aditz inguru tratatzen dira PropBanken. Framesetak ("PropBank adierak") 4,500 inguru dira, eta horrek batez beste, 1.36ko polisemia dakar.

Corpusa

PropBankek Penn Treebankeko nodo sintaktikoei frameseteko rol semantikoak (eta rol adjuntuak) esleitzen dizkie. Rol semantiko bat ez da zertan nodo bakarrera lotuta agertu behar, kasuaren arabera, aditz argumentua nodo batek baino gehiagok osa dezaketelako. PropBank corpusa proposizioen kodeketa bat da, hau da, corpuseko lerro bakoitzean aditz bat eta bere argumentuen posizioak kodetzen dira treebankean duten nodo-posizioaren arabera. Nodo sintaktiko (edo argumentu) horietako bakoitzari dagokion rola esleitzen zaio. Ikus dezagun corpuseko adibide lerro bat:

```
brown/cf/cf01.mrg 13 24 encounter.01 -
16:1*21:0*25:0-ARG1 22:1-ARG0 23:0-ARGM-MOD 24:0-rel 26:1-ARGM-LOC
```

Goiko lerroa PropBankeko proposizio baten kodeketa adibidea da. Zehazki, Brown corpuseko "cf" sekzioko "cf01.mrg" fitxategiko "13". sententziaren analisi zuhaitzari dagokiona. Analisia II.5 irudian ikus daiteke. Sententzia bakarrean aditz bat baino gehiago eta, hortaz, beste hainbeste proposizio aurki ditzakegu. Adibidean kodetzen den proposizioaren gunea (aditza) "24". nodoan dago kokatuta, eta jokatzen duen frameseta "encounter.01" da. II.5 irudian ondo ikusten da nodoak nola kodetzen diren. Nodoak 0tik aurrera zenbatzen dira eta zuhaitz sintaktikoko hostoekin (urdinez irudian) egiten dute bat. Posizioaz gainera, nodoaren sakoneraren beharra dago nodo hostoaren gurasoei ere erreferentzia egin ahal izateko. Horrela, "6:0" identifikatzaileak, 6. nodo hostoari egingo dio erreferentzia 0 sakoneran (irudian `(NN course)`) eta "6:1" identifikatzaileak berriz, nodo horren gurasoari (`(NP (DT the) (NN course))`). Adibideko proposiziora itzuliz, kodeketak

```
( (S-CLF
    (NP-SBJ (PRP It) )
    (VP (VBD was)
      (ADVP-LOC-PRD (RB there) )
      (, ,)
      (PP (IN in)
      (NP
        (NP (DT the) (NN course) )
      (PP (IN of)
        (S-NOM
          (NP-SBJ-2 (-NONE- *) )
          (VP (VBG trying)
            (S
              (NP-SBJ (-NONE- *-2) )
              (VP (TO to)
                (VP (VB prepare)
                  (NP (JJ new) (NNS men) )
                  (PP (IN for)
                    (NP
                      (NP (DT the) (`` ``) (NN culture) (NN shock) (" ") )  16:1-ARG1
                      (SBAR
                        (WHNP-1 (-NONE- 0)  *21:0-ARG1 )
                        (S
                          (NP-SBJ (PRP they) )   22:1-ARG0
                          (VP (MD might)   23:0-ARG1
                            (VP (VB encounter)   24:0-REL
                              (NP (-NONE- *T*-1) )   *25:0-ARG1
                              (PP-LOC (IN in)
                                (NP (JJ remote) (JJ overseas) (NNS posts) )) 26:1-ARGM-LOC
))))))))))))))
```

Irudia II.5: Penn Treebankeko sententzia baten analisi zuhaitza.

dio 16:1[4] nodoan ARG1 rola duen argumentua daukagula (`(NP (DT the) (‘‘ ‘‘) (NN culture) (NN shock) (’’ ’’))`), 22:1 nodoan ARG0 osagaia (`(NP-SBJ (PRP they))`), 23:0 nodoan moduzko argumentu adjuntu bat (`(MD might)`), 24:0 posizioan proposizioko aditza (`(VB encounter)`) eta 26:1 posizioan lekuzko adjuntua (`(PP-LOC (IN in) (NP (JJ ...`

[4]*21:0 eta *25:0 nodoak zuhaitzeko trazei dagozkienak dira. Trazak Penn Treebankeko kategoria hutsak dira eta normalean zuhaitzeko beste osagai bati lotuta doaz (bi traza hauek 16:1 nodoarekin daude konektatuta). SRL sailkatzaileek ez dituzte etiketatzen eta CoNLL-2005 bezalako txapelketetan besterik gabe ezabatu egiten dira. II.5 irudiko analisi zuhaitzean bi traza hauek markatuta agertzen dira gorriz dagokien nodoan.

Frameset head.06
Arg0: Job holder
Arg1: Theme
Arg2: Beneficiary
Ex1: [ArgM-DIS But] [Arg0 **head**] [Arg1 of stock investments] [Arg2 wfor Cigna Corp].

Irudia II.6: NomBankeko head.06 roleseta.

NomBank

PropBankek aditzen argumentu egiturak etiketatzen dituen bezala, Nombank[5] proiektuak izenen argumentu egiturak identifikatu eta markatzeko irizpideak ezartzen ditu, hitzen hiztegi morfologikoak eraikiz eta, PropBanken antzera, hitzek onar ditzaketen rolak zehazten dituzten frameak definituz.

Jarraian datozen adibideetan ikusten den bezala, WSJko izen predikatuen (aditzen nominalizazioak etab.) argumentu egiturak markatzea du helburu NomBank proiektuak.

- $[_{Arg0}$ *house*$]$ $[_{REL}$ *debate*$]$
- *the* $[_{Arg0}$ *parliamentary*$]$ $[_{REL}$ *debate*$]$
- the $[_{ArgM-MNR}$ growing$]$ $[_{REL}$ *debate*$]$ $[_{ArgM-LOC}$ *in Washington*$]$
- *the quality of* $[_{REL}$ *debate*$]$ $[_{ArgM-LOC}$ *in Washington*$]$

NomBankeko izenen frameak PropBankeko aditz frameen antzera definitzen dira, XML erabiliz eta predikatu adiera bakoitzeko onartzen den rol multzoa adibideen laguntzaz definituz. II.6 irudian NomBankeko "head.06" predikatuarentzako NomBanken definitzen den *roleset*a:

"head.06" *roleset*ean, adiera horretarako, "head" izenak onartzen dituen argumentu eta rolak zehazten dira: Arg0 (lanpostu burua), Arg1 (tema) eta Arg2 (onuraduna).

NomBank CoNLL-2008 (Surdeanu *et al.*, 2008) eta CoNLL-2009 (Hajič *et al.*, 2009) ebaluazio saioetan erabili den errekurtsoa izan da eta, hortaz,

[5] `http://nlp.cs.nyu.edu/meyers/NomBank.html`

badira izen argumentuetarako iragarpenak ere egiten dituzten rolak etiketatzeko sistemak. Tesi lan honetan, dena dela, aditz predikatuetan zentratu gara soilik.

II.1.2 FrameNet

Berkeley FrameNet[6] (Fillmore *et al.*, 2004) proiektuak ingeles hizkuntzarako baliabide lexiko baten eraketa du helburu duela urte batzuetatik hona. Izen, adjektibo eta aditzen inguruko informazioa bildu eta haietako (adiera) bakoitza *frame semantiko* delako egituretara lotzen ditu corpus ebidentziez lagunduta. Frame semantiko bat ekintza, gertaera, edota objektu bat eta haren parte hartzaile guztiak deskribatzen dituen egitura kontzeptuala da.

FrameNet datu baseak frame semantikoen eta unitate lexikoen (hitz adieren) deskribapenak gordetzen ditu haien errepresentazio sintaktiko eta semantiko posibleez (*valences*) lagunduta. Frame semantikoek eskematikoki errepresentatzen dituzte egoerak (Fillmore, 1976), haien parte hartzaileak eta joka ditzaketen rol kontzeptualak. FrameNeten aurki dezakegun frame bat, adibidez, **apply-heat** delakoa da. Frame semantiko honek erreferentzia egiten dioten *unitate lexiko* jakin batzuek gordetzen ditu bere baitan (*bake, boil, brown*[7] etab.) eta, jarraian datozen adibideetan ikusiko dugun moduan, `COOK` (sukaldari), `FOOD` (janari), eta `HEATING INSTRUMENT` rolak edo *frame elementuak* baliatuz, egoera edo prozesu baten deskribapena egiten du.

[$_{COOK}$Matilde] **fried** *[$_{FOOD}$ the catfish] [$_{HEAT.-INSTR.}$in a heavy iron skillet].*

Goiko adibidean, *apply-heat* frame semantikoaren instantzia bat ikus daiteke etiketatuta. Frameari erreferentzia egiten dion unitate lexikoa (kasu honetan aditza) *fried* (frijitu) da, eta gauzatzen diren frame elementuak edo rolak goian aipatu ditugun berberak dira (II.1).

FrameNeten frame elementuak frame semantikoekiko menpekoak dira eta horregatik frame elementuen kopurua izugarri handia da. Horri eskuz etiketatutako adibide kopuru eskasa gehitzen badiogu, rolen sakabanaketa handia eta adibide etiketatuen urritasuna dela eta, zailtasunak topa ditzakegu baliabide honi teknika estatistikoak aplikatzeko garaian. Hala eta guztiz ere, datu basea handituz doa eta ez da baztertu behar, noski, etorkizunean FrameNeten

[6] `http://framenet.icsi.berkeley.edu/`

[7] *Bake*, *boil* eta *Brown* elementu lexikoen adierak euskaraz, labean egin, egosi eta txigortu dira hurrenez hurren

Frame elementua	Deskribapena eta Adibidea
`CONTAINER`	Janaria (FOOD) berotzeko erabiltzen den edukiontzia. Adib.: **boile** *potatoes* [$_{CONTAINER}$ *in a medium-sized pan*]
`COOK`	Janaria berotzeko erabakia hartzen duena. Adib.: [$_{COOK}$ *Kevin*] **sauteed** *the garlic in butter*
`FOOD`	Sukaldariak (COOK) berotzen duen entitatea. Adib.: *Suzy usually* **steams** [$_{FOOD}$ *the broccoli*]
`HEATING INSTRUMENT`	Janaria berotzeko erabiltzen den tresna. Adib.: *Jim* **browned** *the roast* [$_{HEAT-INST.}$ *in the oven*]

Taula II.1: FrameNeteko *Apply-heat* frame semantikoa jokatzen duten frame elementuen (rol) deskribapenak eta adibideak. Adibideetan agertzen diren frameko unitate lexikoak (*lexical units*) letra lodiz agertzen dira markatuta.

oinarritutako sailkatzaile lehiakorrak eraikitzeko aukera, izan ere, gogora dezagun, Gildeak eta Jurafsky-ren lan aitzindarian (Gildea eta Jurafsky, 2002) rol multzo honekin lan egiten zuen sailkatzaile baten deskribapena ematen da.

Gaur egun FrameNetek, 10,000 unitate lexiko gordetzen ditu. Haietatik 6,000, 800 bat frame semantikoetan aurki ditzakegu, 135,000 adibidez lagunduta.

II.1.3 VerbNet

VerbNet aditzen lexikoi konputazionala da non portaera sintaktiko eta semantiko berbera duten aditzak hierarkikoki antolatutako klase sistema batean antolatzen diren. Klase hauek Levinen aditz klaseetan oinarritzen dira eta ezaugarri komunak dituzten aditzak eta haien argumentuak gordetzeaz gain, aditzen informazio sintaktikoa eta semantikoa ere ematen dute. Argumentu multzoa klaseko aditzek jokatzen dituzten rol tematikoen eta haien gainean definitutako hautapen murriztapen orokorren zerrenda da. Informazio sintaktikoak rol tematikoak argumentu sintaktikoetara lotzen ditu, eta predikatu semantikoek frame sintaktikoak zehazten duen gertaerako (*event*) parte hartzaileak fase bakoitzean (*start*, *during*, *end* edo *result*) deskribatzen ditu.

Gaur egun, VerbNetek 274 klase nagusi ditu. Guztira 3769 aditz (5257 adiera) biltzen dituzte, haien jokaera 23 rol tematikorekin deskribatuz. Ikus ditzagun VerbNeteko hierarkiako klaseen osagaiak banan-banan:

Rol tematikoak

VerbNeteko klaseetan definitzen diren aditz argumentuak rol tematikoekin etiketatzen dira. VerbNeteko rolek PropBankeko argumentu zenbakituek baino informazio semantiko aberatsagoa ematen dute. Argumentu zenbakituekin gertatzen ez den bezala, rol tematikoen interpretazio semantikoa ez da aditzaren menpekoa eta, horregatik, PropBankeko framesetetan ikusi dugun moduan, VerbNeten ez da beharrezkoa aditz klase bakoitzeko rolen esanahia behin eta berriz definitzea. Guztira, helburu orokorreko 23 rol tematiko definitzen dira:

- **Actor:** komunikazio ekintzak adierazten dituzten klaseetan erabiltzen dira, bi argumentuak (objektu/subjektu) semantikoki simetrikotzat har daitezkeenean (*Actor1*, *Actor2*).
- **Agent:** oro har, bizidun subjektuak dira baina makinak edota borondatea erakusten duen edozer izan daiteke.
- **Asset:** ordainketak eta antzeko diru trukaketak edota aldaketak adierazten dituzten klaseek erabiltzen dute. Rolak "moneta" *currency* hautapen murriztapena du.
- **Attribute:** aldatzen ari den zerbaiten ezaugarria.
- **Beneficiary:** ekintzetatik etekina ateratzen duen entitatea.
- **Cause:** aditz psikologikoekin eta gorputzeko aditzekin lotutako klaseak erabiltzen dute.
- **Location, Destination, Source:** kokapen espazialerako erabiliak
  - **Location:** helburu ezezaguna, jatorria edo lekua adierazteko.
  - **Destination:** norantz jakina edo mugimenduaren amaiera puntua.
  - **Source:** mugimendu baten hasiera puntua.
- **Experiencer:** kontziente den, edota zerbait esperimentatzen ari den parte hartzailea adierazteko.
- **Extent:** klase bakarrean erabiltzen da, indizeen bariazioa edota aldaketa baten gradua adierazten du.

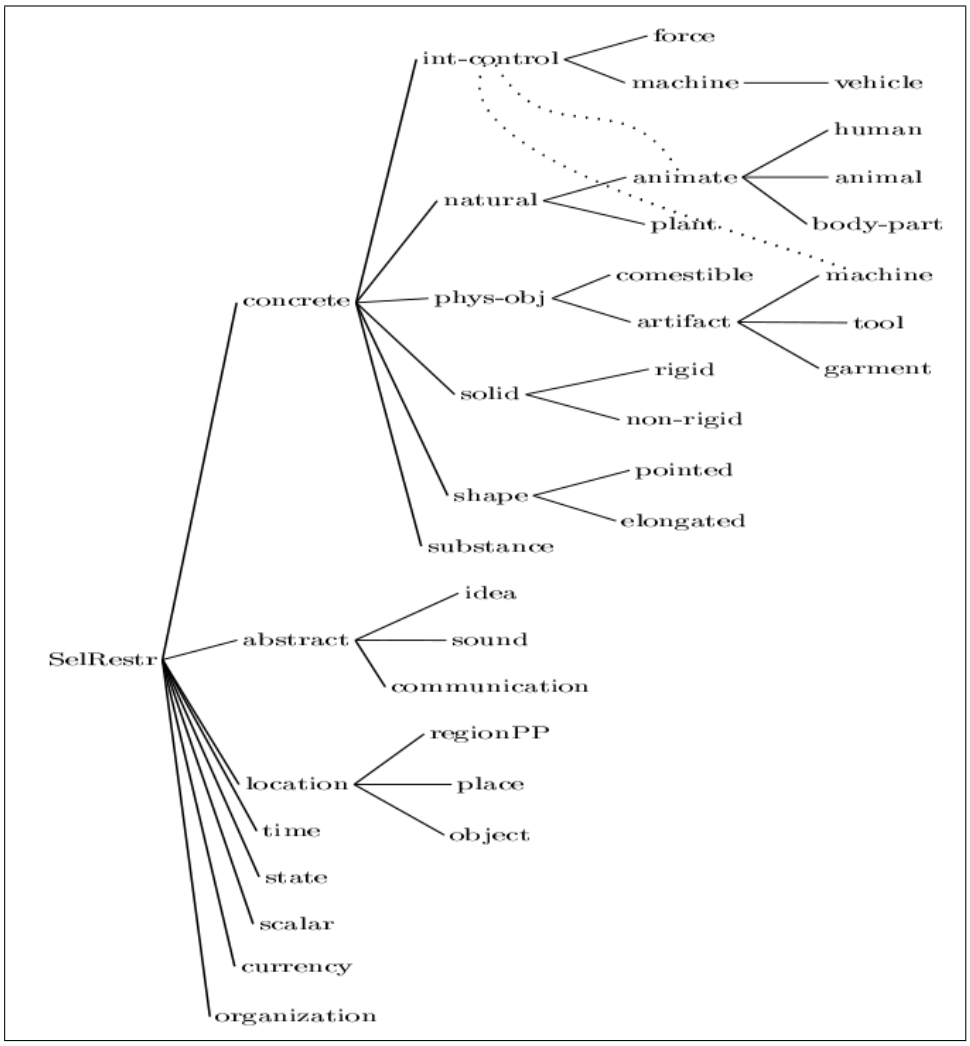
- **Instrument:** tresna bezala erabiltzen diren objektuak. Aldaketak eragiten dituzten objektuak.
- **Material and Product:** *Build* eta *Grow* klaseetan erabiliak.
  - **Material:** eraldaketa baten hasiera puntua.
  - **Product:** eraldaketa baten emaitza.
- **Patient:** modu batera edo bestera ekintza baten ondorioak jaso dituzten parte hartzaileak.
- **Predicate:** osagai predikatiboa erabiltzen duten klaseetan erabilia.
- **Recipient:** transferentzia baten helburua.
- **Stimulus:** *Experiencer* parte hartzaileetan nolabaiteko erantzuna eragiten duena.
- **Theme:** posizio batean edo posizio aldatze baten parte hartzaileak.
- **Time:** *Begin-55.1* klasean denbora adierazteko.
- **Topic:** komunikazio aditzek "hizpide" darabiltena.

VerbNeteko hautapen murriztapenak

VerbNeteko hautapen murriztapenak EuroWordNeteko (Vossen *et al.*, 1997) goiko mailetako kontzeptuetan daude oinarrituta (ikusi II.7 irudia) eta hierarkikoki antolatzen dira, batez ere *is-a* moduko erlazioak baliatuta (WordNeten bezalaxe). Hautapen murriztapenak VerbNeteko zenbait klasetako rol tematikotara lotzen dira modu bitarrean. Horrela, rol tematikoek *[+location,-region]* moduko hautapen murriztapenak izan ditzakete adibidez, rol horrekin etiketatutako aditz argumentuak lekuzko argumentu bat -baina ez lurralde eremu bat- adierazi behar duela zehazteko.

Frame sintaktikoak

VerbNeteko frame sintaktikoek klaseko aditzek baimenduta dituzten errealizazioak (gauzatzeak) deskribatzen dituzte lengoaia sinple eta labur batean.



Irudia II.7: VerbNeteko hautapen murriztapenak (iturria: Kipper *et al.*, 2000).

Klaseko kideen gauzatze iragankorrak (*transitive*) deskribatzen dituzte adibidez, gauzatze iragangaitzak, baimendutako preposizio sintagmak, eta baita Levinek deskribatutako hainbat alternantzia ere.

Frame sintaktikoak rol tematikoen nolabaiteko hautapen murriztapen sintaktikoak dira, rol bakoitzak gustukoen dituen posizio sintaktikoak definitzen dituelako klaseko aditzek baimenduta duten gauzatze bakoitzerako. Horrez gain, aditzaren posizioa zehazten dute rol tematikoekiko eta baita aditzaren

zenbait gauzatze edo alternantzietarako beharrezkoak diren partikula lexikoak eta haien propietateak. Hona hemen frame sintaktiko batzuen adibideak:

1. *Agent V Patient*

   `Kevin hit the window`

2. *Agent V at Patient*

   `Kevin hit at the window`

3. *Agent V Patient[+plural] together*

   `Kevin hit the sticks together`

Lehen adibidean bi rol tematikok eta aditz batek osatutako frame sintaktiko sinplea ikus daiteke. Bigarren adibideko frameak "at" preposizioak gobernatutako preposizio sintagma bat baimentzen du VerbNet klase jakin bateko aditzentzat eta, hirugarrenean, *Patient* rol tematikoan plural kasua eskatzen du bukaeran *together* elementu lexikoa duten gauzatzeetan. Nahiz eta tesi honetan ez diegun VerbNeteko frame sintaktikoei etekinik aterako, badirudi rolen sailkatze automatikoko sistemetan baliagarria gerta daitekeen informazio sintaktiko/semantikoa gordetzen dutela.

Predikatu semantikoak

VerbNeten aditzen informazio semantikoa predikatu semantikoen bidez deskribatzen da (adibidez: *motion* (mugimendua), *contact* (kontaktua), *transfer-info* (informazio trukatzea)...). Predikatu semantikoak lau multzotan banatzen dira:

- **Predikatu orokorrak**: *motion* edota *cause* bezalako predikatuak barneratzen dituzte. Klase gehienetako aditzentzako osagai generikoak dira, baita hizkuntza ezberdinetan ere.

- **Predikatu aldakorrak**: *Pred*, *Prep*, eta *Adv* dira multzo honetako predikatu semantikoak. Beraien esanahia modu unibokoan dago lotuta hizkuntzako hitz multzo jakin bati.

- **Predikatu zehatzak**: aditz adiera berezi batekin lotutako predikatuak. Adibidez, *suffocate* (sua itzali) predikatua *Suffocate-40.7* VerbNet klaseko aditzetan agertzen da soilik.

- **Gertaera askotarako predikatuak**: predikatu hauek gertaeren arteko erlazioak adierazteko erabiltzen dira. Adibidez, *throw a ball across the room*[8] esaldiko *throwing* ebentuaren ostean, pilotaren (*ball*) posizio aldaketa bat gertatzen da. Gertaeren arteko erlazioak (bitarrak guztiak) honako hauek izan daitezke: *Before, After, Meets, Overlaps, Starts, During, Finishes...*

Argumentu semantikoetako predikatuak honako hauek dira:

- ***Event***: gertaera (*event*) adierazten duen E aldagaia edo bere parte bat (*During(E), Before(E)* ...) izan daiteke.

- ***Constant***: Ezaugarri konkretuak dituzten predikatuetan erabiltzen diren argumentuak. Adibidez, *manner* predikatuan *forceful* eta *directedmotion* argumentuak.

- ***ThemRole***: Frame sintaktikoan baimendutako rol tematiko multzoa.

- ***Verb Specific***: klase bereko aditzek modu ezberdinean instantziatzen dituzten argumentuen multzoa.

VerbNet klase baten adibidea

II.2 irudian VerbNeteko *Trasnfer_mesg-31.1.1* (mezua transferitu) klasearen laburpen bat ikus daiteke[9]. Klasean barneratutako aditzak (*Members*), besteak beste, *communicate*, *corroborate*, *demonstrate*, *elucidate* eta *explain*[10] dira. Aditz klase horrek baimentzen dituen frame sintaktikoetako aditzek joka ditzaketen rol tematikoak *Agent*, *Topic*, *Recipient* eta *Source* dira. *Agent* eta *Recipient* argumentuak `ANIMATE` edota `ORGANIZATION` motakoak izan behar dira, alegia "objektu animatu" edota "organizazio" modukoak. Taulako "Frameak" lerroan hiru frame sintaktikoren adibideak jarri dira, klaseko kideen hiru gauzatze sintaktiko ezberdin deskribatzen dituztenak. Predikatu semantikoak, besterik gabe, "informazio transferentzia" gertaera (E)

[8] Pilota gelatik zehar bota.

[9] `http://verbs.colorado.edu/verb-index/vn/transfer_mesg-37.1.1.php`

[10] Aditzen esanahiak euskaraz, komunikatu, baieztatu, frogatu, argitu eta azaldu dira.

Klasea	Transfer_mesg-37.1.1		
Guraso nodoa	–		
Kideak	communicate, corroborate, demonstrate, elucidate, explain...		
Rol tematikoak	Agent Topic Recipient Source		
Hautapen murrizt.	Agent[+ANIMATE |+ORGANIZATION] Recipient[+ANIMATE |+ORGANIZATION]		
Frameak	Izena	Sintaxia	Predikatu semantikoak
	NP V how S	Agent V Topic	transfer_info(during(E), Ag., ?Rec., Top.) cause(Ag., E)
	NP V PP	Agent V from Source	transfer_info(during(E), Ag., ?Rec., Top.) cause(Ag., E)
	NP V	Agent V	transfer_info(during(E), Ag., ?Rec., Top.) cause(Ag., E)

Taula II.2: *Transfer_mesg-37.1.1* VerbNet klasearen errepresentazioa.

Agenteak sortzen duela dio (*cause(Ag., E)*). Horrez gain, transferentzia hori gauzatzen diren rol tematikoen artean gertatzen dela zehazten du, gertaerak irauten duen bitartean (*during(E)*).

SemLink

VerbNet, izatez, errekurtso lexikoa da soilik eta ez du, PropBank eta FrameNetek egiten duten bezala, corpus baliabide sendorik eskaintzen rol tematikoekin etiketatuta. Gabezia horri irtenbidea bilatzeko aukera interesgarri bat SemLink [11] (Loper *et al.*, 2007b) mapaketa corpusak ematen digu. Baliabide honek PropBankeko frameseten eta VerbNeteko rol tematikoen arteko mapaketa bat definitzen du. Zehazki, PropBankeko framesetetan zerrendatzen den argumentu zenbakitu bakoitzari dagokion rol tematikoa esleitzen dio eta, gainera, frameseteko aditzaren adierari ere dagokion VerbNet klasea ematen dio (ikusi II.8).

Argumentu zenbakitu eta rol tematikoen mapaketa honen laguntzaz, SemLink eratu zuen taldeak PropBankeko proposizio guztien %50 etiketatu zuen gutxi gorabehera VerbNeteko rol tematikoekin. Eta bi rol multzo ezberdin horiekin etiketatutako corpus hori hartuko dugu, hain justu, PropBank eta VerbNeten arteko azterketa konparatiboa egiteko abiapuntutzat.

[11] https://verbs.colorado.edu/semlink/

(1)	(2)	(3)	(4)	(5)
wsj/00/wsj_0001.mrg	0 8	join.01;VN=22.1-2		0:2-ARG0[Agent] 7:0-ARGM-MOD 8:0-rel 9:1-ARG1[Patient1] 15:1-ARGM-TMP
wsj/00/wsj_0002.mrg	0 16	name.01;VN=29.3		16:0-rel 0:2*17:0-ARG1[Theme] 18:2-ARG2[Predicate]
wsj/00/wsj_0001.mrg	1 10	publish.01;VN=26.4-1		10:0-rel 11:0-ARG0[Agent]

Irudia II.8: SemLink etiketatzearen adibidea. Lerro bakoitzak aditz bat eta bere argumentuen etiketatzea ematen digu core roletan eta rol tematikotan. (1) Treebankeko fitxategiaren izena, sekzioaz lagunduta. (2) Sententzia eta Terminala: sententzia zenbakiak fitxategiko zenbatgarren esaldia den esaten digu; terminal zenbakiak aditza kokatzen du esaldiko nodo sintaktiko edo terminalen artean 0 tik hasita. (3) Aditza. (4) Aditzaren PropBank framea eta VerbNet klasea. (5) Esaldiko aditzaren eta argumentuen identifikazioa eta etiketatzea. Aditza "rel" bezala agertzen zaigu, eta core argumentuak urdinez eta haien rol tematiko baliokidea kortxete artean. Argumentu bakoitzak analisi zuhaitzean hartzen duen posizioa adierazten du haien aurretik datorren terminal zenbakiak.

II.1.4 Beste hizkuntzetarako errekurtsoak

Nahiz eta ingelesa den rolak etiketatzeko sistemak eraikitzeko errekurtso semantiko gehien dituen hizkuntza, gainontzeko hizkuntzak ere ari dira, batzuk besteak baino azkarrago, beren errekurtso propioak eta sistemak (Hajič *et al.*, 2009) eraikitzen. Besteak beste, badira **txinera** helburu duten sistema lehiakorrak (Xue eta Palmer, 2005; Sun *et al.*, 2009), eta **gaztelera** eta **katalana** helburu dutenak (Màrquez *et al.*, 2007). Datozen urteetan, aurrera egiten ari den Euskal PropBanken (Aldezabal *et al.*, 2010) garapenarekin batera, euskararentzako lehen etiketatzaile semantikoak garatzeko asmoa ere badugu. **Euskarari** dagokionez, zehazki, IXA taldea Euskal PropBank proiektuarekin, *Basque Dependency Treebank* delako corpusa etiketatzen ari da PropBankeko predikatu-argumentu eskema egokitu batekin. Corpus hau 300,000 hitz inguruz osatuta dago eta haietatik, dagoeneko, maiztasun altuko 62 aditzen proposizioak aztertu eta etiketatu dira (42,000 hitz inguru) Aldezabalen (2004) ehun euskal aditzen gaineko azterketa sakona oinarri hartuta.

Jarraian beste hizkuntzetarako garatu diren PropBank moduko errekurtsoen zerrenda ematen da:

- Chinese PropBank

  http://verbs.colorado.edu/chinese/cpb/

- Korean PropBank

  http://www.ldc.upenn.edu/

- AnCora corpus: Spanish and Catalan

  http://clic.ub.edu/ancora/

- Prague Dependency Treebank: Czech

  http://ufal.mff.cuni.cz/pdt2.0/

- Penn Arabic TreeBank: Arabic

  http://www.ircs.upenn.edu/arabic/

Aipatzekoa da FrameNeterako errekurtsoak ere ari direla beste hizkuntzetarako garatzen (http://framenet.icsi.berkeley.edu/).

II.2 Rol Semantikoen Etiketatzaile automatiko baten arkitektura

Orain arte, rolen etiketatze automatikoaren inguruan egindako lan nagusiaren zati handi bat ikasketa automatikoan eta metodo probabilistikoen laguntzaz egin da. Atal honetan rolen etiketatze automatikoaren hurbilpen *gainbegiratuen* errepaso orokor bat egingo dugu, maizen erabiltzen den hurbilpena delako eta arrakasta izan duelako gaia jorratzen duten ikerlarien artean. Azter ditzagun, bada, rolak automatikoki etiketatzeko sistema tipiko horien arkitekturak eta ikasketa ezaugarri garrantzitsuenak.

II.2.1 Rolen etiketatze automatikoa urratsez urrats

Rol semantikoen etiketatze automatikoaren helburua esaldi bat eta p predikatua emanda, p-ren argumentu guztiak identifikatu eta jokatzen duten rol semantikoaren arabera etiketa bat edo beste esleitzea da. Sistema gehienek lau urrats nagusi bereizten dituzte beren arkitekturan:

1. **Inausketa**: sistemen ikasketa errazteko helburuarekin, p predikatuaren argumentu izan ez daitezkeen hitz konbinazioak baztertzean datza.

2. **Argumentuen identifikazioa**: esaldia eta p predikatua emanda, p-ren aditz argumentu kandidatuak esaldiko sintaxiaren laguntzaz identifikatzea.

3. **Argumentuen sailkapena**: identifikazio fasean detektatutako argumentu kandidatuei rol etiketa bat esleitzea aditzarekiko jokatzen duten rol semantikoaren arabera.

4. **Puntuazio globala (*joint scoring*)**: sailkatutako argumentuekin, predikatuaren argumentu egitura orokor egoki bat sortzea.

Urrats hauek rolak etiketatzeko sistema baten eginbeharrekoak islatzen ditu baina, noski, egile eta sistemaren arabera urrats batzuk soberan egon daitezke, besterik gabe jarraitutako estrategia zehatzak ez duelako urrats horren beharrik edota, besterik gabe, ez delako, kasuan kasu, urrats konkreturen bat inplementatzea erabaki. Dena dela, esan bezala, hauek dira inplizituki edo esplizituki sistema gehienek jarraitzen duten arkitektura. Ikus dezagun gertuagotik.

Inausketa

Aditzaren argumentuak jarraituak edo ez-jarraituak izan daitezke. Argumentu jarraituak elkarren ondoan dauden hitzen segida batez osatuta daude eta ez-jarraituak berriz, elkarri jarraian ez dauden hitzen segida biz edo gehiagoz. Ondoren datorren esaldian argumentu jarraien eta ez-jarraien adibideak ikus ditzakegu:

[One troubling aspect of DECs results]$_{Arg1}$*, [analysts]*$_{Arg0}$ **said***, [was its performance in Europe]*[12]$_{C-Arg1}$

Adibide esaldiko predikatua letra lodiz markatuta dagoen "*to say*" (esan) aditza da. Aditz horrek PropBankeko Arg0 eta Arg1 rolekin markatuta dauden bi argumentu ditu, lehena jarraia eta bigarrena, Arg1 eta C-Arg1 etiketekin leku ezberdinetan markatuta dagoena, ez-jarraia.

Rol semantikoen etiketatzearen arazoa, besterik gabe, hitz kateei rol semantikoak esleitzea bezala plantea daiteke, baina, noski, esaldiko hitz kateen

[12]Esaldiaren esanahia euskaraz honako hau da: "DECeko emaitzen aspektu arazotsu bat, esan zuen analistak, Europan izan duen errendimendua da".

konbinazioa hain handia izanda eta kontuan hartuta, gainera, argumentu ez jarraituak ere baditugula, soluzio bideraezintzat jotzen da argumentuen bilaketa espazio handiegi (hitz konbinazio gehiegi) eta desorekatu bat (konbinazio gehienek ez dute argumentuen mugekin bat egiten) suposatuko lukeelako. Horrela, beharrezkoa da hitz kate posibleen (edo argumentu kandidatuen) kopurua txikitzea, argumentuekin bat egingo ez duten hitz kateak filtratu eta ikaste-etiketatze arazoa samurtzeko. Filtratze edo inausketarako esaldiaren zuhaitz sintaktikoan oinarritzen diren erregela heuristikoak erabili oi dira, izan ere, PropBank corpusean behintzat, argumentuen %95ak bat egiten duelako nodo sintaktiko batekin (analisi sintaktiko automatikoa erabiltzen denean, argumentuen %90ak).

Inausketarako erabiltzen den heuristiko ezagunena Xue eta Palmerrek (2004) proposatzen dutena da. Honela dio:

- **Lehen urratsa**: predikatua uneko aldagaitzat hartu eta zuhaitz sintaktikoan bere maila berberean dauden osagai sintaktikoak bildu, koordinatuta dauden horiek izan ezik.

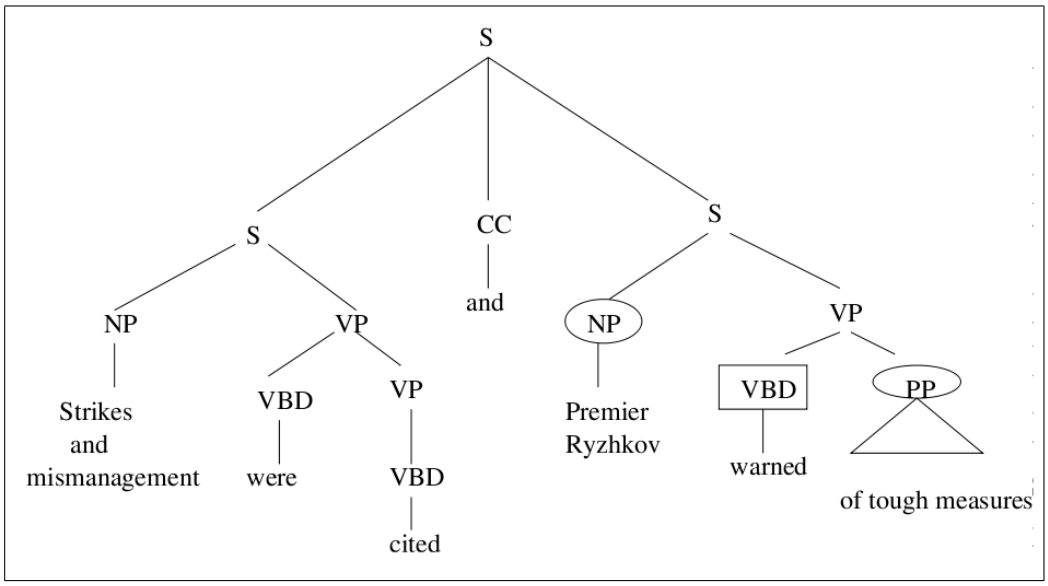
- **Bigarren urratsa**: uneko aldagaia bere gurasoarekin eguneratu, eta errepikatu lehen urratsa, goiko mailako nodo nagusira iritsi arte.

II.9 irudian esaldi koordinatu baten analisi zuhaitza ikus daiteke. "*To warn*" aditzaren (`VBD`) gainean bi urratsetako heuristikoa aplikatuz gero, biribilduta ikus daitezkeen nodoak hautatuko genituzke soilik bere argumentu kandidatu bezala (eta beraz, esaldiko beste hitz kate posible guztiek ez lukete iragazkia gainditukо)

Argumentuen identifikazioa

Inausketa iragan duten argumentu kandidatuak oraindik eta gehiago filtratzeko egiten den urratsa da argumentuen identifikazioa. Bere helburua, beraz, analisi zuhaitzeko nodoei modu definitiboan "argumentu" edo "ez argumentu" estatusa ezartzea da eta, horregatik, ezaugarrien ingeniaritzari dagokionez, normalean, ezaugarri sintaktikoekin entrenatzen diren sailkatzaile bitarrak (`Arg` vs. `Null`) erabiltzen dira inausketa urratsean detektatutako argumentu kandidatuak benetan argumentuak diren ala ez erabakitzeko.

Argumentuen identifikazio prozesuaren albo ondorio onuragarri bat argumentu kandidatuen kopurua murriztea lortzen duela da. Honek eragin zuzena izango du rolak etiketatzeko sistema osoaren errendimenduan eta baita eraginkortasunean ere.



Irudia II.9: Inausketa algoritmoaren aplikazioa esaldi koordinatu batean (iturria: Xue eta Palmer (2004)).

Argumentuen sailkapena

Argumentuen sailkapen prozesua aurreko iragazpen guztiak gainditu dituzten argumentuek (modu independentean) rol posible bakoitza jokatzeko duten probabilitatea zehaztean datza. Horrela, funtzio probabilistiko bat edota ikasketa automatikotik eratorritako puntuazio funtzio bat erabiliz, argumentu kandidatu bakoitzeko rol semantiko posibleen distribuzio probabilistiko bat edo puntuazio multzo bat lortuko dugu.

Argumentuen sailkapena egiten duten sailkatzaileek argumentua edota argumentu horretatik erauzten diren ezaugarriak emanda, n rol posibleetarako distribuzio probabilistikoa ematen dute (sailkatzaile n-tarrak dira). Horrela, arg argumentu kandidatua eta sailkapenerako ezaugarriak erauzteko $feat$ funtzioa emanda, honakoa betetzen da[13]:

[13]Sailkapena egiteko hautatutako metodoa probabilistikoa ez denean (ikasketa automatikoko algoritmoak adibidez), ez da distribuzio probabilistikoa itzultzen, pisuak baizik. Dena dela, pisu hauek, probabilitateetara erraz bihur daitezke behar izanez gero, *softmax* (Bishop, 1995) funtzioaren laguntzaz, adibidez

$$\sum_{r_i \in Roleset} P(r_i|feat(arg)) = 1 \tag{II.1}$$

non *Roleset* multzoa, corpusak baimentzen dituen rolak diren (PropBank corpusaren kasuan adibidez,{`Arg0, Arg1, ..., AM-TMP, AM-LOC ...`})

Ataza honen alderik garrantzitsuenetariko bat ezaugarrien ingeniaritza da. Beharrezkoa da ezaugarri multzo aberats eta esanguratsu bat definitzea ataza modu arrakastatsu batean aurrera eramateko. Puntu honetan, berebiziko garrantzia dute ezaugarri sintaktikoek baina, argumentuen identifikazioan ez bezala, Pradhan eta bestek (2008) argitzen duten moduan, ezaugarri semantikoak funtsezkoak eta garrantzitsuagoak izan daitezke (argumentuaren gunea, aditzaren lema...).

Argumentuen identifikazioa eta sailkapena bi pausotan egin beharrean urrats bakarrean ere integra daitezke ikasi beharreko n rol etiketei "null" etiketa gehituz (argumentu kandidatua argumentua ez dela esateko aukera kontuan hartzeko). Hala ere, bi urrats hauek modu banatuan egiteak bere abantailak ditu, izan ere, sailkatzailearen errendimendu orokorra hobetzeko egokia da ataza bakoitzerako ezaugarri multzo ezberdin bat definitzea, ataza batean eragin positiboa duten ezaugarriek kontrako eragina izan dezaketelako bestean (Pradhan *et al.*, 2005b). Horrez gain, lehenago ere esan dugun bezala, argumentuen identifikazio eta sailkapen prozesuak bereiztea rolak etiketatzeko sistemaren efizientziaren hobetzea dakar, identifikazio fasean argumentu kandidatuen multzoa txikitzeak sailkapen prozesu astunagoaren arintzea baitakar.

Puntuazio globala

Rolak etiketatzeko sistema tipikoek egiten duten laugarren urratsa puntuazio globala edo *joint scoring* delakoa da. Ataza honen helburua argumentuen sailkapen prozesuan esaldiko argumentu bakoitzerako egindako iragarpen lokalak modu trinko batean konbinatzea da, bukaeran esaldiaren baleko argumentu egitura sendoa emateko. Urrats honetan predikatu beraren argumentuen arteko dependetziak, murriztapenak eta beste zenbait aspektu linguistiko ustia daitezke. (Koomen *et al.*, 2005) eta (Punyakanok *et al.*, 2004b) lanetan adibidez, esaldiko etiketatze globalak bete beharreko hainbat murriztapen definitu eta betearazten dira:

- Debekatu gainjarritako edo habiratutako argumentuak.
- Core argumentuak ezin dira errepikatuta azaldu.
- $R{-}rol$ erreferentziazko argumentu rola badugu, orduan rol argumentua ere egotea beharrezkoa da.
- $C - rol$ jarraipen argumentua badugu, orduan rol argumentua egotea beharrezkoa da. rol argumentua gainera, $C - rol$ argumentuaren aurretik azalduko da.
- Predikatuaren arabera, zenbait rol debekatuta egongo dira. Horrela, PropBankeko rolekin ari bagara, "*stalk*" (zelatatu) aditzak adibidez, soilik Arg0 eta Arg1 core argumentuak izango ditu baimenduta.

Beste egile batzuek rolak etiketatzeko oinarrizko sistema batek ematen dituen n soluzio globaletan bilaketak egin eta emaitza hobeak lortzeko saiakera arrakastatsuak egin dituzte. (Toutanova *et al.*, 2005), (Haghighi *et al.*, 2005) eta (Toutanova *et al.*, 2008b) lanetan adibidez, soluzio globalen gaineko *re-ranking*a egiten da SRL sistema batek sortutako n soluzioen artean hoberena aukeratzeko.

Eredu probabilistikoak ere erabili izan dira irteera egituratua lortzeko. Adibidez: eredu generatiboak (Thompson *et al.*, 2003), etiketatzaile sekuentzialak (Màrquez *et al.*, 2005; Pradhan *et al.* 2005c), eta *Conditional Random Fields*-en aplikazioa zuhaitz egituretan (Cohn eta Blunsom, 2005).

Sistemen konbinazioa

Rolak etiketatzeko sistemen arkitektura orokor honen aspektu garrantzitsu bat sistemen konbinazioa da. Sistema gehienek konbinazioaren abantailak erabiltzen dituzte sendotasunean hobetu, estaldura handitu eta analisi sintaktikotik datozen erroreen eragina ahalik eta gehien gutxitzeko. Sistema ezberdinen arkitekturak aztertu eta gero, konbinaziorako ohikoak diren osagaiak identifika ditzakegu:

1. Ezaugarri bereziak dituzten rolak etiketatzeko sistema ezberdinen irteerak (Pradhan *et al.*, 2005c; Surdeanu *et al.*, 2007)

2. Etiketatze sistema beraren hainbat irteera, sarrerako anotazioa (analisi sintaktiko ezberdinak batez ere) edota ikasketa algoritmoko barne parametroen aldaketatik lortzen direnak (Koomen *et al.*, 2005; Toutanova *et al.*, 2005).

Konbinazioa, puntuazio funtzio globalen laguntzaz, esaldiko etiketatze (edo argumentu egitura) posibleen artean onena aukeratzea bezain sinplea izan daiteke. Askotan, soluzio osoak baino, haien zati fidagarrienak konbinatzen dira emaitza bezala esaldiaren etiketatze osoa lortzeko helburuarekin. Konbinazio sistemaren konplexutasunaren arabera beharrezkoa izan daiteke ikasketa automatikoa erabiltzea. Dena dela, konbinazio estrategia sofistikatu eta eraginkorrenak barneratzen dituzten sistemek, 2-3 puntutako hobekuntzak lortzen dituzte normalean, nahiz eta esfortzu konputazional handiaren eraginez sistemen efizientzia orokorra kaltetuta gertatzen den.

Salbuespenak

Orain arte aurreko azpiataletan deskribatu dugun arkitektura tipikoa sistema gehienek eta lehiakorrenek jarraitzen dutena da. Hala ere, arkitektura horretan, egileen arabera, bariazioak aurki ditzakegu azaldutako fase batean ala gehiagotan, eta baita sistemaren etiketatze estrategia orokorrean ere. Horrela, dagokion atalean aipatu bezala, sistema batzuek argumentuen identifikazioa eta sailkapena urrats bakarrean integratzen dute (Màrquez *et al.*, 2005; Surdeanu *et al.*, 2007 laneko "model 3" sistema).

Beste hurbilpen batzuek, adibidez, sententziako predikatu guztiak batera etiketatzen dituzte, predikatuen arteko dependentzia posibleei etekina ateratzeko asmoz. Hala ere, hurbilpen hauen konplexutasuna handia da eta emaitzak ez dira erabakigarriak (Carreras *et al.*, 2004; Surdeanu *et al.*, 2007). Badira hurbilpenak ere sintagmetan oinarritutako sintaxi tradizionalaren ordez dependentzietan oinarritutako analisi sintaktikoa oinarri hartzen dutenak (Johansson eta Nugues, 2007), eta baita analisi sintaktikoa eta rolen etiketatzea urrats batean egiten dutenak ere (Musillo eta Merlo, 2006; Merlo eta Musillo, 2008; Surdeanu *et al.*, 2008; Hajič *et al.*, 2009).

II.3 Ezaugarrien ingeniaritza

Argumentu kandidatuak errepresentatzen dituzten ezaugarriak aurreikustea eginkizun garrantzitsua da rolak etiketatzeko sistema baten eraketa proze-

suan. Aditza eta argumentu kandidatua (≈ nodo sintaktikoa) emanda, hiru motatako ezaugarriak erauzten dira orokorrean argumentuen sailkapen prozesuari aurre egiteko:

1. Argumentu kandidatuaren eta bere testuinguruaren ezaugarriak adierazten dituztenak.
2. Aditz predikatua eta bere testuingurua adierazten dutenak.
3. Argumentu kandidatuaren eta aditzaren arteko erlazio sintaktiko edota semantikoak jasotzen dituztenak.

Gildea eta Jurafskyk (2002) bere lanean definitu zuten rol multzoa gaur egungo sistema moderno gehienek barneratzen dute. Oinarrizko ezaugarri hauek modu honetan egiten dute bat goian aurkeztu dugun sailkapenarekin:

1. Sintagma mota, kandidatuaren izen-gunea eta kategoria.
2. Aditzaren lema, boza eta aditzaren azpikategorizazio patroia.
3. Argumentu kandidatuaren posizioa aditzarekiko ("aurretik" edo "ondoren"), eta bien arteko loturan aurki daitezkeen sintagmen zerrenda (*syntactic path*).

Ezaugarrien ingeniaritza landu duten ikerketek Gildea eta Jurafskyren ezaugarri multzoa ezaugarri esanguratsu berriekin edota existitzen zirenen barianteekin hornitu dute. Ezaugarri multzo handiekin lan egiteko ahalmena duten ikasketa automatikoko algoritmoak erabiliz, zenbait egilek argumentu kandidatua eta bere testuinguruaren adierazpena ezaugarri berriekin hornitu dute. Ezaugarri horietatik aipagarrienak honako hauek dira: argumentuaren lehen eta azken hitzak (eta haien kategoriak), hitz zakuak (argumentu kandidatuan ager daitezkeen adjektiboena, izenena...), eta kandidatua definitzen duten goiko mailetako elementu sintaktikoen zerrenda. Kandidatuaren guraso eta arrebetatik ere orain arte azaldutako ezaugarriak erauzteko saiakerak ere egin dira (Pradhan *et al.*, 2005a; Surdeanu *et al.*, 2007). Beste egile batzuek, aldiz, motibazio linguistiko zehatzak jarraituz ezaugarri berriak ere definitu dituzte. (Surdeanu *et al.*, 2003) lanean adibidez, argumentu kandidatuaren gunearen *content word* delako kontzeptu orokorrago bat eta entitate labelak ere gehitzen dituzten ezaugarri multzoan. Xue eta Palmerrek (2004) frame sintaktiko (*syntactic frame*) ezaugarria diseinatu zuten aditzak eta

argumentuak esaldi osoko egitura sintaktikoan kokatzeko. Ezaugarri hauek guztiek oinarrizko ezaugarri multzoarekiko hobekuntza esanguratsuak ekarri dituzte.

Uneko osagaia eta predikatuaren arteko loturari dagokionez, Gildea eta Jurafskyk proposatutako *syntactic path* delakoaren bariante ugari proposatu dira hainbat lanetan (sakabanaketa gutxitzeko orokorpenak egitea, sintaxi partziala erabiltzea osoa beharrean etab.). Tesi lan honen ekarpenetako bat kandidatua eta aditzaren arteko lotura semantikoa adierazten duten "bateragarritasun" ezaugarrien definizioa (Zapirain *et al.*, 2007; Erk, 2007; Zapirain *et al.*, 2009) eta aplikazioa (Zapirain *et al.*, 2010) da. Ohikoa da ere, orain arte ikusi ditugun ezaugarrien arteko konbinazioak egitea, batez ere, ikasketarako erabilitako metodoak linealak direnean.

Puntuazio orokorra eta konbinazioaz arduratzen diren moduluek ezaugarri multzo aberats bat definitzeko aukera handiagoa dute soluzio kandidatu osoak eta argumentuen arteko dependentziak hartzen baitituzte kontuan. (Toutanova *et al.*, 2008a) lana norantz honetan egindako lan aipagarrienetariko bat da. Egileek soluzio oso hoberena hautatzeaz arduratuko den *ranker* modulua entrenatzeko kandidatuaren argumentu egitura kodetzen duten patroiak erabiltzen dituzte *string*-ekin. Ezaugarri hauen gainean egiten dituzten aldaerek (sakabanaketa gutxitzeko zenbait orokortzerekin) oinarri sistema baten emaitzak modu esanguratsuan hobetzen dituzte. (Pradhan *et al.*, 2005c) eta (Surdeanu *et al.*, 2007) lanetan, oinarrizko SRL sistemen fidagarritasunak kodetzen dituzte ikasketa automatikoan oinarritutako konbinazio sistema bat trebatzeko. Tesi lan honetan antzeko teknika bat erabiltzen dugu V.2.2 kapituluan.

Rolak etiketatzeko sistemek erabili oi dituzten ezaugarri ezagunen deskribapen zehatza egiten da III.2.2 atalean.

II.4 Rol semantikoen etiketatze automatikoaren historia laburra

Semantikoki etiketatutako corpusen agerpenarekin batera, rolak etiketatzeko lehen sistema automatikoak agertu ziren. Gildeak eta Jurafskyk (2002) beren sistema probabilistiko aitzindaria FrameNet ereduan eraiki zutenetik asko izan dira beren sistema propioa garatzeko pausoa eman dutenak. Rolen etiketatze automatikoan iraultza ekarriko zuen CoNLL-2004 (Carreras eta

Màrquez, 2004) eta CoNLL-2005 (Carreras eta Màrquez, 2005) ebaluazio saioak baino lehen, baziren dagoeneko bidea finkatzen hasitako sistema bakan batzuk:

- Aipatutako Gildea eta Jurafskyren lanen antzera, eredu probabilistiko garbiekin eraikitako sistemak (Gildea eta Palmer, 2002; Gildea eta Hockenmaier, 2003)
- Entropia Maximoan oinarritutakoak (Fleischman eta Hovy, 2003)
- Eredu sortzaileetan oinarritutako ereduak (Thompson *et al.*, 2003)
- Erabaki zuhaitzak (Surdeanu *et al.*, 2003, Chen eta Rambow, 2003)
- *Support Vector Machines* delakoak (Hacioglu eta Ward, 2003, Pradhan *et al.*, 2003)

Beste hainbat egilek beren esfortzuak sistemek zuhaitz sintaktiko osoekiko erakusten zituzten dependentziak ahultzera bideratu zituzten (Hacioglu eta Ward, 2003; Pradhan *et al.*, 2003. Bide horretatik jarraitu zuen CoNLL-2004 ebaluazio saioak. Bertan, PropBank corpusarekin lan egiten zuten 10 sistemek hartu zuten parte, eta etiketatzea egiteko sintaxi partzialak eskaintzen dituen baliabide sintaktikoak soilik erabili zituzten. Parte hartzaileek etiketatze ataza modu askotara planteatu zuten, ikasketa automatikoko osagai eta etiketatze estrategia ezberdinak erabiliz. Hala ere, aurretik ezagutzen ziren sistemekin alderatuta, ikasketa algoritmoei dagokienez behintzat ez zen berrikuntza nabaririk ikusi.

Taldeetako bik, entropia maximoa erabili zuten beren sistemak entrenatzeko (Baldewein *et al.*, 2004; Lim *et al.*, 2004). Beste bi taldek (Higgins, 2004; Williams *et al.* 2004) Brill-en errore bidez gidatutako transformazioa erabili zuten, eta beste bik, memorian oinarritutako ikasketa (van den Bosch *et al.*, 2004; Kouchnir, 2004). Gainontzeko lau taldeek, bektoretan oinarritutako ikasketa metodoak erabili zituzten: SVM (Hacioglu *et al.*, 2004; Park *et al.*, 2004), perzeptroia (Carreras *et al.*, 2004) eta SNoW (Punyakanok *et al.*, 2004a).

Etiketatze estrategiei dagokienez, CoNLL-2004 ebaluazio saioko parte hartzaileek laburbilduz honako hauek jarraitu zituzten. Lehen hurbilpenak rolen identifikazioa zuzenean BIO sekuentzien etiketatzearen barnean burutzen du. Bigarrenak, arkitektura tipikoan erakutsi dugun bi fasetako estrategian bezala, *identifikazio* eta *sailkapen* prozesutan banatzen du ataza; eta hirugarrenak, *iragazketa* eta *etiketatzean*.

```
And           CC      *   (S*      (S*         *     -        (AM-DIS*)  (AM-DIS*)
to            TO   (VP*   (S*   (S(VP*         *     -                *   (AM-PNC*
attract       VB      *)    *     (VP*         *     attract       (V*)          *
younger       JJR  (NP*     *     (NP*         *     -            (A1*           *
listeners     NNS     *)    *)       *))))     *     -               *)          *)
,             ,       *     *        *         *     -               *           *
Radio         NNP  (NP*     *     (NP*     (ORG*     -            (A0*        (A0*
Free          NNP     *     *        *         *     -               *           *
Europe        NNP     *)    *        *)        *)    -               *)          *)
intersperses  VBZ  (VP*)    *     (VP*         *     intersperse     *         (V*)
the           DT   (NP*     *  (NP(NP*         *     -               *        (A1*
latest        JJS     *)    *        *)        *     -               *           *
in            IN   (PP*)    *     (PP*         *     -               *           *
Western       JJ   (NP*     *     (NP*    (MISC*)    -               *           *
rock          NN      *     *        *         *     -               *           *
groups        NNS     *)    *        *))))     *     -               *           *)
.             .       *     *)       *)        *     -               *           *
```

Irudia II.10: CoNLL-2005 ebaluazio saiorako prestatutako corpusaren lagina.

ConLL-2005 ebaluazio saioak, 2004ko edizioak bezalaxe, PropBankeko rolen etiketatze automatikoa izan zuen helburu, baina aurreko saioarekiko berrikuntza garrantzitsuak ekarri zituen:

- Sintaxi osoaren ekarpenak ebaluatu ahal izateko, 2004ko edizioan ez bezala, sarrerako informazioan zuhaitz sintaktiko osoak gehitu zituzten.

- Entrenamenduko corpusa nabariki handitu zuten sistemek corpus handietara egokitzeko zuten gaitasuna ikertu eta ikaste kurben bidez entrenamendu corpusen tamaina egokia zehazten saiatzeko.

- Sistemek ezaugarri ezberdinetako corpusetan zuten jokaera aztertzeko PropBank erara etiketatutako domeinuz kanpoko Brown corpusaren lagin txiki bat erabili zuten sistemak testatzeko.

Guztira, etiketatze estrategia ezberdinak inplementatu zituzten hemeretzi sistema aurkeztu ziren ebaluazio saio honetara. Oro har, rolak etiketatzeko sistema tipikoek egiten duten moduan, estrategia gehienak hainbat prozesuren kateaketa bezala planteatu ziren. Zuhaitz sintaktikotik argumentu kandidatu desegokiak kanporatzeko aurreprozesu edo inausketarako, sistema gehienek II.2.1 atalean deskribatu dugun Xue eta Palmerren (2004) bi urratsetako algoritmoa erabiltzen dute, baina badira (Pradhan *et al.*, 2005a) lanean definitzen diren "inausketa leuna" (*soft pruning*) delakoa erabiltzen dutenak ere. Salbuespenak (Màrquez *et al.*, 2005) lanean edota (Lin eta

Smith, 2005) lanean deskribatzen diren sistemak dira. Lehenak egitura sintaktikoak BIO tokenetan sekuentzializatzen ditu eta bigarrenak analizatzaile bat erabiltzen du zuhaitz sintaktikoak, aditza eta bere argumentuak egitura lau batean bilakatzeko.

Argumentuen sailkapen eta identifikazio faseak ere, arkitektura tipikoan deskribatu dugunean azaldu bezala, sistema gehienek bi fase independentetan burutu zituzten, baina badira fase bakarrean burutzen dituztenak ere.

Proposizioen iragarpenetan habiratzeak eta gainjartzeak ekiditeko asmoz, sistema askok neurriak hartu zituzten. Emaitza onenak lortu zituzten sistemek konbinazioa eta puntuazio orokorra erabili zuten zuhaitz sintaktiko ezberdinetatik eratorritako etiketatzeak modu egoki batean, habiratzerik eta gainjartzerik gabe biltzeko.

CoNLL-2005 sistema onenek %80 inguruko F_1 balioak lortu zituzten, eta ordutik hona nekez hobetu dira emaitza horiek nahiz eta hori lortzeko anbizio handiko saiakerak egin diren (Punyakanok *et al.*, 2008; Toutanova *et al.*, 2008a). Ebaluazio saioko sistemen emaitzei erreparatuta ikus daiteke argumentuen identifikazio ataza dela sistemen errendimendu galeraren arrazoi nagusia, izan ere, sistemek %81 inguruko estaldura neurriak lortzen dituzte ataza horretan baina aldiz, zuzen identifikatutako %81 horietatik %95ari, rol zuzena esleitzen diote batez beste. Gainera, sistemek, oro har, emaitza hobeak lortu zituzten *core* motako argumentuetan (>%80 F_1 balioak) argumentu adjuntuetan baino (<%60 F_1balioak). Aditzen frekuentziei so eginez, analisiek erakutsi zuten sistemek maizen agertzen diren aditzen argumentuak maiztasun urriko aditzenak baino hobeto sailkatzeko ahalmena erakusten zutela. Sistema onenaren eta txarrenaren arteko diferentzia 14 puntu absolutu ingurukoak izan zen WSJko test corpusean, baina 9 puntura murriztu zen Brown corpusera salto egitean. Azken corpus honetako test saioan sistema guztiek (baina batez ere WSJ corpusean emaitza onenak lortu zituztenak) 10 puntu inguruko galerak izan zituzten beren F_1 neurrietan. Tesi lan honetan, besteak beste, domeinuz kanpoko korpusetan sistemek jasaten dituzten erorketa aipagarri hauek izango ditugu hizpide.

SemEval-2007an PropBankeko eta VerbNeteko rolekin esperimentuak egiteko ebaluazio saioak proposatu ziren. Haietako batean hitzen adiera desambiguazioa eta rolen etiketatze automatikoa ataza bakarrean egitea planteatu zen (Pradhan *et al.*, 2007b), eta beste batean, VerbNet eta PropBank rolekin etiketatutako 50 aditzen gaineko esperimentuak egin ziren. Tesi lan honen hasieretan ginen orduan eta "VerbNet vs. PropBank" ataza horretan aurkeztu genuen sistemak (Zapirain *et al.*, 2007), lehen postua lortu zuen parte

hartzaile guztien aurretik[14]. Nahiz eta testeko datu basea oso murritza izan, esperimentuek erakutsi zuten zenbait baldintzatan posible zela VerbNeteko eta PropBankeko etiketatzaileekin antzeko errendimendua lortzea.

FrameNeteko rol multzoarekin egindako ebaluazio saio eta esperimentuak, normalean, frame semantikoen azpimultzo batekin egin izan dira. Senseval-2003 ebaluazio saioan adibidez, Gildea eta Jurafskyk (2002) egin zuten lana errepikatu eta hobetzeko asmoarekin (Litkowski, 2004), FrameNeteko 40 frame semantikoekin egindako esperimentuetan parte hartzaileek bi ataza burutu behar izan zituzten: (1) argumentuak emanda, zein frame elementuri (edo roli) zegokion adierazi (argumentuen sailkapena), eta (2) esaldia emanda, argumentuen identifikazioa eta sailkapena egin. Sistema onenek %92ko F_1 neurriak lortu zituzten lehen atazan, eta %83koak bigarrenean.

SemEval-2007 ebaluazio saioan rolen etiketatze automatikoko hainbat ataza proposatu ziren, haien artean FrameNeten oinarritutako bat. Ataza hau Senseval-2003koa baino errealistagoa zen ez baitzen inolako frame azpimultzorik ematen. Atazaren helburua testeko esaldien gainean argumentuak identifikatu eta zegokien rolekin etiketatzea zen. Horrez gain, parte hartzaileek esaldiaren semantika, frameekin eta rolekin etiketatutako argumentuekin eratutako grafoen bidez adierazi behar izan zuten. Esperimentuetarako erabili zen datu multzoak FrameNet lexikoitik kanpo zeuden frame semantikoak zituen. Parte hartzaileen emaitzak orokorrean kaxkarrak izan ziren %60ko prezisio eta %30eko estaldura neurriekin.

CoNLL-2008 (Surdeanu *et al.*, 2008)eta CoNLL-2009 (Hajič *et al.*, 2009) ebaluazio saioek, beste ikuspuntu bat eman zioten ordura arte rolak etiketatzeko bide nagusiari. Ebaluazio saio hauetan, esaldien dependentzia sintaktiko eta semantikoen etiketatze bateratua proposatu zen ingeles hizkuntzarako (2008an) eta beste sei hizkuntzetarako (2009an: katalana, gaztelera, ingelesa, txekiera, alemana, japoniera eta txinera). Sistema parte hartzaile gehienek, dena dela, arkitektura sekuentzial bat jarraitzeko izan ziren diseinatuak (batez ere 2008ko saioan) eta gutxi batzuek besterik ez zuten atazaren antolatzaileek sustatu zuten motibazio bateratuarekin gauzatu. Alegia, esan daiteke bi ebaluazio saiootan, sistema gutxi batzuek jorratu zutela benetan dependentzia sintaktiko eta semantikoen etiketatze bateratua, eta aipatzekoa da ez zirela sistema bateratu hauek izan, hain zuzen, emaitza onenak eskuratu zituztenak. CoNLL-2009 saioan, sistema bateratu onenak hirugarren

[14]Parte hartzaileen kopurua guztira 2 taldetakoa izateak, zalantzarik gabe, txapelketan lortutako rankingari begira behintzat, gauzak erraztu besterik ez zituen egin.

postua lortu zuen sailkapen orokorrean (Gesmundo *et al.*, 2009).

II.4.1 Gaur egungo erronkak

Ikusi dugu, gaur egun, rolak etiketatzeko sistemek errendimendu maila ona erakusten dutela esperimentu kontrolatuetan %80 inguruko F_1 balioak lortuz. Emaitza oso onak dira, baina oraindik ere, egoeraren balorazio kritiko bat eginez, etorkizunean erantzun beharreko hainbat galdera etor dakizkiguke burura (Màrquez *et al.*, 2008):

1. Zein da rolen etiketatze sistema sendoak egiteko beharrezko sintaxi maila egokiena? CoNLL-2004 eta CoNLL-2005 ebaluazio saioetako esperientzia aztertuz, ikusi dugu azaleko sintaxia erabiltzen duten sistemek, sintaxi osoa erabiltzen dutenak baino emaitza kaxkarragoak lortzen dituztela oro har. Dena dela, gogora dezagun, CoNLL-2005eko sistemek sendotasun arazo larriak erakusten zituztela domeinuz kanpoko corpusen gainean test ariketak egiten zituztenean eta, beraz, argitzeko dago erorketa horrek zenbateraino duen zerikusia erabilitako sintaxi mailarekin.

2. Sintaxitik haratago, zein modutara erabil daitezke WordNet, izen entitateak edota FrameNeteko frame semantikoak rolak etiketatzeko sistema sendoagoak lortzeko? Rolen etiketatze automatikoa ataza semantikoa ere bada, baina oraingoz ez da baliabide semantikoek gordetzen duten informazioa (VerbNeteko klaseak, Frame semantikoak, WordNeteko hautapen murriztapenak...) erabiltzeko saiakera seriorik egin.

3. Zenbaterainoko abantailak eskain ditzake metodo ez-gainbegiratu edota ahulki gainbegiratutakoen garapenak? SRL sistema gainbegiratuak entrenatzeko beharrezko corpusak ekoiztea oso garestia da. Hala eta guztiz ere, CoNLL-2005 ebaluazio saioko esperientziak erakutsi digu corpus jakin batzuetan entrenatutako sistemek ez dutela beren lana ongi egiten corpusa aldatzen zaienean. Metodo ez-gainbegiratuetan al dago domeinu arteko bateraezintasunen konponbidea?

4. Zein da corpus ingeniaritzak rolen etiketatze automatikoa hobetzeko jarraitu beharko lukeen bidea? Mugarik ba al du? Sistema ez-gainbegiratuek gehiegizko egokitze arazoak dituzten sistema tradizionalak hobetzen lagunduko al dute?

5. Corpusekiko independentzia maila baxua erakusten duten sistemak hizkuntza batetik "corpusik semantikorik" gabeko beste hizkuntza batera migratzeko aukerak zein dira? Jakina da hizkuntza guztiek ez dutela beren corpus semantikoa garatzeko baliabide ekonomikorik. Baina zein mailataraino da beharrezkoa corpus erraldoi bat sortzea? Atzerriko hizkuntza batean entrenatutako sistemak egokitu al daitezke beste hizkuntzetara "egokitze corpus" txikiago eta merkeen laguntzaz?

6. Zein da rolak etiketatzeko sistemek ikasi beharko lituzketen rolak? Alegia, zein da rol semantikoen multzorik "onena" eta zergatik? PropBankeko core argumentuak? VerbNeteko rol tematikoak? FrameNeteko frame elementuak? Besterik?

7. Zein da rol semantikoen etiketatze automatikoak lengoaia naturalen prozesamenduko beste atazak (itzulpen automatikoa, galdera erantzun sistemak, bilatzaileak...) hobetzeko eskaintzen dituen aukerak?

Galderak egin eta erantzunak bilatzea izan da tesi lan hau eta orokorrean edozein ikerketa lan aurrera bultzatu duen motorra. Dena dela, ez beza irakurleak espero goian planteatutako galdera guztien erantzuna topatuko duenik jarraian datozen kapituluetan. Bizitzeko eman zaigun denborarekin gertatzen den bezala, ikerkuntza egiteko denbora ere finitua izaten da eta tesi lan hau inprentara bidaltzea erabaki genuen egunean, galdera horietako zenbaiten erantzuna sumatu besterik ez genuen egiten.

Zehazki, honako hauetara bideratu ditugu gure lana eta arreta:

- VerbNet eta PropBankeko rolen azterketa enpirikoa. Bi rol multzo hauek nolabait uztartu dituzten lanak (Loper *et al.*, 2007a; Yi *et al.*, 2007) eta ebaluazio saioak (Pradhan *et al.*, 2007b) egin dira dagoeneko. Guk geuk bi rol multzoekin etiketatutako baliabideak aprobetxatuz, gaia sakondu nahi izan dugu bi rol multzoekin lan egiten duten sailkatzaile gainbegiratuak trebatuz eta, besteak beste, haien sendotasuna eta corpus berrietara egokitzeko gaitasuna ebaluatuz.

- Domeinuz kanpoko rolen etiketatze automatikoaren errendimendu baxuaren azterketa eta soluzio posibleen identifikazio eta inplementazioa.

Gai honen inguruan, Pradhan, Ward eta Martinen lanak (2008) rolen etiketatze automatikoko sistemen domeinuz kanpoko portaeraren azterketa sakon bat egitea izan zuen helburu. Guk geuk, jatorri ezberdineko baliabide semantikoak (antzekotasun distribuzionala eta hautapen murriztapenak) modu berritzaile eta eraginkor batean erabiliz, bertan planteatzen diren zenbait hipotesi baieztatu edota baloratzeko saiakera aurkeztuko dugu.

III. KAPITULUA

PropBank eta VerbNeteko rolen azterketa konparatiboa

PropBank eta FrameNet bezalako corpus semantikoen sorrerak rolak automatikoki etiketatzeko prototipoen ugaritzea ekarri du azken bost urteotan. Gildea eta Jurafskyk (2002) beren "Automatic Labeling of Semantic Roles" lan aitzindarian FrameNeten oinarritutako SRL sistema automatiko baten deskribapena egin zutenetik asko izan dira LNP munduko ikerlariak beren interesa eta esfortzua gai zirraragarri eta erlatiboki berri honetara bideratu dutenak. Pizkunde honetan berebiziko garrantzia izan zuten CoNLL-2004 (Carreras eta Màrquez, 2004) eta, batez ere, CoNLL-2005ean Carreras eta Màrquez, 2005 egin ziren ebaluazio saioek. Arestian aipatu bezala, saio haiek PropBanken entrenatutako hogeita hamar bat SRL sistemen arteko analisi konparatiboa eskaini zioten LNPko ikerlari komunitateari, eta oraindik ere ingeleserako SRL sistema berrien errendimendua ebaluatu eta konparatzeko erreferente dira.

Semantikoki etiketatutako corpusen ezaugarriek rolak etiketatzeko sistemak eta hauek aurrera eramateko beharrezko ikerketa baldintzatzen dute. Izan ere, jakina da PropBankeko rol zenbakituen (*Arg0*, *Arg1*,...) propietateak eta VerbNeteko rol tematikoenak (*Agent*, *Theme*, *Actor*, ...) ez direla berdinak eta, hortaz, baliteke, rol multzo ezberdin horiekin lan egingo duten

sistemek ere, besteak beste, propietate, abantaila-desabantaila eta ahalmen, ezberdinak ere izatea.

Gaur egun, PropBank da SRL sistemak garatzeko corpus semantiko erabiliena. PropBank handia da, zabala, oso egokia, edo egokia baino, oso praktikoa sistema automatikoak erosotasunez eraikitzeko. Gainera, sintaxi konputazionalean erreferentzia den Penn Treebank corpusaren WSJ zatiko aditz argumentuak semantikoki etiketatzeak nolabaiteko estandar estatus inplizitua eman dio corpus honi rolekin lan egiten duten ikerlarien artean. Horri guztiari PropBanken oinarritu ziren CoNLL-2004 eta CoNLL-2005eko ekimen arrakastatsuak gehitzen badizkiogu, ez da harritzekoa PropBank corpusarentzat izatea, zalantzarik gabe, lehen postua corpus semantiko erabilienen rankingean.

PropBank corpusari egiten zaizkion kritiketako bat erabiltzen duen rol multzo zenbakituari dagokio. Aurrerago ikusiko dugun moduan, rol zenbakituen esanahia aditzarekiko menpekoa da eta horrek rol zenbakituak oinarri dituzten SRL sistemen orokortzeko gaitasuna eta eramangarritasuna kolokan jar ditzake, batez ere entrenamendu corpusa txikia eta ez adierazgarria denean. Aldiz, hasiera batean behintzat, egokiagoak dirudite teoria linguistikoan hain ezagunak diren rol tematikoak (*Agent*, *Beneficiary*, ...), beren esanahiek aditzarekiko menpekotasunik erakusten ez dutelako, eta rol multzo mugatu eta trinko bat osatzen dutelako. Hori dela eta, teorian behintzat, SRL sistemen eramangarritasuna eta orokortzeko gaitasuna hobetzeko potentziala dutela esaten da.

PropBankeko rolak VerbNeteko rol tematikoetara bihurtzen dituen SemLink mapaketari esker posible da bi rol multzo ezberdin hauekin azterketa enpirikoak eta analisi konparatiboak egitea; eta hori izango da, hain zuzen, atal honen xede nagusia. Zehazki, honakoak izango ditugu aztergai:

- Rol tematikoak eta argumentu zenbakituak ikasketa automatiko edota teknika estatistikoak erabiliz ikasteko aukerak eztabaidatuko ditugu. Ezaugarri eta berezitasun propioak dituzten bi rol multzo izanda, trebatuko ditugun sailkatzaileek ere bere "izaera" eta ezaugarri propioak izango dituzte. Horretarako garrantzitsua izango da bi rol multzoen arteko diferentziak azpimarratzea (III.1 atalean).

- PropBank eta VerbNeteko rol sailkatzaileen sendotasuna aztertzea domeinu aldaketetan eta predikatu ezezagunetan. Bi rol multzo ezberdinetan trebatutako sailkatzaileak edukita posible da sailkatzaile bakoi-

tzaren alde positiboak eta negatiboak konparatiboki aztertzea. Zehazki, bi sailkatzaileak SRL munduan bereziki konplikatuak diren eremutan probatuko ditugu: domeinu aldaketetan eta aditz ezezaguneko testuingurutan (III.3 atalean).

- VerbNeteko rol tematikoak PropBankeko notaziotik abiatuta lortzeko aukerak. Badirudi PropBanken argumentuen rola adierazteko erabiltzen den notazioa ez dela egokiena aplikazio errealen ikuspegitik. Izan ere, argumentu zenbakituen interpretazio semantikoa aditzaren adierari hertsiki lotuta dago eta, askotan, batez ere mundu errealeko aplikazioetan, datu hori ez da eskuragarri egoten. Hori dela eta, semantikoki independenteak diren rol tematikoak interesgarriagoak izan daitezke. Horregatik, rol hauek lortzeko bi bide ezberdinen arteko konparaketa egingo dugu: (1) VerbNeteko rolak zuzenean lortzea VerbNeteko rol sailkatzailearekin eta (2) PropBankeko rolak VerbNeteko rol tematikoetara bihurtzea SemLink erabilita (III.4 atalean)

II kapituluan rol multzo garrantzitsuenen deskribapen zabala eman dugun arren eta tesi lan honen irakurketa ez jarraiak errazteko helburuarekin, hurrengo lerroetan esperimentuetarako erabili ditugun rol multzoen ezaugarriak errepikatuko ditugu berriro ere, bien arteko ezberdintasunetan eta propietate berezietan arreta berezia ipiniz.

III.1 Corpus semantikoak eta rol multzoak: ikuspegi kritikoa

III.1.1 PropBank: teorikoki neutrala

PropBank corpusa Penn Treebank IIko (Marcus *et al.*, 1994) egitura sintaktikoen gainean eraikitako geruza semantikoa da. Zehazkiago, treebankeko Wall Street Journal sekzioko aditzak predikatu-argumentu egiturekin hornitzen ditu, eta horretarako erabiltzen duen rol multzoa "teorikoki neutrala" dela esaten da, alegia, ez dela inolako teoria linguistikorik erabiltzen aditz-argumentu horiek rolekin etiketatzeko irizpide gisa. Esan bezala, PropBankeko rol multzoa *core-argument* izeneko argumentu zenbakituek (*Arg0*, *Arg1*, ..., *Arg5*) osatzen dute. Aditz bakoitzaren *framesetak*, aditzak baimenduta dituen rolak zerrendatzen ditu (adibidea) eta horrez gain rol bakoitzaren semantikaren deskribapen "askea" ere ematen du egituratu gabeko hizkuntzan.

Aditz polisemikoen adiera bakoitzari *frameset* ezberdin bat dagokion arren, argumentu labelak semantikoki kontsistenteak dira aditz adiera jakin baten alternantzia sintaktiko guztientzat. Ikus dezagun adibide batekin:

"[Kevin] broke [the window]$_{Arg1}$"[1]
"[The door]$_{Arg1}$ broke into a million pieces"[2]

Aurreko esaldiek *break.01* (puskatu) aditzaren bi errealizazio sintaktiko ezberdin erakusten dituzte. Bi esaldietako argumentu zenbakituen esanahiak beraz, kontsistenteak dira esalditik esaldira, aditz adiera berdinari dagokion *framesetaren* menpe daudelako. Horregatik, zalantzarik gabe esan dezakegu adibideko Arg1 argumentuek "*broken entity*" (puskatutako entitatea) argumentu semantiko berberari egiten diotela erreferentzia.

Aditzaren adiera eta, ondorioz, frameseta aldatu ahala, argumentu zenbakituen arteko kontsistentzia galdu egiten da neurri handi batean. Beste modu batera esanda, argumentu zenbakituak ez dira bateragarriak aditz batetik bestera. Arg2 etiketa, adibidez, *send* (bidali) aditzaren *Destination* (helburu) argumentua identifikatzeko erabiltzen da baina etiketa zenbaki berak *compose* (konposatu) aditzaren *Beneficiary* (onuradun) argumentua ere adieraz dezake. Mota honetako "desadostasun" ugari topa ditzakegu PropBankeko frameseten artean eta, zentzuzkoa dirudi pentsatzea horrek kolokan jar dezakeela PropBanken entrenatutako sistemen orokortzeko gaitasuna. Izan ere, esan dugun moduan, PropBankeko argumentu zenbakituak aditzaren adierarekiko lotuta egoteak, *Arg0*, *Arg1*, ..., *Arg5* etiketen arteko "muga semantikoak" lausotzea dakar ezinbestean; zer espero daiteke, hortaz, muga horiek ezagutu eta bereizteko sortzen ditugun sistema automatikoez? Argumentu zenbakituen interpretazioa aditzarekiko independente izateak zer mailataraino zailtzen du PropBanken trebatzen diren sistema automatikoen lana? Galdera hauei eta beste batzuei atal honetan zehar bilatuko diegu erantzuna, baina lehenago garrantzitsua da azpimarratzea PropBankeko diseinatzaileen erabakiz argumentu zenbakitu ugarienen kontsistentzia aditzetik aditzera neurri garrantzitsu batean bermatuta dagoela. Argumentu hauek Arg0 eta Arg1 dira eta, hurrenez hurren, *Agent* eta *Theme* rol tematiko orokorrak adierazten dituzte frameset ia gehienetan. Egokitzapen hori, dena dela, ez da erabatekoa. (Loper *et al.*, 2007b) lanaren arabera, Arg0 argumentuek *Agent*

[1]Kevinek leihoa puskatu zuen.
[2]Atea milaka zatitan puskatu zen.

rola jokatzen dute kasuen %85ean, baina halaber, *Experiencer* (%7.2), *Theme* (%2.1) eta *Cause* (%1.9) rolak ere joka ditzakete, framesetaren arabera. Arg1 argumentuei dagokienez, kasuen %47ak *Theme* rola jokatzen dute, baina badira, besteak beste, *Topic* (%23), *Patient* (%10.8) eta *Product* (%2.9) rolak jokatzen dituztenak ere. *Core* argumentuak ez bezala, denborazko (AM-TMP) edota lekuzko (AM-LOC) argumentuak adierazteko erabiltzen diren argumentu-adjuntuak, aditzarekiko independenteak diren rol multzo finitu bat osatzen dute.

III.1.2 VerbNet: rol tematiko orokorrak

VerbNet aditzen lexikoi konputazionala da, non portaera sintaktiko eta semantiko berbera duten aditzak hierarkikoki antolatutako klase sistema batean antolatzen diren. Klase hauek Levinen aditz klaseetan oinarrituta daude, eta ezaugarri komunak dituzten aditzak gordetzeaz gain, haien arteko sintaxia eta semantikaren arteko egokitzapena ere deskribatzen dute rol tematiko edota hautapen murriztapenekin.

VerbNeteko rol tematikoak 23 dira (*Agent, Patient, Theme, Experiencer, Source, Beneficiary, Instrument, ...*) eta PropBankeko argumentuek ez bezala, aditzarekiko independenteak diren rol multzo orokor bat osatzen dute (ikusi II.1.3 atala).

Badirudi abstrakzio maila orokorrago honek PropBankeko argumentu zenbakituak baino egokiagoak egiten dituela rol tematikoak LNPko aplikazioetarako. Izan ere, zaila da, beste ezer jakin gabe, Arg2 bezala etiketatutako aditz-argumentu baten atzean dagoen argumentu semantikoa zein den jakitea; aldiz, ez dirudi interpretazio arazorik egongo litzatekeenik argumentu zenbakituaren ordez *Instument* edota beste rol tematikoren bat izango bagenu.

VerbNet ez da SRL sistemak entrenatzeko baliabide egokia, PropBanken ez bezala, semantikoki etiketatutako esaldiak oso urriak direlako. Hala ere, SemLink izeneko baliabideari esker, posible da VerbNeteko rol tematikoekin etiketatutako corpus zabalak lortu eta SRL sistemak rol hauetarako ere trebatzea.

(1)	(2)	(3)	(4)	(5)
wsj/00/wsj_0001.mrg	0 8	join.01;VN=22.1-2		0:2-ARG0[Agent] 7:0-ARGM-MOD 8:0-rel 9:1-ARG1[Patient1] 15:1-ARGM-TMP
wsj/00/wsj_0002.mrg	0 16	name.01;VN=29.3		16:0-rel 0:2*17:0-ARG1[Theme] 18:2-ARG2[Predicate]
wsj/00/wsj_0001.mrg	1 10	publish.01;VN=26.4-1		10:0-rel 11:0-ARG0[Agent]

Irudia III.1: SemLink etiketatzearen adibidea. Lerro bakoitzak aditz bat eta bere argumentuen etiketatzea ematen digu, core roletan eta rol tematikotan. (1) Treebankeko fitxategiaren izena, sekzioaz lagunduta. (2) Sententzia eta Terminala: sententzia zenbakiak fitxategiko zenbatgarren esaldia den esaten digu; terminal zenbakiak aditza kokatzen du esaldiko nodo sintaktiko edo terminalen artean 0 tik hasita. (3) Aditza. (4) Aditzaren PropBank framea eta VerbNet klasea. (5) Esaldiko aditzaren eta argumentuen identifikazioa eta etiketatzea. Aditza "rel" bezala agertzen zaigu, eta core argumentuak urdinez eta haien rol tematiko baliokidea kortxete artean. Argumentu bakoitzak analisi zuhaitzean hartzen duen posizioa adierazten du haien aurretik datorren terminal zenbakiak.

SemLink

SemLink [3] (Loper *et al.*, 2007b) baliabideak PropBankeko frameseten eta VerbNeteko rol tematikoen arteko mapaketa bat definitzen du[4]. Zehazki, PropBankeko framesetetan zerrendatzen den argumentu zenbakitu bakoitzari dagokion rol tematikoa esleitzen dio eta, gainera, frameseteko aditzaren adierari ere dagokion VerbNet klasea ematen dio.

Argumentu zenbakitu eta rol tematikoen mapaketa honen laguntzaz, SemLink eratu zuen taldeak PropBankeko proposizio guztien %50 etiketatu zuen gutxi gorabehera VerbNeteko rol tematikoekin. Eta bi rol multzo ezberdin horiekin etiketatutako corpus hori hartuko dugu, hain justu, PropBank eta VerbNeten arteko azterketa konparatiboa egiteko abiapuntutzat.

III.2 Esperimentaziorako baliabideak

Kapitulu honetan zehar egingo dugun azterketa konparatiboa ahalik eta modu zorrotz eta orekatuenean aurrera eraman ahal izateko, beharrezkoa da bai

[3] https://verbs.colorado.edu/semlink/

[4] Tesi honetan aintzat hartuko ez diren beste zenbait errekurtso ere lotzen ditu SemLinkek

VerbNeteko rol sailkatzailea eta PropBankekoa corpus baldintza berdinetan trebatuta egotea. Gainera, bi sailkatzailek ikasketarako erabiliko dituzten ezaugarriak ere berdinak edo oso antzekoak izan behar dira, bien arteko errendimenduaren diferentzian rol multzo bakoitzaren izaera berezituaren abantailak eta desabantailak ikus daitezen soilik, azken batean hori baita aztergai izango duguna.

III.2.1 Datu multzoa

Esperimentu hauetan erabiliko dugun corpusa CoNLL-2005eko ebaluazio saioan erabili zen berbera da. SRL sistemak trebatzeko 02-21 PropBank sekzioak erabiliko dira eta sistemen garapen eta testerako, 24 eta 23 sekzioak hurrenez hurren. Sarrerako datuetatik informazio morfologikoa eta Charniak parserraren zuhaitz sintaktikoak hartuko dira kontuan, baina ez entitateak.

Dakigunez, CoNLL-2005eko corpus honetako aditz-argumentuak PropBankeko rol zenbakituekin daude etiketatuta soilik. Horregatik, beharrezkoa gertatzen da SemLink mapaketa erabiltzea rol zenbakitu horiek rol tematikoetara bihurtzeko. Zoritxarrez, gure lanean erabili genuen SemLinken 1.0 bertsioak ez du PropBankeko proposizio eta argumentu guztien itzulpena egiten. Horregatik, corpus homogeneo bat eduki eta esperimentuen emaitzak ahalik eta gutxien baldintzatzeko asmoz, VerbNet rolekin %100ean etiketatuta ez zeuden proposizio guztiak corpusetik atera genituen. Erabaki honen ondorioz, jatorrizko corpusaren %56arekin geratu ginen sistemak entrenatu eta ebaluatzeko. Corpus berri honetan, 50,000 proposizio geratzen zaizkigu sistemen entrenamendurako, eta guztira 1,709 aditz (1,505 aditz-lema) ezberdinek gobernatutako proposizioak aurki ditzakegu. Tamaina eta aberastasunari erreparatuz, uste dugu corpusa egokia dela rol multzoen sendotasuna eta aditzetik aditzera ezagutza orokortzeko gaitasunari buruzko ondorio sendoak ateratzeko.

Domeinuz kanpoko corpusetan ere neurtu nahi izan dugu rol multzoen egokitasuna eta errendimendua. Horretarako domeinuz kanpoko Brown corpusa erabili dugu (testerako). Corpus hau, PropBank proiektuaren barnean etiketatu zuten argumentu zenbakituekin. Hortaz, "*PB vs. VN*" azterketa konparatiboan barneratu ahal izateko, berriro ere, SemLinken laguntzaz, Brown corpuseko argumentu zenbakituen %55a rol tematikoetara bihurtu genuen.

III.2.2 SRL sistema

Erabiliko dugun oinarrizko SRL sistemak etiketatze prozesua Entropia Maximoko Markoven Eredu baten moduan errepresentatzen du (Maximum Entropy Markov Model). Sistemak sintaxi osoa erabiltzen du sarrera testutik sekuentziako osagaiak aukeratzeko, eta aldi berean Begin/Inside/Outside (Hasieran/Barnean/Kanpoan) (BIO) labelekin etiketatzen ditu, errendimendu altuko sailkatzaileak eta ezaugarriak erabiliz. Sistema honek emaitza onak lortzen ditu CoNLL-2005 corpusaren gainean eta baita SemEval-2007ko SRL azpiatazan (Zapirain *et al.*, 2007).

Datuen errepresentazioa

Ikasketa automatikoaren bidez SRLa bezalako arazo bati aurre egin nahi diogunean, ikaste eta etiketatze prozesuak errazteko helburuarekin, sarrera corpusaren errepresentazioa aldatzea komenigarria izaten da. Esperimentu hauetarako erabili dugun sailkatzaileak sarrera testuko egitura sintaktikoetatik BIO tokenak erauzten ditu etiketatu beharreko proposizio guztietarako (Surdeanu *et al.*, 2007) lanean etiketatzaile sekuentzialarentzat egiten den moduan.

Token sekuentzial hauek sententziako perpausetako mugak kontuan hartuz erauzten dira, ondoren, prozesu berean, BIO labelekin etiketatzeko. Token bakoitzari jarri beharreko labela aditz-argumentu batekiko duen posizioak zehazten du: tokena argumentuaren hasieran badago, B labela emango zaio; barnean badago berriz, I; eta tokena ez bada argumentu baten parte, O etiketa. Aurreprozesu honen ondoren, hasierako corpusa token multzo trinko eta erabilerraz batean bilakatzen da. BIO-token sekuentzialen errepresentazio honen beste abantailetako bat da, sistemaren predikzio gainjarriak eta habiratuak ekiditen dituela, eta gai dela bere errepresentazio banatuaren bidez osagai sintaktiko batez baino gehiagoz osatutako argumentuak erraz detektatzeko.

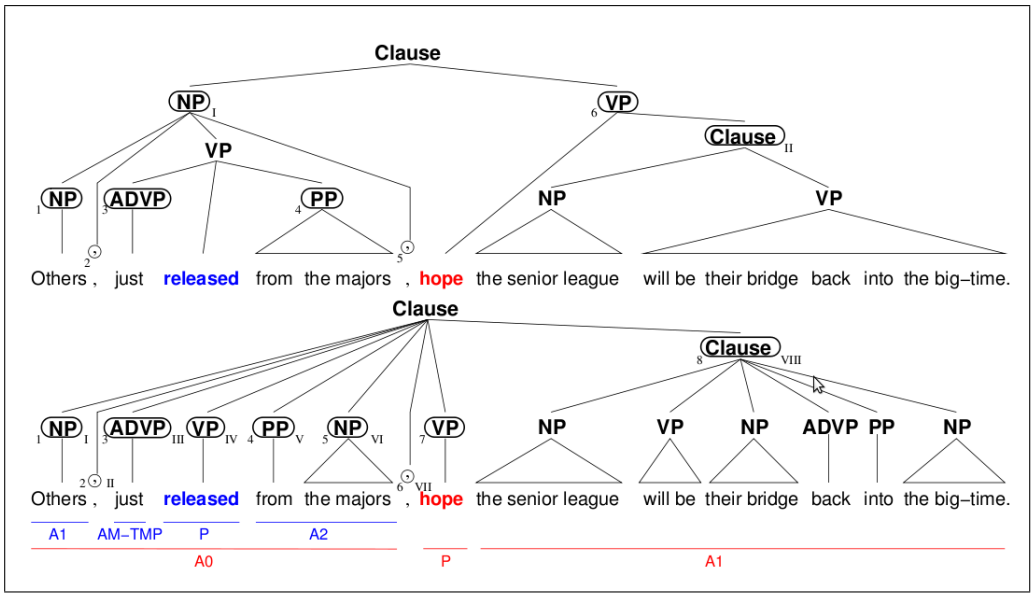
Esan bezala, BIO tokenak perpauseko analisi zuhaitzetik erauzten dira. Hauek dira horretarako jarraitzen diren urratsak:

1. Esaldia gainjartzen ez diren hainbat segmentu ez habiratutan banatzen da, egitura sintaktikoak ezartzen dituen mugen arabera.

2. Segmentu bakoitzean osorik aurki ditzakegun osagai sintaktiko jarrai guztiak (habiratuta eta gainjarrita daudenak izan ezik) token bezala

markatzen dira argumentuaren hasieran (B), tartean (I) edota kanpoan (O) dauden kontuan hartuta.

BIO etiketatzearen adibide bat III.2 eta III.3 iruditan ikus daiteke. III.2 irudian, "*Others, just released from the majors, hope the senior league will be their bridge back into the bigtime*" esaldiaren bi zuhaitz sintaktiko ezberdin ikusten dira, zuhaitz sintaktikoa eratzeko erabili den analizatzaile sintaktiko bakoitzeko bat: goian sintaxi osoa (*full parsing*) eta behean azaleko sintaxia (*partial parsing*).



Irudia III.2: Irudian adibide esaldi baten bi interpretazio sintaktiko ematen dira. Goian analisi zuhaitz osoa ematen da eta behean azalekoa. Esaldiak bi proposizio gordetzen ditu bere baitan, bat aditz nagusi bakoitzeko (*release* eta *hope*). Proposizio bakoitzaren azaleko analisi semantikoa behean ematen da gorriz eta urdinez. Biribilduta agertzen diren nodo sintaktikoak tokenizazio fasean aukeratutakoak dira. *Hope* aditzaren tokenak zenbaki erromatarrez adierazten dira, eta besteak, europarrez. Iturria: Surdeanu et al. (2007)

Esaldiak bi proposizio ditu, bat aditz nagusi bakoitzeko (*release* eta *hope*). III.3 irudian ikusten da interpretazio sintaktikoaren arabera esaldiaren BIO errepresentazioa ere aldatzen dela. Dena dela, bat ala beste, tokenizazioaren

ondoren esaldi osoa geratzen da BIO tokenekin etiketatuta. "*To hope*" aditzaren proposizioan zentratuz soilik, III.2 irudian ikus dezakegu bi argumentu besterik ez dituela: Arg0 ("*Others, just released from the majors,*") eta Arg1 ("*the senior league will be their bridge back into the big-time*"). Argumentu hauen BIO kodeketa III.3 irudian argitzen da, "*hope-PP*" edota "*hope-FP*" zutabeetan. Ikusten denez, sintaxi partziala erabiltzen dugunean (*hope-PP*), token gehiago erauzten dira sintaxi osoa erabiltzen denean baino. Esaterako, Arg0 argumentua kodetzeko sintaxi partziala erabiltzen dugunean, zazpi BIO tokenen segida bezala kodetzen dugu (`NP(I) + ,(II) + ADVP(III) + VP(IV) + PP(V) + NP(VI) + , (VII)`), ez baitago argumentua ordezkatuko duen nodo orokorragorik. Argumentu hori bera, sintaxi osoa erabiltzen dugunean, token bakarrarekin ordezka daiteke zuhaitz egitura osoak horretarako erraztasunak ematen dituelako (`NP(I)`). Sintaxi partziala eta osoa erabiltzearen abantaila eta desabantailen inguruko hausnarketak Màrquez *et al.*, 2005 lanean aurki daitezke.

Ezaugarrien errepresentazioa

Erabiltzen ditugun ezaugarri gehienak gaiko literaturan behin eta berriz agertzen direnak dira. (Gildea eta Jurafsky, 2002; Xue eta Palmer, 2004; Surdeanu *et al.*, 2007).

– **Aditzari dagozkion ezaugarriak:**

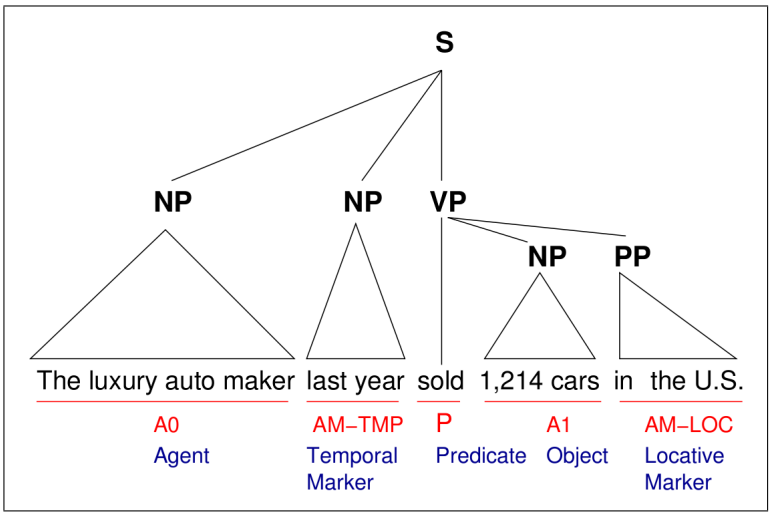
- Aditzaren forma, lema eta PoS labela: III.4 irudiko predikatutik *sold*, *sell* eta `VBD` ezaugarriak erauziko genituzke.
- Chunk mota eta aditz sintagmaren "kardinalitatea": hitz bakarra edo hitz anitzekoa. Adibidez, III.4 irudiko aditza hitz bakarreko VP *chunk* batean dago sartuta.
- Aditzaren boza: 5 motatakoa izan daiteke: *active*. *passive*, *copulative*, *infinitive* eta *progressive*
- Aditza perpaus baten hasieran/bukaeran dagoen (marka bitarra).
- Azpikategorizazio erregela: predikatu nodoaren ama-nodoa eratzen duen erregela. III.4 irudian adibidez, aditzaren ama-nodoa `S` $\rightarrow$ `NP NP VP` erregelak eratzen du.
- Aditzaren PropBank framearen adiera eta VerbNet klasea.

<table>
<tr><th rowspan="2">words</th><th colspan="4">tokens</th></tr>
<tr><th>release–PP</th><th>release–FP</th><th>hope–PP</th><th>hope–FP</th></tr>
<tr><td>1: Others</td><td>1: B_A1</td><td>1: B_A1</td><td>I: B_A0</td><td rowspan="9">I: B_A0</td></tr>
<tr><td>2: ,</td><td>2: O</td><td>2: O</td><td>II: I_A0</td></tr>
<tr><td>3: just</td><td>3: B_AM-TMP</td><td>3: B_AM-TMP</td><td>III: I_A0</td></tr>
<tr><td>4: released</td><td>—</td><td>—</td><td>IV: I_A0</td></tr>
<tr><td>5: from</td><td>4: B_A2</td><td rowspan="3">4: B_A2</td><td>V: I_A0</td></tr>
<tr><td>6: the</td><td rowspan="2">5: I_A2</td><td>VI: I_A0</td></tr>
<tr><td>7: majors</td></tr>
<tr><td>8: ,</td><td>6: O</td><td>5: O</td><td>VII: I_A0</td></tr>
<tr><td>9: hope</td><td>7: O</td><td rowspan="11">6: O</td><td>—</td><td>—</td></tr>
<tr><td>10: the</td><td rowspan="10">8: O</td><td rowspan="10">VIII: B_A1</td><td rowspan="10">II: B_A1</td></tr>
<tr><td>11: senior</td></tr>
<tr><td>12: league</td></tr>
<tr><td>13: will</td></tr>
<tr><td>14: be</td></tr>
<tr><td>15: their</td></tr>
<tr><td>16: bridge</td></tr>
<tr><td>17: back</td></tr>
<tr><td>18: into</td></tr>
<tr><td>19: the</td></tr>
<tr><td>20: big-time</td></tr>
</table>

Irudia III.3: III.2 irudiko esaldiaren bi proposizioen BIO errepresentazioa. Erabilitako egitura sintaktikoaren arabera tokenizazioa ezberdina da. Sintaxi partzialaren tokenizazioa PP (*Partial Parsing*) zutabeek erakusten digute, eta Sintaxi osoarena, FP (*Full Parsing*) zutabeak. Token bakoitza III.2 irudian dagokion osagai sintaktikoaren zenbakiarekin eta rol semantikoarekin dago markatuta.

– **Uneko osagaiari dagozkion ezaugarriak:**

- Mota eta gunea: Collinsek (2003) guneak erauzteko proposatutako erregelen bidez erauziak. Adibidez, lehen elementua PP motako argumentua bada, jarraian datorren NParen argumentuaren gunea erauziko da. Adibidez, III.2 "*in the U.S.*" (EBetan) osagaiaren mota PP da baina bere gunea *U.S.* izango da *in* preposizioa izan beharrean.



Irudia III.4: "*The luxury auto makers sold 1,214 cars in the U.S.*" esaldiaren zuhaitz sintaktikoa eta azaleko errepresentazio semantikoa (gorriz). Urdinez, analisi semantikoaren azalpena. Iturria: Surdeanu et al. (2007)

- Osagaiaren lehen eta azken hitzak eta PoS labelak. Adibidez, III.2 "*in the U.S.*" osagairako *in*/`IN` and *U.S.*/`NNP` ezaugarriak.
- PoS sekuentzia, 5 osagai baino gutxiago baldin baditu. Adibideko osagairako, `IN-DT-NNP`
- Osagaian ageri diren hitzen, adjektibo eta adberbioak (edo *bag-of-words*). Adibidez, III.4 irudiko "*The luxury auto makers*" osagaiaren "izen-zakua" {"*Luxury*", "*auto*", "*makers*"} da.
- TOP sekuentzia: uneko osagaia zabaltzen duen erregela. Adibidez, "*in the U.S.*" osagaiko TOP sekuentzia `IN-PP` izango litzateke.
- TOP sekuentziako 2/3/4-gramak.
- Kategoria nagusia: Gildea eta Jurafskyk (2002) deskribatu zuten moduan. Ezaugarri hau oso erabilgarria da osagai bat sententziako subjektua edo objektua den jakiteko. Adibidez, izen sintagma baten kategoria nagusia `S` (sententzia bada, izen sintagma hori subjektua izango da ziurrenik. Aldiz, bere kategoria nagusia `VP` bada, objektu baten aurrean egongo ginateke ia ziur.

– **Uneko osagaiaren testuinguruari dagozkion ezaugarriak:**

- Uneko osagaiaren aurreko eta ondorengo hitzak eta PoS labelak. Adibidez, III.4 irudiko "*last year*" argumentuaren aurreko hitzetik "*maker*"/`NN` erauziko genuke, eta hurrengotik, "*sold*"/`VBD`.
- Aurreko eta ondorengo osagaietatik uneko osagaiari dagozkion ezaugarri guztiak ere erauzten zaizkie, segmentu berekoak badira bederen.

– **Osagaia eta predikatuaren arteko erlazioari dagozkion ezaugarriak:**

- Posizio erlatiboa, Distantzia (hitzetan eta chunk kopurutan), habiratze maila osagaiarekiko (perpaus kopurutan): Adibidez, irudiko "*in the U.S.*" osagaia predikatuaren ondoren (*after*) agertzen da, 2 hitzetako distantziara edo chunk bakarreko distantziara eta bere habiratzen maila 0 da.
- Posizio erlatibo bitarra (osagaia aditzaren aurretik ala ondoren dagoen).
- Osagai bidea (*Constituent path*):Gildea eta Jurafskyk 2002 definitu zuten bere lanean. Adibidez, III.4 irudiko "*The luxury auto maker*" eta "*sold*" predikatuaren arteko osagai bidea `NP` ↑ `S` ↓ `VP` ↓ `VBD`
- Sintaxi partzialeko bidea (*Partial parsing path*): Carreras et al.-ek 2004 azaltzen dute ezaugarri hau. Adibidez, `NP + PP + NP + S` ↓ `VP` ↓ `VBD` bideak, uneko NP osagaitik aditzera dagoen bidea erakusten du. Eskuinetara `PP` bat eta ondoren `NP` eta `S` bat topatuko genituzke. `S` osagaitik behera `VP` bat eta horren azpian, azkenik, aditza. Sintaxi partzialeko bidea goiko Osagaien bidearen antzekoa da baina goiko mailatara jauzi egiten duten *up* mugimendurik gabe. Horren ordez sintaxi maila berean ezker (-) eskuin (+) mugimenduak baimentzen dira.
- Frame sintaktikoa (*syntactic frame*) Xue and Palmer-ek (2004) deskribatzen duten moduan: frame sintaktikoak sententziaren egituraren deskribapena ematen du aditza eta uneko argumentua oinarri hartuta. Adibidez, "*sold*" predikatua eta "*in the U.S.*" argumentuaren arteko frame sintaktikoa `NP_NP_vp_NP_pp` da (bertan aditza (vp) eta uneko argumentua (pp) letra xehez nabarmenduta agertzen dira eta baita haien posizio erlatiboa ere).

– **Ezaugarrien konbinazioa:** Entropia Maximoko sailkatzaileari (ikusi hurrengo atala) ahalik eta probetxu gehien ateratzeko asmoz, aurretik azaldutako hainbat ezaugarriren arteko konbinazioa ere egiten da, Xue eta Palmerrek (2004) modu arrakastatsuan egin zuten moduan.

- Aditza eta Sintagma mota
- Aditza eta posizio bitarra
- Aditza eta gunea
- Aditza eta PropBank framea
- Aditza eta VerbNet klasea

Entropia Maximoko Markoven Ereduak

Esperimentuetan erabiliko dugun sailkatzaileak etiketatze prozesua Entropia Maximoko Markoven Eredu (MEMM) baten moduan errepresentatzen du. MEMMak etiketatze sekuentzialerako erabiltzen diren eredu diskriminatiboak dira eta $P(s_n|s_{n-1}, o)$ probabilitate lokala modelatzen du, non o obserbazioaren testuingurua den.

MEMM bat emanda, egoera (s) sekuentzia (S) probableena hurrengo espresioak ematen digu:

$$S = argmax \prod_{i=1}^{n} P(s_i \mid s_{i-1}, o)$$

Goiko formulismoak ere balio du proposizio bateko rol sekuentzia modelatzeko: rol etiketak S sekuentziako (edo proposizioko) s egoeretara (*state*) izango genituzke konektatuta, eta o obserbazioak puntu horretan erauzitako ezaugarriak izango lirateke. Rol sekuentzia probableena beraz, Viterbi algoritmoak emango ligukeen egoera sekuentzia probableena izango litzateke. Algoritmo hau da ohiko Markoven Eredu Ezkutuak (*Hidden Markov Models* deskodetzeko erabiltzen dena.

Probabilitate baldintzatu lokal guztiak Entropia Maximoko sailkatzaileak [5] ematen dizkigu.

Hainbat murriztapen hartzen dira kontuan sekuentzia probableena bilatzeko orduan:

[5] http://mallet.cs.umass.edu

1. Core argumentuak ezin dira errepikatuta azaldu sekuentzian (berdin rol tematikoekin)

2. R-X (erreferentzia) edo C-X (jarraipen) moduko rol bat badugu sekuentzian, X rola egotea ere beharrezkoa da

3. I-X moduko token baten aurretik beharrezkoa da B-X ala I-X motako token bat egotea.

4. Proposizioko aditza eta bere PropBank adiera emanda, argumentu zenbakitu batzuk soilik egongo dira baimenduta sekuentzian (aditz batzuek, adibidez, ezin dute Arg2 argumenturik izan).

5. Proposizioko aditza eta bere VerbNet klasea emanda, rol tematiko batzuk soilik egongo dira baimenduta sekuentzian (aditz batzuek, adibidez, ezin dute *Instrument* rolik izan)

III.3 Rol multzoen orokortzeko gaitasuna

Rol tematikoen eta argumentu zenbakituen sailkatzaileen arteko konparaketa egiteko bi ebaluazio ingurune definituko ditugu hasteko: "SemEval" eta "CoNLL"[6]. Lehenengo testuinguruan, SemEval-2007ko SRL atazan egiten zen bezalaxe, sarrerako corpusean eskuragarri dagoen informazio guztia erabiltzen dugu sailkatzaileak entrenatzeko, baita aditzaren PropBank adiera eta VerbNet klasea ere. Azken ezaugarri horiek dira hain zuzen ere, ebaluazio ingurune honi bere berezitasuna ematen diotenak, izan ere, gainerako ezaugarriak ez bezala, hauek eskuz etiketatuak izan direlako eta, beraz, ez genituzkeelako edozein testutatik erauzterik izango ebaluazio erreal batean.

Kontuan izanda, hortaz, aplikazio erreal eta automatikoetan adierari buruzko informaziorik ez genukeela izango, sistemen ebaluaziorako testuinguru errealago bat definitzea beharrezkoa da aurrera eraman nahi dugun azterketa konparatiboak gutxieneko balio bat izan dezan. Helburu horrekin definitu genuen bigarren ebaluazio ingurunea, non, argitzen duen bezala, CoNLL-2005eko ebaluazio irizpide berberak hartzen dituen.

[6]SemEval eta CoNLL ebaluazio inguruneek Semeval-2007 eta CoNLL-2005eko ebaluazio inguruneei egiten diete erreferentzia

Sistemek lehenengo ebaluazio ingurunean lortzen dituzten emaitzak III.1 taulan ikus daitezke. Asmatutakoak, huts egindakoak eta ahaztutako argumentu kopurua ematen da, eta baita balio horiei dagozkien *precision*, *recall* eta *f-score* balio tipikoak ere. F_1 balioentzako esangura tarteak ere erakusten dira sistemen arteko diferentziak estatistikoki esanguratsuak diren ala ez zehazteko balio dutenak [7].

	zuzen	gehiegi	faltan	prezisioa	estaldura	F_1	F_1 core	F_1 adj.
PropBank	6,022	1,378	1,722	81.38	77.76	79.53 ±0.9	82.25	72.48
VerbNet	5,927	1,409	1,817	80.79	76.54	78.61 ±0.9	81.28	71.83

Taula III.1: SemEval testuinguruko emaitzak PropBankeko rol multzorako (goian, PropBank lerroan) eta VerbNeteko rol tematikoetarako. Sailkatzaile bakoitzak guztira asmatutakoak (zuzenak) gehiegizko aurreikuspenak (gehiegi) eta ahaztutakoak (ahaztuak) ikus daitezke balio absolututan eta baita prezisio, estaldura eta F_1 neurri tipikoak ere.

PropBankerako emaitzak apur bat hobeak dira VerbNetekoekin alderatuta. Baliteke honen arrazoia multzo bakoitzaren rol kopuruan egotea. Izan ere, VerbNeteko rolak gehiago dira PropBankekoen aurrean eta horrek VerbNeteko sailkatzailearen zailtasunak areagotu beharko lituzke. Hala ere, kontuan hartuta bien arteko diferentziak ere ez direla horren nabariak eta, esan bezala, atazak zailagoa izan beharko lukeela, VerbNeteko emaitzak nahiko konpetenteak dira, eta ausartu gintezke esatera, ondorioz, rol tematikoek multzo kontsistente eta ikaserrazago bat osatzen dutela, agian aditzetik aditzera egonkorragoak direlako.

CoNLL ebaluazio ingurunean (ikusi III.2 taulan), PropBank sailkatzaileak apur bat behera egiten du, nahiz eta aurreko inguruneko emaitzekin alderatuta, diferentziak ez diren estatistikoki esanguratsuak. Bestalde, rol tematikoen 1.6 puntuko erorketa esanguratsua da eta horrek VerbNeteko sailkatzaileak aditzaren adierarekiko izan dezakeen sentsibilitate handia erakuts dezake, izan ere, gogora dezagun, ebaluazio ingurune honetan sailkatzaileak ez daude eskuz etiketatutako aditz adierak erabiltzeko baimenduta.

Sailkatzaileek aditzaren adiera bi modutan erabiltzen dute "SemEval" ingurunean: (1) ezaugarri arrunt bezala eta (2) rol sekuentzia probableena

[7] *Bootstrap resampling* (Noreen, 1989) bidez kalkulatutako tarteak. Tarte hauetatik haratago dauden sistemen emaitzak positiboki ala negatiboki esanguratsuak dira %95 ziurtasunarekin

	zuzen	gehiegi	faltan	prezisioa	estaldura	F_1	F_1 core	F_1 adj.
PropBank	5,977	1,424	1,767	80.76	77.18	78.93 ±0.9	81.64	71.90
VerbNet	5,816	1,548	1,928	78.98	75.10	76.99 ±0.9	79.44	70.20

Taula III.2: CoNLL testuinguruko emaitzak PropBankeko rol multzorako (goian) eta VerbNeteko rol tematikoetarako. Sailkatzaile bakoitzak guztira asmatutakoak (zuzenak) gehiegizko aurreikuspenak (gehiegi) eta ahaztutakoak (ahaztuak) ikus daitezke balio absolututan eta baita prezisio, estaldura eta F_1 neurri tipikoak ere.

	zuzen	gehiegi	faltan	prezisioa	estaldura	F_1	F_1 core	F_1 adj.
PropBank	5,972	1,434	1,772	80.64	77.12	78.84 ±0.9	81.49	71.50
VerbNet	5,746	1,669	1,998	77.49	74.20	75.81 ±0.9	77.60	71.67

Taula III.3: CoNLL testuinguruko emaitzak, 5. murriztapena kontuan hartu gabe, PropBankeko rol multzorako (goian) eta VerbNeteko rol tematikoetarako. Sailkatzaile bakoitzak guztira asmatutakoak (zuzenak) gehiegizko aurreikuspenak (gehiegi) eta ahaztutakoak (ahaztuak) ikus daitezke balio absolututan eta baita prezisio, estaldura eta F_1 neurri tipikoak ere.

Viterbi bidez bilatzeko murriztapen bezala. Baliteke VerbNet sailkatzailearen emaitzek "CoNLL" ingurunean duten erorketa , hain zuzen, 5. murriztapen horren ahultzeak ekartzen duela. Izan ere, SemEval inguruneko aditzen adierei esker, 5. murriztapen horrek asko laguntzen du sekuentziako rol baimenduen multzoa txikitzen eta, era berean, sailkatzailearen lana errazten. CoNLL ingurunean aldiz, aditz bakoitzak jasan ditzakeen rol tematikoak gehiago dira (%60 batez beste) eta horrek VerbNet sailkatzailearen lana nabariki zailtzea eta emaitzen erortzea dakar.

Hipotesi honen aldeko frogak eskuratzeko beste esperimentu multzo bat egitea lagungarri gerta daiteke. CoNLL ebaluazio ingurunean sailkatzaileak martxan jarri eta 5. murriztapen hori erabiltzea debekatzen badiegu, neurtu dezakegu zenbaterainokoa den mesedegarria sailkatzaile bakoitzarentzat aditzaren (edota aditz adieraren) rol posibleak jakitea. III.3 taulako "CoNLL setting (no 5th)" lerroan ikus daiteke, PropBankeko emaitzen erorketa txikia den bitartean, VerbNetekoek erorketa garrantzitsuak eta estatistikoki esanguratsuak jasaten dituztela.

Datuen beste ikuspegi bat emateko asmoz, sailkatzaileek ebaluazio ingurune bakoitzean argumentuetan eta adjuntuetan orokorrean lortzen dituz-

ten F_1 emaitzak ematen dira emaitza taula guztietako azken bi zutabeetan. PropBanken ikusten da, SemEval ingurunetik CoNLL ingurunera, emaitzak gradualki okertzen direla bai *core* eta baita argumentu adjuntuentzako ere. VerbNeten aldiz, erorketa nabariagoa da rol tematikoetan (3.7 puntu) adjuntuetan baino. Azken hauen erorketa PropBankeko adjuntuen oso antzekoa da. Emaitza hauen arabera, aditzaren adiera eta frameari [8] buruzko informazioak garrantzi handiagoa du VerbNet sailkatzailean PropBankekoan baino, izan ere, informazio horrek core argumentuei eta rol tematikoei eragiten die batez ere, eta ikusi dugu, VerbNeten, PropBanken ez bezala, aditzak jasan ditzakeen rol tematiko posibleak zehazteak asko laguntzen diola sailkatzaileari bere lana egiteko garaian. Honen arrazoietako bat, bi rol multzoen izaera berezituan egon daiteke. Gaur egungo SRL sistemen oinarrizko ezagutza sintaxitik dator batez ere, eta ez argumentuen ezaugarri semantikoetatik. Horrek PropBank sailkatzailearen alde egin dezake, argumentu zenbakituek lotura estua baitute aditz-argumentuen sintaxiarekin, rol tematikoek baino gehiago. Testuinguru sintaktikoari soilik erreparatzen badiogu beraz, errazagoa da Arg0-5 argumentuen artean bereiztea, *Theme* edota *Topic* bezalako rol tematikoen artean bereiztea baino. Bistan da, rol tematiko askoren arteko ezberdintasuna semantikoa dela batez ere, eta nekez lortuko dugula haien artean bereiztea ezaugarri sintaktikoak hartzen baditugu jakintza iturri nagusi bezala.

Emaitzetan gehiago sakontzeko lagungarri gertatzen da III.9 taulan agertzen zaigun rolez rol banatutako ebaluazioa. Bi sailkatzaileek maiztasun handieneko roletan duten errendimendua oso antzekoa da. Arg0-k %88.49 lortzen du, eta bere rol tematiko baliokideek, *Agent* eta *Experiencerrek*, %87.31 eta %87.76 hurrenez hurren. Arg1-ek %79.91 lortzen du eta haren VerbNeteko baliokideek, *Theme*, *Topic* eta *Patient*-ek, %75.46, %85.70 eta %78.64 hurrenez hurren.

Emaitzak aditz maiztasunaren arabera ordenatzen baditugu (III.8. taula) ikus dezakegu bai PropBank eta bai VerbNeten emaitzak okertu egiten direla aditzaren maiztasunak behera egiten duen heinean. Oro har, PropBankeko emaitzak hobeak dira maiztasun tarte guztietan altuenean izan ezik [9].

Sekzio honetan zehar ikusi ditugun emaitzen arabera, VerbNeteko rol tematikoak ahulagoak dira sailkatzailearen sendotasuna eta orokortzeko gai-

[8] Rol framea aditzak baimenduta duen rol multzoa da. Datu hau, PropBankeko framesetan eta VerbNeteko klaseetan ematen da

[9] Maiztasun tarte altuenak aditz bakarra hartzen du kontuan (*say*)

tasunaren aldetik behintzat. Hurrengo puntuetan azterketa sakonago bat egingo dugu, arazo edo fenomeno hau gertuagotik ikusi eta ezagutzeko.

III.3.1 Aditz ezezagunetara orokortzeko gaitasuna

Printzipioz, PropBankeko core rolak (Arg0-5) aditzarekiko menpekoak dira, hau da, aditzaren arabera interpretazio semantiko bat edo beste jasotzen dute. Beste modu batera esanda, aditz bakoitzaren framesetean zehazten da argumentu zenbakitu bakoitzaren interpretazio semantikoa eta, hortaz, ezin daiteke, adibidez, bi aditz ezberdinen Arg2 argumentuek esanahi berbera izango dutenik esan. Dena dela, lehenago aipatu dugun moduan, PropBankeko bi rol usuenak, Arg0 eta Arg1, kontsisteteak dira ia beti aditzetik aditzera, beren interpretazio semantiko orokorra *Agent* eta *Patient* direlako, hurrenez hurren, kasurik gehienetan. Bestalde, VerbNeteko rol tematikoak guztiz independenteak dira aditzarekiko eta hortaz, beren esanahia ez da aldatzen aditzetik aditzera (*Instrument* bat, adibidez, beti da *Instrument* eta ez du bere esanahi orokorra aditzarekin aldatzen). Era berean, PropBanken framesetekin bezala, VerbNeteko aditz bakoitzak baimendutako rolen eta haien arteko konbinazio posibleak zerrendatzen ditu klase bakoitzean.

Azpi-atal honetan egingo ditugun esperimentuekin, rol multzo bakoitzean entrenatutako sistemek aditz ezezagunen aurrean duten jokaera aztertuko dugu. Arazo hau ezin da inolaz ere artifizialtzat hartu oso ohikoa baita testeko datuen artean entrenamenduko adibiderik ez duten aditzak aurkitzea. Hasiera batean aditzarekiko duten independentziagatik espero daiteke VerbNeteko label semantikoen sailkatzaileak sendotasun ona erakustea aditz ezezagunen argumentuak etiketatzeko garaian. Bestalde, ezin dezakegu ahaztu aditzari buruzko informazio semantikoaren gabeziak kontuan hartzeko moduko kalteak eragin ditzakeela VerbNeteko sailkatzailean.

Esperimentu honetarako, aditz ezezagunen test-multzo bat osatu genuen entrenamendu corpuseko 10 aditzen adibideekin. Aditzak ausaz aukeratu ziren eta, noski, esperimentuak hala eskatzen zuenez, beren proposizio guztiak entrenamendu corpusetik ezabatu ziren. Hauek dira aditz ezezagunen testerako aukeratutako aditzak: *allow, approve, buy, find, improve, kill, produce, prove, report* eta *rush*.

Emaitzak III.4 taulako azken lerroan ikus daitezke. Entrenamenduko datu multzoa beste esperimentuetan erabilitakoa baino txikiagoa denez, emaitzak ezin dira beste esperimentuetako emaitzekin zuzenean konparatu.

	zuzen	gehiegi	faltan	prezisioa	estaldura	F_1	F_1 core	F_1 adj.
PropBank	267	89	106	75.00	71.58	73.25 ±4.0	76.21	64.92
VerbNet	207	136	166	60.35	55.50	57.82 ±4.3	55.04	63.41

Taula III.4: Aditz ezezagunen gainean egindako esperimentuen emaitzak, PropBankeko rol multzorako (goian) eta VerbNeteko rol tematikoetarako. Sailkatzaile bakoitzak guztira asmatutakoak (zuzen) gehiegizko aurreikuspenak (gehiegi) eta ahaztutakoak (faltan) ikus daitezke balio absolututan eta baita precision, recall eta F_1 neurri tipikoak ere.

	zuzenak	gehiegi	ahaztuak	precision	recall	F_1	F_1 core	F_1 adj.
PropBank	5,557	1,828	2,187	75.25	71.76	73.46 ±1.0	74.87	70.11
VerbNet	4,679	2,724	3,065	63.20	60.42	61.78 ±0.9	59.19	69.95

Taula III.5: Aditzetik eratorritako ezaugarririk gabe egindako esperimentuen emaitzak, PropBankeko rol multzorako (goian) eta VerbNeteko rol tematikoetarako. Sailkatzaile bakoitzak guztira asmatutakoak (zuzen) gehiegizko aurreikuspenak (gehiegi) eta ahaztutakoak (faltan) ikus daitezke balio absolututan eta baita precision, recall eta F_1 neurri tipikoak ere.

Emaitzek diote PropBankeko sailkatzailearen errendimendua VerbNetekoa baina askoz hobea dela (15 puntuko aldea dago bien artean) eta, hortaz, aditzari buruzko informazioak berebiziko garrantzia duela rol tematikoak egoki markatzen laguntzeko. Azken ondorio hau baieztatzeko helburuarekin gauzatu genuen hurrengo azpiatalean deskribatzen dugun esperimentua.

III.3.2 Aditz ezaugarriekiko sentikortasuna

Esperimentu honekin rol multzoen sentikortasuna neurtu nahi dugu sailkatzaileek aditzetik eratorritako informaziorik erabiltzen ez dutenean. Horretarako, entrenamendu eta testeko datu multzoetatik aditzari erreferentzia egiten dieten ezaugarri guztiak ezabatu genituen: aditzaren forma, lemma, aditzaren PoS labela, eta aditza agertzen den ezaugarri konbinatu guztiak (ikusi III.2.2 atala)

Emaitzak III.5 taulan ikus daitezke. PropBankek 5 puntutako erorketa gogorra jasaten du baina VerbNetek, 15 puntuko jaitsiera dramatikoarekin, oraindik gogorragoa.

Erorketa hauek gertuagotik azter ditzakegu III.9 taulako rol label indibidualen erorketei erreparatuta. “CoNLL setting” eta “no verb features

setting" zutabeen arteko diferentziek erakusten digute rol bakoitzak testuinguru erreal batetik, aditzik gabeko testuinguru batera pasatzean jasoko lukeen erorketa. Emaitzen arabera, badirudi Arg0 eta Arg1 rolek nahiko ongi erantzuten diotela aditzari buruzko informazio gabeziari; ez hain ongi, ordea, Arg2 eta Arg3 rolek. Azken bi rol hauek testeko datu multzoan duten maiztasun absolutua erlatiboki txikia denez, PropBanken erorketa orokorra ez da nabariki kaltetua gertatzen.

VerbNeten erorketak sarraski itxura handiagoa dauka aditzen ezaugarriak galtzen direnean. Rol tematiko usuenei erreparatuz, ikus daiteke *Experiencer*, *Agent* eta *Topic* rolek, 1, 10 eta 12 puntu galtzen dituztela CoNLL testuinguruko emaitzekin alderatuta. Baina beste roletarako erorketak oraindik gogorragoak dira (*Theme* eta *Patient* adibidez, 23 eta 50 puntu erortzen dira hurrenez hurren). Interesgarria da ikustea adjuntu motako rolek emaitza oso antzekoak lortzen dituztela bai PropBankeko sailkatzailearekin eta baita VerbNetekoarekin ere. Horrela, III.5 taulako azken bi lerroek argi eta garbi erakusten dute core argumentuak eta rol tematikoak direla sailkatzaileen errendimendu galeran erantzukizun handiena dutenak.

Badirudi, Arg0 eta Arg1 argumentuek aditzetik aditzera erakusten duten kontsistentzia dela PropBanken sendotasunaren sekretua. Gainera, emaitzen arabera beti ere, bi argumentu hauek nahiko erraz detekta daitezke lexikalizatu gabeko ezaugarri sintaktikoetan oinarrituta, hau da, aditza eta uneko argumentuaren gunea ezagutu gabe. Bestalde, Arg2-5 argumentuen artean bereizteko aditzaren azpikategorizazioari buruzko informazioak garrantzi handiagoa dauka eta ondorioz, informazio honen gabeziak edo kalitate txarrak kalte egiten die.

VerbNeteko rol tematikoen arteko ezberdintasuna aldiz semantikoagoa da askotan sintaktikoa baino, eta badirudi ale xeheko bereizketa horrek kalte egiten diela ebaluazio testuinguruaren baldintzak "gogorrak" direnean. Ez dirudi, adibidez, *Agent* eta *Experiencer* rolen artean bereizteko, sintaxia oso lagungarria denik, ezta *Theme-Topic-Patient* rolen artean bereizteko ere. Aditzaren gaineko informazioaren gabeziak ezaugarri semantiko garrantzitsuak galtzea dakar eta testuinguru horretan zailagoa gertatzen da VerbNeteko sailkatzailearen lana PropBankekoarena baino. Normalean Subjektu/Objektu funtzio sintaktikorik ez duten rolen artean ere (*Recipient, Source, Product, Stimulus...* zaila gertatzen da ezaugarri semantikorik gabe rolen artean bereiztea.

III.4 Mapaketa VNeko rol tematikoetara

Aurrerago aipatu bezala, PropBankeko rolen interpretazioa aditzaren adieraren araberakoa da, eta hori NLPko aplikazioentzako arazo bat izan daiteke, rolari bakarrik erreparatuta ezin baita zehazki jakin zein funtzio semantikoren aurrean gauden. Hori jakinda, egokia izan daiteke esaldiak VerbNeteko rol tematikoekin etiketatzeko bideak aztertzea, izan ere, jakina da dagoeneko, VerbNeteko rol tematikoek aditzaren beharrik ez dutela beren esanahia neurri batean ezagutarazteko, eta horrek, zalantzarik gabe, esportaziorako produktu interesgarriagoan bilakatzen ditu. Bi dira rol tematikoak eskuratzeko aztertutako bideak: (1) esaldia VerbNeteko rol tematikoen sailkatzailearekin zuzenean etiketatzen eta (2) PropBank sailkatzailearen irteera (argumentu zenbakituak) SemLinkeko PropBank-VerbNet mappingaren bidez rol tematikoetara itzultzea.

Lehenengo hurbilketaren emaitzak baditugu dagoeneko III.2 eta III.1 tauletan. Bigarren hurbilketarako besterik gabe PropBankeko rol zenbakituak rol tematikoetara itzuli behar ditugu SemLink erabiliz. SemLinken, rolen arteko mapaketa aditzez aditz egiten denez, beharrezkoa da aditzaren adiera (kasu honetan, bere VerbNet klasea) ezagutzea rolen bihurketa egiteko. Demagun *allow* aditzaren proposizio bat dugula testeko corpusean eta bere VerbNet klasea 29.5 dela. Orduan, SemLinkeko hurrengo sarrera erabiliko genuke proposizioko Arg0 eta Arg1 argumentuak *Agent* eta *Predicate* rol tematikoetara bihurtzeko.

```
<predicate lemma="allow">
   <argmap pb-roleset="allow.01" vn-class="29.5">
      <role pb-arg="1" vn-theta="Predicate" />
      <role pb-arg="0" vn-theta="Agent" />
   </argmap>
</predicate>
```

Proposizio bakoitzeko aditzaren adiera lortzeko bi modu aurreikusi ditugu: (1)sarrerako datuetan eskuz jarrita dagoen VerbNet klasea zuzenean hartzea (alegia, beti VerbNet klase zuzena erabiltzea itzulpena egiteko) edota (2) VerbNet klasea desanbiguatzeko sailkatzaile baten bidez.

Eskuz jarritako VerbNet klaseekin lortutako emaitzak III.6 taulan ikus daitezke. Emaitza hauek SemEval testuinguruan VerbNet sailkatzaileak lortutakoekin alderatzen baditugu, ikus dezakegu 0.5 puntu hobeak eta, hortaz, estatistikoki ez esanguratsuak direla.

Test on WSJ	guztira	core	adj.
PropBank to VerbNet (hand)	79.17 ±0.9	81.77	72.50
VerbNet (SemEval setting)	78.61 ±0.9	81.28	71.84
PropBank to VerbNet (MF)	77.15 ±0.9	79.09	71.90
VerbNet (CoNLL setting)	76.99 ±0.9	79.44	70.88

Taula III.6: VerbNeteko rol tematikoak eskuratzeko hainbat hurbilpenen emaitzak WSJ test corpusaren gainean. "PropBank to VerbNet (hand)" lerroan, mapaketa eskuz etiketatutako VerbNet klaseekin egitean lortzen diren emaitzak ikus daitezke, eta "PropBank to VerbNet (MF)" lerroan berriz, maiztasun maximoko klaseekin lortzen direnak. "VerbNet (SemEval setting)" eta "VerbNet (CoNLL setting)" lerroak (ikusi III.1 eta III.2 taulak), mapaketa eta rol tematikoen zuzeneko sailkapenaren arteko konparaketa errazteko jarri da. Hurbilpen bakoitzaren F_1 balioak azken hiru zutabeetan ikus daitezke. "guztira" zutabean argumentu guztien (core+adj) gaineko emaitzak ikus daitezke; core eta adjuntuekin lortutako F_1 balioak berriz, izen bereko zutabeen azpian.

Lehenengo hurbilpena mundu errealeko baldintzetara pixka bat gerturatzeko, bigarren esperimentuan eskuz etiketatutako VerbNet klasea baztertu eta maiztasun maximoko sailkatzaile bat erabili genuen aditzaren klasea asmatzen saiatzeko [10]. Harrigarria badirudi ere, VerbNet klase usuena hartzearen accuracy-a %97a da. *allow* aditzaren kasuan, maiztasun handieneko klasea 29.5 da beraz goiko adibidean ikusi dugun SemLink sarrera berbera erabiliko genuke aditz horren proposizio guztiak itzultzeko. III.6 taulako 2. lerroan ("PropBank to VerbNet (MF)") aditzaren klase usuena erabiliz lortutako emaitzak erakusten dira (eta baita konparaziotarako lagunduko digun CoNLL settingeko VerbNet sailkatzailearen ("VerbNet (CoNLL setting)") emaitzak ere). Eskuz etiketatutako VerbNet klasea ("PropBank to VerbNet (hand)") erabiltzearekin alderatuta, erorketa nabaria da. Hala ere, ez da diferentzia garrantzitsurik ikusten baldintza berdinetan entrenatutako VerbNet sailkatzailearekiko ("VerbNet (CoNLL setting)").

III.7 taulako azken bi lerroek aztertzen ari garen rol tematikoak lortzeko bi hurbilpenen domeinuz kanpoko emaitzak erakusten dituzte. Diferentziak,

[10] Testeko aditz baten klasea, aditz horrek entrenamendu korpusean maizen hartu duen klase berbera izango da

Test on Brown	guztira	core	adj.
PropBank to VerbNet (MF)	64.79 ±1.0	68.93	55.94
VerbNet (CoNLL setting)	62.87 ±1.0	67.07	54.69

Taula III.7: VerbNeteko rol tematikoak eskuratzeko bi hurbilpenen emaitzak domeinuz kanpoko Brown corpusaren gainean. "PropBank to VerbNet (MF)" lerroan, mapaketa maiztasun maximoko klaseekin egitean lortzen diren emaitzak ikus daitezke. "VerbNet (CoNLL setting)" lerroa (ikusi III.2 taula), mapaketa eta rol tematikoen zuzeneko sailkapenaren arteko konparaketa errazteko jarri da. Hurbilpen bakoitzaren F_1 balioak azken hiru zutabeetan ikus daitezke. "guztira" zutabean argumentu guztien (core+adj) gaineko emaitzak ikus daitezke; core eta adjuntuekin lortutako F_1 balioak berriz, izen bereko zutabeen azpian.

Freq.	PBank	VNet	Freq.	PBank	VNet
0-50	74,21	71,11	500-900	77,97	75,77
50-100	74,79	71,83	> 900	91,83	92,23
100-500	77,16	75,41			

Taula III.8: F_1 balioak entrenamenduko aditzen maiztasunaren arabera.

kasu honetan, garrantzitsuak eta estatistikoki esanguratsuak dira mapaketaren hurbilpenaren alde. VerbNeteko sailkatzaileak ez du, beraz, PropBankekoak domeinuz kanpoko corpusetan erakusten duen sendotasuna berdintzerik lortzen.

III.5 Erlazionatutako lanak

Dakigunaren arabera, rol multzoen arteko konparaketa egiten duten beste bi lan aurkeztu dira gutxienez nazioarteko kongresuetan. Gildea eta Jurafskyk (2002) FrameNeteko frame elementuak *rol tematiko abstraktuen* (hau da, *Agent*, *Theme* edota *Location*) bezalako rol orokor batzuen multzo batera bihurtu zituzten eta ondorio gisa beraien sistemak rol horiek emaitzetan inolako degradaziorik eragin gabe erabil zitzakeela adierazi zuten.

Li, Loper eta Palmerren (2007) azterketa lotuago dago gure lanarekin. Haiek ere PropBank eta VerbNeteko rol multzoak alderatu zituzten, bai-

na beren interesa PropBankeko Arg2 argumentura bideratu zuten. Haien lanaren arabera, sailkatzaileek Arg2 argumentuetan izan oi duten errendimendu eskasa nabariki hobetu daiteke entrenamenduko Arg2 instantziak rol tematikoetan oinarritutako beste azpimultzoetan antolatuz gero (adibidez, Arg2-Instrument, Arg2-Stimulus...). Egileen arabera, VerbNeteko rolak aditzetik aditzera kontsistenteagoak direnez, rol tematikoek entrenamenduko instantzia multzo egokiagoak dituzte PropBankeko rol zenbakituek dutenarekin alderatuta. Horrela, Arg2 bezala, semantikoki heterogeneoak diren argumentuak rol tematikoen arabera berrantolatuz gero, sailkatzaileak arazo gutxiago izango lituzke ikasketa garaian eta horrek Arg2 argumentuaren emaitzak igoko lituzke, domeinu eta aditz ezezagunekin batez ere. Artikuluan, egileek beren ideiak aldeko emaitzekin laguntzen dituzte.

Yi eta besteek egiten dutenaren kontra, guk ez dugu gure azterketa Arg2 argumentura zuzendu, eta besterik gabe bi rol multzoak bere horretan konparatu ditugu inolako nahasketarik gabe.

Lotura estuagoa du gure lanarekin Merlo eta Vanderplasek (2009) beranduago aurkeztu zuten lana. Hauek ere PropBank eta VerbNet rol eskemen arteko azterketa konparatiboa egin zuten baina, guk ez bezala, informazio teorian oinarritutako metodo orokorrago bat jarraituz. Laburbilduz, (Loper *et al.*, 2007b; Yi *et al.*, 2007; Zapirain *et al.*, 2008a, b) lanetan azpimarratzen dena berresten dute entropia, *mutual information*, eta antzeko neurri ezagunak abiapuntu hartuz. Aipatzekoa da, Merlo eta Van der Plasen ondorio batek ez duela guztiz bat egiten guk tesi atal honetan eta (Zapirain *et al.*, 2009) lanean plazaratutako batekin. Egileek, ustekabean, III.2, III.4, eta III.5 tauletako informazioa interpretatzeko beste modu bat topatu eta VerbNeteko rolak, oro-har, rol ezezaguna asmatzeko ataza orokorrean (CoNLL testuinguruan) PropBankeko rolak baino "hobeak" izan daitezkeela adierazten dute, guk kontrakoa esan genuen bitartean. Ondorio horretara iristeko errorearen ratioaren erredukzioa (ERR) izan zuten kontuan. Haien kalkuluen arabera, VerbNeteko sailkatzailearen ERRa txikiagoa da rolak etiketatzeko "CoNLL testuinguru" orokorrean (ikusi III.2). Merlo eta Van der Plasen metodoa eta ondorioa oso interesgarriak iruditzen zaizkigun arren, badirudi apur bat haratago jotzerik ere badagoela planteatzen duten arrazonamenduan: sistema baten ERRa erreferentzia puntu edo oinarri sistema (*baseline*) batekiko kalkulatzen da, eta beraz, balio erlatibo bat dela esan daiteke (abiadura eta denbora magnitude fisikoak bezala). Egileek beren kalkuluen erreferentzia puntu bezala eskema bakoitzaren rolen maiztasunean oinarritutako sailkatzaile sinple bat hartzen dute. Haien erabakiz, sailkatzaile horrek

	CoNLL setting				No verb features	
	PBank		VNet		PBank	VNet
	corr.	F_1	corr.	F_1	F_1	F_1
Overall	5977	78.93	5816	76.99	73.46	61.78
Arg0	1919	88.49			84.02	
Arg1	2240	79.81			73.29	
Arg2	303	65.44			48.58	
Arg3	10	52.63			14.29	
Actor1			44	85.44		0.00
Actor2			10	71.43		25.00
Agent			1603	87.31		77.21
Attribut.			25	71.43		50.79
Cause			51	62.20		5.61
Experien.			215	87.76		86.69
Location			31	64.58		25.00
Patient1			38	67.86		5.71
Patient			208	78.64		25.06
Patient2			21	67.74		43.33
Predicate			83	62.88		28.69
Product			44	61.97		2.44
Recipient			85	79.81		62.73
Source			29	60.42		30.95
Stimulus			39	63.93		13.70
Theme			1021	75.46		52.14
Theme1			20	57.14		4.44
Theme2			21	70.00		23.53
Topic			683	85.70		73.58
ADV	132	53.44	129	52.12	52.67	53.31
CAU	13	53.06	13	52.00	53.06	45.83
DIR	22	53.01	27	56.84	40.00	46.34
DIS	133	77.78	137	79.42	77.25	78.34
LOC	126	61.76	126	61.02	59.56	57.34
MNR	109	58.29	111	54.81	52.99	51.49
MOD	249	96.14	248	95.75	96.12	95.57
NEG	124	98.41	124	98.80	98.41	98.01
PNC	26	44.07	29	44.62	38.33	41.79
TMP	453	75.00	450	73.71	73.06	73.89

Taula III.9: CoNLL ebaluazio ingurunean lortutako emaitzak. Erreferentziazko rolak, aditzak eta 10 aldiz baino gutxiago agertzen diren rolak ez dira erakusten (horien artean Arg4 eta Arg5). Azken bi zutabeak CoNLL ebaluazio ingurunearen barruko emaitzak erakusteko erabiltzen dira ere, baina aditzetatik eratorriko ezaugarriak erabiltzen ez direnean.

beti maiztasun maximoa duen rola itzuliko du, bai PropBanken (Arg1) eta baita VerbNeten (*Agent*) ere. Noski, VerbNeteko rolen distribuzioa PropBankeko rolena baino homogeneoagoa denez, *baseline* oinarri sistema honen errendimenduan diferentziak ikusiko dira eskema batetik bestera. Zehazki, PropBanken %52ko F_1 balioa eskuratzen du eta VerbNeten, %33koa. Hala

ere, gure ustez, zalantzak egon litezke ERRa kalkulatzeko kontuan hartutako oinarri sistema hau egokiena den ala ez erabakitzeko orduan. Egia da, hasiera batean behintzat, metodoak "bidezkoa" dirudiela, izan ere VerbNeteko rol tematikoen kopurua *core* rolena baino askoz handiagoa da eta zentzuzkoa dirudi pentsatzeak VerbNeteko rolekin lan egiten duen edozein *baseline* sistemak *core* rolekin baino emaitza kaxkarragoak atera beharko lituzkeela. Tamalez, horrek ez du esan nahi bi sistemen arteko diferentzia Merlo eta Van der Plasek proposatzen dutena izango denik, eta beraz, gure ustez, ERRa kalkulatzeko kontuan hartu diren erreferentzia puntuak, kalkuluak eta gainontzeko ondorioak ere alda litezke. Gure ustez, VerbNeteko *baseline* oinarri sistemaren zailtasuna ez da Merlo eta Van der Plasek proposatzen duten sistemarena bezain handia. Egia da, VerbNeten rol gehiago daudela eta sailkatzaileak horien guztien artean erabakitzeko lana hartu behar duela, baina egia da baita ere, VerbNeteko rol guztiak ezin direla aditz guztien argumentuekin azaldu eta beraz, atazaren zailtasuna ez dela hasiera batean pentsa litekeen bezain altua. Honek esan nahi du, egileek hautatutako oinarri sistema ez dela, agian, egokiena, eta merezi duela, gutxienez, kontuan hartzea beste oinarri sistema "bidezkoagoak" ere ezarri daitezkeela ERRa kalkulatu eta ondorioak ateratzeko.

III.6 Ondorioak

Kapitulu honetan, emaitza lehiakorrak erakusten dituen sailkatzaile baten jokaera aztertu dugu bi rol multzo ezberdinekin lan egiten duenean (PropBankeko rol zenbakituak eta VerbNeteko rol tematikoak). Sailkatzailearen errendimendua hainbat ebaluazio inguruneetan ebaluatu dugu, besteak beste, aditzari buruzko informazioa gehitu edo kentzen den inguruneetan, aditz ezezagun edota maiztasun baxuko aditzekin, eta domeinuz kanpoko corpusetan. Esperimentuen emaitzen arabera, ikusi dugu sailkatzaileak PropBankeko rolekin lan egiten duenean errendimendu hobea erakusten duela aipatu ingurune guztietan eta, beraz, bere sendotasuna VerbNeteko rolak erabiltzen dituenean baino handiagoa dela.

Bestalde, kontuan hartuta NLPko aplikazioen ikuspegitik VerbNeteko rolak interesgarriagoak direla, bi modutara saiatu gara rol tematikoak eskuratzen: (1) Sailkatzailea VerbNeteko roletan zuzenean trebatzen eta (2) PropBankeko sailkatzailearen irteera SemLinkeko mapaketa erabiliz rol tematikoetara itzuliz. Bi hurbilpenek antzeko emaitzak ematen dizkigute WSJko

testean. Domeinuz kanpoko Brown corpusean aldiz, emaitzen arteko diferentziak estatistikoki esanguratsuak dira mapaketaren hurbilpenaren alde.

IV. KAPITULUA

Ezaugarri lexikalak orokortzen: Hautapen Murriztapenak rolen sailkatze automatikorako

Aurreko kapituluetan ikusi dugun moduan, SRLko sistema tipikoek etiketatze prozesua bi fasetan egin oi dute: argumentuen *identifikazioa* eta argumentuen *sailkapena*. Lehena, ataza sintaktiko bat da batez ere, eta aditz baten argumentuak heuristiko sinpleen bidez (Xue eta Palmer, 2004), edota sailkatzaile berezituen laguntzaz identifikatzean datza. Bigarren ataza ataza semantikoagoa da eta bertan, aurreko fasean identifikatutako argumentu bakoitzari rol semantiko bat esleitzen zaio. Zeregin horretarako oso lagungarriak izan daitezke sailkatzaileak erabiltzen dituen ezaugarri lexiko-semantikoak, batez ere aditzetik eta sailkatu beharreko argumentuaren gunetik eratortzen direnak (ikusi III.2). Ezaugarri hauekin gertatzen den sakabanaketa (*sparsity*) fenomenoaren eraginez, (corpus txikietan batez ere), gaur egungo SRL sistemek entrenamenduko lexikora gehiegi egokitzeko (*overfitting*) joera erakusten dute eta, neurri handi batean horregatik, domeinuz kanpoko corpus edota datu sorta berrietara zailtasun handiekin moldatzen dira (Pradhan *et al.*, 2005b).

SRL sistemek domeinuz kanpoko ingurunetara egokitzeko erakusten dituzten zailtasunen adibide garbi bat CoNLL-2004 eta CoNLL-2005eko ebaluazio saioek eman ziguten. Bertan ikusi ahal izan genuen rol etiketatzaileek oso emaitza onak ematen zituztela domeinu barruko datu multzoak etiketatzeko garaian, baina, aldiz, beren errendimendua larriki kaltetua gertatzen zela testeko corpusak domeinuz kanpokoak zirenean. Hau da, sistemak WSJ domeinuan entrenatu eta probatzen baziren, %80 inguruko F-score neurriak lortzen ziren; baina, aldiz, sistema horiek berak domeinuz kanpoko Brown corpusa etiketatzeko erabiltzen baziren, 10 puntutako galera larriak izaten zituzten batez beste.

Aurreko kapituluan SRL sistemen errendimendu galera honen arrazoiak aztertu ditugu. Pradhan *et al.*-ek (2008) erakutsi zuten errendimenduaren degradazioa *argumentuen sailkapen* prozesuan gertatzen zela batez ere eta, gainera, ezaugarri lexikalen sakabanaketa proposatu zuten galeraren arrazoi nagusienetako bat bezala. Egile berberek (2007b), ezaugarri ezberdinek SRLko ataza bakoitzean duten eragina aztertu ondoren, ikusi zuten aditza eta argumentuaren gunea bezalako ezaugarri lexikalek berebiziko garrantzia zutela argumentuen sailkapen prozesuan. Hori guztia kontuan hartuz, pentsa daiteke atributu lexikalen sakabanaketak edota beste degradazioren batek negatiboki eragingo lukeela argumentuen sailkapen prozesuan eta, ondorioz, baita SRL sistemen errendimendu orokorrean ere.

Atal honetan, hautapen murriztapenak erabiliko ditugu SRL sistemen dependentzia lexikala gutxitzen saiatzeko. Hautapen murriztapenek aditzek "gustuko" dituzten argumentuei buruzko ezaugarri semantikoak isla ditzakete eta horregatik, maiztasun baxuko edota ezezagunak diren ezaugarri lexikalak[1] orokortu eta ezaugarrien sakabanaketa gutxitzeko erabilgarriak izan daitezke. Horrela, gorago aipatu ditugun arrazoiengatik, pentsa dezakegu hautapen murriztapenek ezaugarri lexikoen kalitatea handitu eta argumentu sailkatzaileen errendimenduan positiboki eragiteko ahalmena dutela.

Ezaugarri lexikalek argumentuen sailkapen prozesuan duten garrantziaren adibide bat hurrengo esaldiekin jar daiteke:

[1]Ezaugarri bat ezezaguna dela diogu, entrenamendurako datu multzoan agertu ez bada. Era berean maiztasun baxuko ezaugarriak entrenamendurako corpusean maiztasun urria duten ezaugarriak dira.

JFK was assassinated (in Dallas)$_{Location}$[2]
JFK was assassinated (in November)$_{Temporal}$[3]

Bi esaldiek egitura sintaktiko berbera partekatzen dute, eta beraz ezaugarri lexikalek ("Dallas" eta "November" hitzek) berebiziko garrantzia dute bi argumentu adjuntuen artean bereiztu ahal izateko (Dallas-Location, November-Temporal). Baina, zer gertatzen da etiketatu beharreko testu batean, adibidez, "in Texas" (Texasen) edota "in Autumn" (udazkenean) bezalako espresio berriekin egiten badugu topo? Zer gertatzen da espresio horietako bat bera ere ez bada entrenamendurako datuen artean azaldu? Ba orduan SRL sistemak ezingo lituzkeela argumentu horien ezaugarri lexikoak kontuan hartu sailkapena egiteko; eta ezaugarri horiek dira, hain zuzen, aurrerago ikusiko dugunaren arabera, argumentuen sailkapen prozesuan gehien eragiten duten ezaugarriak.

Hautapen murriztapenek ezaugarri lexikoen kalitatea hobetzeko gaitasuna izan lezakete. Atal honetan erakutsiko dugu aditzek beren rolak betetzeko "izaera semantiko" konkretua duten argumentuak aukeratzeko joera erakusten dutela, alegia, rol bakoitzerako beren hautapen murriztapen konkretuak –beren preferentziak– badituztela, eta, hortaz, posible dela aditza eta argumentuaren gune semantikoa ezagututa bien arteko rola ere zein izan daitekeen jakitea. Horrela, "*to assassinate*" aditzaren bi argumentutan "Dallas" eta "November" izen guneak aurkitzen baditugu, gehienok pentsatuko genuke "Dallas" gunea lekuzko argumentuarena dela, eta "November" gunea berriz, denborazko argumentuarena. Izan ere, "*to assassinate*" aditzaren hautapen murriztapenak ezagutzen ditugu, eta badakigu bi gune horiek nekez topatuko luketela lekua beste rol batekin etiketatutako argumenturen batean. Munduari buruz dugun ezagutzak erraztu egiten ditu horrelako asoziazioak. Era berean, aditzek roletarako dituzten hautapen murriztapenak automatikoki sortuz, rol sailkatzaileek beste ezagutza iturri semantiko bat izango lukete beren inferentziak egiteko garaian.

Beraz, WordNet eta antzekotasun distribuzionaleko neurrietatik eratorritako hautapen murriztapenak erabiliz, posible litzateke ezaugarri lexikoen arteko antzekotasun semantikoak bilatu eta topatzea. Horrela, sailkatzaileek III.2 atalean ikusi ditugun ezaugarri lexikoez gain, antzekotasun semantiko horiek ere erabili ahalko lituzkete argumentuak modu sendoago batean sailkatzeko. Goian jarritako adibidearekin jarraituz, antzekotasun neurriak

[2]JFK Dallasen hil zuten.
[3]JFK maiatzean hil zuten.

baliatuz, “Autumn” eta “November” guneen arteko antzekotasun semantikoa neurtu eta, agian, oso antzekoak direla esateko gai izango ginateke. Ziur aski, gainera, hitzok aditzaren argumentu jakin bateko hautapen murriztapen berbereko kide izan daitezkeela pentsatu eta ondorioz gune ezagunaren rol berbera (*Temporal*) esleituko genioke gune ezezaguneko argumentuari. Hori posible balitz, sailkatzaileek ez lukete guztiz galduko “Autumn” edota “Texas” bezalako ezaugarri lexiko ezezagun baina garrantzitsuen eragina, eta zentzu horretan argumentu sailkatzaileen sendotasuna eta errendimendua hobetzeko aukerak izango genituzke.

Hurrengo ataletan, WordNet eta antzekotasun distribuzionalean oinarritutako hautapen murriztapenen eredu ezberdinak modu automatikoan lortzeko hainbat metodo sakonki aurkeztu eta rolen sailkatze automatikoko ataza erreal batean ebaluatuko ditugu.

IV.1 Aurrekariak

Hautapen murriztapenen sorkuntza automatikoa erlatiboki gai zaharra da eta modu askotara eta arazo askoren gainean aplikatu izan da urteetan zehar. Baina atal honen helburua ez da hautapen murriztapenen historiaren errebisio sakon bat egitea izango eta soilik gure lanarekin lotutako erreferentziak aipatuko ditugu.

Resnikek (1993) sintaxian ohikoak diren zenbait anbiguotasun arazo konpontzeko (izen konposatuak, koordinazioa, PP-attachment) WordNetetik eratorritako hautapen murriztapenak erabili zituen. Resniken ereduan, p aditzak gobernatutako r posizio sintaktiko batean, w_0 izen gune posiblearentzako hautapen murriztapena, $S_{Res}(p, r, w_0)$, ondoren datorren bezala formulatzen da:

$$S_{Res}(p, r, w_0) = \frac{\max_{c_0 \ni w_0} P(c_0|p, r) log \frac{P(c_0|p,r)}{P(c_0)}}{\sum_{c \in C} P(c|p, r) log \frac{P(c|p,r)}{P(c)}} \tag{IV.1}$$

Zenbakitzaileak w_0 gunea eta c_0 klase semantikoaren arteko egokitze maila neurtzen du. Izendatzaileak $P(C)$ eta $P(C|p, r)$ distribuzioen arteko entropia erlatiboa gordetzen du, non C klase semantiko guztien multzoa den. Klase semantiko bakoitzaren probabilitatea bere kideen (c) maiztasuna erabiliz estimatzen da. Tesi honetako IV.2.1 atalean, Resniken ereduaren berrinplementazio bat (Agirre eta Martinez, 2001) erabiliko dugu WordNeten

oinarritutako HM ereduetako bat modelatzeko.

Brockman eta Lapatak (2003) Resniken hautapen murriztapenak, Lin eta Aberen (1998) MDL metodoa, Clark eta Weir-en (2001) hurbilpena eta maiztasunetan oinarritutako beste hainbat ad-hoc metodo erabili zituzten (*aditz,funtzio-sintaktiko,izen-gune*) hirukoteentzako pisuak eman, eta gizaki batek egingo lituzkeen balorazioen pisuekin alderatzeko. Resniken hautapen murriztapenen emaitzak hoberenak izan ziren klaseetan oinarritutako metodoen artean, baina ez ziren probabilitate baldintzatuko eredu lexiko sinple batek ematen zituenak baino askoz hobeak. Maiztasunen kontaketa sinpleen emaitzak ere oso onak izan ziren.

Antzekotasun distribuzionala ere erabili izan da anbiguotasun sintaktikoaren arazoa gainditzen saiatzeko. Pantel eta Linek (2000) oso emaitza onak lortu zituzten Linen antzekotasun neurria (Lin, 1998) erabiliz. Egileek, nahiz eta esplizituki adierazi ez zuten, hautapen murriztapenen eredu distribuzional bat erabiltzen ari ziren ezaugarrien sakabanaketaren eragina leundu eta anbiguotasun sintaktikoa gutxitzeko. IV.4 irudian, besteak beste, egileek erabilitako antzekotasun formula ikus daiteke (sim_{Lin}). Linen antzekotasun neurriarekin osatutako thesaurusa publikoa da[4]. Antzekotasun distribuzionaleko xehetasunak IV.2.2 atalean ematen dira.

Hautapen murriztapenak Rol Semantikoetara aplikatzeko ideia apur bat berriagoa da eta, guk dakigula, Gildea eta Jurafsky (2002) dira lehenak eta bakarrak. Egileek beren lan aitzindari honetan, besteak beste, argumentuen izen guneak orokortzen saiatu eta gune ezezagunentzako irtenbideak bilatu zituzten FrameNeteko roletan oinarritutako SRL sistema bat hobetzeko asmoz. Hiru teknika azpimarra daitezke haien lanean:

- *Clustering* distribuzionala: egileek, rol semantiko jakin batekin etiketatuta egon daitezkeen guneak topatzeko, antzeko egoeratan azaltzen diren izenen *clustering* bat egiten dute corpus zabal batetik abiatuta. Rol jakin batekin etiketatuta agertzen diren hitzak, *cluster* edo multzo berekoak izan zitezkeenaren ideia jarraitzen dute egileek. Horrela NP argumentuetako guneak haien cluster klaseekin ordezten dituzte, guneen estaldura areagotu eta sakabanaketa gutxitzeko.

- WordNeten hierarkia semantikoaren erabilera: metodo hau *clustering* distribuzionalaren antzekoa da baina kasu honetan, argumentu guneak automatikoki sortutako *cluster* edo multzotan sailkatu beharrean,

[4] `http://www.cs.ualberta.ca/~lindek/downloads.htm`

WordNeteko hierarkiaren barnean kokatzen dituzte. Test multzoan gune ezezagun batekin topo egiten dutenean, hau ere hierarkian kokatu eta hiperonimia erlaziotatik gora egiten dute harik eta gune ezagun bat (edo haren ahaideren bat) topatu arte (McCarthy, 2000). Gune ezezagunak, orduan, topatutako adibidearen rola hartzen du.

- *Bootstrapping* etiketatu gabeko datuetatik: gune lexikoen estaldura hobetzeko egileek egiten duten azken saiakera, besterik gabe entrenamendu multzoa datu berriekin zabaltzea da. Datu horiek, hori bai, ez dute *gold standard* rolik. Horren ordez, etiketatzea jatorrizko corpusarekin entrenatutako sailkatzailearekin egiten dute.

Teknika hauen erabilerarekin sailkatzailearen errendimenduan oso hobekuntza txikiak lortu zituzten, gurea bezalako lanei ateak irekita utziz. Hurrengo ataletan, esperimentu zerrenda osagarri bat burutuko dugu, FrameNet baino rol multzo erabiliago baten gainean lan eginez (PropBank), Gildea eta Jurafsky-k baino hautapen murriztapenen sorta aberatsago bat erabiliz, eta domeinuz kanpoko analisia ere eskainiz.

Badira SRLaren testuinguruan kokatuko genituzkeen beste hainbat lan baina, oro har, sasi-atazak ebaztera bideratutako lanak izan dira. Erk-ek (2007), Pantel eta Linek (2000) aurkeztutakoa gogorarazten duen hautapen murriztapen eredu distribuzional bat aurkeztu zuen. Eredu honek w_0 hitza, p aditzak gobernatutako r rola duen argumentu batean egoteko joera jarraian azaltzen den bezala kalkulatzen zuen:

$$SP_{erk}(p, r, w_0) = \sum_{w \in Seen(p,r)} sim(w_0, w) \cdot weight(p, r, w) \qquad \text{(IV.2)}$$

non $sim(w_0, w)$ hitz baten (w_0) eta hitz ezagun baten (w) arteko antzekotasuna den, $Seen(p, r)$ p predikatua eta r rola duten entrenamenduko argumentuen guneak diren, eta $weight(p, r, w)$ w hitzaren pisua den.

Erken lanean oinarrizko eredua antzekotasun distribuzionaleko hainbat neurrirekin instantziatzen zen, besteak beste Linen antzekotasunarekin eta baita Jaccard eta kosinuaren neurriekin ere (ikusi IV.4 irudia). Instantzia hauek guztiak desanbiguazioko sasi-ataza batean ebaluatzen ziren, non helburua rol bati hobekien egokitzen zitzaizkion hitzen artean aukeratzea zen.

Gaur egun, hautapen murriztapenekin erabiltzeko aproposak diren hainbat antzekotasun neurri daude eskuragarri. Padó eta Lapatak (2007) adibi-

dez, antzekotasun distribuzionaleko eredu multzo zabal baten gaineko azterketa aurkeztu zuten. Antzekotasun eredu guztiak eraikitzeko baliatu zuten softwarea eskuragarri dago sarean[5].

Ikusiko dugun moduan, gure azterketan goian aipatutako teknikak eta antzekoak erabiliko ditugu, baina aurreko lanek ez bezala, hautapen murriztapenak ataza erreal batean aplikatzeko moldatuko ditugu eta ez gara sasi-ataza batean zentratuko ereduen ebaluazioa egiteko garaian.

IV.2 Hautapen Murriztapenen ereduak

Hurrengo azpi-ataletan tesi lan honetarako eraiki ditugun hautapen murriztapenen deskribapen zehatza egingo dugu. Erabiltzen duten errekurtsoaren arabera, bi multzotan banatuko ditugu:(1) WordNeteko hautapen murriztapenak eta (2) distribuzionalak. Jarraian, aditz-argumentuak rolekin etiketatzeko eredu baten proposamena egingo dugu eta eredu hori izango da esperimentazio fasean, hautapen murriztapenak bata bestearekin konparatu eta guztien ebaluazio orokorra egiteko erabiliko duguna.

Hautapen murriztapen distribuzionalekin alderatzen baditugu, WordNeten oinarritutako hautapen murriztapenen desabantaila nagusietako bat eskuz eraikitako errekurtsoen (WordNet) beharra dutela da. Muga honek negatiboki eragingo du hautapen murriztapen sistemaren estalduran eta baita bere orokortzeko gaitasunean ere, jakina baita eskuz eraikitako errekurtsoetan eta batez ere WordNet bezalako taxonomia batean, maiz nabaritzen direla hitzak faltan.

Ondoren azalduko ditugun ereduekin aditzek beren rolak jokatzeko baliatzen dituzten argumentuentzako hautapen murriztapenak modelatuko ditugu. Horrela, p aditza (edo preposizioa, ikusi IV.2.3 atala) eta r rola emanda hautapen murriztapenak definituko ditugu. Hautapen murriztapen horietako bakoitzak p aditzaren gidaritzapean, r rolarekin ikusitako argumentu guneak gordeko ditu, edozein argumentu nominal emanda, bere w_0 gunea zenbateraino egokitzen den zehazteko. Hori da oinarria. Hortik abiatuta, WN eta HM distribuzionalek egokitzapen hori modu batera ala bestera egingo dute. WordNeteko hautapen murriztapenek adibidez, argumentu gunearen synsetak erabiliko dituzte egokitzapena egiteko, eta eredu distribuzionalek, ikusiko dugun moduan, hainbat antzekotasun neurri.

[5] `http://www.coli.uni-saarland.de/~pado/dv.html`

IV.2.1 WordNeten oinarritutako Hautapen Murriztapen Ereduak

WordNet klaseak erabiltzen dituzten ereduei dagokienez, dagoeneko aurkeztutako Resniken eredua ez ezik (IV.1 ekuazioa), WordNet sakoneran eta synseten maiztasunean oinarritutako metodo propio eta sinpleago baten inplementazioa ere egin dugu (SP_{wn}) azpiatal honetan.

Resniken ereduan oinarritutako HMak

Resniken ereduan oinarritutako hautapen murriztapenak kalkulatzeko Agirre eta Martinezek (2001) aurkeztu zuten Resniken ereduaren berrinplementazioa erabili dugu. Konkretuki, egileek *word-to-class* izendatzen duten WordNeteko eredua moldatu dugu v aditzak r rolerako gustuko dituen argumentu guneen hautapen murriztapenak eratzeko. *Word-to-class* ereduarekin, egileek, p aditzaren r erlazio gramatikalean, s_i hitz adiera (synseta) zenbateraino egokitzen den adierazten dute. Horretarako, $W(s_i|p,r)$ pisu funtzioa definitzen dute:

$$W(s_i|p\,r) = \sum_{s \in hypers(s_i)} P(s_i|s) \times P(s|p\,r) = \sum_{s \in hypers(s_i)} \frac{\hat{f}r(s_i,s)}{\hat{f}r(s)} \times \frac{\hat{f}r(s\,p\,r)}{fr(p\,r)} \tag{IV.3}$$

Formulako maiztasunen kalkulua IV.1 irudiko kontaketen bidez egiten da. Bi motako maiztasunak ikus daitezke formuletan: $fr(s_i)$ eta $\hat{fr}(s_i)$. Lehena maiztasunaren kontaketa zehatza da (corpusean s_i synsetari dagokion hitza agertzen delako eta konta dezakegulako) eta bigarrena, estimatua (corpusean s_i synsetari dagokion hitzik ez dugulako eta bere hiperonimoen agerpenekin estimatu behar dugulako).

IV.4 formulan s synsetaren estimazioa bat egiten da bere hiponimoen agerpenean oinarrituz. Zatitzailea ondoko baldintza betetzeko aplikatzen den normalizatzailea da:

$$\sum_{s \supseteq s_i} P(s_i|s) = 1$$

IV.5 formulak bi synseten maiztasuna estimatzen du. Bi kasu daude: (1) lehena bigarrenaren arbasoa bada, maiztasuna 0 da, bestela (2) maiztasuna IV.4 formulan azaltzen den bezala estimatzen da.

IV.6 formulak *(synset, erlazio, aditz)* hiruko ezagunen (corpusean agertzen dira) kontaketatik abiatuz, hiperonimo guztientzako hirukoen maiztasuna estimatzen ditu. Berriro ere, zatitzailea hurrengoa ziurtatzeko ezartzen den normalizatzailea da:

$$\sum_{s} P(s|r\,p) = 1$$

$$\hat{f}r(s) = \sum_{s_i \in hyponims(s)} \frac{1}{hypers(s_i)} \times fr(s_i) \tag{IV.4}$$

$$\hat{f}r(s_i, s) = \begin{cases} \sum_{sj \in hypers(si)} \frac{1}{classes(s_j)} \times fr(s_j) & if\ s_i \in hyponims(s) \\ 0 & bestela \end{cases} \tag{IV.5}$$

$$\hat{f}r(s\,p\,v) = \sum_{s_i \in hypers(s)} \frac{1}{classes(s_i)} \times fr(s_i\,p\,v) \tag{IV.6}$$

Irudia IV.1: Maiztasunen estimazioa

Laburbilduz, IV.3 ekuazioari erreparatuz, egileek s_i hitz baten r erlaziorako egokitzapena (ala egokitzapenaren pisua) bere hiperonimoen ($hypers(s_i)$) egokitzapenera ($P(s|p\,r)$) ere baldintzatzen dute. Honek lexiko ezezagunaren arazoa gainditzeko aukera bat eman diezaguke, izan ere, hitza bera ezezaguna izan arren, haren hiperonimoak ezagunak izan daitezke. Ondorioz, kontuan hartuz hiperonimoen egokitzapena ere neur daitekeela, argumentu gune nominalak (ezezagunak barne) aditz/rol bikote batean zenbateraino egoki daitezkeen estima dezakegu ondoren agertzen den moduan:

$$SP_{AM}(p, r, w_0) = \sum_{w \in hypers(w_0)} W(w|p, r) \tag{IV.7}$$

```
n#00019671 7.956 communication "something that is communicated between people or groups"
n#04949838 4.257 message content subject_matter substance "what a communication that ..."
n#00018916 3.848 relation "an abstraction belonging to or characteristic of two entities"
n#00013018 3.574 abstraction "a concept formed by extracting common features from examples"
```

Irudia IV.2: `write-Arg1` hautapen murriztapeneko lehen 4 synsetak, haien pisua eta WordNeteko jatorrizko deskribapena. Deskribapenen itzulpen laburtua lerroz lerro: (1) *komunikazioa*, (2) *mezuaren edukia*, (3) *erlazioa*, (4) *abstrakzioa*.

```
n#00002086 5.875 life_form organism being living_thing "any living entity"
n#00001740 5.737 entity something "anything having existence (living or nonliving)"
n#00009457 4.782 object physical_object "a physical (tangible and visible) entity;"
n#00004123 4.351 person individual someone somebody mortal human soul "a human being;"
```

Irudia IV.3: `write-Arg0` hautapen murriztapeneko lehen 4 synsetak, haien pisua eta WordNeteko jatorrizko deskribapena. Deskribapenen itzulpen laburtua lerroz lerro: (1) bizi-forma, (2) entitatea (3) objektua (4) gizaki.

non $hypers(w_0)$, w_0 argumentu gunea eta haren hiperonimo guztien zerrenda den (adiera guztietarako).

IV.2 eta IV.3 iruditan, *to write* aditzak Arg1 eta Arg0 argumentuetarako gustukoen dituen synset zerrendaren zati bat ikus daiteke. Preferentzia horiek ikusita, ziurra da "writer" (idazle) gunea eta haren hiperonimoak gertuago egongo direla Arg0-rako preferentzietatik, besteetatik baino (kontrakoa gertatzen da "letter" (gutuna) gunearekin).

Lan honetan erabilitako hautapen murriztapenak sarean daude eskuragarri[6].

WordNeten oinarritutako HM prototipoa

Intuizioz esan genezake WorNeteko synset orokorrenak ongi egokitzen direla aldi berean hautapen murriztapen batean baino gehiagotan. Adibidez, $< entity >$ delako synset orokorra hitz askoren superklase bezala agertzen zaigu WordNeten. Synseta, beraz, egokia litzateke adibidez, *to break* aditzaren *Agent*, *Patient* edota *Instrument* argumentuen hautapen murriztapenak orokortzeko. Bistan da, ondorioz, egokitzapen horrek ez duela, kasu honetan behintzat, rolen artean bereizketak egiteko askorik laguntzen, eta dagoeneko

[6] `http://ixa2.si.ehu.es/know2/index.php/Inventario_recursos#semantic_lexicons`

azalduta geratu den moduan, hori da, hain zuzen ere, rol semantikoen etiketatze automatikoaren helburu garrantzitsuetako bat. Gure susmoa da synset espezifikoagoak hobeak direla hautapen murriztapenak modelatzeko (adibidez, badirudi $< tool >$ bezalako synset espezifikoagoek modu zehatzago batean mugatuko lituzketela *Instrument* argumentuen eskakizun semantikoak). Gauzak horrela, gure iritziz synset espezifikoen erabilera hobestea justifikatuta dago, hauek direlako azken batean hautapen murriztapenetako informazio semantiko esanguratsuena modu zehatzago baten islatuko dutenak.

WordNeten oinarritzen diren gure hautapen murriztapenak synseten multiset bezala modelatzen ditugu. M multiset[7] hauetan, entrenamenduko corpusean aditz eta rol berbera partekatzen duten argumentuen guneak gordetzen dira.

$$M(p,r) = \biguplus_{w \in Seen(p,r)} hyp(w) \tag{IV.8}$$

$Seen(p,r)$, p aditza eta r rola duten argumentuen guneak dira eta $hypers(w)$ funtzioak w argumentu gunearen synset eta hiperonimo guztiak ematen ditu. Multiseten edukia, itxura aldetik, IV.2 eta IV.3 iruditako synset zerrenden oso antzekoa da. Ezberdintasuna synseten lehentasunean dago. Multisetetan lehentasuna synsetaren WordNet sakonera[8] eta maiztasunaren arabera antolatzen da, ondoren azaltzen den moduan.

Bitez s WordNeteko synseta, $d(s)$ s synsetak WordNeten duen sakonera, eta $1_{SP_{mul(p,r)}}(s)$[9] funtzioa. $a, b \in SP_{mul}(p,r)$ synseten arteko orden partziala definitzen dugu ondoren azaltzen den moduan: $ord(a) > ord(b)$ baldin $d(a) > d(b)$, edo $d(a) = d(b) \wedge \mathbf{1}_{SP_{mul}(p,r)}(a) > \mathbf{1}_{SP_{mul}(p,r)}(b)$.

Gauzak horrela, p aditzaren gidaritzapean agertu den argumentu baten gunea eta multisetaren arteko egokitzapena, biek komunean duten sakonera gehieneko synsetak ematen digu. Beraz, aurretik aurkeztutako notazio berbera erabilita, $SP_{wn}(p,r,w_0)$ hautapen murriztapena honela definituko dugu:

$$SP_{wn}(p,r,w_0) = \arg \max_{s \in hypers(w_0) \cap M(p,r)} ord(s) \tag{IV.9}$$

[7]Multisetak elementu errepikatuak baimentzen dituzten multzoak dira.

[8]WordNet taxonomian, kontzeptuak gero eta orokorragoak dira sakonera gutxitu ahala

[9]$1_{SP_{mul(p,r)}}(s)$ funtzioak s synsetak $SP_{mul(p,r)}$ multisetean duen maiztasuna adierazten du.

$$sim_{Jac}(w, w_0) = \frac{|T(w) \cap T(w_0)|}{|T(w) \cup T(w_0)|}$$

$$sim_{cos}(w, w_0) = \frac{\sum_{i=1}^{n} \vec{T_i}(w)\vec{T_i}(w_0)}{\sqrt{\sum_{i=1}^{n} \vec{T_i}(w)^2}\sqrt{\sum_{i=1}^{n} \vec{T_i}(w_0)^2}}$$

$$sim_{Lin}(w, w_0) = \frac{\sum_{(r,v)\in T(w)\cap T(w_0)} I(w, r, v) + I(w_0, r, v)}{\sum_{(r,v)\in T(w)} I(w, r, v) + \sum_{(r,v)\in T(w_0)} I(w_0, r, v)}$$

Irudia IV.4: Azterketa honetan erabilitako antzekotasun distribuzionaleko neurriak. *Jac* eta *cos* dira, hurrenez hurren, Jaccarden eta kosinuaren antzekotasun neurriak adierazteko erabiltzen diren izenak. $T(w)$, w hitzarekin batera agertzen diren hitzen zerrenda da (w hitzaren antzeko hitzak adibidez, thesaurus batek ematen dituenak), $\vec{T_i}(w)$, w hitzarekin batera agertzen diren hitzen bektoreko i. osagaiak duen pisua da, eta Linen formulari dagokionez, $I(w, r, v)$, w eta r, v osagaien arteko *mutual information* (Hindle, 1990) balioa da.

IV.2.2 Hautapen Murriztapen Distribuzionalak

Esperimentuetan erabiliko ditugun Hautapen Murriztapen Distribuzionalak ez dira eskuz sortutako baliabideetatik erauzten, automatikoki eraikitako baliabide konputazionaletik baizik. Zehazki, antzekotasun distribuzionaleko bi thesaurus ezberdin erabili ditugu HMak sortzeko:

1. Linen (Lin, 1998) Antzekotasun distribuzionaleko thesaurusa. Zenbait kazetaritza-corpusen gainean Linen antzekotasun formula aplikatuta sortzen da.

2. British National Corpusetik ad hoc erauzitako thesaurusa. Horretarako Padó eta Lapataren (2007, 179. orr.) softwarea erabili genuen egileek aipatzen duten parametrizazio optimoarekin: hitzean oinarritutako espazioa, testuinguru ertaina eta 2,000 oinarri elementu. Antzekotasun neurriak, Jaccard, Lin eta kosinuaren formularekin kalkulatu genituen.

Thesaurusak hitz bakoitzaren antzeko hitzak zerrendatzen ditu pisu batez lagunduta (gero eta pisu handiagoa, gero eta antzekoagoak izango dira hitzak jatorrizko hitzarekiko). Bi hitzen arteko antzekotasuna neurtzeko, beraz, aski da thesaurusean hitz horietako bat bilatu eta, adibidez, antzeko

hitzen zerrendan beste hitzak duen pisuari erreparatzea. Hiru dira hautapen murriztapen distribuzionaletarako probatu ditugun antzekotasun motak:

1. Linen baliabidearekin gertatzen den bezala, thesaurusa asimetrikoa bada, w_1 eta w_2 hitzak emanda, beren antzekotasuna (edo sim) ezberdina izango da thesaurusa atzitzeko erabili den hitzaren arabera. Hau da, $sim(w_1, w_2) \neq sim(w_2, w_1)$. Gauzak horrela, sim^{th} bi antzekotasunen arteko batez bestekoa bezala definitzen dugu[10].

2. Hitzen arteko antzekotasuna neurtzeko beste aukera bat bigarren mailako antzekotasuna erabiltzea da. Bigarren mailako antzekotasuna kalkulatzeko, hitzen thesauruseko sarrerak erabiltzen dira jatorrizko hitzen ordez. Horretarako IV.4 irudiko $T(w)$ eta $\vec{T_i}(w)$ osagaien definizioa aldatu behar da soilik. Bigarren mailako antzekotasunean, $T(w)$ diogunean, w hitzaren antzeko hitzak (alegia, thesauruseko w sarrera) esan nahi dugu eta $\vec{T_i}(w)$ diogunean $sim(w, i)$. Antzekotasun mota honi sim^{th2} esango diogu hemendik aurrera. Jaccard (sim^{th2}_{Jac} edota kosinuaren sim^{th2}_{cos} formuletan aplikatu dugu bigarren mailako antzekotasuna.

3. BNCko thesaurusa ez da asimetrikoa, eta bertan IV.4 irudiko kosinu eta Jaccarden formulak erabili dira bere horretan. Corpus honetarako beraz, sim_{Jac} eta sim_{cos} definitu ditugu antzekotasun neurri bezala

Gauzak horrela, hautapen murriztapen distribuzionaletarako egokitzapen funtzioak definituko ditugu, bat antzekotasun neurri bakoitzeko. Horrela, orain arteko izendegi berbera erabilita, w_0 izen gunea p aditzeko r rolera zenbateraino egokitzen den jakiteko funtzioa, Erk-ek bezala, honela definitzen dugu:

$$S_{sim}(p, r, w_0) = \sum_{w \in Seen(p,r)} sim(w_0, w) \cdot freq(p, r, w) \qquad \text{(IV.10)}$$

non $sim(w_0, w)$ hitz baten (w_0) eta hitz ezagun baten (w) arteko antzekotasuna den (ikusi aukera posibleak IV.1 taulan), $Seen(p, r)$ p predikatua eta r rola duten entrenamenduko argumentuen guneak diren, eta $freq(p, r, w)$, w hitza p aditzaren gidaritzapean r rolarekin etiketatutako argumentu guneetan zenbatetan agertu den.

[10] w_1 eta w_2 hitzak emanda, $sim^{th} = (sim(w_1, w_2) + sim(w_2, w_1))/2$.

	Antzekotasun neurria	Thesaurusa
sim^{th}	Lin	Lin
sim_{cos}	kosinua	BNC
sim_{Jac}	Jaccard	BNC
sim^{th2}_{cos}	kosinua (2 maila)	Lin
sim^{th2}_{Jac}	Jaccard (2 maila)	Lin

Taula IV.1: Antzekotasun distribuzionaleko neurriak adierazteko sinboloak eta kalkulatzeko erabili diren thesaurusak

WordNeten oinarritutako ereduek ez bezala, antzekotasun distribuzionaleko teknikek ez dute eskuz eratutako errekurtsoen beharrik hitzen arteko antzekotasuna neurtu ahal izateko. Corpusak behar dituzte, besterik ez; eta corpus horiek dira antzekotasun distribuzionaleko ereduaren hitz estaldura -alegia zein hitzen arteko antzekotasunak emateko gai den- eta orokortzeko gaitasuna finkatuko dutenak. Honek domeinu berrietara egokitzeko erraztasuna ematen die eredu hauei, jatorrizko corpusak corpus zehatzagoekin aberastuz, eraldatuz edota besterik gabe, ordeztuz.

IV.2.3 Preposizioentzako Hautapen Murriztapenak

Aditzen hautapen murriztapen ereduez gain, preposizioen hautapen murriztapenekin ere esperimentuak egitea egokia da. Litkowski eta Hargraves-en lanak (2006) erakutsi zuen preposizioek ere, aditzek bezala, HMak gordetzen zituztela bere baitan, bereziki argumentu adjuntuetan agertzen zirenean. Zoritxarrez, aditz argumentuak sailkatzerako garaian ezin daiteke besterik gabe jakin sailkatu beharreko argumentu hori corea edo adjuntua den eta, beraz, alferrik egindako lana izango litzateke adjuntu edo *core* rolentzako hautapen murriztapen eredu bereziak prestatzen ibiltzea, inoiz ez baikenuke zehazki jakingo zein HM multzo aplikatu beharko genukeen argumentu jakin baten gainean. Horren ordez, hautapen murriztapenak argumentuaren funtzio sintaktiko ezberdinen arabera zatitzea ideia ona izan daiteke. Gure datu multzoen gainean beharrezko neurketak egin eta gero, ikusi genuen NP argumentuen %5ari soilik egokitzen zitzaiola adjuntu motako rola; alegia, NP motako argumentuek *core* izateko joera handia zutela, eta hortaz argumentuak beren funtzio sintaktikoaren arabera banatzea, “core vs. adjuntu” ba-

```
write-Arg0:   Angrist anyone baker ball bank Barlow Bates ...
write-Arg1:   abstract act analysis article asset bill book ...
write-Arg2:   bank commander hundred jaguar Kemp member ...
write-AM-TMP: month
write-AM-LOC: paper space
...
```

Irudia IV.5: `write-ROLE` (*aditz-rol*) motako hautapen murriztapen zerrenda. "write-Arg0" lerroak, adibidez, *write* aditzeko Arg0 (NP) argumentutan topatutako guneen zerrenda ematen du.

naketaren hurbilpen onargarria izan zitekeela[11]. Azkenean, funtzio sintaktiko bakoitzerako hautapen murriztapen berezituak prestatu genituen:

- *Aditz-rol* motako hautapen murriztapenak (izen sintagma (NP) motako argumentuak tratatzeko). Ikusi IV.6 irudia. Adibidez: `take-Arg0`, `say-Arg1`, `lose-AM-TMP`...
- *Preposizio-rol* motako hautapen murriztapenak (preposizio sintagma (PP) motako argumentuak tratatzeko). Ikusi IV.5 irudia. Adibidez: `in-AM-TMP`, `in-AM-LOC`, `at-AM-LOC`...

IV.5 eta IV.6 tauletan *aditz-rol* eta *preposizio-rol* motako hautapen murriztapenen adibideak ikus daitezke hurrenez hurren. Derrigorrezkoa da argitzea, IV.2.1 sekzioan aurkeztu ditugun WordNet HM metodoek ez dituztela tauletako hautapen murriztapenak erabiliko. Horietatik abiatuta, metodoaren arabera beti ere, synset zerrenda bat eratu eta antolatuko dute modu batera ala bestera (Resniken oinarritutako hautapen murriztapenek adibidez, aipatu bezala, IV.3 eta IV.2 iruditan agertzen direnen itxura izango dute). IV.2.2 sekzioko HM distribuzionalek aldiz, metodoaren arabera, thesaurus bat edo beste eta antzekotasun neurri ezberdinak aplikatuko dituzte IV.10 ekuazioari jarraituz. IV.5 eta IV.6 tauletako hitz zerrenda, ekuazio horretako $Seen(p, r)$ osagaiak izango dira (adibidez, $Seen(from, A0)$, $Seen(write, A1)$...).

Proposatutako aukeraz gain, egon litezke argumentuetako preposizioei etekina ateratzeko beste hainbat aukera ere. Guk kontuan hartu genituen *prep-aditz-rol* motako hautapen murriztapenak baina aukera hori gure datu

[11] PP argumentuetan balantza orekatuago dago: %45a coreak dira.

```
from-Arg0:   Abramson agency association barrier cut ...
from-Arg1:   accident ad agency appraisal arbitrage  ...
from-Arg2:   academy account acquisition activity ad ...
from-Arg3:   activity advertising agenda airport ...
from-Arg4:   europe Golenbock system Vizcaya west
from-AM-LOC: agency area asia body bureau orlando ...
from-AM-TMP: april august beginning bell day dec. half ...
from-AM-CAU: air design experience ...
...
```

Irudia IV.6: `from-ROLE` (*preposizio-rol*) motako hautapen murriztapen zerrenda. "from-Arg0" lerroak, adibidez, *from* preposizioak gobernatutako Arg0 (PP-NP) argumentutan topatutako guneen zerrenda ematen du.

multzorako oso egokia ez zela iruditu zitzaigun, orokortzeko gaitasuna mugatua zuten hautapen murriztapen oso espezifikoak eratzen zirelako. Datu multzoaren tamaina dela eta, ez dira *prep-aditz-rol* hiruko berdina duten argumentu asko topatzen (adibide gutxi batzuk zorte pixka batekin) eta horrelako baldintzetan oso zaila orokorpenak egitea.

NP eta PP argumentuak aditz-rol moduko HMekin soilik tratatzearen aurrean, preposizio sintagmentzako hautapen murriztapen berezituak erabiltzearen abantailak IV.4 esperimentazio atalean argituko dira. Dena dela, teorikoki behintzat, esan genezake ideia ona dela NP eta PPen artean ezberdindu eta goian proposatzen den hautapen murriztapenen banaketa egitea, asko direlako aditzarekiko independenteak diren rolekin -edo rol adjuntuekin- etiketatuta dauden preposizio sintagmak. Eta beraz, pentsa genezake PP argumentuak tratatzeko HMak eratzeko garaian preposizioak aditzak baino diskriminatzeko ahalmen handiagoa izan dezakeela, besterik gabe, esan bezala, adjuntuak aditzarekiko independenteak direlako eta, hortaz, aditzak ez dituelako, preposizioak bezala, PP argumentuen eskakizun edo preferentzia lexikalak definituko (Gildea eta Jurafsky, 2002).

IV.3 Argumentuen sailkapena HM ereduekin

Esperimentazio fasearekin hasi baino lehen, beharrezkoa da hautapen murriztapenentzako ebaluazio ingurune bat definitzea. Egile ezberdinek (Erk,

2007; Padó *et al.*, 2007) ebaluazio sistema propioak proposatu dituzte, baina guk dakigula, gu gara lehenak hautapen murriztapenak zuzenean argumentuak sailkatzeko ataza erreal batean aplikatzen. Horrela, aditz argumentuen mugak soilik markatuta dituen esaldi bat emanda, atazaren helburua argumentu bakoitzari bere rola esleitzea izango da. Esleipen hori egin ahal izateko hautapen murriztapenak sailkatzaile moduan erabiliko ditugu, alegia, aditza eta argumentu baten gunea soilik emanda, hautapen murriztapenak baliatu beharko ditugu argumentuaren rola asmatzen saiatzeko. Beste modu batera esanda, esaldiko argumentu gune bakoitzari hobekien egokitzen zaion rola itzultzea izango da eginkizuna, eta hautapen murriztapen horiek izango dira hain zuzen argumentu gunearen eta rolaren arteko egokitzapena neurtzeko erabiliko direnak. Beraz, p predikatua eta w_0 argumentu gunea emanda, rolen hautaketarako R erregela honela definituko dugu:

$$R(p, w_0) = \arg \max_{r \in Roles(p)} SP(p, r, w_0) \tag{IV.11}$$

$SP(p, r, w_0)$-k w_0 gunea r rolera zenbateraino egokitzen den adierazten du (ikusi IV.7,IV.9 eta IV.10 ekuazioak). Kontuan hartu SP_{wn} denean $argmin$ erabili behar dugula $argmax$ beharrean.

IV.4 Esperimentazio ingurunea

Atal honetan, hautapen murriztapenak ebaluatuko ditugu argumentuak sailkatzeko ataza erreal batean, IV.3 puntuan adierazi den bezalaxe. Horrez gain, aditz-rol eta preposizio-rol motako hautapen murriztapen erabilera konbinatua aditz-rol hautapen murriztapenak soilik erabiltzearekin alderatuko dugu.

Esperimentu hauetan erabilitako datu multzoa, berriro ere, CoNLL-2005eko antolatzaileek prestatutako berbera da (ikusi III.2). Datu multzo honek PropBankeko hainbat sekzio multzokatzen ditu eta baita Brown Corpuseko zati bat ere. PropBankeko 02-21 sekzioak hautapen murriztapenak sortzeko erabiltzen dira eta 23. sekzioa testerako. Brown Corpusa domeinuz kanpoko testa egiteko erabiltzen da baina bere tamaina txikia dela eta SemLinkeko instantziekin aberastu genuen PropBank eta VerbNeten arteko azterketa konparatiborako egin genuen modu berberean (ikusi III.2). Lan honen helburua argumentuen sailkapena (eta ez identifikazioa) denez, argumentuen identifikatzeko PropBankeko gold informazioa erabili da.

Hautapen murriztapenak argumentu ez nominalekin lan egiteko gai ez

direla jakinik, soilik izen gunea duten NP eta PP argumentuak izan dira kontuan esperimentu hauek egiteko. Entrenamenduko sekzioetatik 140,000 argumentu nominal inguru erabiltzen dira hautapen murriztapenak modu gainbegiratuan erauzteko, eta testerako 8,000 argumentu inguru, bai domeinu barneko testerako eta baita domeinuz kanpokorako ere. Hautapen murriztapen bakoitzaren errendimendua *prezisioa*, *estaldura* eta F_1 neurri estandarren arabera erakusten da.

Aipatzekoa da HM ereduetako batek ere ez duela inolako rol predikziorik egiten argumentuaren gunea hitz ezezaguna denean (alegia, hitza taxonomian edota thesaurusean agertzen ez denean). Gertaera hau WordNeten oinarritutako ereduetan gertatzen da eredu distribuzionaletan baino gehiagotan hitzen estaldura txikiagoa izaten delako eskuz eratutako errekurtsoetan.

Konparaketarako helburuarekin "baseline" eredu bat definitu dugu ezaugarri lexikalek, bere horretan, argumentuak klasifikatzeko duten ahalmena neurtu eta hautapen murriztapenek ekarriko lituzketen hobekuntzak hobeto identifikatzen laguntzeko. (p, w_0) aditz eta argumentu gune bikote bat emanda, baseline ereduak w_0 gunea duten argumentuek p aditzaren gobernupean kasu gehienetan hartu izan duten rola bueltatzen du. Hau da.

$$R(p, w_0) = \arg \max_{r \in Roles(p)} freq(p, r, w_0) \tag{IV.12}$$

non $freq(p, r, w_0)$ entrenamendurako (p, r, w_0) hirukoaren maiztasuna den. "Baseline" Eredu honi "lexikal" esango diogu diogu hemendik aurrera. Maiz agertzen ez diren argumentuen guneekin berdinketak gerta litezke maiztasunarekin rol batekin ala beste batekin. Kasu horretan, maiztasun altueneko rola itzultzen da.

IV.5 Emaitzak eta analisia

Hautapen murriztapenen eredu bakoitzak rolak sailkatzeko atazan lortzen duen emaitza IV.3 taulan ikus daiteke. Taula honek bi zutabe nagusi ditu: (1) Verb-role SPs izeneko bat eta (2) Preposition-Role and Verb-Role SPs izeneko beste bat. Lehenengo zutabeak argumentu guztiak aditz-rol motako HMekin etiketatzean lortzen diren emaitzak erakusten ditu, eta bigarrenak, NP eta PP motako argumentuak hurrenez hurren aditz-rol eta preposizio-rol motako hautapen murriztapenekin tratatzeak ematen dituen emaitzak bistaratzen ditu. Gorago aurreikusi dugun moduan, emaitzei erreparatuz gero,

	Verb-Role SPs					
	WSJ-test			Brown		
	prec.	rec.	F_1	prec.	rec.	F_1
lexical	**70.75**	26.66	39.43	**59.39**	05.51	10.08
SP_{Res}	45.07	37.11	40.71	36.34	27.58	31.33
SP_{wn}	55.44	45.58	50.03	41.76	31.58	35.96
$SP_{sim_{Jac}}$	48.85	46.38	47.58	42.10	34.34	37.82
$SP_{sim_{cos}}$	53.13	50.44	51.75	43.24	35.27	38.85
$SP_{sim^{th2}_{Jac}}$	61.76	**58.63**	**60.16**	51.97	**42.39**	**46.69**
$SP_{sim^{th2}_{cos}}$	61.12	58.12	59.63	51.92	42.35	46.65

Taula IV.2: Aditz-rol motako Hautapen murriztapenentzako emaitzak WSJ (ezkerrean) eta Brown corpusentzat

bistan geratzen da argumentuen funtzio sintaktikoaren araberako hautapen murriztapenak erabiltzeak (IV.3) hobetu egiten dituela emaitzak argumentuak sailkatzeko atazan, bai WSJ-n (WSJ-test zutabea) eta baita domeinuz kanpoko corpusean ere (Brown zutabea); zenbait kasutan, zutabetik zutabera, aldea 10 puntu baino handiagokoa da.

IV.2 eta IV.3 tauletako "Lexical" izeneko lerroak baseline sistemaren emaitzak ematen ditu. Hurrengo bi lerroak WordNeten oinarritzen diren metodoei dagozkie, hau da Resnik (SP_{Res}) eta WordNet (SP_{wn}) metodoei. Jarraian datozenak metodo distribuzionalenak dira: BNC corpusetik eratorritako thesaurusa abiapuntu hartuta erauzitako $SP_{sim_{Jac}}$ eta $SP_{sim_{cos}}$ hautapen murriztapen ereduak eta Linen thesaurusetik erauzitako sakonera bikoitzeko $SP_{sim^{th2}_{Jac}}$ eta $SP_{sim^{th2}_{cos}}$ ereduak.

"Lexical" metodoak prezisio neurri onenak ematen ditu kasu eta datu multzo guztietan. Honek, Pradhanek (2008) beste bide batzuetatik erakutsi zuen bezala, agerian uzten du ezaugarri lexikalen garrantzia argumentuen sailkapen prozesuan. Gogora dezagun metodo honek argumentu guneak eta aditzaren lema soilik erabiltzen dituela rola aurkitzeko garaian (inolako aldaketa edota orokortzerik gabe) eta horiek aski zaizkiola prezisio neurri oso altuak erakusteko. Hala eta guztiz ere, bistan da, prezisio altua ematen duen metodo honek ere *recall* edota estaldura neurri oso baxuak ere ematen dituela aldi berean. Estalduraren erorketak hain zuzen, ezaugarri lexikalen arazo garrantzitsu bat uzten digu agerian: ezagutza orokortzeko gaitasun eza eta

	Preposition-Role and Verb-Role SPs					
	WSJ-test			Brown		
	prec.	rec.	F_1	prec.	rec.	F_1
lexical	**82.98**	43.77	57.31	**68.47**	13.60	22.69
SP_{Res}	63.47	53.24	57.91	55.12	44.15	49.03
SP_{wn}	65.70	63.88	64.78	60.08	48.10	53.43
$SP_{sim_{Jac}}$	61.83	61.40	61.61	55.42	53.45	54.42
$SP_{sim_{cos}}$	64.67	64.22	64.44	56.56	54.54	55.53
$SP_{sim^{th2}_{Jac}}$	70.82	**70.33**	**70.57**	62.37	**60.15**	**61.24**
$SP_{sim^{th2}_{cos}}$	70.28	69.80	70.04	62.36	60.14	61.23

Taula IV.3: Aditz-rol eta preposizio-rol hautapen murriztapenen emaitzak WSJ (ezkerrean) eta Brown corpusentzat

ezaugarrien sakabanaketa larria. Esan dugun moduan, ezaugarri lexikalak bere horretan garrantzitsuak dira argumentuen rola zein izan daitekeen asmatzeko; baina horretarako beharrezkoa da ezaugarri horiek entrenamendu multzoan agertzea eta hori, domeinuz kanpoko estaldura baxuak erakusten duen moduan, ez da beti posible izaten. WSJ corpusean, esaterako, argumentuen guneen %48a lexiko ezezaguna da eta Brown corpusean aldiz, %80a.

"Lexical" metodoaren estaldura gainditzeko beharrezkoa da ezaugarri lexikal ezagun eta ezezagunen arteko konexio semantiko bat egitea, azken hauek, nolabait, sailkatze prozesurako erabilgarriak izan daitezen. Adibide batekin azaltzeagatik, garrantzitsua da, gauzak ongi egin nahi baditugu behintzat, *to drive [a car]* (kotxea gidatu) eta *to drive [a truck]* (kamioia gidatu) bezalako argumentuen arteko antzekotasuna kontuan hartzea, ezaugarri lexikalen sakabanaketak eta ezaugarri lexikal ezezagunek estalduran eragiten dituzten sarraskiak ahalik eta neurri handienean ekiditeko. Eta hori da, hain zuzen ere, hautapen murriztapenak egitera datozena. Hautapen murriztapenek ezaugarri lexikal ezezagunak lexiko ezagunarekin[12] "lotzen" dituzte antzekotasun neurrien laguntzaz, eta gai dira entrenamenduko datuetan ikusi ez den lexikotik abiatuta, ezagutza semantiko erabilgarria erauzi eta argumentuak hobeto sailkatzeko. Emaitzek hori diote behintzat. Prezisioaren

[12]Lexiko ezaguna, entrenamendu multzoko argumentuen gune nominalek osatzen dute. Lexiko hau rol konkretuetara egoten da lotuta, nahiz eta ez den modu unibokoan izaten. Horrela, lexiko ezezaguna ezagunarekin lotzen dugunean, roletara ere lotzen dugu.

galera txiki baten truk, 30 puntutako irabaziak ere ikusten dira hautapen murriztapenak erabiltzen dituzten ereduen estalduran eta baita F_1ean ere.

Adibideekin jarraituz, kontuan hartu hurrengo argumentu guneak: *doctor*, *men*, *tie*, *shoe*.[13] Gune horietako bat bera ere ez da agertu entrenamenduko multzoan *to wear* aditzaren argumentu baten gune bezala, eta beraz ezaugarri lexikoek ez lukete adibide praktikorik izango argumentu gune ezezagun horiei rol zuzena esleitzeko garaian. Hautapen murriztapenekin aldiz, gai gara *doctor* eta *men* guneak dituzten argumentuei Arg0 rola esleitu eta *tie* eta *shoe* guneak Arg1 motako argumentuenak direla erabakitzeko.

Hautapen murriztapenak

Hautapen murriztapenen eredu bakoitzari begiratuta ikus daiteke WordNeten oinarritutako S_{Res} eta S_{wn} direla emaitza kaxkarrenak ematen dituztenak, batez ere domeinuz kanpoko corpusetan. Metodo hauen erorketa nagusiki estalduran ematen da, ziur aski, WordNetek berak duen hitzen estaldura baxuak eraginda. WordNeteko metodoek WSJ corpuseko argumentu guneen %85erako iragarpenak egiten dituzten bitartean, metodo distribuzionalek %99rako egiten dute. Brown corpusean, WordNeteko metodoak %81era jaisten dira eta distribuzionalak %95era. (errore ortografikoak dituzten lemetarako ez da iragarpenik egiten, ezta aditz ezezagunetarako. Metodo batek argumentu baten iragarpena egin dezan honako baldintzak bete behar dira:

- Argumentuaren aditza ezaguna izan behar da. Hau da, testeko argumentu baten aditza ez bada entrenamenduko adibideen artean agertu, ezin da aditz ezezagun horren argumentuentzako iragarpenik egin, besterik gabe bere hautapen murriztapenak ezin izan ditugulako entrenamendu fasean ikasi. WSJ corpusean aditz ezezagunak ia %1a dira eta Brown corpusean berriz %4. Adibideak: *excise* (ezabatu), *fry* (frijitu), *croak* (korroka egin), *mash* (purea egin) etab.

- Aditz argumentuaren guneak ezaguna izan behar du WordNeten (WordNeteko metodoekin ari bagara) ala thesaurusean (Lin edo BNC thesaurusean, metodo distribuzionalekin ari bagara). Thesaurusak erraldoiak dira eta ez da batera ohikoa testeko argumentu guneak ez aurkitzea (zenbakiak, laburdurak, izen propioak eta antzeko berezitasunak barrabarra aurki daitezke). WordNeteko metodoek ordea, zailtasun handia-

[13] doktore, gizonak, gorbata, zapata.

goak izaten dituzte hitz guztiak taxonomian topatu eta iragarpenak egiteko. WordNeten ageri ez diren hitzen artean, besteak beste, laburdurak (*Inc.*, *Corp*, ...) eta marka izenak (*Texaco*, *Sony*, ...) topa ditzakegu.

Antzekotasun motak

Bigarren mailako antzekotasun distribuzionala erabiltzen duten ereduak dira, oro har, emaitza onenak ematen dituzten HM metodoak, bai prezisio eta baita estalduraren aldetik ere. Guk dakigula, gu izan gara lehenak antzekotasun metodo hau hautapen murriztapenen modelatzera aplikatzen eta badirudi lehen mailako ereduen alternatiba sendoa dela. Emaitzetan ikusten denez, aplikatutako ereduen errendimendu erlatiboa mantendu egiten da bi domeinutako datu multzoetan eta horrek erabilitako metodoen sendotasunaren aldeko ideia bat eman diezaguke.

WordNeteko hautapen murriztapenak

Aipatzekoa da gure WordNeteko metodo sinpleak ere Resniken eredua gainditzea lortzen duela bai WSJ corpusean eta baita domeinuz kanpo ere. Resniken ereduak maiztasun handieneko rolak aurreikusteko joera izaten du, baina gure metodoak rol multzo zabalago baterako iragarpenak egiten ditu. Resniken ereduko pisu sistemaren eraginez, maiztasun urriko rolen pisuak ez dira maiztasun oso altuko rolen pisuekin lehiatzeko gai, eta horregatik maiztasun handieneko rolen aldeko aurreikuspenak egitera jotzen du gehienetan, maiztasun urriko rolak (adjuntuak batez ere) neurri handi batean albo batera utziz. Gure metodoa, zehazki, nolabaiteko diskriminazio hori gainditzeko diseinatu da, synset espezifikoei oso maiz agertzen direnei baino pisu handiagoa emanez (ikusi IV.2.1 sekzioa), eta emaitzek, partzialki den arren, bide zuzena hautatu dugula erakutsi digute. Resniken eta WordNeteko prototipoaren errendimenduaren azterketa konparatiboa egiteko IV.4 eta IV.5 taulak kontsulta daitezke. Lehenengoan bi HM ereduek WSJ test corpusean lortzen dituzten emaitzak ikus daitezke eta bigarrenean, Brown corpusekoak. Tauletan, oro-har, WordNeteko prototipoak rol gehienetarako eta batez ere adjuntuetarako aurreikuspen hobeak egiten dituela ikusten da.

Thesaurusak: BNC eta Lin

BNC corpusaren gaineko antzekotasunerako softwarea (Padó eta Lapata, 2007) eta Linen thesaurusaren arteko lehiari dagokionez, badirudi biekin

	Resnik HM eredua			WordNet HM eredua		
	prec.	rec.	F_1	prec.	rec.	F_1
Arg0	66.29	45.59	54.03	73.51	51.88	60.83
Arg1	75.94	66.9	71.14	62.65	84.91	72.1
Arg2	46.21	32.42	38.11	54.06	38.16	44.74
Arg3	16.36	32.92	21.86	38.46	30.48	34.01
Arg4	6.75	11.9	8.62	39.28	26.19	31.42
AM-ADV	20.24	30.84	24.44	50.00	31.77	38.85
AM-CAU	20.31	41.93	27.36	52.94	58.06	55.38
AM-DIR	15.49	45.83	23.15	16.66	8.33	11.11
AM-DIS	59.72	82.69	69.35	85.41	78.84	82.00
AM-LOC	51.26	31.56	39.07	74.66	70.00	72.25
AM-MNR	23.45	28.57	25.76	42.7	30.82	35.8
AM-PNC	28.00	18.91	22.58	42.85	24.32	31.03
AM-TMP	74.21	68.07	71.01	89.21	57.26	69.75
Adj.	48.61	49.73	49.16	74.21	54.16	62.62
Core	67.72	54.34	60.30	64.98	66.28	65.62
Guztira	63.47	53.24	57.91	65.70	63.88	64.78

Taula IV.4: Resnik (ezkerrean) eta WordNeteko hautapen murriztapenen errendimendua WSJ corpusean. Emaitzak (goitik behera) rolez rol ikus daitezke. "Adjuntuak guztira" lerroan argumentu adjuntuak soilik kontuan hartuz orokorrean lortzen diren emaitzak ikus daitezke. "Coreak guztira" lerroan, berdin, baina *core* argumentuentzat.

antzeko emaitzak eskuratzen direla, $SP_{sim_{Lin}}$ eta $SP_{sim^{th}_{Lin}}$ lortzen dituzten emaitzei erreparatzen badiegu. Lehenak WSJ corpusean duen nagusitasuna, thesaurusa eratzeko erabili zen jatorrizko corpusaren izaeragatik izan daiteke. Gogora dezagun Linen thesaurusa kazetaritza testutatik eratortzen dela eta WSJ, hain zuzen, kazetaritza corpusa dela. Era berean, ikus daiteke Linen thesaurusean oinarritutako metodoak Brown corpusean erorketa gogorra jasaten duela, eta BNCko metodoaren azpitik geratzen dela sailkatuta. Pentsa dezakegu beraz, metodo distribuzionalak erraz egoki daitezkeela corpus batera edo bestera jatorrizko corpusaren izaera soilik aldatuta. Horrela, etiketatu beharreko corpusaren arabera, thesaurus bat edo beste aukera dezakegu, antzekotasunen izaera komeni ahala aldatzen joateko.

	Resnik HM eredua			WordNet HM eredua		
	prec.	rec.	F_1	prec.	rec.	F_1
Arg0	50.78	38.55	43.83	52.30	39.39	44.93
Arg1	73.08	55.49	63.08	72.44	60.71	66.06
Arg2	27.91	21.86	24.52	28.66	30.82	29.70
Arg3	3.57	11.90	5.49	3.17	4.76	3.80
Arg4	17.33	23.63	20.00	14.00	12.72	13.33
ArgM-ADV	15.43	11.90	13.44	35.93	10.95	16.78
ArgM-CAU	14.00	29.16	18.91	30.76	33.33	32.00
ArgM-DIR	32.06	28.76	30.32	43.75	14.38	21.64
ArgM-DIS	23.33	37.83	28.86	43.75	37.83	40.57
ArgM-LOC	53.26	34.90	42.17	64.75	46.03	53.81
ArgM-MNR	33.03	26.42	29.36	54.45	39.28	45.64
ArgM-PNC	26.47	23.68	25.00	32.00	15.78	20.68
ArgM-TMP	61.78	67.23	64.39	74.32	70.08	72.14
Adjuntuak guztira	41.33	36.37	38.69	61.00	41.03	49.06
Coreak guztira	59.20	46.19	51.89	60.16	50.00	54.61
Guztira	55.12	44.15	49.03	60.08	48.10	53.43

Taula IV.5: Resnik (ezkerrean) eta WN hm sistemen errendimendua Brown corpusean

Antzekotasun neurriak

Antzekotasun neurriei dagokienez, badirudi kosinuak emaitza hobeak ematen dituela lehen mailako antzekotasuna erabiltzen dugunean. Bigarren mailako antzekotasunean ordea, Jaccard da, gutxigatik, nagusi. Antzekotasun softwarean oinarrituta thesaurus oso bat sortzeko karga konputazional astuna dela eta, ezin izan ditugu bigarren mailako antzekotasun neurriak probatu BNCko datuen gainean.

Atal honetan erakutsi diren emaitzak ezin dira SRL sistema oso batekin alderatu. Gogora dezagun gure helburua ez dela SRL sistema horien prezisio eta estaldura neurri lehiakorren pare jartzea izan, baizik eta ezaugarri lexikoak eta hautapen murriztapenak testuinguru erreal batean konparatzea.

IV.6 Ondorioak

Enpirikoki erakutsi dugu, WordNeten eta antzekotasun distribuzionalean oinarritutako hautapen murriztapenak gai direla SRL sistema tipikoetan hain garrantzitsuak diren ezaugarri lexikalak orokortu eta, beren izaera propioa dela eta, ezaugarri horiek pairatzen duten halabeharrezko sakabanaketaren arazoa leuntzeko. Horretarako, CoNLL-2005eko esperimentazio ingurune estandarrean planteatutako sailkapen prozesu errealista bat definitu eta hainbat hautapen murriztapen metodo ezberdin jarri ditugu lehian. Metodo guztiek erakutsi dute beren gaitasuna ezaugarri lexikal tipikoen baitan gordetako ezagutza semantikoa orokortu eta neurri batean ala bestean, argumentuen sailkapena hobetzeko. Esperimentu hauek kontuan hartzeko igoerak erakutsi dituzte azaldutako metodo guztien *recall* eta F_1 neurrietan.

Hautapen murriztapenak eratzeko metodoen artean WordNeten eta antzekotasun distribuzionalean oinarritutakoak aztertu ditugu. WordNet metodoek prezisio neurri onak erakutsi dituzte, baina distribuzionalekin alderatuta estaldurarekin lotutako hainbat gabezia ere erakutsi dute. Hautapen murriztapenak sortzeko bigarren mailako antzekotasun metodo berri bat proposatu da. Metodo honek orokorrean emaitzarik onenak eman ditu neurri eta corpus guztietan.

Esperimentu hauen emaitzei erreparatuta pentsa daiteke hautapen murriztapenetan oinarritutako lexikoaren orokortze teknikak oso erabilgarriak izan daitezkeela SRL sistema oso baten ezaugarri lexikoak sendotu eta bere errendimendu orokorra hobetzeko, batez ere entrenamenduko adibideak urriak direnean edota sistema domeinuz kanpoko corpus baten gainean aplikatu nahi denean. Kontuan hartuz SRL sistemek domeinuz kanpoko corpusetan izaten dituzten errendimendu galerak guztiz justifikatuta dago noranzko horretan aurrerapenak ekarri ditzaketen bideak ikertzea, eta hauxe izango da, hain zuzen, hurrengo ataletan egingo duguna.

V. KAPITULUA

Hautapen Murriztapenak Argumentuen Sailkapenerako

Aurreko kapituluan aipatu dugu SRL sistemek aditz argumentuak sailkatzeko (argumentuei rola esleitzeko) darabiltzaten ezaugarri garrantzitsuenak esaldiko lexikotik eratortzen zirela, hau da, aditzaren forma/lema eta argumentuaren gunetik. Ikusi dugu, bestalde, ezaugarri horiek beren mugak ere bazituztela eta, bereziki, sakabanaketa larria erakusten zutela domeinuz kanpoko corpusetan. Sakabanaketa hau gainditu eta ezaugarri lexikoen kalitatea hobetzeko asmoz, hautapen murriztapenetan oinarritutako hainbat metodo aurkeztu ditugu aurreko atalean (IV.2). Argumentuen sailkapenerako ataza baten gainean egindako "in vitro" esperimentuek erakutsi digute hautapen murriztapenak gai direla ezaugarri lexikoen sailkatze ahalmena modu esanguratsuan hobetu eta, ustez behintzat, rolen etiketatze automatikoko sistema osoak "in vivo" hobetzeko.

Kapitulu honetan, hautapen murriztapenak erabiliko ditugu artearen egoeran dagoen *SwiRL* (Surdeanu eta Turmo, 2005; Surdeanu *et al.*, 2007) rol etiketatzailearen lexikoa orokortu eta bere errendimendua hobetzeko. Alde batetik, aurreko kapituluan azaldu ditugun hautapen murriztapenak bananbanan *SwiRL* izeneko rol sailkatzaile batean integratzeko bidea aztertuko dugu; bestetik, kontuan hartuta hautapen murriztapen bakoitzak bere be-

rezitasun propioen eraginez errendimendu hobea ala okerragoa erakutsiko duela rol batean edo bestean, hautapen murriztapenak erabiltzen dituzten hainbat *SwiRL* sailkatzaileren arteko konbinazioa egingo dugu, sailkatzaile horietako bakoitzaren abileziak "metasailkatzaile" bakarrean biltzeko asmoz.

V.1 Hautapen Murriztapenak SRL sistema batean integratzen

Esperimentu hauetarako *SwiRL*[1] etiketatzailea hautapen murriztapenek eskaintzen diguten ezagutzarekin hornitu dugu.

V.1.1 *SwiRL*

SwiRL artearen egoeran dagoen rolen sailkatze automatikorako sistema oso bat da. Bere berezitasunetako bat, esaldi bateko aditz argumentu bakoitza osagai sintaktiko bakarrarekin lotzen duela da, nahiz eta hau ez den beti posible izaten, adibidez, esaldirako ematen den analisi sintaktikoa okerra delako edo besterik gabe, sintaxia zuzena izanda ere, aditz argumentua osagai sintaktiko batean baino gehiagotan zatituta dagoelako (Surdeanu eta Turmo, 2005). Esaterako, Charniak-en parserrak ematen dituen analisietan aditz argumentuen %90 soilik lotzen da osagai sintaktiko bakarrarekin. Oso interesgarria da egileek arazo horiek gainditzeko planteatzen dituzten irtenbideak baina ez dagokigu guri hemen horiek guztiak azaltzea. Izan ere, arazo "sintaktiko" horiek *argumentuen identifikazio* faseari lotutakoak dira, eta guk hautapen murriztapenen laguntzaz *argumentuen sailkapena* da hobetu nahi duguna. Horregatik, ataza horretan lortuko ditugun hobekuntzak ahalik eta hobekien isolatu eta aztertzeko *SwiRL*ek argumentuak identifikatzeko darabilen modulua desaktibatu eta zuzenean eskuz identifikatutako gold argumentuak erabiliko ditugu sailkapena egiteko.

SwiRL sailkatzailearen aldeko apustua egiteko zehazki honako ezaugarri hauek hartu genituen kontuan:

- ConLL-2005 txapelketako sistema ez konbinatu onena izan zen. Horrek sailkatzailearen maila altuaren ideia bat eman diezaguke.

[1] `http://www.surdeanu.name/mihai/swirl/`

- Rolen etiketatze prozesu osoa bi fase ezberdinetan planteatzen du: (1) *argumentuen identifikazioa* eta (2) *argumentuen sailkatzea*. Zatiketa honek analisia errazten du, sailkatze prozesurako planteatuko ditugun hobekuntzak identifikazio fasean sor daitezkeen interferentziarik gabe aztertu ahal izango ditugulako.

- *SwiRL* kode irekiko sailkatzaile bat da eta, beraz, aldaketak egin eta ezagutza iturri berriak gehitzeko (ezaugarriak, analizatzaileak...) oso aproposa da.

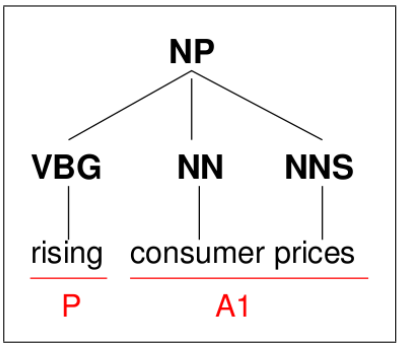
- Argumentu osoekin lan egiten duen sailkatzailea da.

Hautapen murriztapenak rol sailkatzaile "tradizional" batean integratzeko, oro har, III kapituluan aurkeztutako BIO sailkatzaileen ordez, argumentu osoekin lan egiten duen sailkatzaile bat erabiltzeri egokia deritzogu.

SwiRL vs. BIO

Argumentu zatituek (III.2.2 sekzioan aurkeztu ditugun BIO argumentuak adibidez), esaldiko sintaxiaren arabera, argumentu bat token segida baten bidez irudika dezakete. Horrela, token horietako bakoitzak bere izen-gune propioa izango luke eta, hortaz, IV.3 atalean ikusi dugun bezalaxe, hautapen murriztapen ereduek rol aurreikuspenak egingo lituzkete token horietako bakoitzarentzat. Hau arazo bat izan daiteke, izan ere, argumentuko token bakoitzerako aurreikuspenak eginez, hainbat aurreikuspen izango genituzke argumentu bakarrerako. Aurreikuspen horiek, noski, kontraesankorrak izan litezke eta horrek hautapen murriztapenen fidagarritasunaren kontra egingo luke. Gauzak horrela, hautapen murriztapenak sailkatzaile batean integratzeko, *SwiRL* bezala argumentu osoekin lan egiten duten sailkatzaileak hobesten ditugu, sailkatzaile hauek argumentuen benetako izen-guneekin lan egiten saiatzen direlako eta, hortaz, gune horietatik abiatuta, hautapen murriztapenek aurreikuspen fidagarriagoak (eta bakarrak) egiteko aukera izango luketelako. Ikus dezagun arrazonamenduaren adibide bat:

V.1 irudian *rising* predikatua eta *consumer prices* bere Arg1 argumentua ikus ditzakegu. Argumentu osoekin lan eginez gero, *consumer prices* osorik islatuko genuke ikasketa/etiketatze adibide bakarrean. Argumentu adibide indibidual horren aditz ezaugarria *rise* (igo) izango litzateke eta bere gunea, *prices* (prezioak). Hautapen murriztapenei kontsulta eginez gero,



Irudia V.1: *rising consumer prices* (kontsumo prezioen igoera) esaldiaren zuhaitz sintaktikoa eta bere azaleko interpretazio semantikoa

($R(rise, prices) = Arg1$) Arg1 rola aurreikusiko lukete argumenturako. Honaino arazorik ez. Baina zer gertatuko litzateke argumentu osoekin lan egin ordez BIO argumentuekin lan egingo bagenu? Litekeena da (analisi zuhaitzaren arabera) irudiko aditz argumentua bi tokenetan banatuta azaltzea: B-Arg1 + I-Arg1. Bi tokenek, noski, *rise* aditza izango lukete buruzagi, baina, zalantzarik gabe, ezberdintasunak erakutsiko lituzkete argumentuaren guneari dagokionez. B tokenak *consumer* (kontsumitzaile) izango luke gune bezala eta I tokenak berriz, *prices* (prezio) hitz forma. Arazoa token hauen gainean hautapen murriztapenek egingo lituzketen rol aurreikuspenetan dago. Ikusi dugu dagoeneko *rise-prices* bikoterako aurreikuspena Arg1 dela, baina berdin gertatzen al da B tokenari dagokion *rise-consumer* bikotearekin? Tamalez, *consumer* gunea gertuago dago Arg0 (*Agent*) argumentuen guneetatik Arg1 argumentuen (*Patien/Theme*) guneetatik baino. Horrela, hautapen murriztapenen aurreikuspena B tokenerako Arg0 argumentu "kontraesankorra" izango litzateke (I tokenerako aurreikuspena Arg1 zen). Pentsa dezakegu beraz, hautapen murriztapenek hobeto egin dezaketela lan argumentu osoekin zatituekin baino, azken hauekin argumentu berbererako aurreikuspen ezberdinak egiteko aukera dagoelako eta litekeena delako baita ere, aurreikuspen ezberdin eta anbiguo horien eraginez aurreikuspenen fidagarritasun orokorra gutxitu eta hautapen murriztapenen errendimendu orokorra kaltetua gertatzea.

V.1.2 Ezaugarriak

Ikasketarako ezaugarriei dagokienez, *SwiRL*-ek ezaugarri tradizional guztiak modu zabalean inplementatzen ditu (Surdeanu *et al.*, 2007). Oro har, *SwiRL*-ek erabiltzen dituen ezaugarriak bost multzotan bana daitezke: (a) argumentu kandidatuaren barne egituraren berri ematen duten ezaugarriak, (b) argumentuaren testuingurutik erauzitako ezaugarriak, (c) helburu aditzaren propietateak deskribatzen dituzten ezaugarriak, (d) predikatuaren testuingurutik erauzitako ezaugarriak, eta (e) predikatua eta argumentuaren arteko distantzia neurtzen duten ezaugarriak.

V.1.3 SwiRLen ikasketa eredua

SwiRL sailkatzailea Adaboost ikasketa algoritmoa (Freund eta Schapire, 1995) oinarri hartuta garatu da. Adaboostek, pisu bidezko bozketaren eskema erabiliz, hainbat sailkatzaile edo erregela sinple konbinatzen ditu sailkatzaile sendo eta indibidual bat sortzeko. Gure kasuan, erregela sinple horiek 3 sakonerako erabaki zuhaitzak izango dira. Sailkatzaile horiek sekuentzialki ikasten dira pisudun adibideetatik abiatua. Pisu horiek dinamikoki egokitzen dira ikasketa iterazio bakoitzean, aurretik ikasitako erregelen arabera.

One-vs-all moduko sailkatzaile bat eraikitzea izan da helburua. Lehenago ere esan dugun moduan, sailkatzaileak ez du argumentuak identifikatzen ikasiko sailkapen prozesura bideratuko baitugu gure ikerketa. Horrela, gold argumentuak bakarrik sailkatuko ditugu datu multzoko 24 rol usuenak erabiliz.

V.2 HMen integrazioa *SwiRL*en

Hautapen murriztapenak rolen sailkapen prozesuan integratzeko bi modu aztertuko ditugu guztira. Lehena, bistakoena, *SwiRL*ek entrenatzeko erabiltzen duen ezaugarri multzoa Hautapen Murriztapenek egiten dituzten aurreikuspenekin zabaltzean datza. Bigarrena, emaitza onenak emango dizkiguna, apur bat sofistikatuagoa da eta hainbat *SwiRL* ereduren eta Hautapen murriztapenen arteko meta-sailkatzaile bat egitean datza.

V.2.1 SwiRL-i ezaugarriak gehitzen

Aurreko atalean ikusi ditugun hautapen murriztapenak *SwiRL*en integratzen ditugunean "*SwiRL*-HM" sistemak sortzen ditugu. *SwiRL*-ek entrenatzeko edo testerako erabiltzen duen argumentu (adibide) bakoitzaren gainean Hautapen Murriztapenak aplika ditzakegu, aditzetik eta argumentuaren gunetik abiatuta argumentuaren rola aurreikusteko (ikusi IV.3 atala). *SwiRL* eta HMak integratzeko bide bat, beraz, HMen aurreikuspen horiek *SwiRL*ek entrenatzeko eta sailkatzeko erabiltzen dituen adibideetan ezaugarri multzoan txertatzea da. Modu horretan, *SwiRL*ek, bere lana egiteko garaian, aditza eta argumentuaren gunetik erauzten diren ezaugarri lexikalez gain, sakabanaketarako joera txikiagoa izango duten HM aurreikuspenak ere izango ditu eskuragai. Izatez, HMen aurreikuspenak ezaugarri semantikoak dira, dakigun moduan, argumentu gunearen eta aditzaren arteko erlazio semantikoari buruzko informazioa ematen digutelako. Hortaz, argumentuen sailkapen prozesuan, bai aditzak eta baita guneak ere duten pisua eta garrantzia kontuan izanda, pentsa liteke HMen laguntzaz aditza eta gunetik eratortzen ditugun ezaugarri berri hauek ere, neurri txikiagoan behar bada, beren eragina izango dutela argumentuen sailkapenean. Era berean, aurreko ataleko HM ereduen errendimenduak gora egin ahala, suposa dezakegu hobeak izango direla *SwiRL*-HM sistema integratuaren emaitzak, HM ezaugarriek pistak ematen dituztelako sailkatu beharreko adibide bakoitzaren rolari buruz, eta, jakina, pista horiek gero eta hobeak izateak sailkatzailearen lana erraztea ekar dezakeelako.

Inplementatuta ditugun HM eredu bakoitzeko *SwiRL*-HM sistema bat eratu dugu. Hauek dira eredu bakoitza sortzeko burututako pausoak:

1. *SwiRL*ek entrenatzeko erabiltzen dituen adibideen gainean HMak aplikatu adibide horietako bakoitzaren rolaren aurreikuspen bat izateko. Horretarako adibide bakoitzaren ezaugarrien artean aditza eta argumentuaren gunea erabiltzen dira. Honela, a_n edozein argumenturen aurreikuspena, IV.3 atalean aurkeztu dugun formulak ematen digu

$$R(p, w_0) = \arg \max_{r \in Roles(p)} S(p, r, w_0) \tag{V.1}$$

p eta w a_n argumentuarekin lotutako aditza eta gunea dira eta S unean uneko HMa. Entrenamenduko argumentuen HM iragarpenak lortzeko

beharrezkoa da *cross-validation* teknika aplikatzea. Gogoratu, hautapen murriztapen ereduak entrenamenduko datuetatik sortzen direla, eta hortaz ez dela zilegi adibide horiek hautapen murriztapen berdinekin tratatzea, emaitzak, testeko adibideetarako lortuko genituzkeenekin alderatuta, oso distortsionatuta aterako liratekeelako. Horregatik, entrenamenduko korpusa n=5 zatitan zatitzen dugu. Lehen n-1 zatiekin hautapen murriztapenak eratu eta n. zatiko argumentuen rolak iragartzeko erabiltzen ditugu. Prozesu hau n aldiz errepikatzen dugu harik eta entrenamenduko argumentu guztiak etiketatuta ditugun arte.

2. Behin entrenamenduko argumentu guztien HM iragarpenak egin direla, argumentu bakoitzak jaso duen iragarpena bere ezaugarrien artean txertatuko da.

3. Entrenamenduko datuak unean uneko ikasketa automatikoko metodoarekin (ikusi V.1.3 atala) ikasten dira. Emaitza bezala lortzen dugun rol sailkatzaile eredua *SwiRL*-HM eredua izango da.

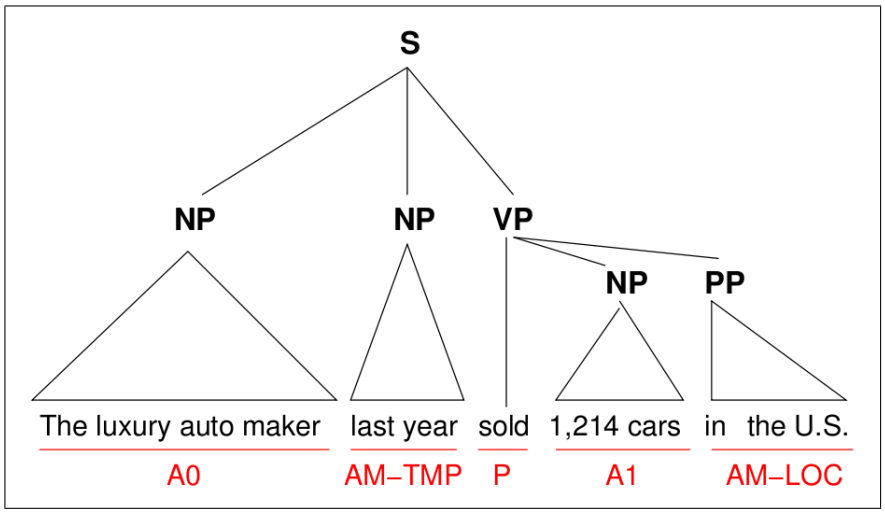
Guztira 6 *SwiRL* -HM sailkatzaile sortu ditugu dagokion atalean aurkeztuko diren esperimentuak egiteko, sailkatzaile bat HM eredu bakoitzeko (SP_{wn}, SP_{Res}, $SP_{sim^{th}}$, $SP_{sim_{Jac}}$, $SP_{sim_{cos}}$, $SP_{sim^{th2}_{Jac}}$, $SP_{sim^{th2}_{cos}}$).

Entrenamenduko adibideekin ikasitako *SwiRL*-HM ereduak testeko argumentuen gainean aplikatzeko, besterik gabe, goian azaldutako prozedura berbera erabili behar da. Alegia, testeko adibide bakoitzerako hautapen murriztapenen iragarpenak kalkulatu behar dira. Iragarpen horiek ezaugarrien artean txertatu ondoren, posible da entrenamenduko datuekin ikasitako *SwiRL*-HM eredu bakoitza testeko adibideak sailkatzeko erabiltzea.

Adibidea: Ikus ditzagun, urratsez urrats, V.2 irudiko esalditik abiatuta, HMak *SwiRL* sailkatzailean gehitu eta *SwiRL*-HM eredu bat sortzeko jarraitu beharreko prozesuaren xehetasunak:

1. Esaldiko egitura sintaktikoaren arabera, lehenik eta behin, *SwiRL*-ek aditza eta haren lau argumentu identifikatuko lituzke (irudian gorriz).

2. Argumentu horientzako adibide bana sortu eta V.1.2 atalean aipatutako ezaugarri guztiak ateratzea izango litzateke hurrengo pausoa (ikusi V.3 irudia).



Irudia V.2: "*The luxury auto maker...*" esaldiaren zuhaitz sintaktikoa eta azaleko analisi semantikoa (gorriz).

3. Lehendik eratutako hautapen murriztapenak erabiliz, argumentu horietako bakoitzerako aurreikuspenak egingo ditugu ondoren. Horretarako, lehenago aipatutako $R(p, w_0)$ funtzioa erabili besterik ez dugu egin behar. p osagaiak, NP argumentuen kasuan, aditzaren lema hartuko du (*sell*) eta PP argumentuen kasuan berriz, preposizioaren balioa (*in*). w_0 osagaiari dagokionez, argumentu bakoitzerako bere gunearen lema hartuko du. PP argumentuen kasuan, barnean gordeta dagoen NP osagaiaren gunearen lema hartuko da. Horrela, argumentu bakoitzerako, hautapen murriztapen ereduak hurrengo iragarpenak egingo lituzke:

```
R(sell,maker)=Arg0
R(sell,year)=AM-TMP
R(sell,car)=Arg1
R(in,U.S.)=AM-LOC
```

4. Hautapen murriztapenen bidez argumentu guztien iragarpenak egin eta gero (entrenamendu[2] eta testeko argumentu guztientzako), iragarpen bakoitza dagokion argumentuaren ezaugarrien artean txertatuko dugu "HM:" etiketaz lagunduta (V.4 irudian, urdinez).

[2]Gogoratu entrenamenduko adibideen aurreikuspenak egiteko *cross-validation* teknika erabili behar dela

Arg0 type:chunk elem:pNP head:POS:N head:maker vtarget:sell vform:sold ...
AM-TMP type:chunk elem:pNP head:POS:N head:year vtarget:sell vform:sold ...
< Predicate >
Arg1 type:chunk elem:pNP head:POS:N head:cars vtarget:sell vform:sold ...
AM-LOC type:chunk elem:pPP head:POS:IN head:in head:POS:N head:U.S. vtarget:sell

Irudia V.3: Jatorrizko *SwiRL* sistemaren entrenamenduko hainbat adibide, V.2 irudiko esalditik eratorriko liratekeen bezala. Argumentuak goitik behera agertzen dira ordenatuta. Gorriz labelak (entrenamenduko datuak direla gogoratu) eta beltzez zenbait ezaugarri tipiko: mota (*type*), elementua (*elem*), gunearen PoS marka (*head:POS:*), gunea (*head*), aditza (*vtarget*) ...

Arg0 type:chunk elem:pNP head:POS:N head:maker vtarget:sell vform:sold ... **HM:Arg0**
AM-TMP type:chunk elem:pNP head:POS:N head:year vtarget:sell vform:sold ... **HM:AM-TMP**
< Predicate >
Arg1 type:chunk elem:pNP head:POS:N head:cars vtarget:sell vform:sold ... **HM:Arg1**
AM-LOC type:chunk elem:pPP head:POS:IN head:in head:POS:N head:U.S. vtarget:sell ... **HM:AM-LOC**

Irudia V.4: *SwiRL*-HM sistema baten entrenamenduko adibideak. Hauetan hautapen murriztapenen ezagutza dago integratuta ezaugarri moduan (urdinez). Irudiko argumentu adibideak V.2 eta V.3 irudietatik eratorri dira atal honetan deskribatutako urratsak jarraituz.

Aurkeztu dugun adibide honetan, aipatzekoa da aukeratu dugun hautapen murriztapen ereduak asmatu egiten duela bere iragarpen guztietan. Hau, noski, ez da ohikoa. Erabiltzen den HM ereduaren arabera, aurreko kapituluan ikusi dugun bezala, emaitza hobeak ala kaxkarragoak lortuko ditugu, baita kontraesankorrak ere. Hala ere, adibidea oso egokia (erraza) da hautapen murriztapenetarako eta %100eko asmatze tasa lortzen da.

SwiRL-HMak sortzeko beste aukera bat

Puntu honetaraino ikusi dugu *SwiRL*-HM ereduak sortzeko aukeratu dugun bidea, *SwiRL* sailkatzaileari hautapen murriztapenen aurreikuspenak txertatzea dela. Txertatze hau, esan bezala, banan-banan egin dugu, HM bakoitzeko *SwiRL*-HM bakarra eratuz. Alternatiba gisa azter daiteke HMen aurreikuspen guztiak *SwiRL* bakarrean txertatzea. Horrela, hainbat *SwiRL*-HM izan ordez, bakarrarekin soilik aritzeko. Zoritxarrez, gure ustez behintzat, *SwiRL*-HM bakarra izatea edo ez, ez da kontuan hartu beharreko gauza

bat. Bestalde, alternatiba honetan proposatzen den lez, hautapen murriztapenek argumentu baten gainean egiten dituzten aurreikuspen guztiak argumentuaren ezaugarrien artean txertatzeak gutxienez arazo bat ekar lezake: aurreikuspena egin duen HMaren arabera, txertatutako aurreikuspen asko kontraesankorrak izan daitezke eta informazioaren interferentzia horiek sailkatzailearen lana zail dezakete esperimentazio fasean ikusiko dugun moduan (ikusi V.5 irudia).

Demagun, adibidez, hiru hautapen murriztapenen iragarpenak (HM_1, HM_2 eta HM_3) *SwiRL*-HM bakarrean integratu nahi ditugula. Hona hemen, hautapen murriztapen bakoitzak V.3 irudiko argumentuetarako egingo lituzketen iragarpenak:

R_{HM_1}(sell,maker)=Arg0	R_{HM_2}(sell,maker)=Arg1	R_{HM_3}(sell,maker)=Arg0
R_{HM_1}(sell,year)=AM-TMP	R_{HM_2}(sell,year)=AM-TMP	R_{HM_3}(sell,year)=AM-LOC
R_{HM_1}(sell,car)=Arg1	R_{HM_2}(sell,car)=Arg0	R_{HM_3}(sell,car)=Arg0
R_{HM_1}(in,U.S.)=AM-LOC	R_{HM_2}(in,U.S.)=AM-LOC	R_{HM_3}(in,U.S.)=AM-LOC

Hautapen murriztapenaren arabera, ikusten da argumentu beraren gainean egiten diren iragarpenak kontraesankorrak izan daitezkeela. Horrela, iragarpen horiek guztiak *SwiRL* bakarrean integratu nahiko bagenitu, ikasketako moduluak zailtasunak izango lituzke interferentzia horien guztien artean bere lana egiteko (ikusi V.5 irudia). Hau, noski, mota honetako *SwiRL*-HM sailkatzaileekin esperimentuak egin eta emaitzak ikusi ondoren atera dugun ondorio bat besterik ez da. Izan ere, erabiltzen den ikasketa moduluaren arabera, ustezko interferentzia hauek modu egoki batean tratatuak izateko aukerak egon litezke. Dena dela, *SwiRL* sistemako ikasketa modulua ez da arazo honekin modu arrakastatsu batean borrokatzeko gai; eta horregatik, seguruagoa dirudi HM bakoitzeko *SwiRL*-HM bakarra sortu eta ondoren guztiak, HMen aurreikuspenekin batera, meta-sailkatzaile batekin konbinatzea. Horrela, *SwiRL*-HM bakoitzak dagokion HMaren propietateak bereganatuko lituzte (beste HMen ezaugarrien interferentziarik gabe) eta metasailkatzaileak, azkenik, *SwiRL*-HM bakoitzaren onurak identifikatu eta biltzeko eginkizuna izango luke.

V.2.2 Sailkatzaileen arteko meta-sailkatzailea

SwiRL-HMen irteerak eta hautapen murriztapenen aurreikuspenak konbinatzeko metasailkatzaile bitar bat erabiliko dugu. Lehenago ere, SRL sistema ezberdinak konbinatzeko erabili izan da teknika hau (Surdeanu *et al.*, 2007)

Arg0 type:chunk elem:pNP head:POS:N head:maker... **HM_1:Arg0 HM_2:Arg1 HM_3:Arg0**
AM-TMP type:chunk elem:pNP head:POS:N head:year... **HM_1:AM-TMP HM_2:AM-TMP HM_3:AM-LOC**
< Predicate >
Arg1 type:chunk elem:pNP head:POS:N head:cars... **HM_1:Arg1 HM_2:Arg0 HM_3:Arg1**
AM-LOC type:chunk elem:pPP head:POS:IN head:in... **HM_1:AM-LOC HM_2:AM-LOC HM_3:AM-LOC**

Irudia V.5: *SwiRL* sistemak entrenatzeko erabiltzen dituen argumentuen adibideak. Urdinez, adibide bakoitzari, HM ereduen hainbat iragarpen (kontraesankor) txertatu zaizkio. *SwiRL*eko ikasketa modulua ez da horrelako interferentziekin modu arrakastatsuan lan egiteko gai.

eta izaera ezberdineko aurreikuspenak modu erraz eta eraginkor batean konbinatzeko bidea ematen du.

Meta-sailkatzailearen sarrera *SwiRL*-HM eta HM indibidualen irteeretatik eratortzen da. Zehazki, irteera horietako bakoitzetik, *datum* delako bat sortuko da meta-sailkatzailearentzat. *SwiRL*-HMak eta HM ereduak bezalako sailkatzaile arruntentzat *datum*-ak aditz argumentuak diren bitartean, meta-sailkatzaile bitarraren *datum*-ak argumentu horientzat aurreikuspenak egin dituzten sailkatzaileen aurreikuspenak izango dira. Horrela, $6 + 6$ oinarri sistema konbinatu nahi baditugu (6 SwiRL-HM eredu eta 6 HM), aditz argumentu bakoitzeko, meta-sailkatzaileak 12 *datum* izango ditu. Meta-sailkatzailea entrenatzeko *datum* horietatik 0 ala gehiago zuzenak izango dira ($+1$ labela izango dutenak), *datum* horiei dagozkien jatorrizko sistemek, kasuan kasu, aurreikuspen zuzenak eman badituzte; eta gainerako datumak (-1 labela) okerrak izango dira. Meta-sailkatzaileak ez du, beraz, jatorrizko aditz argumentuaren rola aurreikusten ikasiko. Horren ordez, jatorrizko argumentuaren gainean aurreikuspenak egin dituzten oinarri sistemetatik, zeinek egin duen bere lana hobekien esango digu. Alegia, testeko *datum*-a emanda, meta-sailkatzailearen lana datum hori zuzena ($+1$) ala okerra (-1) den erabakitzea izango da.

Adibidea (datumak): Demagun bi *SwiRL*-HM sistema eta HM eredu bat konbinatu nahi ditugula meta-sailkatzailearekin: `SwiRL-HM`$_1$, *`SwiRL`*`-HM`$_2$ eta `HM`. Entrenamendu multzoko argumentu bakoitzeko hiru datum sortuko dira beraz, hiru direlako konbinatu nahi ditugun oinarri sistemak. Demagun orain, argumentuetako baten rola Arg2 dela. Demagun baita ere, `SwiRL-HM`$_1$ sistemak eta `HM` ereduak Arg2 rola aurreikusi dutela argumentu horretarako

+1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**
-1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**
+1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**

Irudia V.6: Argumentu berari dagozkion hiru datum, bat sailkatzaile bakoitzeko. Lehen datuma *SwiRL*-HM$_1$ sailkatzaileari dagokio, bigarrena *SwiRL*-HM$_2$ sailkatzaileari eta azkena, HM ereduari. Datum bakoitzaren labela (+1/-1) gorriz agertzen da. Labelaren ondoren, ezaugarri zerrendako 1) puntuko ezaugarriak (beltzez) eta haien balioak (urdinez) ditu erantsita.

eta `SwiRL-HM`$_2$ sistemak berriz, AM-LOC lekuzko rol adjuntu okerra. Hortaz, argumentu horretarako, esan bezala, hiru datum sortuko genituzke. Haietatik bik (`SwiRL-HM`$_1$ eta `HM`) aurreikuspen zuzena eman dutenez, dagozkien datumek +1 labela izango dute meta-adibidean, eta `SwiRL-HM`$_2$ sistemari dagokionak berriz, -1 labela. Datum bat hortaz oinarrizko sailkatzaile (*SwiRL*-HM ala HM) bati egongo da lotuta eta, funtsean, labelak sailkatzaile horrek lana ongi ala gaizki egin duen kodetuko du.

Meta-ezaugarriak

Metasailkatzailea entrenatu eta *datum* berriak sailkatu ahal izateko beharrezkoak dira ezaugarriak. Guztira hiru motatakoak izan daitezke *datum* bakoitzerako erauzi ditugunak:

1. Sarrerako ereduek (*SwiRL*-HM sailkatzaileek eta HM ereduek) argumenturako ematen dituzten rol aurreikuspen guztiak. Ezaugarri hauek, sistema bakoitzak egiten dituen aurreikuspenak ezagutu eta beren fidagarritasuna neurtzeko bidea irekitzen diote meta-sailkatzaileari. Goiko adibidearekin jarraituz, datum bakoitzean sailkatzaile bakoitzaren aurreikuspenak gehitu beharko genituzke (ikusi V.6 irudia).

2. Datumarekin lotuta dagoen oinarrizko sailkatzailearen rol berbera proposatu duten sailkatzaile kopurua. Honek meta-sailkatzaileari bozketa moduko erregelak ikasten laguntzen dio. Rola sistema bakarrak proposatu badu "kop>0" ezaugarria gehituko da datumean; bi sistemak egin badute, "kop>0" eta "kop>1" etab. Adibidearekin jarraituz, ezaugarri hauek gehitu ondoren, datum zerrendak V.7 irudiko itxura izango luke.

+1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**	**kop>0**	**kop>1**
-1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**	**kop>0**	
+1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**	**kop>0**	**kop>1**

Irudia V.7: V.6 irudiko adibideari 2. puntuko ezaugarriak gehitu ondoren (urdinez), datumek izango luketen itxura

+1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**	**kop>0**	**kop>1**	**SwiRL-HM$_1$+HM**
-1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**	**kop>0**		**SwiRL-HM$_2$**
+1	SwiRL-HM$_1$:**ARG2**	SwiRL-HM$_2$:**AM-LOC**	HM:**ARG2**	**kop>0**	**kop>1**	**SwiRL-HM$_1$+HM**

Irudia V.8: V.7 irudiko adibideari 3. puntuko ezaugarriak gehitu ondoren, datumek izango luketen itxura

3. Datumarekin lotuta dagoen oinarrizko sailkatzailearen rol berbera proposatu duten sailkatzaileen zerrenda. Ezaugarri honek, sailkatzaileen arteko konbinazioei etekina ateratzen laguntzen dio meta-sailkatzaileari. V.8 irudian ezaugarri hau erantsi ondoren, adibideko datumek izango luketen itxura ikus daiteke.

V.8 irudiko lehen eta azken datumak berdinak dira nahiz eta jatorrizko sailkatzaile ezberdinei dauden lotuta (gogoratu `SwiRL-HM`$_1$ eta `HM` ereduei lotuta daudela hurrenez hurren). Dena dela, ohar moduan aipatu behar da, nolanahiko sinplifikazioak egiteko tentazioa ekidin beharra dagoela, meta-sailkatzailea trebatzeko erabiliko dugun ikasketa ereduak kontuan hartzen baititu adibide errepikatuak.

Ikasketa eredua

Meta-sailkatzaile bitarra inplementatzeko *Support Vector Machines* (SVM) erabili dugu. SVMak oso aproposak gerta daitezke sailkatzaile bitarrak sortzeko, eta zalantzarik gabe, beren sendotasuna eta efizientzia dela eta, gero eta erabiliagoak dira Lengoaia Naturalaren Prozesamenduan ikasketa automatikoaren bidez izaera oso ezberdineko atazak ebazteko. Asko idatzi da SVMeei buruz 90. hamarkadan publikoki aurkeztu zituztenetik (Cortes eta Vapnik, 1995; Vapnik, 1999) eta ez diogu, beraz, ikasketa automatikoko teknika konplexu honi buruzko azalpenak emateari ekingo . Sarean hainbat eta

hainbat SVM inplementazio topa ditzakegu. Guk meta-sailkatzailea trebatzeko SVM-Light[3] pakete ezaguna erabili dugu, kernel polinomial koadratikoa erabiliz eta C parametroa 0.001ean ezarriz (garapenerako corpusaren gainean *hill-climbing* motako doitze esperimentuak egin ondoren).

Gogora dezagun meta-sailkatzailea trebatzeko erabiltzen diren datuma eta ezaugarriak hainbat sailkatzaileren irteeretatik erauzten direla. Horregatik, entrenamendu argumentuen gainean oinarrizko sailkatzaileen aurreikuspenak ere behar ditugunez, beharrezkoa da entrenamenduko aurreikuspen horiek *cross-validation* teknika erabiliz lortzea. Horretarako, entrenamenduko corpusa n zatitan banatzen dugu. Aurreikuspenak egin behar dituzten *SwiRL*-HM sistemak eta HM ereduak lehen n-1 zatiekin entrenatuz, n. zatia modu "garbi" batean etiketatzea lortzen da (ez baita egokia, jakina, etiketatzea entrenamenduko adibideetan egitea). Behin entrenamenduko n zatiak eredu eta sistema guztien aurreikuspenekin etiketatu direla, posible da guztia bildu eta meta-sailkatzailea entrenatzea. Testeko corpusa etiketatzeko erabiltzen diren sistemak eta ereduak entrenamendu corpus osotik abiatuta ikasten dira *cross-validation* teknikarik erabili gabe.

Meta-sailkatzaileak eratzeko beste hainbat aukera

Orain arte azaldu dugun meta-sailkatzailea sinplea baina arrakastatsua da. Dena dela, ez da tesi lan honetan kontuan hartu dugun aukera bakarra. Bi multzotan bana ditzakegu aztertu ditugun meta-sailkatzaileak:

- **Gainbegiratu gabeko meta-sailkatzaileak**: multzo honetan bozketan oinarritutako meta-sailkatzaileak ditugu. Guztira, bi bozketa sistema planteatu ditugu:

  1. **Bozketa sinplea**: *SwiRL*-HM eta HM ereduen arteko bozketa. Testeko argumentu bakoitzeko, sistema bakoitzak (*SwiRL*-HM edo HM) bere kandidatuaren (iragarpenaren) aldeko bozka ematen eta meta-sailkatzaileak boto gehien jaso duen labela itzultzen du. Sistema hau oso sinplea da baina errendimendu baxuko HM ereduek kalte egiten diote emaitza globalari, besterik gabe, kalitate baxuko boto asko ematen dituztelako.

  2. **Bozketa ponderatua**: sistema bakoitzaren botoen kalitatea berdina ez denez, bakoitzari pisu bat ematen zaio bozketan. Pisu hori,

[3] `http://svmlight.joachims.org`

botoa eman duen sistemaren prezisio orokorraren araberakoa da (garapeneko corpusaren gainean lortutako prezisio indizeen araberakoa). Modu horretan, meta-sailkatzaileak sistema indibidual onenen botoak lehenetsiko ditu, kaxkarrenen aurretik.

- **Meta-sailkatzaile gainbegiratuak**: bozketan oinarritzen direnak ez bezala, sailkatzaile hauek gainbegiratutako informazioan oinarritzen dira (zuzenean ala zeharka) beren sailkatze ereduak sortzeko.

  1. ***SwiRL*-HM moduko meta-sailkatzailea**: aukera hau V.2.1 puntuan aurkeztu dugunaren oso antzekoa da baina oinarri sisteman, hautapen murriztapenen iragarpenak ezaugarri moduan txertatzeaz gain, *SwiRL*-HM sistemen iragarpenak ere txertatzen dira behar bezala kodetuta. V.2.1 atalean deskribatzen dugun sistema bezala, hau ere *one vs. all* moduko Adaboost sailkatzailea da.

  2. **Meta-sailkatzaile "arrunta"**: V.2.2 atalean sakonki aztertu dugun 3 ezaugarritako meta-sailkatzaile arrunta. SVM sailkatzaile bitar moduan ikasten da.

  3. **Meta-sailkatzaile "aberastua"**: meta-sailkatzaile arruntaren berdina da baina SRLko ezaugarri tipikoekin (adibidez, aditzaren lema, argumentuaren gunea...) aberastua. Horrela, meta-sailkatzaileak datumetan, V.2.2 atalean aipatutako 3 metaezaugarriez gain, beste zenbait ezaugarri (askoz ere sakabanatuagoak bestalde) izango lituzke iragarpenak egiteko. SVM sailkatzaile bitar moduan ikasten da.

V.3 Esperimentazio ingurunea

Esperimentu hauetarako, berriro ere, CoNLL-2005eko SRL atazan eskaini zen datu multzoa erabili dugu. HMak, *SwiRL*-HM ereduak eta meta-sailkatzailea beraz, 02-21 sekzioetatik ikasi dira. Testerako 23. sekzioa erabili dugu eta parametroak doitzeko garapen corpuserako berriz, 00 sekzioa. Domeinuz kanpoko corpus bezala, aurreko ataletan bezala, SemLinken atzigarri dagoen Brown corpusa erabili dugu.

corpus	Meta-adibideak	pos. (neg.)	1 mota	2 mota	3 mota
train	1,824,498	%62 (%48)	%56	%33	%11
test (WSJ)	104,689	? (?)	%51	%39	%10
test (Brown)	105,911	? (?)	%52	%37	%11

Taula V.1: Meta-sailkatzailea entrenatu eta testatzeko datuak corpusaren arabera ordenatuta. "Meta adibideak" zutabeak, meta-corpus bakoitzak duen adibide kopurua azaltzen du. "Pos (neg)" zutabean, adibideetatik zenbat diren positiboak (+1) adierazten da; parentesien artean, negatiboak (−1). Azken 3 zutabeetan, adierazten da meta-sailkatzaileak erauzten dituen ezaugarri guztietatik zenbat diren 1 motakoak, 2 motakoak edota 3 motakoak (ikusi V.2.2 atala).

Gure esperimentuen helburua argumentuen *sailkatze prozesua* hobetzera zentratuko denez, datu multzoko gold argumentuekin egingo dugu lan. Horrez gain, HMek gune nominalekin soilik lan egin dezaketela kontuan hartuta, mota horietako argumentuekin zentratuko gara soilik. Guztira, entrenatu eta HMak modu gainbegiratuan sortzeko 141,000 argumentu geratzen dira. Garapena eta testerako 5,000 eta 8,000 argumentu, hurrenez hurren, eta azkenik domeinuz kanpoko Brown corpuserako 8,100 argumentu. Meta-sailkatzaileak entrenamenduko corpus estandarreko argumentu bakoitzeko, 12 datum ikasten ditu, bat konbinatzen den sistema bakoitzeko. Beraz entrenamendu eta testeko adibide multzoa 12 aldiz handiagoa izango da meta-sailkatzailearentzat (ikusi V.1 taulako "Meta-adibideak" eta "pos (neg)" zutabeak). Tamainak, noski, eragin zuzena du ikasteko edota etiketatzeko denboran. Horrela *SwiRL*-HM ereduak ordu gutxi batzutan ikasten diren bitartean, meta-sailkatzailearen datuak ikasteko 170 ordutako lan konputazionala behar izan dugu QuadCore Xeon X5460 (3160MHz, 32GB) makina batean.

Meta-sailkatzaileak erauzten dituen ezaugarrien distribuzioari dagokionez, V.2.2 atalean azaldu ditugun "1 motako" ezaugarriak dira gehienetan azaltzen zaizkigunak. Ikusi V.1 taula xehetasunetarako.

V.4 Emaitzak eta analisia

V.2 taulan azaldutako bi konbinazio metodoen errendimendua konparatzen da *SwiRL* sailkatzailearenarekin. Emaitzak core (Core zutabean) eta adjuntuentzat (Adj zutabean) ematen dira eta baita guztiak batera ere (All).

SwiRL-HM sistema bakoitzari dagozkion emaitzak taulako "$+SP_*$" lerroetan ematen dira. Gogora dezagun, *SwiRL*-HM sailkatzaileak *SwiRL* oinarri sistemari HMetatik erauzitako ezaugarriak gehituz sortzen direla. Hortaz, ezaugarri horien jatorrizko HMaren arabera "$+SP_*$" lerro bat izango dugu. Esperimentuetan erabilitako HMak IV.2 atalean ikusitako berberak dira eta bakoitzetik erauzitako ezaugarri semantikoen errendimendua IV.3 taulan ikus daiteke. Guztira 6 hautapen murriztapen eredutatik erauzi ditugu *SwiRL* oinarri sistemari banan-banan txertatu beharreko ezaugarriak. Horrela, taulan ikus daitezkeen 6 *SwiRL*-HM ereduak eratu ditugu :

- $SwiRL + SP_{Res}$: WordNeten oinarritutako Resniken hautapen murriztapenen aurreikuspenak barneratzen dituen *SwiRL*-HM sistema.
- $SwiRL + SP_{wn}$: WordNeten gainean lan egiteko eratu genuen HM sistema prototipikoaren aurreikuspenak barneratzen dituena.
- $Swirl + SP_{sim_{Jac}}$: BNC corpusetik erauzitako thesaurusaren gainean Jaccarden antzekotasun neurria aplikatuz erauzten diren ezaugarriak txertatuta dituen *SwiRL*-HM sistema (antzekotasun zuzena erabiltzen duena, lehen mailakoa).
- $Swirl+SP_{sim_{cos}}$: Aurrekoaren antzekoa, kosinuaren antzekotasuna erabiltzen duena Jaccardenaren ordez.
- $Swirl + SP_{sim^2_{Jac}}$: Linen antzekotasun thesaurusean 2. mailako antzekotasuna Jaccard neurriaren bidez kalkulatzen duen HMaren aurreikuspenak txertatuta dituen *SwiRL*-HM sistema.
- $Swirl+SP_{sim^2_{cos}}$: aurrekoaren antzekoa, baina kosinua erabiltzen duena Jaccard beharrean.

Emaitzen taulan ikus daiteke *SwiRL*-HM ereduak orokorrean oinarrizko *SwiRL* sistema baino hobeak direla bai domeinu barnean (WSJ) eta baita domeinuz kanpo ere (Brown). Hala ere, hobekuntzak txikiak dira eta estatistikoki ez esanguratsuak. Bestalde, meta-sailkatzaileari dagokionez, "meta" lerroan ikus daitekenenez, emaitzak hobeak eta estatistikoki esanguratsuak dira %99ko ziurtasunarekin [4]. Domeinu barnean errore murrizketa %14.07 da eta domeinuz kanpo, %11.67.

[4] Ausaz aukeratutako 100 adibideren gaineko t-probaren arabera

Azkenik, core eta adjuntuei dagokienez, bietan ikus daiteke hobekuntza oinarri sistematik meta-sailkatzailearen emaitzetara. Aipatzekoa da, domeinuz kanpo, adjuntuen igoera (5.64 puntu absolutu) core argumentuena (1.78 puntu) baino nabariki handiagoa dela. Espero zitekeen zerbait da hau, izan ere, core argumentuen gehiengoa Arg0 eta Arg1 roletan banatzen da eta hauen artean ezberdintzeko erabilgarriagoak izan daitezke ezaugarri sintaktikoak (adibidez, objektu vs. subjektu modukoak). Bestalde, argumentu adjuntuen rola bereizten laguntzeko ezaugarri sintaktikoak ez dira coreetarako bezain esanguratsuak eta hortaz, kasu hauetan sistemak HMek ematen diguten informazioari atarramendu handiagoa ateratzen ikasten du.

SwiRL oinarri sistemaren eta meta-sailkatzailearen emaitzak rolez rol aztertzen baditugu, argi ikusten da hobekuntzak nabariak direla azkenaren alde ia rol guztietarako. Meta-sailkatzailea gai da beraz, *SwiRL*-HM sistemen berezitasunak eta bertuteak ezagutu eta haien iragarpenen artean "seinale positiboa" topatzeko. Horren froga dira V.3 eta V.4 tauletan ikusten diren emaitzak. Lehen taulan, *SwiRL* oinarri sistemak eta meta-sailkatzaileak WSJ corpusean lortzen dituzten emaitzak ikus daitezke eta, bigarrenean, Brown corpusean lortzen dituztenak. Emaitzen tendentzia gorakorra nabaria da oinarri sistematik meta-sailkatzailera, baina badira salbuespen batzuk. AM-TMP argumentu adjuntu garrantzitsuarekin emaitzak hobeak dira WSJ corpusean *SwiRL* sailkatzailearentzat (gutxigatik bada ere) baina ikusten denez, gauzak erabat aldatzen dira Brown corpusean, meta-sailkatzailearen onerako. Oro har, core argumentu garrantzitsuetan (maizen agertzen direnetan, alegia, Arg0, Arg1, Arg2 eta Arg3 roletan) hobea da meta-sailkatzailea bi corpusetan (Arg4 rolean erorketa jasaten du Brown corpusean estaldura galera baten ondorioz). Eta gauza berbera gertatzen da rol adjuntu garrantzitsuekin (AM-TMP, AM-DIS, AM-CAU, AM-LOC...)

Meta-sailkatzaile "alternatiboen" errendimendu eskasaren analisia

V.2.2 atalean aipatu ditugun meta-sailkatzaile alternatiboek ez dute emaitza lehiakorrik ematen. Garapen corpusaren gainean soilik egindako esperimentutan ikusi genuen **bozketa sinpleko** emaitzak bereziki txarrak direla eredu guztiak konbinatzen direnean (*SwiRL* oinarri sistemaren azpitik geratzen da, distantzia handira). Lehenago iradoki dugun bezala, bozketa sinplean errendimendu oso ezberdineko sailkatzaileak (*SwiRL*-HM eta HM ereduak)

	WSJ-test			Brown		
	Core	Adj	All	Core	Adj	All
SwiRL	93.25	81.31	90.83	84.42	57.76	79.52
SwiRL+SP_{Res}	93.17	81.08	90.76	84.52	59.24	79.86
SwiRL+SP_{wn}	92.88	81.11	90.56	84.26	59.69	79.73
SwiRL+$SP_{sim_{Jac}}$	93.37	80.30	90.86	84.43	59.54	79.83
SwiRL+$SP_{sim_{cos}}$	93.33	80.92	90.87	85.14	60.16	80.50
SwiRL+$SP_{sim^2_{Jac}}$	93.03	82.75	90.95	85.62	59.63	80.75
SwiRL+$SP_{sim^2_{cos}}$	93.78	80.56	91.23	84.95	61.01	80.48
Meta	**94.37**	**83.40**	**92.12**	**86.20**	**63.40**	**81.91**

Taula V.2: SwiRL-HM ezberdinen prezisio, estaldura eta F_1 neurriak WSJ (ezkerrean) eta Brown (eskuinean) corpusetan.

	SwiRL			Meta-sailkatzailea		
	prec.	rec.	F_1	prec.	rec.	F_1
A0	93.61	96.66	95.11	95.45	96.97	96.21
A1	93.28	94.47	93.87	93.92	95.81	94.86
A2	85.97	82.63	84.27	84.63	86.94	85.77
A3	77.61	63.41	69.79	84.61	67.07	74.82
A4	86.84	78.57	82.50	87.17	80.95	83.95
AM-ADV	58.51	51.4	54.72	64.63	49.53	56.08
AM-CAU	61.11	70.96	65.67	73.52	80.64	76.92
AM-DIR	46.15	25.00	32.43	70.01	29.16	41.17
AM-DI	84.31	82.69	83.49	93.61	84.61	88.88
AM-EXT	50.00	12.50	20.00	50.00	12.50	20.00
AM-LOC	85.19	80.93	83.01	84.78	85.31	85.04
AM-MNR	55.81	54.13	54.96	66.35	53.38	59.16
AM-PNC	51.85	37.83	43.75	64.02	43.24	51.61
AM-TMP	93.57	95.94	94.74	93.11	95.94	94.50

Taula V.3: *SwiRL* oinarri sistema (ezkerrean) eta meta-sailkatzailearen emaitzak rolez rol, WSJ test corpusean.

konbinatzen dira inolako kontrol edo murriztapenik gabe. Sailkatzaile bakoitzak bere kandidatuaren alde egiten du eta halabeharrez kandidatu okerrak gehiegitan gailentzen zaizkie kandidatu zuzenei.

Antzeko zerbait gertatzen da baita ere **bozketa ponderatuko** meto-

	SwiRL			Meta-sailkatzailea		
	prec.	rec.	F_1	prec.	rec.	F_1
Arg0	87.62	89.28	88.44	90.25	90.73	90.49
Arg1	84.31	90.58	87.33	86.46	92.12	89.20
Arg2	52.74	56.81	54.70	52.78	59.49	55.93
Arg3	36.36	19.04	25.00	41.66	23.80	30.30
Arg4	59.37	34.54	43.67	60.71	30.90	40.96
AM-ADV	45.13	24.28	31.57	53.53	25.23	34.30
AM-CAU	64.70	45.83	53.65	78.57	45.83	57.89
AM-DIR	64.70	45.83	53.65	71.08	40.41	51.52
AM-DIS	52.63	27.02	35.71	60.00	32.43	42.10
AM-LOC	67.77	61.24	64.34	68.58	71.52	70.02
AM-MNR	47.39	38.92	42.74	58.22	46.78	51.88
AM-PNC	51.72	39.47	44.77	51.85	36.84	43.07
AM-TMP	78.96	78.06	78.51	84.48	83.76	84.12

Taula V.4: *SwiRL* oinarri sistema (ezkerrean) eta meta-sailkatzailearen emaitzak rolez rol, Brown corpusean.

doarekin, nahiz eta kasu honetan emaitzak ez diren horren txarrak. Sistema bakoitzak bere botoa garapen korpusaren gainean lortutako prezisioaren arabera ponderatzen du bozketan beste sailkatzaileek baino eragin handiagoa izaten saiatzeko. Doitze txiki honek bozketa sinplearekiko emaitzak nabari hobetzen ditu, nahiz eta *SwiRL* oinarri sistemarekiko hobekuntza esanguratsurik ez erakutsi.

Meta-sailkatzaile gainbegiratuei dagokionez, berriro ere garapen corpusarekin soilik egindako esperimentuek erakutsi zuten planteatutako alternatibak ez zirela ezagutza behar bezala orokortzeko gai (V.2.2 atalean sakonki aztertu dugun hiru ezaugarritako "**meta-sailkatzaile arruntaren**" kasuan izan ezik). Guztiek, aipatutako salbuespenak izan ezik, ez dute garapeneko corpusean oinarri sistema gainditzea lortzen:

- ***SwiRL*-HM motako meta-sailkatzailearen** kasuan, badirudi erabiltzen den ikasketa eredua ez dela gai hautapen murriztapenek beste ezaugarri guztien artean integratzen duten ezagutza (askotan) kontraesankorrari etekinik ateratzeko. Honek noski ez du esan nahi gaizki pentsatutako eredua denik edota adibide horietan txertatzen diren kontraesan horiek direnik emaitza ahul horien errudun, baizik eta ez dio-

gula (guk ala Adaboosten oinarritzen den ikasketa moduluak) etekinik ateratzen jakin.

- **Meta-sailkatzaile "aberastuak"**, gogora dezagun, hiru ezaugarriko meta-sailkatzaile arruntari SRLko ezaugarri tipikoak gehitzen dizkio emaitzak hobetzeko asmoz, baina honek ere ez du ezer positiborik lortzen. Garapeneko corpusaren gainean egindako esperimentuek erakutsi zuten aditzaren lema bezalako ezaugarri bat soilik gehitzeak emaitzak *SwiRL* oinarri sistemarekiko okertu besterik ez zituela egiten. Meta-ezaugarrien alboan (ikusi V.2.2 atala) aditzaren lema edota argumentu gunea bezalako ezaugarriak oso sakabanatuta daude, eta badirudi sailkatzailea nahastu besterik ez dutela egiten.

Meta-sailkatzaile "alternatiboekin" egin ditugun esperimentuak ez dira arrakastatsuak izan. Aitortu beharra dago gertaera honek ez gaituela ustekabean harrapatu (bozketan oinarritutako metodoen erabilera arinagatik batik bat) baina esperimentuok tesi lan honen emaitza partzial negatiboen zerrendan bere tokia topa dezakete, nahiz eta, noski, ez diogun atea ixten metodo hauekin (ala hauen bertsio sofistikatuagoekin) errendimendu altuko sailkatzaileak lortzeari.

V.5 Ondorioak

Rolak modu egokian sailkatzeko SRL sistema tipiko gehienek erabiltzen dituzten ezaugarri sintaktikoez gain, beharrezkoak dira ezaugarri semantikoak. Hala ere, aditza eta argumentuen gune nominaletatik erauzten diren ohiko ezaugarriak ez dira nahikoa eta beharrezkoa da horien sakabanaketa gainduko duten beste ezaugarri semantiko osagarriak garatzea. Guk ezaugarri horiek tesi lan honetarako aurkeztutako Hautapen Murriztapenak erabiliz sortu ditugu eta erakutsi dugu horiek baliatuz rolen sailkatze automatikoa modu esanguratsuan hobetzerik badagoela.

Hautapen murriztapenak artearen egoeran dagoen sailkatzaile baten hainbat barianteetan integratu ditugu ezaugarrietan oinarritutako teknika erraz bat erabiliz. Ondoren, meta-sailkatzaile bitar sinple baten laguntzaz, bariante horien eta hautapen murriztapenen gaitasunak identifikatu eta aprobetxatu ditugu NP eta PP argumentuen sailkapen ataza errealista baten emaitzak domeinu barnean eta domeinuz kanpo nabarmen hobetzeko.

Oro-har, puntako SRL sistema baten sailkatze ahalmena hobetu dugu ezaugarri lexiko-semantikoen sakabanaketaren albo ondorio tipikoak hautapen murriztapenen bidez gutxituz.

VI. KAPITULUA

Ondorioak eta etorkizuneko lanak

Lengoaia naturalaren prozesamenduan zeresan handia eman du azken urteotan rolen etiketatze automatikoak. Tesi lan honetan ikusi dugu esperimentu kontrolatuetan ingeles hizkuntzarako sistema onenek %80 inguruko F_1 neurriak emateko gai direla PropBankeko corpusaren gainean lan egiten dutenean. Hala ere, errendimendu ikusgarri hauetatik haratago, ikusi dugu oraindik ere argitu gabeko zenbait puntuk inguratzen dutela ataza, eta zenbaitzuetan argi apur bat egiten saiatu gara.

Domeinuz kanpoko corpusen rolen etiketatze automatikoa izan dugu ikerlan honen abiapuntu nagusi, izan ere, emaitza lehiakorrenak ematen dituzten sistemek ere erorketa aipagarriak jasaten dituzte mota honetako corpusetan. Arazo hau hobeto ulertzeko helburuarekin bere sustraietara bideratu dugu gure ikerketa, hipotesiak planteatuz eta esperimentu sorta egokien bidez, ekarpenak egiteko.

PropBank eta VerbNet arteko konparaketa

PropBank corpusean oinarritutako sailkatzaileekin egin dugu lan batik bat corpus egokia eta oso erabilia delako mota honetako sistemak eraikitzeko.

Hala ere, PropBankek erabiltzen duen rol multzoari egiten zaizkion kritikak eta erakusten dituen mugak kontuan hartuz, erabilera konputazionalerako egokia den beste rol multzo batekin, VerbNeteko rol tematikoekin, ere trebatu ditugu sistemak, izaera ezberdineko bi rol multzo hauekin lan egiten duten sailkatzaileak kondizio esperimental berezietan alderatzeko asmoz. Lan honen berri, III kapituluan ematen dugu. Bertan entropia maximoko markoven eredutan oinarritutako sailkatzaile sekuentzial eta lehiakor bat trebatzen dugu rol multzo bakoitzean, ondoren, bere errendimendua testeko ingurune ezberdinetan ebaluatzeko. Sailkatzaileak rol multzo bakoitzarekin lortzen dituen emaitzak aztertuz honakoak ondorioztatu ditugu:

- VerbNeteko sailkatzailearen sentsibilitatea handiagoa aditzetik eratortzen diren ezaugarriekiko: aditza eta aditzetik eratorritako ezaugarriak oso garrantzitsuak dira rolak sailkatzakeko garaian. Horregatik, ikasketa eta testeko corpusetatik aditzaren inguruko ezaugarri hauek ezabatzen ditugunean, normala da sailkatzaileen errendimendua jaistea. Hala ere, ikusi ahal izan dugu aditzaren inguruko informazioak batez ere VerbNeteko sailkatzailean duela eragina, PropBankekoak ez bezala, VerbNeteko sailkatzaileak errendimendu galera esanguratsuak erakusten baititu aditzetik eratorritako ezaugarriak ezabatzen direnean. Gertaera honen arrazoia sinplea izan daiteke: PropBankeko rol zenbakituak "sintaktikoagoak" diren bitartean, rol tematikoak semantikoagoak dira. VerbNeteko sailkatzaileak sintaktikoki patroi oso antzekoak jokatzen dituzten rol semantiko ezberdinen artean bereizteko gaitasuna erakutsi beharra du, eta zailtasunak topatzen ditu aditzaren inguruko informaziorik topatzen ez duenean. Izan ere, VerbNeteko rol tematikoen artean diferentzia sintaktikoez gain, diferentzia semantiko finak ere badira (`Patient/Theme`, `Agent/Actor`) eta askotan, rol baten ala bestearen aldeko apustua egitea aditzak ematen digun informazio semantikoaren araberakoa izaten da. Zenbait aditzek, esaterako, ez dute `Agent` rolik onartzen baina aditzaren informaziorik gabe, eta bestelako informazio semantikorik gabe, ezinezkoa gertatzen zaio rol tematikoen sailkatzaileari hori jakitea.

- PropBankeko rol sailkatzaileak hobeto sailkatzen ditu aditz ezezagunak: Askotan ikasketa corpusean agertu ez diren aditzen argumentuak sailkatu behar izaten dituzte rol sailkatzaileek. Aditz hauek onartzen duten rol multzoari buruzko informazioa zein den jakitea, beraz, ikasketa corpusetik kanpo lortutako ezagutza iturriren batean kontsultatu

beharreko zerbait da. Iturri horien laguntzarik gabe, VerbNeteko sailkatzaileak, PropBankekoak baino zailtasun handiagoak erakusten ditu etiketatze lana modu egokian egiteko eta ondorioz, ikasketa corpusak zehazten duen ezagutza orokortzeko ahalmen mugatuagoa erakusten du. Hau azaltzeko proposatu dugun arrazoietako bat, VerbNeteko rol tematikoen arteko diferentzia semantiko finak eta zenbait aditzentzako bereziki definitzen diren rolak dira. Rol tematikoen artean diferentzi semantiko finak zehazten dira definizioz eta horregatik, rol horien aukeraketan berebiziko garrantzia du aditzak, berak definitzen baititu askotan sintaktikoki oso antzekoak izan daitezkeen bi rolen artean zein aukeratu behar den. PropBanken kasuan ez da antzekorik gertatzen, aditz guztiek rol etiketa berdinekin egiten dutelako lan eta, horregatik, bere errendimendua ez da VerbNeteko sailkatzailearena bezain beste degradatzen aditz ezezagunekin lan egiten duenean.

- PropBankeko rolen sailkatzaileak hobeto etiketatzen ditu domeinuz kanpoko adibideak: PropBank eta VerbNeteko rol sailkatzaileek domeinuz kanpoko adibideekin aritzeko abilezia neurtu genuen Brown corpuseko zenbait adibiderekin esperimentuak eginez. Ikusi genuen, zalantzarik gabe, mota horretako adibideekin hobea zela, kasu honetan ere, PropBankeko sailkatzailea VerbNetekoa baino. Honen arrazoia aurreko puntuetan azaldutakoak eman liezaguke, domeinuz kanpoko corpusetan, ohikoagoa izaten delako aditz ezezagunak edota maiztasun gutxikoak aurkitzea eta, esan bezala, mota horretako adibideekin VerbNeteko sailkatzaileak errendimendu galera garrantzitsuak izaten ditu.

Gaur egungo sailkatzaileek rolen etiketatzea gauzatzeko erauzten dituzten ezaugarriekin eta erabiltzen dituzten ikasketa teknikekin ez dirudi erraza denik VerbNeteko sailkatzaile eraginkorrak sortzea. Hala ere, sailkatzaile hauek uneko mugak onartzen ditugun bezala, onartzen dugu baita ere, LNPko aplikazioen aldetik behintzat, rol tematikoak PropBankeko argumentu zenbakituak baino interesgarriagoak izan daitezkeela eta, horregatik, tesi lan honetan rol tematikoak lortzeko bi bide ezberdin proposatu ditugu:

- VerbNeteko rol tematikoen sailkatzailea zuzenean erabili: rol tematikoekin etiketatutako corpusa abiapuntu hartuz, sailkatzaile bat entrenatu eta testeko adibideak zuzenean sailkatzaile horrekin etiketatzea.

- PropBankeko rolak SemLink erabiliz rol tematikoetara itzultzea: PropBankeko rol zenbakituekin etiketatutako corpusa abiapuntu hartuz, sailkatzaile bat entrenatu eta testeko adibideak rol zenbakituekin etiketatzean datza. Ondoren, SemLinkek PropBankeko roletatik rol tematikoak lortzeko eskaintzen digun mapaketaren laguntzaz (eta aditzaren VerbNet klasea desanbiguatu ondoren), rol zenbakituak rol tematikoetan bihurtzen dira.

Ikusi dugu domeinuz kanpoko adibideekin mapaketaren hurbilpenak ia modu esanguratsuan gainditzen duela VerbNeteko sailkatzailea, nahiz eta WSJ corpusean bi hurbilpenek emaitza antzekoak lortzen dituzten,.

Hautapen murriztapenak eta haien aplikazioa

Gainontzeko kapituluetan gure arreta zehazki rolak etiketatzeko sistemek osagai lexikalekiko (argumentuen gunea eta aditzekiko) erakusten duten gehiegizko dependentzia aztertu eta soluzioak proposatzera bideratu dugu. Ikusi dugun moduan, osagai lexikoak etiketatzerako garrantzitsuak diren arren, ezaugarri multzo handiegia osatzen dute eta, ondorioz, sistemek beren sakabanaketa handia sufritzeko joera erakusten dute, domeinuz kanpoko corpusetan batez ere.

IV atalean, WordNeten eta antzekotasun distribuzionalean oinarritutako hautapen murriztapenak diseinatu ditugu. Hautapen murriztapen hauek, rolak etiketatzeko ataza erreal batean jarri ditugu lehian, sistemek darabiltzaten osagai lexikoak modu eraginkorrean orokortzeko gaitasuna erakuts zezaten.

Guztira bi motako antzekotasun neurritan oinarritutako hautapen murriztapenak diseinatu ditugu:

- WordNeten oinarritutako hautapen murriztapenak: bi metodo inplementatu ditugu: Resniken hautapen murriztapenak (IV.2.1 atala) eta synseten sakoneran oinarritutako metodo sinple bat (IV.2.1).

- Antzekotasun distribuzionalean oinarritutako hautapen murriztapenak: erabiltzen duten thesaurusaren (BNC-n oinarritutako bat edo Linen thesaurusa) eta antzekotasun neurriaren arabera, hainbat metodo definitu ditugu (ikusi IV.2.2 kapituluan). Emaitza onenak Linen thesaurusarekin eta Jaccard eta kosinuaren 2. mailako antzekotasun neurriekin lortu ditugu.

WordNeten oinarritutako antzekotasun neurriek estaldura arazoak erakutsi dituzte batez ere antzekotasun distribuzionalean oinarritzen direnekin alderatuta (IV atala). Dena dela, erakutsi dugu aditza eta argumentuen guneak ematen diguten informazioa osa daitekeela hautapen murriztapenen laguntzarekin, eta, hortaz, argumentuen guneek ezaugarri forma hartzen dutenean erakusten duten sakabanaketaren ondorio negatiboak gutxitzeko aukerak egon badaudela.

Hautapen murriztapen diseinuari dagokionez, gure esperimentuetarako egokiak diruditen bi multzo ezberdin probatu ditugu: (1) *aditz-rol* motako hautapen murriztapenak eta (2) *aditz-rol* eta *preposio-rol* motako hautapen murriztapenen erabilera konbinatua (argumentuaren sintagma motaren arabera). Orokorrean, bigarren erabilerak lehenak baino emaitza hobeak eman dizkigu argumentuen sailkapen prozesuan eta, horrela, preposizioak modu bereizian tratatzeak argumentu adjuntuak hobeto sailkatzea ekar dezakeela erakutsi dugu (ikusi IV.2.3 atala).

Lexikoaren orokortzean egindako saiakeren benetako balioa egiaztatzeko hautapen murriztapenak eta antzekotasun distribuzionaleko neurriak artearen egoeran dagoen sailkatzaile batean integratu ditugu lehen aldiz gure ikerketa arloan (V. kapitulua). Hautapen murriztapenek, erabiltzen dituzten antzekotasun metodoen edota thesaurusen arabera, zenbait orokortze hobeto ala okerrago egiten dituztela ikusi dugu. Horregatik, hasteko, hautapen murriztapen bakoitza sailkatzaile bakarrean integratu dugu, beste hautapen murriztapenen interferentziak erabat ezabatzeko. Jarraian, sailkatzaile horien gaitasunak meta-sailkatzaile batean biltzen saiatu gara, sailkatzaileen konbinaziorako ikasketa automatikoko metodoak erabiliz. Sailkatzaile indibidualekin eta meta-sailkatzaile konbinatuarekin egindako esperimentuek erakutsi dute hautapen murriztapenak lexikoa orokortzeko aproposak izateaz gain, rolak etiketatzeko sistemak sendotu eta hobetzeko ere gai direla. Horrela, sistema meta-konbinatuak modu esanguratsuan hobetzen du hautapen murriztapenik erabiltzen ez duen jatorrizko sistema. Ikusi dugu gainera, hobekuntza horiek domeinu barneko corpusekin nahiz domeinuz kanpokoekin gertatzen direla eta, beraz, hautapen murriztapenen integrazioak modu egokian hornitzen dituela sailkatzaileak informazio semantiko erabilgarriarekin.

VI.1 Etorkizuneko lanak

Ikuspegi konputazional batetik, rolak etiketatzeko sistemen garapen prozesua laburra izan da eta, ondorioz, asko dira etorkizunean ataza honen inguruan egin beharreko lanak eta erantzun beharreko galderak. Tesi lan honetan aurkeztutako ikerkuntzari dagokionez, etorkizunari begira, jarraian zehazten diren lerroak jorratu nahiko genituzke:

1. Hautapen murriztapenek rolak etiketatzeko sistema osoetan izan dezaketen eragin positiboaren azterketa sakonagoa. Tesi lan honetan ikusi dugu hautapen murriztapenak modu egokian erabiliz gai garela rolak sailkatzeko sistemak modu esanguratsuan hobetzeko. Hala ere, oraindik ez daude erabat argi hobekuntza horren alderdi guztiak eta uste dugu zentzu horretan egindako azterketa kualitatiboek gure ikerketa hedatu eta lerro berriak identifikatzeko balioko duela.

2. Tesi lan honetan hautapen murriztapenak sortzeko deskribatu ditugun tekniken sinpletasuna gure ikerketaren alderdirik interesgarrienetarikoa da gure ustez. Sinpletasuna interesgarria dela diogu, alde batetik esperimentuak berregiteko erraztasunak ematen dituelako eta bestetik, ez duelako, bestelako filtratze tekniken edo antzekoen bidez, zeharkako ezagutza (edo kutsadura) iturriren eraginik sartzen esperimentuetan. Hautapen murriztapenak gordinik probatu ditugu, gehigarririk gabe, rolak etiketatzeko sistemetan duten eragina ahalik eta modu garbienean irits dakigun. Hala ere, hautapen murriztapenak eratzeko bide "sofistikatuagoei" ala ez hain gordinei ez zaie atea itxi behar, noski, eta etorkizunean bide berri hauek aztertzea izango da gure helburuetako bat.

3. Esperimentuetan erabili ditugun hautapen murriztapenak PropBankeko roletarako modelatu ditugu. Uste dugu tesi lan honetan argi utzi dugula ikerlerro honetan PropBankeko rol multzoa dela erabiliena zalantzarik gabe, eta alderdi horretatik behintzat justifikatuta dagoela erabaki hori. Hala eta guztiz ere, III. kapituluan batez ere aipatu dugun moduan, badira aplikazioetarako interesgarriagoak izan daitezkeen rol multzoak, eta hortaz, haientzako ere hautapen murriztapenak sortu eta ebaluatzea etorkizunean kontuan hartzeko helburua izan behar da.

4. VerbNeteko rol tematikoek mundu errealeko aplikaziotarako erakuts dezaketen egokitasunarekin jarraituz, beharrezkoa da rol tematikoen sailkatzaile eraginkorragoak sortzeko esfortzuak bideratzea. Posible al da VerbNeteko sailkatzaileekin PropBankeko rol zenbakituen sailkatzaileek erakusten duten errendimendu maila erdietsi edo hobetzea?

5. Gure beste lehentasunetako bat euskarazko testuak etiketatzeko sistema bat sortzea izango da, eta horretarako eraketa prozesuan den EusPropBank (Aldezabal *et al.*, 2010) corpusa hartuko dugu oinarri. Euskararako sistema honen garapenak planteatzen duen erronka nagusia eta erakargarriena, sailkapenerako sistema gure hizkuntzaren berezitasunetara egokitzea da. Egokitzapen honek, bestalde, tesi lan honetan beste hizkuntzetarako garatu ditugun errekurtso semantikoak probatzeko aukera emango digu, hizkuntzen arteko errekurtso linguistikoen trukaketan ikertuz eta lan eginez. Ez dago zalantzarik norantz honetan egiten diren ikerketak oso baliagarriak izan daitezkeela errekurtso handirik gabeko hizkuntzentzako.

Bibliografia

Agirre E. eta Martinez D. Learning class-to-class selectional preferences. *Proceedings of the 2001 workshop on Computational Natural Language Learning (CoNLL-2001)*, 1–8, Toulouse, France, 2001.

Aldezabal I. *Aditz azpikategorizazioaren azterketa: 100 aditzen azterketa zehatza, Levin (1993) oinarri harturik eta metodo automatikoak baliatuz.* Doktoretza-tesia, Euskal Herriko Unibertsitatea, 2004.

Aldezabal I., Aranzabe M., de Ilarraza A.D., Estarrona A., eta Uria L. Eus-PropBank: Integrating Semantic Information in the Basque Dependency Treebank. *Proceedings of the 11th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2010)*, 60–73, Iasi, Romania, 2010.

Baldewein U., Erk K., Pado S., eta Prescher D. Semantic Role Labeling With Chunk Sequences. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, 98–101, Boston, MA, USA, 2004.

Bishop C.M. *Neural Networks for Pattern Recognition.* Oxford University Press, Inc., New York, NY, USA, 1995.

Boas H.C. Bilingual framenet dictionaries for machine translation. *Proceedings of the Third International Conference on Language Resoucres and Evaluation (LREC)*, 1364–1371, Las Palmas de Gran Canaria, Spain, 2002.

Briscoe T. eta Carroll J. Automatic Extraction of Subcategorization from

Corpora. *Proceedings of the 5th ACL Conference on Applied Natural Language Processing (ANLP)*, 356–363, Washington, D.C., USA, 1997.

Brockmann C. eta Lapata M. Evaluating and combining approaches to selectional preference acquisition. *Proceedings of the 10th Conference of the European Chapter of the Association of Computational Linguistics (EACL-2003)*, 27–34, Budapest, Hungary, 2003.

Burchardt A., Reiter N., Thater S., eta Frank A. A Semantic Approach To Textual Entailment: System Evaluation and Task Analysis. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 10–15, Prague, 2007.

Carreras X. eta Màrquez L. Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, 89–97, Boston, MA, USA, 2004.

Carreras X. eta Màrquez L. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 152–164, Ann Arbor, MI, USA, 2005.

Carreras X., Màrquez L., eta Chrupała G. Hierarchical Recognition of Propositional Arguments with Perceptrons. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, 106–109, Boston, MA, USA, 2004.

Chen J. eta Rambow O. Use of deep linguistic features for the recognition and labeling of semantic arguments. *Proceedings of the 2003 conference on Empirical methods in natural language processing (EMNLP-2003)*, 41–48, 2003.

Cohn T. eta Blunsom P. Semantic Role Labelling with Tree Conditional Random Fields. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 169–172, Ann Arbor, MI, USA, 2005.

Collins M. Head-Driven Statistical Models for Natural Language Parsing. *Comput. Linguist.*, 29(4):589–637, 2003.

Connor M., Gertner Y., Fisher C., eta Roth D. Baby SRL: Modeling Early Language Acquisition. *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL-2008)*, 81–88, Manchester, England, 2008.

Connor M., Gertner Y., Fisher C., eta Roth D. Minimally Supervised Model of Early Language Acquisition. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, 84–92, Boulder, Colorado, 2009.

Copestake A. eta Flickinger D. An open-source grammar development environment and broad-coverage English grammar using HPSG. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, 591–600, Athens, Greece, 2000.

Cortes C. eta Vapnik V. Support-Vector Networks. *Machine Learning*, 20 (3):273–297, 1995.

Division S.C., Clark S., eta Weir D. Class-Based Probability Estimation using a Semantic Hierarchy. *Computational Linguistics*, 28(2):187–206, 2001.

Dowty D. Thematic proto-roles and argument selection. *Language*, 67(3): 547–619, 1991.

Erk K. A Simple, Similarity-based Model for Selectional Preferences. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL-2007)*, 216–223, Prague, Czech Republic, 2007.

Fillmore C. The case for case. In Bach E. eta Harms R.T., editors, *Universals in Linguistic Theory*, 1–88. Holt, Rinehart and Winston, 1968.

Fillmore C.J. Frame Semantics and the Nature of Language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Lnaguage and Speech*, 280(1):20–32, 1976.

Fillmore C.J., Ruppenhofer J., eta Baker C.F. FrameNet and Representing the Link between Semantic and Syntactic Relations. *Frontiers in Linguistics, volume I of Language and Linguistics Monograph Series B*, 19–59. Institute of Linguistics, Academia Sinica, Taipei, 2004.

Fleischman M. eta Hovy E. A maximum entropy approach to FrameNet tagging. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-2003)*, 22–24, Edmonton, Canada, 2003.

Frank A., Krieger H.U., Xu F., Uszkoreit H., Crysmann B., Jrg B., eta Schfer U. Question Answering from Structured Knowledge Resources. *Journal of Applied Logic*, 5(1):20–48, 2007.

Freund Y. eta Schapire R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *Proceedings of the Second European Conference on Computational Learning Theory (EuroColt-1995)*, 23–37, London, UK, 1995.

Fung P., Wu Z., Yang Y., eta Wu D. Learning Bilingual Semantic Frames: Shallow Semantic Parsing vs. Semantic Role Projection. *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, 75–84, Skovde, Sweden, 2007.

Gesmundo A., Henderson J., Merlo P., eta Titov I. A Latent Variable Model of Synchronous Syntactic-Semantic Parsing for Multiple Languages. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009): Shared Task*, 37–42, Boulder, Colorado, 2009.

Gildea D. eta Jurafsky D. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288, 2002.

Gildea D. eta Hockenmaier J. Identifying Semantic Roles Using Combinatory Categorial Grammar. *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, EMNLP-2003*, Sapporo, Japan, 2003.

Gildea D. eta Palmer M. The necessity of parsing for predicate argument recognition. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL-2002)*, 239–246, Philadelphia, Pennsylvania, 2002.

Giménez J. eta Màrquez L. Linguistic Features for Automatic Evaluation of Heterogenous MT Systems. *Proceedings of the Second Workshop on Statistical Machine Translation*, 256–264, Prague, Czech Republic, 2007.

Giménez J. eta Màrquez L. A Smorgasbord of Features for Automatic MT Evaluation. *Proceedings of the Third Workshop on Statistical Machine Translation*, 195–198, Columbus, Ohio, 2008.

Goldberg A.B., Zhu X., Dyer C.R., Eldawy M., eta Heng L. Easy as ABC? Facilitating Pictorial Communication via Semantically Enhanced Layout. *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL-2008)*, 119–126, Manchester, England, 2008.

Grimshaw J. *Argument Structure*. MIT Press, 1990.

Gruber J. *Studies in Lexical Relations.* Doktoretza-tesia, MIT, Cambridge, MA, 1965.

Habash N., Dorr B.J., eta Traum D. Hybrid Natural Language Generation from Lexical Conceptual Structures. *Machine Translation*, 18(2):81–128, 2003.

Hacioglu K., Pradhan S., Ward W., Martin J.H., eta Jurafsky D. Semantic Role Labeling by Tagging Syntactic Chunks. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, 110–113, Boston, MA, USA, 2004.

Hacioglu K. eta Ward W. Target word detection and semantic role chunking using support vector machines. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-2003)*, 25–27, Edmonton, Canada, 2003.

Haghighi A., Toutanova K., eta Manning C.D. A joint model for semantic role labeling. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 173–176, Ann Arbor, Michigan, 2005.

Hajič J., Ciaramita M., Johansson R., Kawahara D., Martí M.A., Màrquez L., Meyers A., Nivre J., Padó S., Štěpánek J., Straňák P., Surdeanu M., Xue N., eta Zhang Y. The CoNLL-2009 Shared Task: Syntactic and Semantic Dependencies in Multiple Languages. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009): Shared Task*, 1–18, Boulder, Colorado, 2009.

Higgins D. A transformation based approach to argument labeling. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, USA, 2004.

Hindle D. Noun classification from predicate-argument structures. *Proceedings of the 28th annual meeting on Association for Computational Linguistics (ACL-1990)*, 268–275, Pittsburgh, Pennsylvania, 1990.

Hirst G. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, 1987.

Jackendoff R. *Semantic Interpretation in Generative Grammar.* Cambridge University Press, 1972.

Johansson R. eta Nugues P. LTH: Semantic Structure Extraction using Nonprojective Dependency Trees. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 227–230, Prague, Czech Republic, 2007.

Kipper K., Dang H.T., eta Palmer M. Class Based Construction of a Verb Lexicon. *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, USA, 2000.

Koomen P., Punyakanok V., Roth D., eta tau Yih W. Generalized inference with multiple semantic role labeling systems. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005*, 181–184, 2005.

Kouchnir B. A Memory-Based Approach for Semantic Role Labeling. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, USA, 2004.

Levin B. *English Verb Classes and Alternations: A Preliminary Investigation.* The University of Chicago Press, Chicago, 1993.

Levin B. eta Rappaport Hovav M. *Argument Realization.* Cambridge University Press, Cambridge, 2005.

Li H. eta Abe N. Generalizing Case Frames Using a Thesaurus and the MDL Principle. *Computational Linguistics.*, 24(2):217–244, 1998.

Lim J.H., Hwang Y.S., Park S.Y., eta Rim H.C. Semantic Role Labeling Using Maximum Entropy model. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, USA, 2004.

Lin C.S.A. eta Smith T.C. Semantic Role Labeling via Consensus in Pattern-Matching. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 185–188, Ann Arbor, Michigan, 2005.

Lin D. Automatic Retrieval and Clustering of Similar Words. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL-1998)*, 768–774, Montreal, Quebec, Canada, 1998.

Litkowski K.C. Senseval-3 task: Automatic Labeling of Semantic Roles. *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, 9–12, Barcelona, Spain, 2004.

Litkowski K. eta Hargraves O. Coverage and inheritance in the preposition project. *Prepositions '06: Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, 37–44, Trento, Italy, 2006.

Loper E., Yi S., eta Palmer M. Combining Lexical Resources: Mapping Between PropBank and VerbNet. *Proceedings of the 7th International Workshop on Computational Semantics*, Tilburg, The Netherlands, 2007a.

Loper E., Yi S.T., eta Palmer M. Combining Lexical Resources: Mapping between PropBank and VerbNet. *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands, 2007b.

M. Connor C.F. Y. Gertner eta Roth D. Starting from Scratch in Semantic Role Labeling. *Proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL-2010)*, Uppsala, Sweden, 2010.

Marcus M., Kim G., Marcinkiewicz M.A., MacIntyre R., Bies A., Ferguson M., Katz K., eta Schasberger B. The Penn Treebank: annotating predicate argument structure. *Proceedings of the workshop on Human Language Technology (HLT-94)*, 114–119, 1994.

Màrquez L., Comas P.R., Giménez J., eta Català N. Semantic Role Labeling as Sequential Tagging. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 193–196, Ann Arbor, MI, USA, 2005.

Màrquez L., Carreras X., Litkowski K.C., eta Stevenson S. Semantic Role Labeling: an Introduction to the Special Issue. *Computational Linguistics*, 34(2):145–159, 2008.

Màrquez L., Villarejo L., Martí M.A., eta Taulé M. SemEval-2007 task 09: multilevel semantic annotation of Catalan and Spanish. *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*, 42–47, Prague, Czech Republic, 2007.

McCarthy D. Using Semantic Preferences to Identify Verbal Participation in Role Switching Alternations. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (NAACL-2000)*, 256–263, Seattle, Washington, 2000.

Melli G., Wang Y., Liu Y., Kashani M.M., Shi Z., Gu B., Sarkar A., eta Popowich F. Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 Summarization Task. *Proceedings of the HLT/EMNLP Document Understanding Workshop (DUC)*, Vancouver, B.C., Canada, 2005.

Merlo P. eta Stevenson S. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3): 373–408, 2001.

Merlo P. eta Musillo G. Semantic Parsing for High-Precision Semantic Role Labelling. *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL-2008)*, 1–8, Manchester, United Kingdom, 2008.

Merlo P. eta Van Der Plas L. Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP-2009)*, 288–296, Suntec, Singapore, 2009.

Meyers A., Reeves R., Macleod C., Szekely R., Zielinska V., Young B., eta Grishman R. The NomBank Project: An Interim Report. *Proceedings of the HLT-NAACL-2004. Workshop: Frontiers in Corpus Annotation*, 24–31, Boston, MA, USA, 2004.

Musillo G. eta Merlo P. Accurate Parsing of the Proposition Bank. *Proceedings of the Human Language Technology: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, 101–104, New York City, NY, USA, 2006.

Narayanan S. eta Harabagiu S. Question Answering based on Semantic Structures. *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, 693–701, Geneva, Switzerland, 2004.

Noreen E.W. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons, 1989.

Padó S. eta Lapata M. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

Padó S., Padó U., eta Erk K. Flexible, Corpus-Based Modelling of Human Plausibility Judgements. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL-2007)*, 400–409, Prague, Czech Republic, 2007.

Palmer M., Gildea D., eta Kingsbury P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–105, 2005.

Pantel P. eta Lin D. An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words. *Proceedings of the 38th Annual Conference of the Association of Computational Linguistics (ACL-2000)*, 101–108, Hong Kong, China, 2000.

Park K.M., mi Park K., sook Hwang Y., eta chang Rim H. Two-Phase Semantic Role Labeling based on Support Vector Machines. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, USA, 2004.

Pradhan S., Hacioglu K., Krugler V., Ward W., Martin J., eta Jurafsky D. Support Vector Learningfor Semantic Argument Classification. *Machine Learning*, 60(1):11–39, 2005a.

Pradhan S., Loper E., Dligach D., eta Palmer M. SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 87–92, Prague, Czech Republic, 2007a.

Pradhan S., Ward W., eta Martin J.H. Towards Robust Semantic Role Labeling. *Computational Linguistics*, 34(2), 2008.

Pradhan S., Hacioglu K., Krugler V., Ward W., Martin J.H., eta Jurafsky D. Support Vector Learning for Semantic Argument Classification. *Machine Learning*, 60(1-3):11–39, 2005b.

Pradhan S., Hacioglu K., Ward W., Martin J.H., eta Jurafsky D. Semantic Role Parsing: Adding Semantic Structure to Unstructured Text. *Proceedings of the Third IEEE International Conference on Data Mining (ICDM-2003)*, page 629, Washington, DC, USA, 2003.

Pradhan S., Hacioglu K., Ward W., Martin J.H., eta Jurafsky D. Semantic role chunking combining complementary syntactic views. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 217–220, Ann Arbor, Michigan, 2005c.

Pradhan S., Loper E., Dligach D., eta Palmer M. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 87–92, Prague, Czech Republic, 2007b.

Punyakanok V., Roth D., eta Yih W. The Importance of Syntactic Parsing and Inference in Semantic Role Labeling. *Computational Linguistics*, 34 (2):257–287, 2008.

Punyakanok V., Roth D., tau Yih W., Zimak D., eta Tu Y. Semantic role labeling via generalized inference over classifiers. *Proceedings of the 8th Conference on Natural Language Learning (CoNLL-2004)*, 130–133, Boston, MA, USA, 2004a.

Punyakanok V., Roth D., Yih W.t., eta Zimak D. Semantic role labeling via integer linear programming inference. *Proceedings of the 20th international conference on Computational Linguistics (COLING-2004)*, page 1346, Geneva, Switzerland, 2004b.

Pustejovsky J. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995.

Resnik P. Semantic classes and syntactic ambiguity. *Proceedings of the workshop on Human Language Technology*, 278–283, Morristown, NJ, USA, 1993.

Schulte im Walde S. Clustering Verbs Semantically According to their Alternation Behaviour. *Proceedings of the 18th International Conference on Computational Linguistics (COLING-2000)*, 747–753, Saarbrücken, Germany, 2000.

Shi L. eta Mihalcea R. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. *Proceedings of the Sixth International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2005)*, 100–111, Mexico City, Mexico, 2005.

Sun W., Sui Z., Wang M., eta Wang X. Chinese semantic role labeling with shallow parsing. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, 1475–1483, Singapore, 2009.

Surdeanu M., Harabagiu S., Williams J., eta Aarseth P. Using Predicate-Argument Structures for Information Extraction. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, 8–15, Sapporo, Japan, 2003.

Surdeanu M., Johansson R., Meyers A., Màrquez L., eta Nivre J. The CoNLL-2008 Shared Task on Joint Parsing of Syntactic and Semantic Dependencies. *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL-2008)*, 159–177, Manchester, England, 2008.

Surdeanu M., Màrquez L., Carreras X., eta Comas P.R. Combination Strategies for Semantic Role Labeling. *Journal of Artificial Intelligence Research (JAIR)*, 29:105–151, 2007.

Surdeanu M. eta Turmo J. Semantic Role Labeling Using Complete Syntactic Analysis. *Proceedings of the 9th International Conference on Computational Natural Language Learning (CoNLL-2005)*, Ann Arbor, Michigan, USA, 2005.

Swier R. eta Stevenson S. Exploiting a Verb Lexicon in Automatic Semantic Role Labelling. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-2005)*, 883–890, Vancouver, B.C., Canada, 2005.

Tatu M. eta Moldovan D. A Semantic Approach to Recognizing Textual Entailment. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP-2005)*, 371–378, Vancouver, British Columbia, Canada, 2005.

Thompson C.A., Levy R., eta Manning C. A Generative Model for Semantic Role Labeling. *Proceedings of 14th European Conference on Machine Learning (ECML-2003)*, 397–408, Dubrovnik, Croatia, 2003.

Toutanova K., Haghighi A., eta Manning C. A Global Joint Model for Semantic Role Labeling. *Computational Linguistics*, 34(2):161–191, 2008a.

Toutanova K., Haghighi A., eta Manning C.D. Joint learning improves semantic role labeling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL-2005)*, 589–596, Ann Arbor, Michigan, 2005.

Toutanova K., Haghighi A., eta Manning C.D. A global joint model for semantic role labeling. *Computational Linguistics*, 34(2):161–191, 2008b.

van den Bosch A., Canisius S., Canisius E., eta Hendrickx I. Memory-Based Semantic Role Labeling: Optimizing features, algorithm, and output. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, USA, 2004.

Vapnik V. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

Vossen P., Peters W., eta Dez-orzas P. The Multilingual design of the EuroWordNet Database. *Proceedings of the International Joint Conference on*

Artificial intelligence (IJCAI-97). Workshop Multilingual Ontologies for NLP Applications, 23–29, 1997.

Williams K., Dozier C., eta McCulloh A. Learning transformation rules for semantic role labeling. *Proceedings of the Eigth Conference on Computational Natural Language Learning (CoNLL-2004)*, Boston, MA, USA, 2004.

Wu D. eta Fung P. Can Semantic Role Labeling Improve SMT? *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT 2009)*, 218–225, Barcelona, Spain, 2009a.

Wu D. eta Fung P. Semantic Roles for SMT A Hybrid Two-Pass Model. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (ACL-2009)*, 13–16, Boulder, Colorado, 2009b.

Xue N. eta Palmer M. Calibrating Features for Semantic Role Labeling. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, 88–94, Barcelona, Spain, 2004.

Xue N. eta Palmer M. Automatic Semantic Role Labeling for Chinese verbs. *Proceedings of the 19th International Joint Conference on Artificial intelligence (IJCAI-2005)*, 1160–1165, Edinburgh, Scotland, 2005.

Yi S.T., Loper E., eta Palmer M. Can Semantic Roles Generalize Across Genres? *Proceedings of the Human Language Technology Conferences/North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2007)*, Rochester, NY, USA, 2007.

Yih W. eta Toutanova K. Automatic Semantic Role Labeling. *Tutorial of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL-2006)*, New York, NY, USA, 2006.

Zapirain B., Agirre E., eta Màrquez L. UBC-UPC: Sequential SRL Using Selectional Preferences. An approach with Maximum Entropy Markov Models. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 354–357, Prague, Czech Republic, 2007.

Zapirain B., Agirre E., eta Màrquez L. A Preliminary Study on the Robustness and Generalization of Role Sets for Semantic Role Labeling. *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2008)*, 219–230, Haifa, Israel, 2008a.

Zapirain B., Agirre E., eta Màrquez L. Robustness and Generalization of Role Sets: PropBank vs. VerbNet. *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL-08)*, 550–558, Columbus, Ohio, 2008b.

Zapirain B., Agirre E., eta Màrquez L. Generalizing over lexical features: selectional preferences for semantic role classification. *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP-2009)*, 73–76, Suntec, Singapore, 2009.

Zapirain B., Agirre E., Màrquez L., eta Surdeanu M. Improving Semantic Role Classification with Selectional Preferences. *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010), Los Angeles, California.*, 2010.