Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling

http://www.lsi.upc.edu/∼srlconll

Xavier Carreras and **Lluís Màrquez**

TALP Research Center

Technical University of Catalonia

Ann Arbor, June 2005

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;

Goal of the Shared Task

- Machine Learning–based systems for SRL
- 3 months to develop a system
- CoNLL-2004 Shared Task Revisited. Novelties:
  - ⋆ Several levels of syntactic annotations
  - ⋆ Training data substantially enlarged
  - ⋆ Cross-corpora evaluation on a portion of the Brown corpus

Data: PropBank 1.0

- Proposition Bank corpus (Palmer, Gildea and Kingsbury, 2004)
- WSJ part of Penn TreeBank corpus enriched with predicate–argument structures
- Types of Arguments (i.e. Semantic Roles):
  - ⋆ Numbered arguments (`A0–A5, AA`):
    Verb-specific roles. Their semantics depends on the verb.
  - ⋆ Adjuncts (`AM-`):
    cause, direction, temporal, location, manner, negation, etc.
  - ⋆ References (`R-`)
  - ⋆ Verbs (`V`)

Data: Training/Dev./Test

- We moved to the "full parsing standard partition":
  - ⋆ Training : WSJ sections 02–21
  - ⋆ Development: WSJ section 24
  - ⋆ Test: WSJ section 23

- + 3 sections of "PropBanked" Brown corpus, for testing.
  $\longrightarrow$ many thanks to the PropBank team for providing fresh data

Data: Counts

	Train.	Devel.	tWSJ	tBrown
Sentences	39,832	1,346	2,416	426
Tokens	950,028	32,853	56,684	7,159
Propositions	90,750	3,248	5,267	804
Verbs	3,101	860	982	351
Arguments	239,858	8,346	14,077	2,177

Data: Most Frequent Core Arguments

	Train.	Devel.	tWSJ	tBrown
A0	**61,440**	**2,081**	**3,563**	**566**
A1	**84,917**	**2,994**	**4,927**	**676**
A2	**19,926**	**673**	**1,110**	**147**
A3	**3,389**	**114**	**173**	**12**
A4	**2,703**	**65**	**102**	**15**
A5	68	2	5	0
AA	14	1	0	0
R-A0	**4,112**	**146**	**224**	**25**
R-A1	**2,349**	**83**	**156**	**21**
R-A2	291	5	16	0

Data: Most Frequent Adjuncts

	Train.	Devel.	tWSJ	tBrown
AM-ADV	**8,210**	**279**	**506**	**143**
AM-CAU	**1,208**	**45**	**73**	**8**
AM-DIR	**1,144**	**36**	**85**	**53**
AM-DIS	**4,890**	**202**	**320**	**22**
AM-EXT	628	28	32	5
AM-LOC	**5,907**	**194**	**363**	**85**
AM-MNR	**6,358**	**242**	**344**	**110**
AM-MOD	**9,181**	**317**	**551**	**91**
AM-NEG	**3,225**	**104**	**230**	**50**
AM-PNC	**2,289**	**81**	**115**	**17**
AM-TMP	**16,346**	**601**	**1,087**	**112**
R-AM-LOC	214	9	21	4
R-AM-MNR	143	6	6	2
R-AM-TMP	719	31	52	10

Problem Setting

In a sentence:

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

Evaluation

SRL is a "recognition" task:

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

An argument is correct iff its spanning and label are correct

Closed and Open Challenges

- Closed Challenge :
  - ⋆ Make use of **any** annotation of the TreeBank and PropBank training sections to develop a system.
  - ⋆ Or use **any** existing tool developed with such data.
  - ⋆ We provided annotations predicted by state-of-the-art analyzers.
- An open setting was also proposed but again . . .

Input Annotations Provided

- Pos Tagger of (Giménez and Màrquez, 2003)
- Chunker and Clauser of (Carreras and Màrquez, 2003)
- Full parser of (Collins 1999)
- Full parser of (Charniak 2000)
- Named Entity Extractor of (Chieu and Ng, 2003)
- Any other tool developed with WSJ sections 02-21 was welcome.

Participant Teams 1

- Wanxiang **Che**, Ting Liu, Sheng Li, Yuxuan Hu and Huaijun Liu. **Harbin Institute of Technology**.
- Trevor **Cohn** and Philip Blunsom. **University of Melbourne**.
- Alessandro **Moschitti**, Ana-Maria Giuglea, Bonaventura Coppola and Roberto Basili. **University of Rome “Tor Vergata”, ITC-Irst, Universtiy of Trento**.
- Aria **Haghighi**, Kristina Toutanova and Christopher Manning. **Stanford University**.
- Richard **Johansson** and Pierre Nugues. **Lund University**.
- Chi-San **Lin** and Tony C. Smith. **Waikato University**.
- Lluís **Màrquez**, Pere Comas, Jesús Giménez and Neus Català. **Technical University of Catalonia**.

Participant Teams$_2$

- Necati Ercan **Ozgencil** and Nancy McCracken. **Syracuse University**.
- Tomohiro **Mitsumori**, Masaki Murata, Yasushi Fukuda, Kouichi Doi and Hirohumi Doi. **Nara Institute of Science and Technology, National Institute of Information and Communications Technology, Sony-Kihara Research Center Inc.**
- Kyung-Mi **Park** and Hae-Chang Rim. **Korea University**.
- Simone Paolo **Ponzetto** and Michael Strube. **EML Research gGmbH, Germany**.
- Sameer **Pradhan**, Kadri Hacioglu, Wayne Ward, James H. Martin and Daniel Jurafsky. **University of Colorado, Stanford University**.
- Vasin **Punyakanok**, Peter Koomen, Dan Roth and Wen-tau Yih. **University of Illinois at Urbana-Champaign**.

Participant Teams $_3$

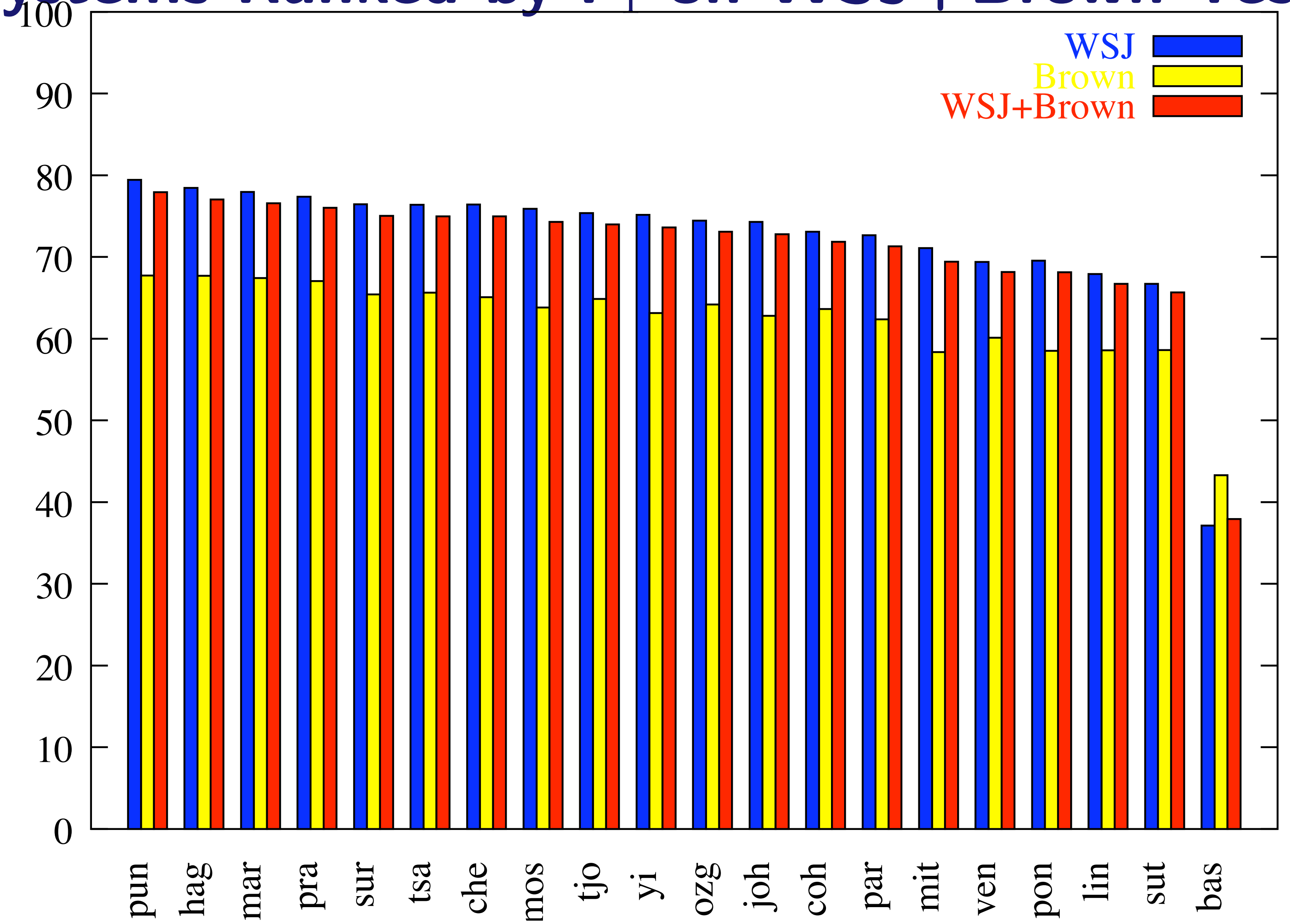
- Mihai **Surdeanu** and Jordi Turmo. **Technical University of Catalonia**.
- Charles **Sutton** and Andrew McCallum. **University of Massachusets**.
- Erik **Tjong Kim Sang**, Sander Canisius, Antal van den Bosch and Toine Bogers. **University of Amsterdam, Tilburg University**.
- Tzong-Han **Tsai**, Chia-Wei Wu, Yu-Chun Lin and Wen-Lian Hsu. **Academia Sinica, Taiwan**.
- Sriram **Venkatapathy**, Akshar Bharati and Prashanth Reddy. **Language Technologies Research Center, IIIT, India**.
- Szu-ting **Yi** and Martha Palmer. **University of Pennsylvania**.

Baseline System

Same as in CoNLL-2004. Developed by Erik Tjong Kim Sang. Six heuristic rules that make use of PoS and Chunks:

- 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.

Systems Ranked by F_1 on WSJ+Brown Test



Outline of the Session

- Introduction
- **System Presentations:**
  - ⋆ Surdeanu and Turmo: *SRL Using Complete Analysis*.
  - ⋆ Pradhan et al.: *Semantic Role Chunking Combining Complementary Syntactic Views*.
  - ⋆ Haghighi et al.: *A Joint Model for Semantic Role Labeling*.

  coffee break
  - ⋆ Punyakanok et al.: *Generalized Inference with Multiple Semantic Role Labeling Systems.*
- Overview and Evaluation of Systems
- Discussion, with Spotlight Notes

Outline of the Session

- Introduction
- System Presentations
- **Overview of Systems:**
  - **System Properties**
  - Evaluation
- Discussion, with Spotlight Notes

System Properties: Learning Method

punyakanok	SNoW	ozgencil	SVM
haghighi	MaxEnt	johansson	RVM
marquez	AdaBoost	cohn	Tree-CRF
pradhan	SVM	park	MaxEnt
surdeanu	AdaBoost	mitsumori	SVM
tsai	MaxEnt,SVM	venkatapathy	MaxEnt
che	MaxEnt	ponzetto	DT
moschitti	SVM	lin	CPM
tjongkimsang	MaxEnt,SVM,MBL	sutton	MaxEnt
yi	MaxEnt		

System Properties: Explored Syntax

punyakanok	n-cha,col	ozgencil	cha
haghighi	n-cha	johansson	cha
marquez	cha,upc	cohn	col
pradhan	cha,col/chunk	park	cha
surdeanu	cha	mitsumori	chunk
tsai	cha	venkatapathy	col
che	cha	ponzetto	col
moschitti	cha	lin	cha
tjongkimsang	cha	sutton	n-bikel
yi	cha,AN,AM		

System Properties: SRL Strategy

	pre	label	embed	glob	post
punyakanok	x&p	i+c	defer	yes	no
haghighi	?	i+c	dp-prob	yes	no
marquez	seq	bio	!need	no	no
pradhan	?	c/bio	?	no	no
surdeanu	prun	c	g-top	no	yes
tsai	x&p	c	defer	yes	no
che	no	c	g-score	no	yes
moschitti	prun	i+c	!need	no	no
tjongkimsang	prun	i+c	!need	no	yes
yi	x&p	i+c	defer	no	no
ozgencil	prun	i+c	g-score	no	no
johansson	softp	i+c	?	no	no
cohn	x&p	c	g-top	yes	no
park	prun	i+c	?	no	no
mitsumori	no	bio	!need	no	no
venkatapathy	prun	i+c	frames	yes	no
ponzetto	prun	c	g-top	no	yes
lin	gt-para	i+c	!need	no	no
sutton	x&p	i+c	dp-prob	yes	no

System Properties: System Combination

		comb	type
1	punyakanok	n-cha+col	ac-ILP
2	haghighi	n-cha	re-rank
3	marquez	cha+upc	s-join
4	pradhan	cha+col→chunk	stack
	...		
6	tsai	ME+SVM	ac-ILP
	...		
9	tjongkimsang	ME+SVM+MBL	s-join
10	yi	cha+AN+AM	ac-join
	...		
	...		
19	sutton	n-bikel	re-rank

Features

- Generally, all systems implemented the standard features for SRL.
- The key works are:
  - ⋆ (Gildea and Jurafsky, 2002)
  - ⋆ (Surdeanu et al., 2003)
  - ⋆ (Pradhan et al., 2003, 2005)
  - ⋆ (Xue and Palmer, 2004)
  - ⋆ CoNLL-2004 Shared Task

Features: Sources

	synt	ne
punyakanok	cha,col,upc	+
haghighi	cha	·
marquez	cha,upc	+
pradhan	cha,col,upc	+
surdeanu	cha	+
tsai	cha,upc	+
che	cha	+
moschitti	cha	·
tjongkimsang	cha	+
yi	cha,an,am	·
ozgencil	cha	·
johansson	cha,upc	+
cohn	col	·
park	cha	·
mitsumori	upc,cha	+
venkatapathy	col	+
ponzetto	col,upc	+
lin	cha	·
sutton	bik	·

Features on the Argument Candidate

	at	aw	ab	ac	ai	pp	sd
punyakanok	+	h	+	t	+	+	·
haghighi	+	h	+	p,s	·	+	+
marquez	+	h	+	t	+	·	+
pradhan	+	h,c	+	p,s,t	+	+	·
surdeanu	+	h,c	+	p,s	+	·	+
tsai	+	h	+	p,s,t	·	·	·
che	+	h	+	·	·	+	·
moschitti	+	h	+	p	+	+	·
tjongkimsang	+	·	+	p,t	·	+	·
yi	+	h,c	·	p,s	·	+	·
ozgencil	+	h	·	p	·	·	·
johansson	+	h	·	·	·	·	·
cohn	+	h	+	p,s	·	+	·
park	+	h,c	·	p	·	·	·
mitsumori	+	·	+	t	·	·	+
venkatapathy	+	h	+	·	·	·	·
ponzetto	+	h	+	·	+	·	·
lin	+	h	+	·	·	·	·
sutton	+	h	+	p,s	·	·	·

Features on the Verb

	v	sc
punyakanok	+	+
haghighi	+	+
marquez	+	+
pradhan	+	+
surdeanu	+	+
tsai	+	+
che	+	+
moschitti	+	+
tjongkimsang	+	+
yi	+	+
ozgencil	+	+
johansson	+	+
cohn	+	+
park	+	+
mitsumori	+	·
venkatapathy	+	·
ponzetto	+	·
lin	+	·
sutton	+	·

Features on the Verb-Arg Relation

	rp	di	ps	pv	pi	sf
punyakanok	+	c	+	·	+	+
haghighi	+	t	+	+	·	·
marquez	+	w,c	+	+	·	+
pradhan	+	c,t	+	+	+	+
surdeanu	+	w,t	+	+	+	·
tsai	+	w	+	·	·	·
che	+	t	+	+	·	·
moschitti	+	t	+	+	·	+
tjongkimsang	+	w,t	+	+	+	·
yi	+	w	+	·	·	+
ozgencil	+	·	+	+	·	·
johansson	+	·	+	+	·	·
cohn	+	w	+	·	+	+
park	+	·	+	·	+	·
mitsumori	+	c,t	·	+	·	·
venkatapathy	+	·	+	·	·	·
ponzetto	·	w,c,t	·	·	+	·
lin	+	w	·	·	·	·
sutton	+	·	+	·	·	·

Outline of the Session

- Introduction
- System Presentations
- **Overview of Systems:**
  - ⋆ System Properties
  - ⋆ **Evaluation**
- Discussion, with Spotlight Notes

Evaluation of Input Analyzers

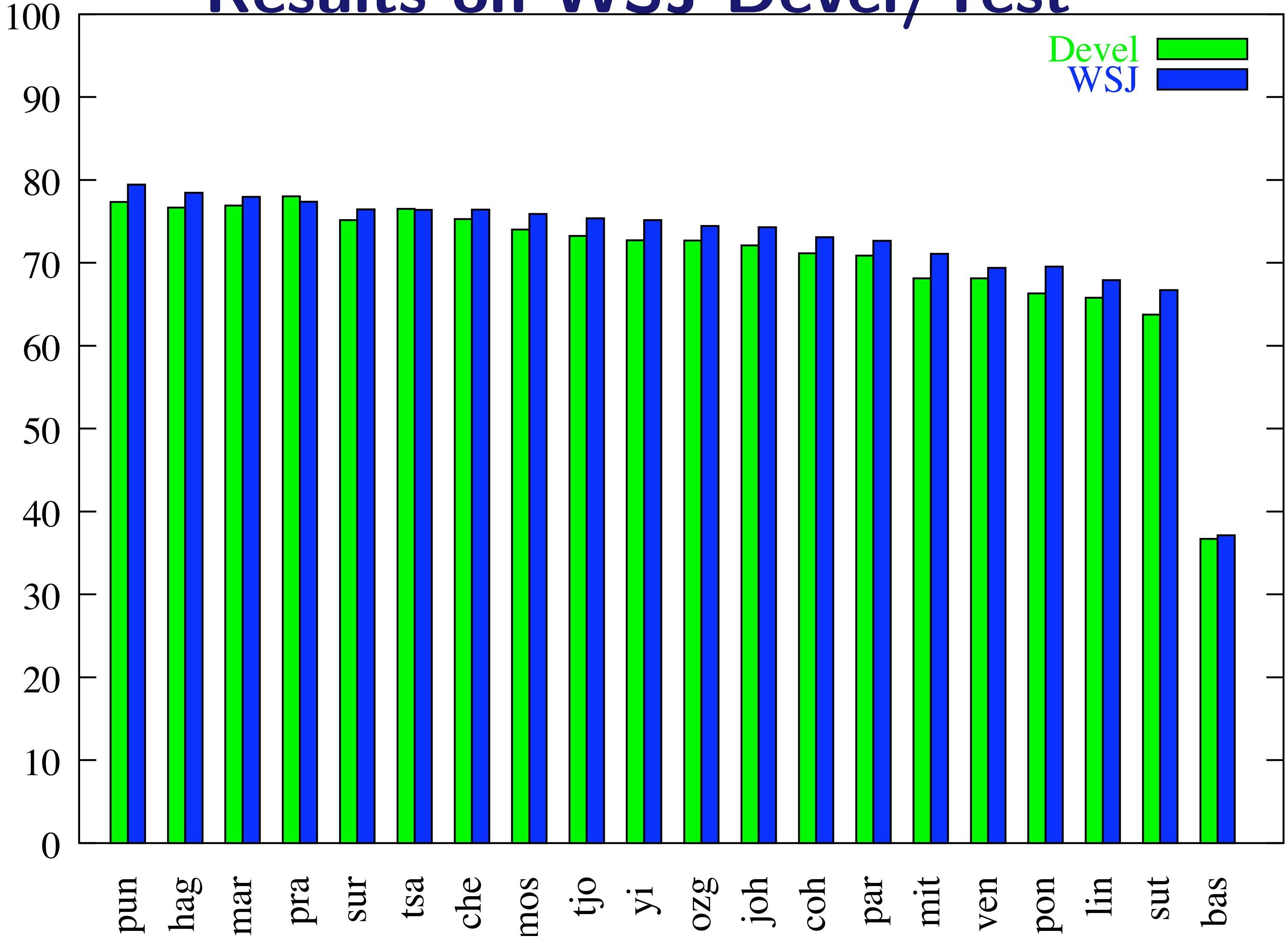
PoS Taggers :

	Dev.	tWSJ	tBrown
UPC	97.13	97.36	94.73
Charniak	92.01	92.29	87.89

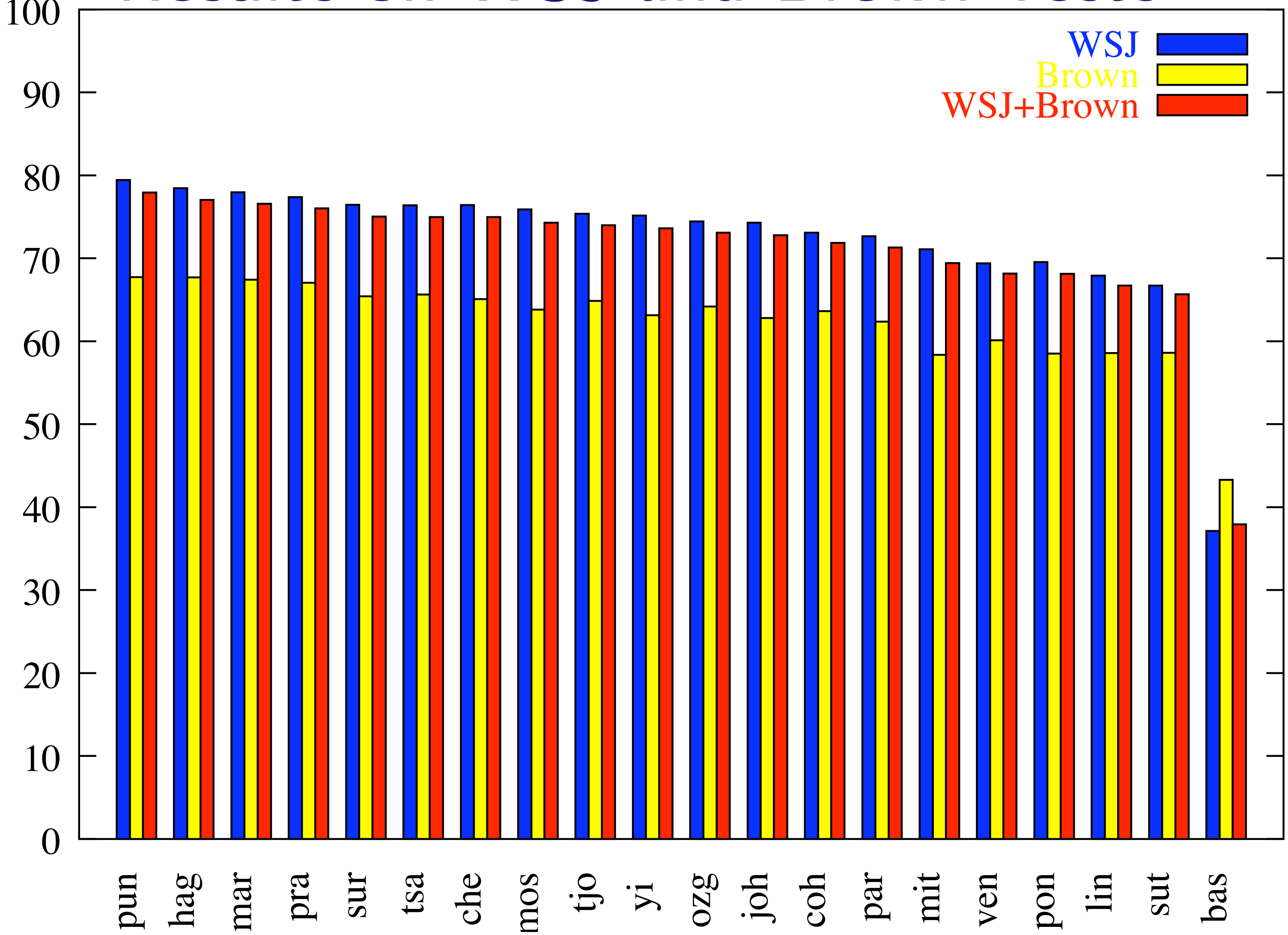
Syntactic Parsers:

	Devel.			Test WSJ			Test Brown		
	P(%)	R(%)	F_1	P(%)	R(%)	F_1	P(%)	R(%)	F_1
Chunker	94.66	93.17	93.91	95.26	94.52	94.89	92.64	90.85	91.73
Clauser	90.38	84.73	87.46	90.93	85.94	88.36	84.21	74.32	78.95
Collins	85.02	83.55	84.28	85.63	85.20	85.41	82.68	81.33	82.00
Charniak	87.60	87.38	87.49	88.20	88.30	88.25	80.54	81.15	80.84

Results on WSJ Devel/Test



Results on WSJ and Brown Tests



Top-10 Systems : Results on WSJ and Brown Tests

	WSJ			Brown		
	Prec.	Rec	F_1	Prec.	Rec	F_1
punyakanok	82.28	76.78	**79.44**	73.38	62.93	**67.75**
haghighi	79.54	77.39	**78.45**	70.24	65.37	**67.71**
marquez	79.55	76.45	**77.97**	70.79	64.35	**67.42**
pradhan	81.97	73.27	77.37	73.73	61.51	67.07
surdeanu	80.32	72.95	76.46	72.41	59.67	65.42
tsai	82.77	70.90	76.38	73.21	59.49	65.64
che	80.48	72.79	76.44	71.13	59.99	65.09
moschitti	76.55	75.24	75.89	65.92	61.83	63.81
tjongkimsang	79.03	72.03	75.37	70.45	60.13	64.88
yi	77.51	72.97	75.17	67.88	59.03	63.14

Top-10 Systems : Core Arguments on WSJ Test

	A0	A1	A2	A3	A4	R-A0	R-A1
punyakanok	**88.05**	**79.91**	68.16	**64.31**	**77.25**	87.67	73.01
haghighi	**88.31**	**78.51**	**70.26**	62.71	73.85	**91.16**	**81.79**
marquez	**86.69**	**78.13**	**68.46**	**64.67**	**74.35**	87.61	75.80
pradhan	86.57	77.97	65.47	60.27	73.30	**91.24**	**79.73**
surdeanu	86.14	75.83	65.55	**65.26**	73.85	86.15	73.03
tsai	86.56	76.89	61.55	59.50	68.60	**88.79**	**79.49**
che	85.81	76.10	**69.09**	62.59	73.58	83.66	74.25
moschitti	82.67	75.63	67.56	61.04	73.63	82.07	68.02
tjongkimsang	83.64	74.34	63.99	58.67	70.83	82.89	72.55
yi	81.04	76.37	65.03	58.90	**74.75**	86.04	74.20

Top-10 Systems : Adjuncts on WSJ Test

	ADV	CAU	DIS	LOC	MNR	MOD	NEG	TMP
pun	**59.73**	53.97	77.95	**60.33**	**59.22**	97.40	**97.61**	**77.44**
hag	54.47	58.21	**78.54**	**58.59**	**57.65**	**98.47**	**97.84**	70.90
mar	**55.56**	**64.62**	76.54	58.20	53.61	95.81	**98.91**	**78.21**
pra	52.71	55.65	70.98	57.27	52.70	95.41	96.92	77.23
sur	51.27	51.95	74.16	57.66	52.65	95.63	96.98	76.18
tsa	54.81	49.56	77.18	50.33	54.79	**97.82**	95.95	70.45
che	53.29	51.28	74.36	58.45	54.49	97.43	96.93	72.07
mos	55.28	57.36	78.07	**60.99**	**59.02**	95.51	96.94	**78.03**
tjo	**57.14**	57.63	**80.45**	56.15	57.78	97.20	97.17	75.32
yi	55.36	**58.91**	**78.59**	56.38	55.22	95.31	95.59	75.61
bas	0.00	0.00	0.00	0.00	0.00	88.71	91.84	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: Results on WSJ test

	Precision	Recall	F_1	Acc
punyakanok	86.78	80.98	**83.78**	**94.82**
haghighi	83.49	81.24	**82.35**	**95.26**
marquez	85.01	81.69	**83.32**	93.58
pradhan	86.86	77.64	81.99	94.37
surdeanu	83.81	76.12	79.78	**95.84**
tsai	87.54	74.98	80.77	94.56
che	85.57	77.40	81.28	94.05
moschitti	82.23	80.83	81.52	93.09
tjongkimsang	83.90	76.47	80.01	94.19
yi	82.41	77.58	79.92	94.06

Verbs grouped by Frequency

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

	0	1–20	21–100	101–500	501–1000
Verbs	34	418	359	149	18
Props.	37	568	1,098	1,896	765
Args.	70	1,049	2,066	3,559	1,450

- Then, we evaluate performance of A0-A5 arguments:

Verbs grouped by Frequency : WSJ test results

	0	1–20	21–100	101–500	501–1000
punyakanok	**68.80**	**75.73**	**80.43**	**81.03**	**79.70**
haghighi	**71.94**	**76.05**	**80.07**	**81.70**	**80.31**
marquez	**73.38**	**73.34**	**79.13**	79.22	**79.08**
pradhan	52.99	72.53	78.16	79.27	78.33
surdeanu	64.66	71.26	77.45	**79.29**	77.12
tsai	66.67	73.17	77.52	79.04	77.15
che	64.62	73.26	78.84	79.06	76.10
moschitti	63.83	69.33	76.03	77.53	75.99
tjongkimsang	65.62	70.80	76.35	76.22	74.09
yi	24.44	70.02	74.66	77.17	75.03

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,.5]	(.5, 1]	(1, 1.5]	(1.5, 2]	(2, 2.8]	$2.8 <$H
Verbs	2,647	105	280	76	37	19	3
Props.	2,823	717	776	384	166	333	68
Args.	5,314	1,343	1,420	743	340	641	79

Verbs by Ambiguity : WSJ Test Results

	H= 0	(0,.5]	(.5, 1]	(1, 1.5]	(1.5, 2]	(2, 2.8]	2.8 <H
punyakanok	**83.83**	**80.51**	**79.21**	**81.21**	**79.31**	**71.24**	**63.69**
haghighi	**82.50**	**81.33**	**79.50**	**80.97**	**79.50**	**75.38**	62.89
marquez	**82.38**	79.87	77.81	78.20	**79.27**	68.59	**68.71**
pradhan	81.47	79.60	77.23	**79.11**	75.14	**73.01**	**67.97**
surdeanu	79.62	**80.16**	**78.13**	77.61	76.48	70.70	52.78
tsai	81.12	78.54	76.71	78.32	75.76	69.08	64.00
che	80.43	79.69	76.95	78.41	75.61	69.70	58.28
moschitti	78.51	77.80	75.94	77.70	73.96	68.07	62.58
tjongkimsang	78.87	77.24	74.53	74.18	72.83	65.62	56.98
yi	78.34	78.30	74.27	73.49	75.25	69.98	56.60

Outline of the Session

- . . .
- Spotlight Notes :
  - ⋆ Trevor Cohn, *Applying Tree-CRFs to SRL*
  - ⋆ Szu-ting Yi, *Learning SRL-specific syntactic parsers*
  - ⋆ Lluís Màrquez, *Partial vs. Full Parsing*
  - ⋆ B. Coppola and A. Moschitti, *SRL decomposed in four steps*
  - ⋆ Antal van den Bosch, *The effect of Levenshtein-based argument labeling correction*
  - ⋆ Charles Sutton, *Experiments on Reranking Parse Trees with a SRL system*
  - ⋆ Nancy McCracken, *The effect of increasing the amount of training data*
  - ⋆ Wen-Lian Hsu, *Argument Classifier Combination*
  - ⋆ X. Carreras, *System Combination, or willing to see 80% in WSJ test*

Conclusion

- In CoNLL-2004, SRL systems working on partial parsing achieved **$\mathbf{F}_1$ at ~70** in performance.
- This year:
  - ⋆ We considered full syntactic parses
  - ⋆ We enlarged the training data (5 times more)
- 19 systems contributed, achieving **$\mathbf{F}_1$ at ~80**.

Conclusion

- We also evaluated on a portion of the Brown Corpus:
  - ⋆ SRL performance went down below $\sim$ **70**
  - ⋆ All the analyzers in the pipeline suffered a big drop
- Is our current NLP pipeline really robust?

Open Questions

- What syntactic structures are needed for SRL?
  Current approach :
  - ⋆ Use many SRL system working on different syntactic structures
  - ⋆ Combine SRL systems
- How semantic resources (e.g., WordNet) should be used?

Thank you very much for your attention!