

CoNLL-2004 Shared Task: Semantic Role Labeling

Xavier Carreras and Lluís Màrquez

TALP Research Center

Technical University of Catalonia

UIUC, June 9th, 2004

Semantic Role Labeling (SRL)

- Analysis of propositions in a sentence
- Recognize constituents which fill a semantic role

[A_0 He] [$AM-MOD$ would] [$AM-NEG$ n't] [V **accept**] [A_1 anything of value] from [A_2 those he was writing about] .

Roles for the predicate **accept** (PropBank frames scheme):

V : verb; A_0 : acceptor; A_1 : thing accepted; A_2 : accepted-from;
 A_3 : attribute; $AM-MOD$: modal; $AM-NEG$: negation;

Existing Systems

- On the top of a full syntactic tree:
most systems use Collins' or Charniak's parsers
Best results ~ 80 (F_1 measure)
See (Pradhan et al. NAACL-2004)
- On the top of a chunker:
(Hacioglu et al., 2003) and (Hacioglu, NAACL-2004)
Best results ~ 60 (F_1 measure)

Goal of the Shared Task

- Machine Learning–based systems for SRL
- Use of only shallow syntactic information and clause boundaries (partial parsing)
- An open setting was also proposed but...
- Very hard time constraints

Problem Setting

In a sentence:

- N target verbs. Marked as input
- Output: N chunkings representing the arguments of each verb
- Arguments may appear discontinuous (unfrequent)
- Arguments do not overlap

Evaluation

SRL is a “recognition” task:

- **precision**: percentage of predicted arguments that are correct
- **recall**: percentage of correct arguments that are predicted
- $F_{\beta=1} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision} + \text{recall})}$

An argument is correct **iff** its spanning and label are correct

Data: PropBank

- Proposition Bank corpus (PropBank)
(Palmer, Gildea and Kingsbury, 2004)
- Penn Treebank corpus enriched with
predicate–argument structures
- Verb senses from VerbNet. A *roleset* for each sense.
- February 2004 version

Types of Arguments

- Numbered arguments (A0–A5, AA):
Arguments defining verb-specific roles. Their semantics depends on the verb and the verb usage in a sentence.
- Adjuncts (AM–):
cause, direction, temporal, location, manner, negation, etc.
- References (R–)
- Verbs (V)

Data Sets₁

WSJ sections: 15-18 training, 20 validation, 21 test

	Training	Devel.	Test
Sentences	8,936	2,012	1,671
Tokens	211,727	47,377	40,039
Propositions	19,098	4,305	3,627
Distinct Verbs	1,838	978	855
All Arguments	50,182	11,121	9,598

Data Sets₂

	Training	Devel.	Test
A0	12,709	2,875	2,579
A1	18,046	4,064	3,429
A2	4,223	954	714
A3	784	149	150
A4	626	147	50
A5	14	4	2
AA	5	0	0
R-A0	738	162	159
R-A1	360	74	70
R-A2	49	17	9
R-A3	8	0	1
R-AA	1	0	0

Data Sets₃

	Training	Devel.	Test
AM-ADV	1,727	352	307
AM-CAU	283	53	49
AM-DIR	231	60	50
AM-DIS	1,077	204	213
AM-EXT	152	49	14
AM-LOC	1,279	230	228
AM-MNR	1,337	334	255
AM-MOD	1,753	389	337
AM-NEG	687	131	127
AM-PNC	446	100	85
AM-PRD	10	3	3
AM-REC	2	1	0
AM-TMP	3,567	759	747
R-AM-ADV	1	0	0
R-AM-LOC	27	4	4
R-AM-MNR	4	0	1
R-AM-PNC	1	0	1
R-AM-TMP	35	6	14

Input Information

From previous CoNLL shared tasks:

- PoS tags
- Base chunks
- Clauses
- Named-Entities

Annotation predicted by state-of-the-art linguistic processors

Example

The	DT	B-NP	(S*	0	-	(A0*	*
San	NNP	I-NP	*	B-ORG	-	*	*
Francisco	NNP	I-NP	*	I-ORG	-	*	*
Examiner	NNP	I-NP	*	I-ORG	-	*A0)	*
issued	VBD	B-VP	*	0	issue	(V*V)	*
a	DT	B-NP	*	0	-	(A1*	(A1*
special	JJ	I-NP	*	0	-	*	*
edition	NN	I-NP	*	0	-	*A1)	*A1)
around	IN	B-PP	*	0	-	(AM-TMP*	*
noon	NN	B-NP	*	0	-	*AM-TMP)	*
yesterday	NN	B-NP	*	0	-	(AM-TMP*AM-TMP)	*
that	WDT	B-NP	(S*	0	-	(C-A1*	(R-A1*R-A1)
was	VBD	B-VP	(S*	0	-	*	*
filled	VBN	I-VP	*	0	fill	*	(V*V)
entirely	RB	B-ADVP	*	0	-	*	(AM-MNR*AM-MNR)
with	IN	B-PP	*	0	-	*	*
earthquake	NN	B-NP	*	0	-	*	(A2*
news	NN	I-NP	*	0	-	*	*
and	CC	I-NP	*	0	-	*	*
information	NN	I-NP	*S)S)	0	-	*C-A1)	*A2)
.	.	0	*S)	0	-	*	*

Participant Teams₁

- Ulrike **Baldewein**, Katrin Erk, Sebastian Padó and Detlef Prescher. Saarland University, University of Amsterdam
- Antal **van den Bosch**, Sander Canisius, Walter Daelemans, Iris Hendrickx and Erik Tjong Kim Sang. Tilburg University, University of Antwerp
- Xavier **Carreras** and Lluís Màrquez and Grzegorz Chrupała. Technical University of Catalonia, University of Barcelona.
- Kadri **Hacioglu**, Sameer Pradhan, Wayne Ward, James H. Martin and Daniel Jurafsky. University of Colorado, Stanford University.
- Derrick **Higgins**. Educational Testing Service.

Participant Teams₂

- Beata **Kouchnir**. University of Tübingen
- Joon-Ho **Lim**, Young-Sook Hwang, So-Young Park and Hae-Chang Rim. Korea University
- Kyung-Mi **Park**, Young-Sook Hwang and Hae-Chang Rim. Korea University
- Vasin **Punyakanok**, Dan Roth, Wen-Tau Yih, Dav Zimak and Yuancheng Tu. University of Illinois
- Ken **Williams**, Christopher Dozier and Andrew McCulloh. Thomson Legal and Regulatory

Learning Algorithms

- Maximum Entropy (baldewein, lim)
- Transformation-based Error-driven Learning (higgins, williams)
- Memory-Based Learning (vandenbosch, kouchnir)
- Support Vector Machines (hacioglu, park)
- Voted Perceptron (carreras)
- SNoW (punyakanok)

SRL Architectures

	prop-treat	labeling	granularity	glob-opt	post-proc
hacioglu	separate	seq-tag	P-by-P	no	no
punayakanok	separate	filt+lab	W-by-W	yes	no
carreras	joint	filt+lab	P-by-P	yes	no
lim	separate	seq-tag	P-by-P	yes	no
park	separate	rec+class	P-by-P	no	yes
higgins	separate	seq-tag	W-by-W	no	yes
vandenbosch	separate	class+join	P-by-P	part.	yes
kouchnir	separate	rec+class	P-by-P	no	yes
baldewein	separate	rec+class	P-by-P	yes	no
williams	separate	seq-tag	mixed	no	no

- Nobody performed verb sense disambiguation

Features

- Highly inspired by previous work on SRL (Gildea and Jurafsky, 2002; Surdeanu et al., 2003; Pradhan et al., 2003)
- Feature Types:
 - ★ *Basic*: local context, window based (words, POS, chunks, clauses, named entities)
 - ★ *Internal Structure* of a candidate argument
 - ★ Properties of the target *verb predicate*
 - ★ *Relations* between verb predicate and the constituent
- Importance of lexicalization and “path”–based features

Types of Features

	sy	ne	al	at	as	aw	an	vv	vs	vf	vc	rp	di	pa	ex
hacioglu	+	+	+	-	-	+	-	+	+	-	+	+	+	+	+
punyakankok	+	+	+	+	+	+	-	+	-	+	+	+	-	+	+
carreras	+	-	-	-	+	+	-	+	-	-	-	+	-	+	+
lim	+	-	-	-	-	+	+	+	-	-	-	+	-	+	-
park	+	-	-	-	-	-	-	+	-	-	+	+	+	+	+
higgins	+	+	-	-	-	-	+	+	-	-	-	+	+	+	-
vandenbosch	+	+	-	-	-	-	-	+	+	-	-	+	+	-	-
kouchnir	+	-	+	-	+	+	-	+	-	+	-	+	+	-	-
baldewein	+	+	+	+	+	+	-	+	+	-	-	+	+	-	-
williams	+	+	-	-	-	-	-	-	-	-	-	+	-	-	-

Baseline System

Developed by Erik Tjong Kim Sang. Six heuristic rules.

- Tag *not* and *n't* in target verb chunk as **AM-NEG**.
- Tag modal verbs in target verb chunk as **AM-MOD**.
- Tag first NP before target verb as **A0**.
- Tag first NP after target verb as **A1**.
- Tag *that*, *which* and *who* before target verb as **R-A0**.
- Switch **A0** and **A1**, and **R-A0** and **R-A1** if the target verb is part of a passive VP chunk.

Results on Test

	Precision	Recall	F ₁
hacioglu	72.43%	66.77%	69.49
punayakanok	70.07%	63.07%	66.39
carreras	71.81%	61.11%	66.03
lim	68.42%	61.47%	64.76
park	65.63%	62.43%	63.99
higgins	64.17%	57.52%	60.66
vandenbosch	67.12%	54.46%	60.13
kouchnir	56.86%	49.95%	53.18
baldewein	65.73%	42.60%	51.70
williams	58.08%	34.75%	43.48
baseline	54.60%	31.39%	39.87

Comparative Analysis

- Detailed Results
- Recognition + Classification Performance
- Results per Argument Size
- Results per Argument-Verb Distance
- Results per Verb Frequency
- Results per Verb Polisemy

Core Roles: test results

	A0	A1	A2	A3	A4	A5	R-A0	R-A1	R-A2
hac	81.37	71.63	49.33	51.11	66.67	0.00	85.43	71.54	50.00
pun	79.38	68.16	46.69	34.04	65.22	0.00	78.96	57.97	36.36
car	79.05	66.96	43.28	31.22	62.07	0.00	78.10	57.14	36.36
lim	77.42	66.00	49.07	41.77	54.55	0.00	80.81	60.27	40.00
par	76.38	66.14	46.57	42.32	51.76	0.00	81.73	61.02	50.00
hig	70.67	62.72	45.52	40.00	39.64	0.00	79.61	62.07	36.36
van	74.95	60.83	40.41	37.44	62.37	0.00	78.46	55.56	36.36
kou	65.49	54.48	30.95	19.71	36.07	0.00	76.77	58.27	47.06
bal	66.76	53.37	37.60	22.89	27.69	0.00	0.00	0.00	0.00
wil	56.24	49.05	0.00	0.00	0.00	0.00	65.61	0.00	0.00
bas	57.65	34.19	0.00	0.00	0.00	0.00	74.86	33.33	0.00

Adjuncts: test results

	ADV	CAU	DIR	DIS	LOC	MNR	MOD	NEG	PNC	TMP
hac	44.91	32.35	32.18	64.56	40.89	38.94	95.43	93.89	23.64	56.82
pun	37.69	39.53	37.78	58.61	34.05	40.60	93.70	90.71	27.40	58.30
car	43.00	38.36	32.84	60.74	27.81	31.06	96.40	92.31	21.49	54.60
lim	40.15	40.00	35.44	54.73	35.32	32.62	90.43	87.16	35.82	49.73
par	44.74	27.85	20.00	57.41	28.34	39.22	94.17	91.43	33.10	48.39
hig	36.13	48.10	27.27	55.42	23.67	34.00	93.60	93.08	19.30	44.12
van	7.71	0.00	17.65	54.27	26.16	27.04	93.21	80.87	8.51	41.90
kou	14.83	0.00	27.37	53.18	13.37	31.28	91.58	91.83	11.11	38.04
bal	21.46	3.57	25.71	39.25	22.22	21.20	83.08	74.77	18.52	35.35
wil	0.00	0.00	0.00	0.00	0.00	0.00	72.35	60.36	0.00	11.68
bas	0.00	0.00	0.00	0.00	0.00	0.00	90.71	92.12	0.00	0.00

Split Arguments

- Split Arguments: difficult but not very frequent.
- 3 systems did not treat them.
- Occurrences:

training :	525
devel. :	104
test :	108

Split Arguments: test results

	Precision	Recall	F ₁
hacioglu	71.64	48.00	57.49
punayakanok	58.33	28.00	37.84
carreras	0.00	0.00	0.00
lim	80.00	16.00	26.67
park	61.54	24.00	34.53
higgins	47.92	23.00	31.08
vandenbosch	21.95	9.00	12.77
kouchnir	43.75	21.00	28.38
baldewein	0.00	0.00	0.00
williams	0.00	0.00	0.00
baseline	0.00	0.00	0.00

Recognition + Labeling

- We evaluate the performance of recognizing argument boundaries (correct argument = correct boundaries).
- For each system, we also evaluate classification accuracy on the set of recognized arguments.
- Clearly, all systems suffer from recognition errors.

Recognition + Labeling: test results

	Precision	Recall	F ₁	Acc
hacioglu	78.61	72.47	75.42	92.14
punyakankok	77.82	70.04	73.72	90.05
carreras	79.22	67.41	72.84	90.65
lim	75.43	67.76	71.39	90.71
park	73.64	70.05	71.80	89.13
higgins	70.72	63.40	66.86	90.73
vandenbosch	75.48	61.23	67.61	88.96
kouchnir	66.52	58.43	62.21	85.49
baldewein	75.13	48.70	59.09	87.48
williams	70.62	42.25	52.87	82.24
baseline	66.51	38.24	48.56	82.10

Confusion Matrix (Hacioglu)

	-NONE-	A0	A1	A2	A3	ADV	DIS	LOC	MNR	TMP
-NONE-	–	332	805	289	42	60	45	71	49	138
A0	448	2060	58	8	0	0	0	0	0	1
A1	861	77	2446	33	4	0	0	0	1	4
A2	283	5	57	352	3	3	1	0	4	3
A3	64	3	5	8	69	0	0	1	0	0
ADV	141	3	3	1	0	119	8	4	8	16
DIS	49	0	1	0	0	7	133	1	3	18
LOC	129	0	0	0	0	1	0	83	5	10
MNR	125	0	4	6	1	11	3	9	81	12
TMP	311	1	9	4	1	16	9	7	8	379

Arguments grouped by Size

- Size of an argument = length at chunk level
(words outside chunks count as 1 chunk)

	$s = 1$	$2 \leq s \leq 5$	$6 \leq s \leq 10$	$11 \leq s \leq 20$	$20 < s$
Args.	5,549	2,376	996	507	70

- Verbs and split arguments are not considered.
- Arguments of size 1 are the easiest.
- No aggressive degradation as the size increases.

Arguments grouped by Size: test results

	$s = 1$	$2 \leq s \leq 5$	$6 \leq s \leq 10$	$11 \leq s \leq 20$	$20 < s$
hac	76.78	56.93	63.26	64.05	52.35
pun	74.17	51.81	57.11	60.98	59.02
car	73.67	52.67	59.39	60.48	57.14
lim	72.33	51.40	58.92	60.72	49.62
par	72.38	49.25	56.89	59.49	50.00
hig	69.81	45.73	52.26	56.81	45.38
van	69.24	47.70	45.48	51.11	43.30
kou	65.03	35.88	36.05	39.06	28.57
bal	59.67	33.27	44.23	51.37	45.71
wil	53.39	15.15	28.62	47.04	50.49
bas	54.22	2.46	0.00	0.00	0.00

Argument-Verb Distance

- $\text{distance}(a, v)$ = number of chunks from a to v
(words outside chunks count as 1 chunk)

	0	1	2	3-5	6-10	11-15	16+
Args	4,703	1,948	1,171	1,186	377	89	24

- Verbs and split arguments are not considered.
- Performance decreases progressively as distance increases.

Argument-Verb Distance: test results

	0	1	2	3-5	6-10	11-15	16+
hac	78.21	66.35	66.08	53.55	38.99	26.67	24.49
pun	76.18	64.85	61.75	52.53	30.07	12.68	14.46
car	75.24	61.46	63.27	51.77	36.56	28.00	29.17
lim	73.59	62.79	63.01	50.59	32.44	22.89	11.59
par	73.27	62.54	60.98	47.78	29.76	15.49	14.29
hig	71.55	59.18	57.06	41.08	22.74	13.61	13.64
van	69.87	58.03	56.39	37.33	14.74	0.00	6.67
kou	66.19	50.04	46.18	28.28	7.24	2.23	0.00
bal	63.53	44.45	44.15	30.29	13.60	2.13	0.00
wil	55.86	31.18	39.49	20.80	3.58	0.00	0.00
bas	46.73	40.77	34.22	19.94	1.04	0.00	0.00

Verbs grouped by Frequency

- We group verbs by their frequency in the training data:

	0	1–5	6–20	21–100	101–300	450	1821
Verbs	133	277	252	170	20	1 (have)	1 (say)
Props.	147	376	631	1,369	586	97	418
Args.	265	740	1,256	2,709	1,158	192	838

- Then, we evaluate performance of A0-A5 arguments:
 - ★ The more frequent, the better.
 - ★ But systems perform not so bad on unseen verbs!

Verbs grouped by Frequency: test results

	0	1-5	6-20	21-100	101-300	450	1821
hac	60.90	62.98	73.08	69.26	73.32	82.08	92.29
pun	58.19	60.92	67.18	66.47	70.53	81.08	91.29
car	62.34	59.20	65.37	65.47	69.80	84.38	90.91
lim	57.73	57.33	67.08	65.11	66.87	83.77	90.20
par	57.70	57.80	64.89	64.76	69.10	79.17	88.86
hig	54.58	52.62	60.13	60.61	64.97	79.79	85.34
van	49.27	56.04	60.70	61.85	61.00	78.85	86.28
kou	40.95	44.82	52.41	52.85	55.57	68.59	79.37
bal	0.00	38.99	51.28	51.88	58.03	71.93	83.78
wil	44.05	46.49	46.89	41.25	45.05	55.47	75.61
bas	43.70	47.46	43.74	42.24	41.47	58.29	35.38

Verbs grouped by Sense Ambiguity

- For each verb:
 - ★ We compute the distribution of senses in the data.
 - ★ Then, we calculate the entropy of the verb sense.
- We group verbs by the entropy of the verb sense, and evaluate A0-A5 of each group.

	H= 0	0 <H≤ .8	.8 <H≤ 1.5	1.5 <H≤ 2.0	2.0 <H
Verbs	617	95	109	23	9
Props.	2,058	824	451	145	145
Args.	4,064	1,631	882	304	280

Verbs by Sense Ambiguity: test results

	H= 0	0 <H≤ .8	.8 <H≤ 1.5	1.5 <H≤ 2.0	2.0 <H
hac	76.18	73.72	64.40	61.03	56.13
pun	72.62	68.97	65.12	62.09	57.79
car	72.12	67.95	62.55	59.14	60.31
lim	71.24	67.65	61.58	62.96	54.24
par	70.90	67.33	61.96	58.66	52.42
hig	67.14	62.99	58.00	50.94	51.21
van	68.01	62.06	56.47	56.01	46.15
kou	59.42	54.74	47.51	43.88	42.80
bal	57.24	55.97	46.80	48.32	51.84
wil	51.83	45.96	41.49	44.44	36.21
bas	41.53	45.07	41.97	48.63	40.75

Concluding Remarks

- 10 systems participated in the 2004 Shared Task on Semantic Role labeling.
- The best system was developed by the team of the University of Colorado, and performs a BIO tagging along chunks with Support Vector Machines.
- Its performance on test data is 69.49 in F-measure.
- Detailed evaluations show general superiority of the best system over competing ones.

Concluding Remarks

- Performance of systems is moderate, and far from acceptable figures for real usage.
- Systems rely only on partial syntactic information: chunks and clauses.
 - ★ Full parsing: $F_1 = \sim 80$
 - ★ Chunking+clauses (CoNLL-2004): $F_1 = \sim 70$
 - ★ Chunking: $F_1 = \sim 60$
- Do we need full syntactic structure?