Universitat Oberta de Catalunya

DIGITHUM
Les humanitats en l'era digital

http://digithum.uoc.edu

Etiquetatge automàtic de rols semàntics amb un sistema d'aprenentatge basat en memòria*

Roser Morante
Investigadora al grup de recerca CNTS de la Universitat d'Anvers
Roser.Morante@ua.ac.be

Data de presentació: gener de 2008
Data d'acceptació: febrer de 2008
Data de publicació: maig de 2008

Citació recomanada:

MORANTE, Roser (2008). "Etiquetatge automàtic de rols semàntics amb un sistema d'aprenentatge basat en memòria". *Digithum*, núm. 10 [article en línia]. DOI: http://dx.doi.org/10.7238/d.v0i10.504

Resum

En aquest article presentem un sistema d'etiquetatge automàtic de rols semàntics, el principal component del qual és un classificador basat en memòria. El sistema s'ha entrenat amb el corpus Cast3LB-CoNLL-SemRol. Els atributs codifiquen informació de sintaxi de dependències. Els resultats obtinguts (F_1 0,86) són comparables amb els dels sistemes existents (F_1 entorn de 0,86), que utilitzen informació de sintaxi de constituents.

Paraules clau

etiquetatge automàtic de rols semàntics, aprenentatge basat en memòria, TiMBL

Abstract

In this paper we present a semantic role labelling system. The main component of the system is a memory-based classifier. The system has been trained with the Cast3LB-CoNLL-SemRol. The features encode information from dependency syntax. The results (F_1 0.86) are comparable with state-of-the-art results (F_1 around 0.86) from systems that use information from constituent syntax.

Keywords

semantic role labelling, memory-based learning, TiMBL

* Aquest treball ha estat possible gràcies a l'ajut postdoctoral EX2005-1145 del Ministeri d'Educació i Ciència atorgat al projecte Técnicas semiautomáticas para el etiquetado de roles semánticos en corpus del español.

Introducció

En aquest article presentem un sistema d'etiquetatge automàtic de rols semàntics (EARS). L'EARS és una tasca de processament de llenguatge natural (PLN) que consisteix a identificar els arguments dels verbs d'una frase i assignar-los un rol semàntic. Els rols semàntics són entitats simbòliques que descriuen conceptualment la funció que tenen els participants en un esdeveniment des del punt de vista de la situació del món real. La noció de *rol semàntic* és deguda originalment al treball del lingüista Charles Fillmore (1968). Per exemple, a la frase (1)[1] el verb és *ha obtingut, La UOC* és l'entitat que ha obtingut alguna cosa i *el Premio Nacional Empresa Flexible 2007* és la cosa obtinguda. Tant *La UOC* com *el Premio Nacional Empresa Flexible 2007* són participants en l'esdeveniment *obtenir* i reben un rol semàntic. Un predicat amb un determinat significat assigna determinats rols als participants en l'esdeveniment que denota, independentment de la forma sintàctica de la frase. Encara que la mateixa situació del món real s'expressi amb una altra estructura sintàctica, com ara la construcció passiva de la frase (2), els rols semàntics assignats són els mateixos.

(1) [La UOC] ha obtingut [el Premio Nacional Empresa Flexible 2007].

(2) [El Premio Nacional Empresa Flexible 2007] ha estat obtingut [per la UOC].

El sistema que presentem etiqueta rols semàntics de manera automàtica. Des del punt de vista del PLN, un sistema d'aquest tipus pot ser útil per a integrar-lo en sistemes d'extracció d'informació, de pregunta-resposta o de resum automàtic de textos, ja que permet saber qui fa què, a qui, quan, on, amb quin propòsit i en quines circumstàncies.

Per a l'anglès, els resultats de referència en EARS són els aconseguits pels sistemes participants en les competicions CoNLL Shared Task 2004 (Màrquez *et al.*, 2004) i 2005 (Carreras *et al.*, 2005). En aquests casos el corpus que s'utilitza és el PropBank (Palmer *et al.*, 2005). Per a una frase etiquetada amb constituents superficials com la frase (3), l'anotació que es fa a PropBank és la que es mostra a (4):

(3) He wouldn't accept anything of value from those he was writing about. ('No acceptaria res de valor d'aquells sobre els qui escrivia.')

(4) [$_{A0}$ He] [$_{AM\text{-}MOD}$ would] [$_{AM\text{-}NEG}$ n't] [$_{V}$ **accept**] [$_{A1}$ anything of value] from [$_{A2}$ those he was writing about].

Segons el lexicó de PropBank, els rols que assigna el verb *accept* ('acceptar') són els següents:

(5) **V:** verb
A0: acceptor ('qui accepta')
A1: thing accepted ('el que és acceptat')
A2: accepted-from ('acceptat de')
A3: attribute ('atribut')
AM-MOD: modal ('modal')
AM-NEG: negation ('negació')

Per al català i el castellà, els resultats de referència són els dels sistemes participants en la competició SemEval 2007, en concret la tasca 9 Multilevel Semantic Annotation of Catalan and Spanish (Màrquez *et al.*, 2007b).

El sistema que presentem en aquest article s'ha desenvolupat per a l'espanyol, fent servir el corpus Cast3LB-CoNLL-SemRol (Morante, 2006). La diferència fonamental entre aquest sistema i els sistemes esmentats anteriorment és que es basa en sintaxi de dependències, en comptes de basar-se en sintaxi de constituents. La frase (1) es representaria en sintaxi de constituents com es mostra a (6), i en sintaxi de dependències, com es mostra a (7):

(6)

Paraula	Constituent
La	(S (sn-SUJ (espec.fs*)
UOC	(grup.nom.fs))
ha	(gv*
obtingut	*)
el	(sn-CD (espec.ms*)
Premio Nacional Empresa Flexible 2007	(grup.nom.ms)))

(7)

Núm.	Paraula	Depèn de	Funció
1	La	2	ESP
2	UOC	4	SUJ
3	ha	4	AUX
4	obtingut	0	ROOT
5	el	6	ESP
6	Premio Nacional Empresa Flexible 2007	4	CD

1. Obtinguda a la pàgina web de la UOC (www.uoc.edu) consultada el 06-12-07.

La característica bàsica del sistema és el mètode d'aprenentatge automàtic aplicat. Es tracta d'un mètode basat en memòria (*Memory-based learning*) (Daelemans *et al.*, 2005), desenvolupat pels grups d'investigació ILK,[www1] de la Universitat de Tilburg, i CNTS,[www2] de la Universitat d'Anvers.

L'article s'organitza de la manera següent: a la secció 1 presentem el corpus Cast3LB-CoNLL-SemRol, amb el qual hem entrenat el sistema; a la secció 2 introduïm el mètode d'aprenentatge basat en memòria; a la secció 3 descrivim el sistema d'EARS, els resultats del qual apareixen a la secció 4; i finalment, a la secció 5 apuntem algunes conclusions.

1. Descripció del corpus

El corpus Cast3LB-CoNLL-SemRol és una versió revisada i augmentada del corpus Cast3LB (Civit, 2003) de l'espanyol utilitzat en la competició CoNLL Shared Task 2006[www3] (Buchholz *et al.*, 2006), dedicada a l'anàlisi sintàctica de dependències. La revisió del corpus ha consistit a corregir errors[2] causats per la conversió del corpus original en format de constituents al format CoNLL de dependències, i a fer algunes adaptacions sintàctiques per a tractar fenòmens com ara la coordinació i la subordinació. Addicionalment, el corpus s'ha anotat amb els rols semàntics següents: ARG0, ARG1, ARGM, Attributive, Benefactive, Cause, Company, Concessive, Condition, Consequence, Destination, Extent, Instrument, Location, Manner, Means, Opposition, Origin, Predicative, Purpose, Quantity, Result, Source, State, Temporal i Topic. A Morante (2006) es poden trobar detalls sobre les anotacions sintàctiques i semàntiques.

El corpus conté 89.199 paraules en 3.303 frases, de les quals 11.023 són formes verbals corresponents a 1.443 lemes verbals. El format és el mateix que l'utilitzat en la competició CoNLL Shared Task 2006: cada paraula d'una frase es representa en una línia i les frases se separen amb una línia en blanc. Cada paraula es representa en nou camps que contenen informació morfològica, de dependències i de rols. A la taula 1 es mostra la representació de la frase (8):

(8) Asimismo defiende la financiación pública de la investigación básica y pone de manifiesto que las empresas se centran más en la I+D con objetivos de mercado.

La columna 1 conté la posició de la paraula a la frase; la columna 2, la paraula; la columna 3, el lema; la columna 4, la categoria sintàctica; la columna 5, el tipus de categoria sintàctica; la columna 6, els trets morfològics (nombre, persona, temps, mode, etc.); la columna 7, la posició de la paraula de la qual la paraula en qüestió depèn sintàcticament; la columna 8, la funció sintàctica; i la columna 9, el rol semàntic.

Cal indicar que les anotacions de les dades en recursos creats amb la finalitat d'entrenar sistemes computacionals no sempre reflecteixen anàlisis lingüístiques correctes. Per exemple, a la taula 1, *de manifiesto* s'analitza com a complement circumstancial del verb *poner*, perquè en aquest corpus no s'ha pres en consideració l'anotació d'expressions multiparaula. Un altre exemple seria l'assignació de la funció de complement directe al verb *centrarse*. Aquest tipus de fenòmens són habituals en processament del llenguatge natural quan es treballa a gran escala amb exemples reals.

2. Aprenentatge basat en memòria

Els mètodes d'aprenentatge automàtic (*machine learning*) faciliten l'adquisició automàtica de coneixement a partir de dades. L'aprenentatge basat en memòria, *memory-based learning* (MBL) (Daelemans *et al.*, 2005), és un mètode simbòlic d'aprenentatge automàtic que es basa en la idea que el comportament intel·ligent pot ser el resultat d'establir analogies, en comptes de ser el resultat d'aplicar un conjunt de regles abstractes. En aquest sentit contrasta amb el processament basat en regles. La hipòtesi principal del MBL és que a partir de representacions d'experiències anteriors emmagatzemades a la memòria es pot fer una extrapolació a noves situacions. Els dos principis generals d'aquest mètode són:

(i) L'aprenentatge consisteix a emmagatzemar representacions d'experiències a la memòria.

(ii) Per a resoldre un problema nou es fan servir solucions de problemes similars vists anteriorment.

Aquest mètode s'ha aplicat a diverses tasques de PLN i s'ha comprovat que és un mètode adequat, perquè en descriure problemes de PLN només es poden establir algunes generalitzacions, ja que en el llenguatge natural les irregularitats i les excepcions són molt freqüents.

Les tasques de PLN es poden solucionar emmagatzemant primer en memòria exemples anotats del problema en qüestió i aplicant després un raonament basat en mesures de similitud per tal de solucionar nous exemples. Els algoritmes MBL, que són descendents de l'algoritme *k-nn* (Cover *et al.*, 1967), prenen un conjunt d'exemples com a *input* i produeixen un classificador que pot classificar nous exemples que no s'havien vist anteriorment.

2. Per exemple, un error de conversió consistia a posar el nucli del sintagma nominal com a fill de l'adjectiu pronominal que el modifica.

[www1]: <http://ilk.uvt.nl>.

[www2]: <http://www.cnts.ua.ac.be/cnts/>.

[www3]: <http://nextens.uvt.nl/~conll/>.

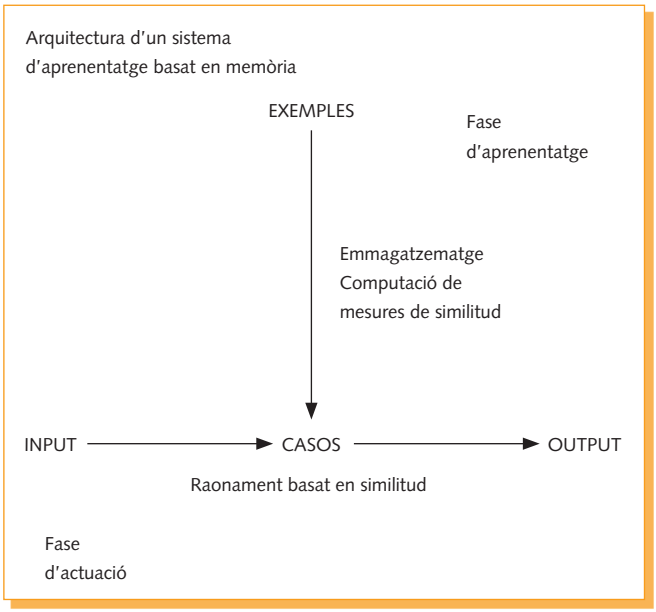
Taula 1. Exemple d'una frase del corpus Cast3LB-CoNLL-SemRol

Núm.	Paraula	Lema	Cat.	Tipus cat.	Trets morfològics	Dep.	Funció	Rol semàntic
1	Asimismo	asimismo	r	rg	_	2	MOD	_
2	defiende	defender	v	vm	nre=s|per=3| mod=i|tmp=p	0	ROOT	_
3	la	el	d	da	nre=s|gèn=f	4	ESP	_
4	financiación	financiación	n	nc	nre=s|gèn=f	2	CD	ARG1
5	pública	pública	a	aq	nre=s|gèn=f	4	CN	_
6	de	de	s	sp	for=s	4	CN	_
7	la	el	d	da	nre=s|gèn=f	8	ESP	_
8	investigación	investigación	n	nc	nre=s|gèn=f	6	_	_
9	básica	básico	a	aq	nre=s|gèn=f	8	CN	_
10	y	y	c	cc	_	2	CTE	_
11	pone	poner	v	vm	nre=s|per=3| mod=i|tmp=p	10	CDO	_
12	de	e	s	sp	for=s	11	CC	ARG_ST
13	manifiesto	manifiesto	n	nc	gèn=m|nre=s	12	_	_
14	que	que	c	cs	_	18	_	_
15	las	el	d	da	gèn=f|nre=p	16	ESP	_
16	empresas	empresa	n	nc	gèn=f|nre=p	18	SUJ	ARG1
17	se	él	p	p0	per=3	18	_	_
18	centran	centrar	v	vm	nre=p|per=3| mod=i|tmp=p	11	CD	ARG1
19	más	más	r	rg	_	20	_	_
20	en	en	s	sp	for=s	18	CREG	ARG_LOC
21	la	el	d	da	nre=s|gèn=f	22	ESP	_
22	I+D	I+D	n	np	_	20	_	_
23	con	con	s	sp	for=s	18	CC	ARG_PRP
24	objetivos	objetivo	n	nc	gèn=m|nre=p	23	_	_
25	de	de	s	sp	for=s	24	CN	_
26	mercado	mercado	n	nc	gèn=m|nre=s	25	_	_
27	.	.	F	Fp	_	2	PUNC	_

A la figura 2 reproduïm la representació d'un sistema MBL de Daelemans *et al.* (2007). Té dos components: el d'aprenentatge, basat en memòria, i el d'actuació (*performance*), que aplica mesures de similitud. L'aprenentatge consisteix a afegir nous exemples a la memòria per tal d'entrenar el sistema. Del conjunt d'exemples que s'ha fet servir en aquesta fase en diem *corpus d'aprenentatge*. Els algoritmes MBL no abstrauen ni reestructuren les dades, raó per la qual s'anomenen *lazy learners* ('aprenents ganduls'). Aquests aprenents contrasten amb els *eager learners* ('aprenents ansiosos'), que tracten d'abstraure teories a partir de les dades, ignorant les excepcions i els casos infreqüents.

En la fase d'actuació s'utilitza el producte del component d'aprenentatge per a fer projeccions d'*input* a *output*. La majoria de tasques de PLN es poden plantejar com una projecció

Figura 1. Arquitectura general d'un sistema de MBL (Daelemans *et al*., 2007, pàg. 19)



complexa entre diferents tipus de representacions. Per exemple, a la tasca d'etiquetatge de rols es fa una projecció d'una representació morfosintàctica a una representació semàntica. A la taula 1 s'observa que les columnes 1-8 contenen la representació d'*input,* i la columna 9, la representació d'*output*.

Generalment les projeccions d'*input* a *output* consisteixen a fer una classificació. Durant la classificació el sistema rep un exemple no vist anteriorment i mesura la similitud entre el nou exemple *X* i els exemples *Y* emmagatzemats en memòria fent servir una de les mesures de similitud disponibles. L'extrapolació es fa assignant al nou exemple la categoria més freqüent dintre del conjunt d'exemples més similars que el sistema ha trobat. Els exemples més similars reben el nom de *k-nearest neighbors* ('nombre *k* de veïns més propers').

En aquest tipus de sistemes un exemple, o *instància*, és un vector de *n* parells d'atribut-valor, més un camp que conté la *classe* d'aquest vector.

Segons Daelemans *et al.* (2005), l'aprenentatge automàtic és fonamentalment un paradigma de classificació: si tenim una representació d'un *input* en termes de parells atribut-valor, és a dir, un vector d'atributs, el sistema ha de generar una etiqueta de classe. Quan l'aprenentatge és supervisat aquesta etiqueta es pren d'un conjunt d'etiquetes conegut *a priori*. Si un algoritme s'entrena amb un nombre suficient d'*exemples*, és a dir, vectors d'atributs amb la seva classe, aquest algoritme pot induir un classificador, que farà la projecció de vectors d'atributs a classes.

El sistema que presentem actua d'aquesta manera. Les classes són els rols semàntics esmentats a la secció 1. Amb la informació disponible a les columnes 1-8 (taula 1) es creen els vectors d'atributs per a representar el que seran els exemples del corpus d'entrenament i de test. Un possible vector per a representar el sintagma *la financiación pública de la investigación básica* de la taula 1 seria el que es mostra a (9), que conté 31 atributs:

(9)

núm. atribut	1	2	3	4	5	6	7	8
nom atribut	**nucli**	**n.lema**	**n.cat**	**n.gèn**	**n.nom**	**n.nom.propi**	**n.cat-2**	**n.cat-1**
valor atribut	financiación	financiación	n	f	s	f	v	d

núm. atribut	9	10	11	12	13	14	15
nom atribut	**n.cat+1**	**n.cat+2**	**s.últim.lema**	**s.última.cat**	**s.lema1**	**s.cat1**	**s.lema2**
valor atribut	a	s	básico	aq	el	da	financiación

núm. atribut	16	17	18	19	20	21	22	23
nom atribut	**s.cat 2**	**s.lema3**	**s.cat3**	**s.lema4**	**s.cat 4**	**s.lema5**	**s.cat 5**	**s.funció**
valor atribut	nc	público	aq	de	sp	el	da	CD

núm. atribut	24	25	26	27	28	29
nom atribut	**s.posició.rel**	**s.categories**	**verb**	**verb lema**	**verb cat**	**verb pers**
valor atribut	post	d-n-a-s-d-n-a	defiende	defender	vm	3

núm. atribut	30	31
nom atribut	**verb nombre**	**nombre fills verb**
valor atribut	s	3

Els atributs 1-10 contenen informació relativa al nucli del sintagma, és a dir, la paraula *financiación*: forma, lema, categoria, gènere, nombre, atribut boleà que indica si el nucli és nom propi, i categoria de les dues paraules anteriors i les dues paraules següents.

Els atributs 11-25 contenen informació sobre el sintagma: lema i categoria de l'última paraula, lema i categoria de les paraules 1-5, funció sintàctica, posició relativa del sintagma en relació al verb, i cadena de caràcters amb totes les categories de les paraules que formen el sintagma.

Els atributs 26-31 contenen informació relativa al verb del qual depèn el sintagma: forma verbal, categoria, persona, nombre, i nombre de fills que té el verb.

Per a construir els corpus d'entrenament i de test, tots els exemples s'han de convertir en vectors d'atributs-valor amb exactament el mateix nombre i tipus d'atributs. Addicionalment, al corpus d'entrenament el vector d'atributs va seguit per la classe de cada vector. En el cas del sintagma *la financiación pública de la investigación básica* la classe és *ARG1*.

Per a fer els experiments que presentem hem utilitzat TiMBL (Daelemans *et al.*, 2007),[3] un programa que implementa diversos algoritmes basats en memòria.

3. Descripció del sistema d'etiquetatge automàtic de rols semàntics

El component principal del sistema és un classificador basat en memòria. La tasca d'etiquetatge es planteja en tres fases:

- Fase 1 de preprocessament, que consisteix a identificar els candidats que podrien tenir assignat un rol semàntic. El sistema busca un verb i els límits de la frase per tal de trobar els sintagmes que depenen del verb. Cada sintagma es converteix en un exemple. Per a l'oració de (8), els candidats obtinguts per a cada verb són els següents:

Candidat	Verb
asimismo	defiende
financiación	defiende
manifiesto	pone
centran	pone
empresas	centran
se	centran
más	centran
I+D	centran
objetivos	centran

- Fase 2 de classificació, que consisteix a assignar un rol semàntic fent servir un algoritme. L'algoritme utilitzat és el classificador IB1, en la seva implementació a TiMBL, versió 6 (Daelemans *et al.*, 2007). Es tracta d'un algoritme supervisat inductiu dissenyat per a aprendre tasques de classificació. Es basa en l'algoritme de classificació *k-nn*. En l'algoritme IB1 la similitud es defineix en termes de la distància per atribut entre un exemple test i els exemples memoritzats. Els paràmetres de l'IB1 utilitzats han estat la mesura de similitud Jeffrey Divergence, assignació de pes als atributs usant la mesura Gain Ratio, 11 *k-nearest neighbors*, i distància lineal inversa per a pesar el vot de classe dels *k-nearest neighbors*. Es poden trobar més detalls sobre aquests paràmetres a Daelemans *et al.* (2007).
- Fase 3 de postprocessament, en què es corregeixen algunes assignacions de rols per a cada sintagma, és a dir, algunes de les prediccions del sistema, tenint en compte totes les prediccions que el classificador ha fet per a la frase. Per exemple, si el sistema ha predit dos subjectes a la mateixa frase, una de les prediccions es modifica.

El sistema s'ha desenvolupat fent experiments de *10-fold crossvalidation*. La selecció dels atributs s'ha fet començant amb un conjunt bàsic i afegint-n'hi gradualment de nous. Els atributs per a cada exemple són els següents:

- Relatius al sintagma: forma, categoria, lema, gènere i nombre del nucli; tres atributs boleans que indiquen si el nucli és un nom propi, una partícula temporal i una partícula locativa; categoria de les dues paraules anteriors i de les tres següents al nucli; categoria i lema de les tres primeres paraules del sintagma; preposició, si n'hi ha; funció, posició relativa al verb (pre, post), posició relativa al verb del sintagma següent, cadena de les categories de totes les paraules, cadena de paraules (noms, adjectius, adverbis, verbs).
- Relatius al verb: distància al sintagma, forma verbal, lema, categoria, dues paraules anteriors i posteriors, concordança de gènere amb el sintagma, tres atributs boleans que indiquen si el verb és causatiu, pronominal o passiu.
- Relatius a la frase: nombre de sintagmes amb funció de complement circumstancial, nombre de constituents, posició relativa al verb dels constituents amb determinades funcions sintàctiques, cadenes de funcions dels sintagmes.

Cada exemple tindria un format similar al que es mostra a (9), amb un total de 49 atributs.

3. TiMBL es pot descarregar a <http://ilk.uvt.nl/timbl>.

4. Resultats

Les mesures d'avaluació utilitzades són les mesures estàndard *precision*, *recall* i F_1, que es calculen a partir dels paràmetres següents:

tp: nombre de casos detectats correctament
fp: nombre de casos detectats incorrectament
fn: nombre de casos no detectats

Recall mesura el percentatge de detecció de casos segons la fórmula següent: tp / (tp + fn); *precision* mesura com és d'acurat el sistema segons la fórmula tp / (tp + fp); finalment, F_1 es calcula amb la fórmula (2 * prec * rec) / (prec + rec).

El sistema obté resultats de 0,86 F_1, 0,88 *precision* i 0,84 *recall*. A la taula 2 es mostren els resultats del sistema avaluat sobre el corpus del test. S'observen grans diferències entre rols. Els rols que apareixen més de 100 vegades (5,06%) obtenen resultats per sobre del 0,80. Dos rols que apareixen més de 100 vegades també obtenen resultats de més de 0,80: ARG_PRED, predicatiu, i ARG_BEN, beneficiari. Això és degut al fet que els dos rols tenen

Taula 2. Resultats del sistema per rol semàntic

Rol semàntic	Total	Correctes	Falsos negatius	Falsos positius	*Precision*	*Recall*	F_1
ARG_RES	1	0	1	0	0,00	0,00	0,00
ARG_MEANS	2	0	2	0	0,00	0,00	0,00
ARG_SRC	2	1	1	1	0,50	0,50	0,50
ARG_OP	4	1	3	1	0,50	0,25	0,33
ARG_ST	5	1	4	0	1,00	0,20	0,33
ARG_CONS	6	1	5	0	1,00	0,17	0,28
ARG_CONC	9	5	4	1	0,83	0,55	0,66
ARG_COND	9	5	4	0	1,00	0,55	0,71
ARG_COMP	14	10	4	2	0,83	0,71	0,77
ARG_INSTR	14	4	10	1	0,80	0,28	0,42
ARG_OR	21	11	10	1	0,92	0,52	0,67
ARG_EXT	22	12	10	1	0,92	0,54	0,68
ARG_TOP	23	5	18	3	0,63	0,22	0,32
ARG_PRP	33	23	10	6	0,79	0,69	0,74
ARG_CAU	34	29	5	14	0,67	0,85	0,75
ARG_DEST	36	24	12	16	0,60	0,66	0,63
ARG_PRED	45	41	4	4	0,91	0,91	0,91
ARGM	53	12	41	12	0,50	0,23	0,31
ARG_BEN	78	69	9	11	0,86	0,88	0,87
ARG_MNR	88	55	33	30	0,65	0,62	0,63
ARG_LOC	124	97	27	10	0,91	0,78	0,84
ARG_ATR	140	140	0	2	0,98	1,00	0,99
ARG_TMP	192	166	26	40	0,80	0,86	0,83
ARG0	285	265	20	27	0,91	0,93	0,92
ARG1	735	683	52	40	0,94	0,93	0,94
Global	**1.975**	**1.660**	**315**	**223**	**0,88**	**0,84**	**0,86**

marques que els caracteritzen. El rol beneficiari sol anar introduït per la preposició *a* i el predicatiu concorda amb gènere i nombre amb subjecte o objecte directe.

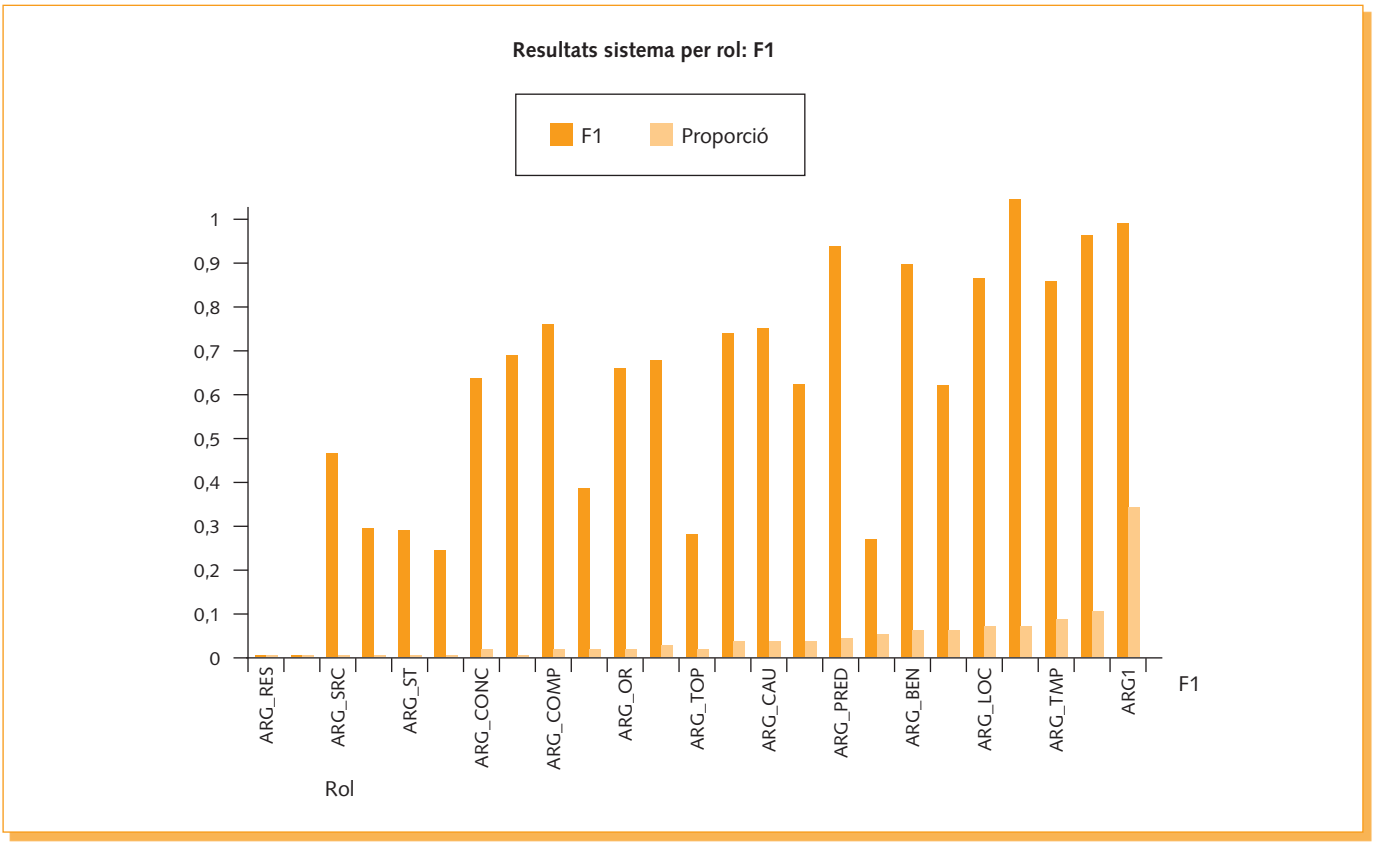
Els resultats globals són comparables als resultats de referència actuals (Màrquez *et al.*, 2007a; Morante *et al.*, 2007; Surdeanu *et al*., 2008) que són entorn del 0,86. Ara bé, el sistema que presenten utilitza informació de sintaxi de dependències, mentre que els altres sistemes fan servir informació de sintaxi de constituents i corpus diferents, però de grandària similar.

A la figura 2 es mostren els resultats gràficament. Per a cada rol s'indica la F_1 i la seva proporció normalitzada a 1 en el corpus. El gràfic mostra que la F_1 varia més en els rols amb una proporció de menys de 0,05 (5%).

5. Conclusions

En aquest article hem presentat un sistema d'etiquetatge automàtic de rols semàntics, el principal component del qual és un classificador basat en memòria. El sistema utilitza informació de sintaxi de dependències. Els resultats obtinguts són comparables als dels sistemes existents, que fan servir informació de sintaxi de constituents. Per tant, es pot concloure que utilitzar informació de sintaxi de dependències no fa la tasca més difícil, encara que tampoc no la facilita especialment. Això pot ser degut al fet que els corpus amb sintaxi de constituents contenen informació de sintaxi més complexa i es poden definir atributs més rics.

Figura 2. F_1 per rol comparat amb la proporció per rol



Bibliografia

BUCHHOLZ, S; MARSI, E. (2006). «CoNLL-X shared task on multilingual dependency parsing». A: *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. Nova York: ACL. Pàgs. 149-164.

CARRERAS, X.; MÀRQUEZ, LL. (2005). «Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling». A: *Proceedings of the Ninth Conference on Computational Natural Language Learning*. Ann Arbor: ACL. Pàg. 152-164. <http://www.lsi.upc.es/%7Esrlconll/st05/papers/intro.pdf>

CIVIT, M. (2003). *Guía para la anotación sintáctica de Cast3LB: un corpus del español con anotación sintáctica, semántica y pragmática*. Barcelona: CliC–UB (X-TRACT-II WP-03-06 i 3LB-WP-02-01).

COVER, T. M.; HART, P. E. (1967). «Nearest neighbor pattern classification». *Institute of Electrical and Electronics*

Engineers Transactions on Information Theory. Núm. 13, pàg. 21-27.

DAELEMANS, W.; VAN DEN BOSCH, A. (2005). *Memory-based language processing.* Cambridge: Cambridge University Press. 189 pàg.

DAELEMANS, W.; ZAVREL, J.; VAN DER SLOOT, K. [*et al.*] (2007). *TiMBL: Tilburg Memory-Based Learner, version 6.0, reference guide.* Tilburg: ILK (Technical Report Series 04-02). 57 pàg.

FILLMORE, C. J. (1968). «The case for case». A: E. BACH, R. T. HARMS (eds.). *Universals in Linguistic Theory*. Londres: Holt, Rinehart. Pàg. 1-88.

MÀRQUEZ, LL.; CARRERAS, X. (2004). «Introduction to the CoNLL-2004 Shared Task: Semantic Role Labeling». A*: Proceedings of the 8th Conference on Computational Natural Language Learning*, CoNLL-2004. Boston: ACL. Pàg. 89-97. <http://www.lsi.upc.edu/~srlconll/systems/results.html>

MÀRQUEZ, LL.; PADRÓ, LL.; SURDEANU, M. [*et al.*] (2007a). «UPC: Experiments with Joint Learning within SemEval Task 9». A: *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*. Praga: ACL. Pàg. 426-429.

MÀRQUEZ, LL.; VILLAREJO, L.; MARTÍ, M. A. [*et al.*] (2007b). «Semeval-2007 Task 09: Multilevel semantic annotation of Catalan and Spanish». A: *SemEval 2007. Proceedings of the 4th International Workshop on Semantic Evaluations*. Praga: ACL. Pàg. 42-47.

MORANTE, R. (2006). *Semantic role annotation in the Cast3LB-CoNNL-SemRol corpus.* Tilburg: ILK (Technical Report Series 06-03). 119 pàg.

MORANTE, R.; VAN DEN BOSCH, A. (2007). «Memory-based semantic role labelling». A: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2007)*. Borovets (Bulgària). Pàg. 388-394.

PALMER, M.; GILDEA, D.; KINGSBURY, P. (2005). «The Proposition Bank: An Annotated Corpus of Semantic Roles». *Computational Linguistics*. Vol. 31. Núm. 1. Pàg. 71-106.

SURDEANU, M.; MORANTE, R.; MÀRQUEZ, LL. (2008). «Analysis of joint inference strategies for the semantic role labeling of Spanish and Catalan». *Lecture Notes in Computer Science*. Berlín/Heidelberg: Springer. Vol. 4919/2008. Pàg. 206-218.

Roser Morante
Investigadora al grup de recerca CNTS de la Universitat d'Anvers
Roser.Morante@ua.ac.be

Llicenciada en Lingüística i en Filologia Espanyola per la Universitat de Barcelona. És doctora en Lingüística per la Universitat de Barcelona i doctora per la Universitat de Tilburg. Actualment treballa d'investigadora al grup de recerca CNTS de la Universitat d'Anvers en un projecte de mineria de textos aplicada al domini biomèdic.
Més informació sobre l'autora a: http://ilk.uvt.nl/~roser.