Semantic Role Labeling
CS 224N
Christopher Manning

Slides mainly from a tutorial from **Scott Wen-tau Yih** and **Kristina Toutanova** (Microsoft Research), with additional slides from **Sameer Pradhan** (BBN) as well as **Dan Jurafsky** and myself.

1

Syntactic Variations versus Semantic Roles

Yesterday, Kristina hit Scott with a baseball

Scott was hit by Kristina yesterday with a baseball

Yesterday, Scott was hit with a baseball by Kristina

With a baseball, Kristina hit Scott yesterday

Yesterday Scott was hit by Kristina with a baseball

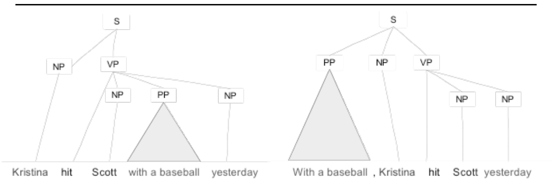
The baseball with which Kristina hit Scott yesterday was hard

Kristina hit Scott with a baseball yesterday

Agent, hitter | Patient, Thing hit | Instrument | Temporal adjunct

2

Syntactic Variations (as trees)



3

Semantic Role Labeling – *Giving Semantic Labels to Phrases*

- [$_{\text{AGENT}}$ John] **broke** [$_{\text{THEME}}$ the window]
- [$_{\text{THEME}}$ The window] **broke**
- [$_{\text{AGENT}}$Sotheby's] .. **offered** [$_{\text{RECIPIENT}}$ the Dorrance heirs] [$_{\text{THEME}}$ a money-back guarantee]
- [$_{\text{AGENT}}$ Sotheby's] **offered** [$_{\text{THEME}}$ a money-back guarantee] to [$_{\text{RECIPIENT}}$ the Dorrance heirs]
- [$_{\text{THEME}}$ a money-back guarantee] **offered** by [$_{\text{AGENT}}$ Sotheby's]
- [$_{\text{RECIPIENT}}$ the Dorrance heirs] will [$_{\text{ARM-NEG}}$ not] be **offered** [$_{\text{THEME}}$ a money-back guarantee]

4

Some typical semantic roles

Thematic Role	Definition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The beneficiary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

5

Some typical semantic roles

Thematic Role	Example
AGENT	*The waiter* spilled the soup.
EXPERIENCER	*John* has a headache.
FORCE	*The wind* blows debris from the mall into our yards.
THEME	Only after Benjamin Franklin broke *the ice*...
RESULT	The French government has built a *regulation-size baseball diamond*...
CONTENT	Mona asked *"You met Mary Ann at a supermarket?"*
INSTRUMENT	He turned to poaching catfish, stunning them *with a shocking device*...
BENEFICIARY	Whenever Ann Callahan makes hotel reservations *for her boss*...
SOURCE	I flew in *from Boston*.
GOAL	I drove *to Portland*.

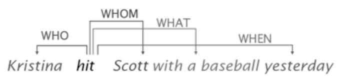
6

What is SRL good for?
Question Answering

Q: What was the name of the first computer system that defeated Kasparov?
A: [PATIENT Kasparov] was defeated by [AGENT Deep Blue] [TIME in 1997].

Q: When was Napoleon defeated?
Look for: [PATIENT Napoleon] [PRED **defeat-synset**] [ARGM-TMP *ANS*]

More generally:



- Who hit Scott with a baseball?
- Whom did Kristina hit with a baseball?
- What did Kristina hit Scott with?
- When did Kristina hit Scott with a baseball?

What is SRL good for?
Applications as a simple meaning rep'n

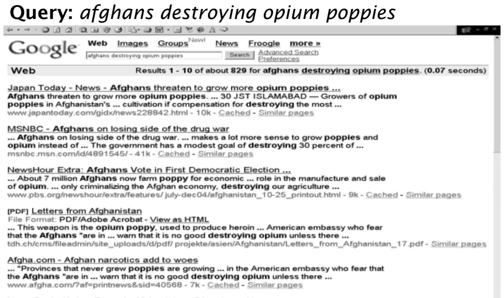
- Machine Translation

English (SVO)	Farsi (SOV)
[AGENT The little boy]	[AGENT pesar koocholo] boy-little
[PRED **kicked**]	[THEME toop germezi] ball-red
[THEME the red ball]	[ARGM-MNR moqtam] hard-adverb
[ARGM-MNR hard]	[PRED **zaad-e**] hit-past

- Document Summarization
  - Predicates and Heads of Roles summarize content
- Information Extraction
  - SRL can be used to construct useful rules for IE

Application: Semantically precise search

Query: *afghans destroying opium poppies*



Diathesis alternations

John *broke the window.*
AGENT THEME
John *broke the window with a rock.*
AGENT THEME INSTRUMENT
The rock *broke the window.*
INSTRUMENT THEME
The window broke.
THEME
The window was broken by John.
THEME AGENT

Doris *gave the book to Cary.*
AGENT THEME GOAL
Doris *gave Cary the book.*
AGENT GOAL THEME

Problems with semantic roles

- It's very hard to produce a formal definition of a role
- There are all sorts of arbitrary role splits
- Intermediary instruments (1-2) vs. enabling instruments (3-4):
  1. The cook opened the jar with the new gadget
  2. The new gadget opened the jar
  3. Sally ate the sliced banana with a fork
  4. *The fork ate the sliced banana

Solutions to the difficulty of defining semantic roles

- Ignore semantic role labels, and just mark arguments of individual verbs as 0, 1, 2
  - PropBank
- Define semantic role labels for a particular semantic domain
  - FrameNet

PropBank

- A corpus of labeled sentences (Penn Treebank *WSJ*)
- The arguments of each verb are labeled with numbers rather than names, though there are verb frame files:

(19.29) **agree.01**
Arg0: Agreer
Arg1: Proposition
Arg2: Other entity agreeing
Ex1: [$_{Arg0}$ The group] *agreed* [$_{Arg1}$ it wouldn't make an offer unless it had Georgia Gulf's consent].
Ex2: [$_{ArgM-TMP}$ Usually] [$_{Arg0}$ John] *agrees* [$_{Arg2}$ with Mary] [$_{Arg1}$ on everything].

(19.30) **fall.01**
Arg1: Logical subject, patient, thing falling
Arg2: Extent, amount fallen
Arg3: start point
Arg4: end point, end state of arg1
Ex1: [$_{Arg1}$ Sales] *fell* [$_{Arg4}$ to $251.2 million] [$_{Arg3}$ from $278.7 million].
Ex2: [$_{Arg1}$ The average junk bond] *fell* [$_{Arg2}$ by 4.2%].

13

Proposition Bank (PropBank) Frame Files

- hit.01 "strike"
  - A0: agent, hitter; A1: thing hit; A2: instrument, thing hit by or with

  [$_{A0}$ *Kristina*] ***hit*** [$_{A1}$ *Scott*] [$_{A2}$ *with a baseball*] *yesterday*.

  AM-TMP
  Time

- look.02 "seeming"
  - A0: seemer; A1: seemed like; A2: seemed to

  [$_{A0}$ *It*] ***looked*** [$_{A2}$ *to her] like* [$_{A1}$ *he deserved this*].

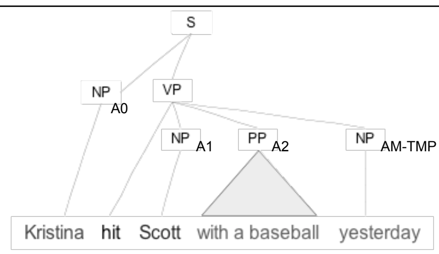
- deserve.01 "deserve"
  - A0: deserving entity; A1: thing deserved; A2: in-exchange-for

  It looked to her like [$_{A0}$ *he*] ***deserved*** [$_{A1}$ *this*].

  Proposition:
  A sentence and a target verb

14

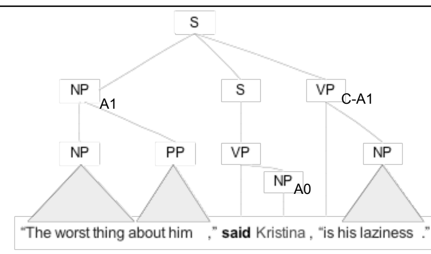
Proposition Bank (PropBank) Add a Semantic Layer



[$_{\mathbf{A0}}$ *Kristina*] ***hit*** [$_{\mathbf{A1}}$ *Scott*] [$_{\mathbf{A2}}$ *with a baseball*] [$_{\mathbf{AM-TMP}}$ *yesterday*].

15

Proposition Bank (PropBank) Add a Semantic Layer – Continued



[$_{\mathbf{A1}}$ *The worst thing about him*] ***said*** [$_{\mathbf{A0}}$ *Kristina*] [$_{\mathbf{C-A1}}$ *is his laziness*].

16

Proposition Bank (PropBank) Final Notes

- Current release (Mar 4, 2005): Proposition Bank I
  - Verb Lexicon: 3,324 frame files
  - Annotation: ~113,000 propositions
    http://www.cis.upenn.edu/~mpalmer/project_pages/ACE.htm
- Alternative format: CoNLL-04,05 shared task
  - Represented in table format
  - Has been used as standard data set for the shared tasks on semantic role labeling
    http://www.lsi.upc.es/~srlconll/soft.html

17

CoNLL format

1. lie("he",…)
2. leak("he", "information obtained from … he supervised")
3. obtain(X, "information", "from a wiretap he supervised")
4. supervise("he", "a wiretap")

He	-	(A0*)	(A0*)	*	*
is	-	*	*	*	*
also	-	*	*	*	*
accused	-	*	*	*	*
of	-	*	*	*	*
lying	lie	(V*)	*	*	*
under	-	(AM-LOC*	*	*	*
oath	-	*)	*	*	*
and	-	*	*	*	*
of	-	*	*	*	*
leaking	leak	*	(V*)	*	*
information	-	*	(A1*	(A1*)	*
obtained	obtain	*	*	(V*)	*
from	-	*	*	*	*
a	-	*	*	(A2*	(A1*
wiretap	-	*	*	*	*)
he	-	*	*	*	(A0*)
supervised	supervise	*	*)	*)	(V*)
.	-	*	*	*	*

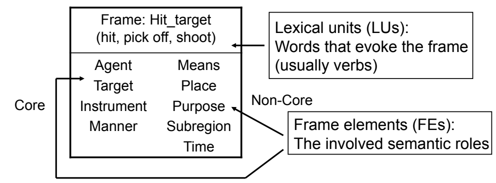
18

Other Corpora

- Chinese PropBank http://www.cis.upenn.edu/~chinese/cpb/
  - Similar to PropBank, it hadds a semantic layer onto Chinese Treebank
- NomBank http://nlp.cs.nyu.edu/meyers/NomBank.html
  - Label arguments that co-occur with nouns in PropBank
  - [$_{A0}$ ***Her***] [$_{REL}$ ***gift***] ***of*** [$_{A1}$ ***a book***] [$_{A2}$ ***to John***]

19

FrameNet [Fillmore et al. 01]



[$_{Agent}$ *Kristina*] ***hit*** [$_{Target}$ *Scott*] [$_{Instrument}$ *with a baseball*] [$_{Time}$ *yesterday*].

http://framenet.icsi.berkeley.edu

20

FrameNet

- A frame is a semantic structure based on a set of participants and events
- Consider the "change_position_on_scale" frame

VERBS:	dwindle	move	soar	escalation	shift
advance	edge	mushroom	swell	explosion	tumble
climb	explode	plummet	swing	fall	
decline	fall	reach	triple	fluctuation	**ADVERBS:**
decrease	fluctuate	rise	tumble	gain	increasingly
diminish	gain	rocket		growth	
dip	grow	shift	**NOUNS:**	hike	
double	increase	skyrocket	decline	increase	
drop	jump	slide	decrease	rise	

21

Roles in this frame

Core Roles	
ATTRIBUTE	The ATTRIBUTE is a scalar property that the ITEM possesses.
DIFFERENCE	The distance by which an ITEM changes its position on the scale.
FINAL_STATE	A description that presents the ITEM's state after the change in the ATTRIBUTE's value as an independent predication.
FINAL_VALUE	The position on the scale where the ITEM ends up.
INITIAL_STATE	A description that presents the ITEM's state before the change in the ATTRIBUTE's value as an independent predication.
INITIAL_VALUE	The initial position on the scale from which the ITEM moves away.
ITEM	The entity that has a position on the scale.
VALUE_RANGE	A portion of the scale, typically identified by its end points, along which the values of the ATTRIBUTE fluctuate.
Some Non-Core Roles	
DURATION	The length of time over which the change takes place.
SPEED	The rate of change of the VALUE.
GROUP	The GROUP in which an ITEM changes the value of an ATTRIBUTE in a specified way.

22

Examples / Quiz question

- [$_{ITEM}$ Oil] *rose* [$_{ATTRIBUTE}$ in price] [$_{DIFFERENCE}$ by 2%].
- [$_{ITEM}$ It] has *increased* [$_{FINAL\_STATE}$ to having them 1 day a month].
- [$_{ITEM}$ Microsoft shares] *fell* [$_{FINAL\_VALUE}$ to 7 5/8].

- [Colon cancer incidence] *fell* [by 50%] [among men].
- a steady *increase* [from 9.5] [to 14.3] [in dividends]
- a [5%] [dividend] *increase*

- Give the roles for the 3 items with blanks
  - Core: Attribute, Difference, Final_State, Final_Value, Initial_State, Initial_Value, Item, Value_Range
    - An Item has a scalar Attribute which moves in the Value_Range portion of a scale
  - Some Non-Core: Duration, Speed, Group
    - The Item changes its Attribute within a Group for a certain Duration or at a Speed

23

Problems with FrameNet

- Example sentences are chosen by hand
  - Not randomly selected
  - Complete sentences not labeled
- Since TreeBank wasn't used
  - No perfect parses for each sentence
- Still ongoing (that's good and bad)

24

Some History

- Fillmore 1968: The case for case
  - Proposed semantic roles as a shallow semantic representation

- Simmons 1973:
  - Built first automatic semantic role labeler
    - Based on first parsing the sentence

25

FrameNet vs PropBank -1

FRAMENET ANNOTATION:

$[_{\text{Buyer}}$ Chuck$]$ *bought* $[_{\text{Goods}}$ a car$]$ $[_{\text{Seller}}$ from Jerry$]$ $[_{\text{Payment}}$ for \$1000$]$.

$[_{\text{Seller}}$ Jerry$]$ *sold* $[_{\text{Goods}}$ a car$]$ $[_{\text{Buyer}}$ to Chuck$]$ $[_{\text{Payment}}$ for \$1000$]$.

PROPBANK ANNOTATION:

$[_{\text{Arg0}}$ Chuck$]$ *bought* $[_{\text{Arg1}}$ a car$]$ $[_{\text{Arg2}}$ from Jerry$]$ $[_{\text{Arg3}}$ for \$1000$]$.

$[_{\text{Arg0}}$ Jerry$]$ *sold* $[_{\text{Arg1}}$ a car$]$ $[_{\text{Arg2}}$ to Chuck$]$ $[_{\text{Arg3}}$ for \$1000$]$.

26

FrameNet vs PropBank -2

FRAMENET ANNOTATION:

$[_{\text{Goods}}$ A car$]$ was *bought* $[_{\text{Buyer}}$ by Chuck$]$.

$[_{\text{Goods}}$ A car$]$ was *sold* $[_{\text{Buyer}}$ to Chuck$]$ $[_{\text{Seller}}$ by Jerry$]$.

$[_{\text{Buyer}}$ Chuck$]$ was *sold* $[_{\text{Goods}}$ a car$]$ $[_{\text{Seller}}$ by Jerry$]$.

PROPBANK ANNOTATION:

$[_{\text{Arg1}}$ A car$]$ was *bought* $[_{\text{Arg0}}$ by Chuck$]$.

$[_{\text{Arg1}}$ A car$]$ was *sold* $[_{\text{Arg2}}$ to Chuck$]$ $[_{\text{Arg0}}$ by Jerry$]$.

$[_{\text{Arg2}}$ Chuck$]$ was *sold* $[_{\text{Arg1}}$ a car$]$ $[_{\text{Arg0}}$ by Jerry$]$.

27

Information Extraction versus Semantic Role Labeling

Characteristic	IE	SRL
Coverage	narrow	broad
Depth of semantics	shallow	shallow
Directly connected to application	sometimes	no

- The goal of SRL is to provide a general capability for semantic relation and argument identification, not custom ones for particular applications

28

Overview of SRL Systems

- Definition of the SRL task
  - Evaluation measures
- General system architectures
- Machine learning models
  - Features & models
  - Performance gains from different techniques

29

Subtasks

- Identification: $\boxed{2^{\{1,2,\ldots,m\}} \longmapsto \{NONE, ARG\}}$
  - Very hard task: to separate the argument substrings from the rest in this exponentially sized set
  - Usually only 1 to 9 (avg. **2.7**) substrings have labels ARG and the rest have NONE for a predicate
- Classification: $\boxed{2^{\{1,2,\ldots,m\}} \longmapsto L \setminus \{NONE\}}$
  - Given the set of substrings that have an *ARG* label, decide the exact semantic label
- Core argument semantic role labeling: (easier)
  - Label phrases with core argument labels only. The modifier arguments are assumed to have label NONE.

30

Evaluation Measures

Correct: [$_{A0}$ The queen] **broke** [$_{A1}$ the window] [$_{AM\text{-}TMP}$ yesterday]
Guess: [$_{A0}$ The queen] broke the [$_{A1}$ window] [$_{AM\text{-}LOC}$ yesterday]

Correct	**Guess**
{The queen} →A0	{The queen} →A0
{the window} →A1	{window} →A1
{yesterday} ->AM-TMP	{yesterday} ->AM-LOC
all other → NONE	all other → NONE

- Precision ,Recall, F-Measure {*tp*=1,*fp*=2,*fn*=2} *p*=*r*=*f*=1/3
- Measures for subtasks
  - Identification (Precision, Recall, F-measure) {tp=2,fp=1,fn=1} *p*=*r*=*f*=2/3
  - Classification (Accuracy) acc = .5 (labeling of correctly identified phrases)
  - Core arguments (Precision, Recall, F-measure) {tp=1,fp=1,fn=1} *p*=*r*=*f*=1/2

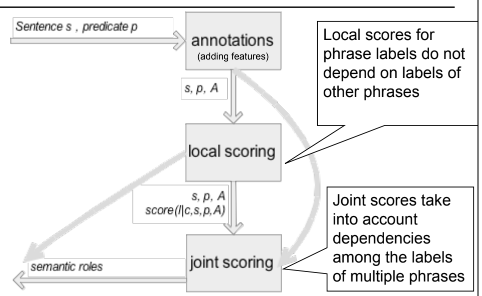
31

What's the problem with these evaluations?

- Approximating human evaluations is dangerous
  - Humans don't always agree
  - Not clear if it's good for anything
  - Sometimes called the "match-a-linguist" task
- What's a better evaluation?

32

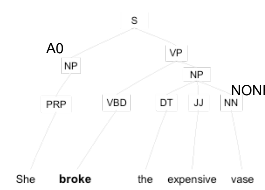
Basic Architecture of a Generic SRL System



33

SRL architecture: Walk the tree, labeling each parse tree node

- Given a parse tree *t*, label the nodes (phrases) in the tree with semantic labels

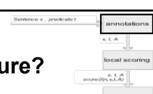


Alternative approach: labeling chunked sentences..

[$_{NP}$Yesterday] , [$_{NP}$Kristina] [$_{VP}$hit] [$_{NP}$Scott] [$_{PP}$with] [$_{NP}$a baseball].

34

Why this parse-tree architecture?



- Semantic role chunks tend to correspond to syntactic constituents
- Propbank:
  - 96% of arguments = 1 (gold) parse tree constitutent
  - 90% of arguments = 1 (Charniak) parse tree constituent
  - Simple rules can recover missing 4-10%

  FrameNet,
  87% of arguments = 1 (Collins) parse tree constituent

  Why?
  they were labeled from parse trees
  by humans trained in syntax

35

Parsing Algorithm

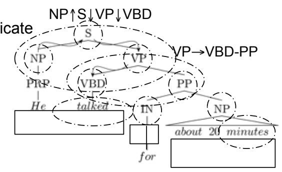
- Use a syntactic parser to parse the sentence
- For each predicate (non-copula verb)
  - For each node in the syntax tree
    - Extract a feature vector relative to the predicate
    - Classify the node
  - Do second-pass informed by global info

Slide from Sameer Pradhan

36

Baseline Features [Gildea & Jurafsky, 2000]

- Predicate (verb)
- Path from constituent to predicate
- Phrase type (syntactic)
- Position (before/after)
- Voice (active/passive)
- Head Word
- Sub-categorization



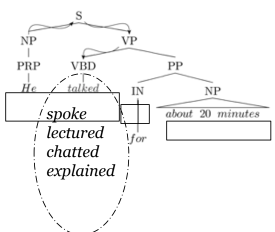
Slide from Sameer Pradhan

Pradhan et al. (2004) Features

- Predicate cluster
- Noun head and POS of PP constituent
- Verb sense
- Partial path
- Named entities in constituent (7) [Surdeanu et al., 2003]
- Head word POS [Surdeanu et al., 2003]
- First and last word in constituent and their POS
- Parent and sibling features
- Constituent tree distance
- Ordinal constituent position
- Temporal cue words in constituent
- Previous 2 classifications

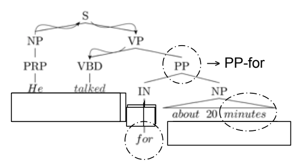
Slide from Sameer Pradhan

Predicate cluster, automatic or WordNet



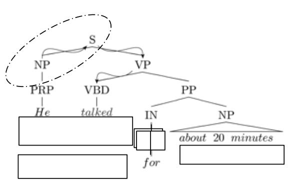
Slide from Sameer Pradhan

Noun Head and POS of PP



Sameer Pradhan

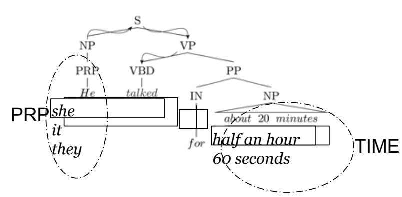
Partial Path



Sameer Pradhan

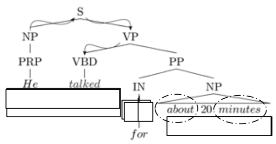
Named Entities and Head Word POS

[Surdeanu et al., 2003]

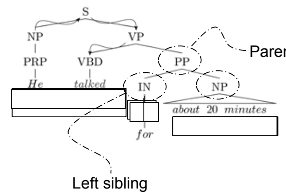


Sameer Pradhan

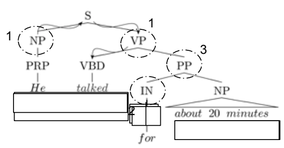
First and Last Word and POS



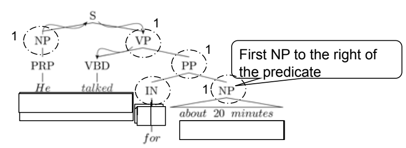
Parent and Sibling features



Constituent tree distance



Ordinal constituent position

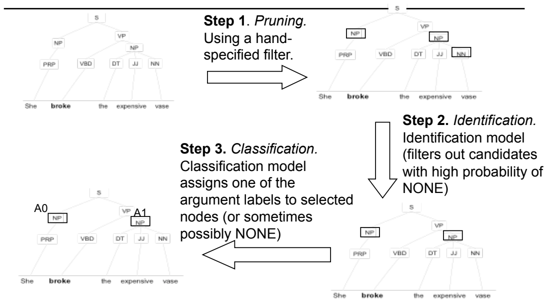


Temporal Cue Words (~50)

time	years;ago
recently	night
days	hour
end	decade
period	late

Previous 2 classifications

Combining Identification and Classification Models



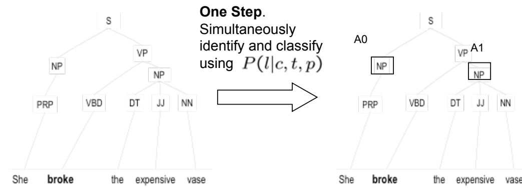
49

Combining Identification and Classification Models – Continued

$P(l|c,t,p) = P_{ID}(Id(l)|\Phi(c,t,p)) * P_{CLS}(l|Id(l),\Phi(c,t,p))$

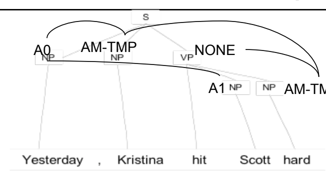
or

$P(l|c,t,p) = P(l|\Phi(c,t,p))$



50

Joint Scoring Models



- These models have scores for a whole labeling of a tree (not just individual labels)
  - Encode some dependencies among the labels of different nodes

$$P_{JOINT}(l_1, \ldots, l_n | n, t, p)! = \prod_i P(l_i | n_i, t, p)$$

51

Combining Local and Joint Scoring Models

- Tight integration of local and joint scoring in a **single probabilistic model** and exact search [Cohn&Blunsom 05] [Màrquez et al. 05],[Thompson et al. 03]
  - When the joint model makes strong independence assumptions
- **Re-ranking** or approximate search to find the labeling which maximizes a combination of local and a joint score [Gildea&Jurafsky 02] [Pradhan et al. 04] [Toutanova et al. 05]
  - Usually exponential search required to find the exact maximizer
- Exact search for **best assignment by local model satisfying hard joint constraints**
  - Using Integer Linear Programming [Punyakanok et al 04,05] (worst case NP-hard)

52

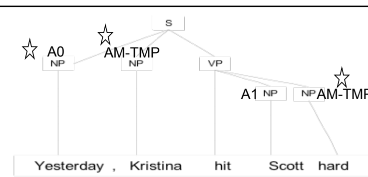
Joint Scoring: Enforcing Hard Constraints

- Constraint 1: Argument phrases do not overlap

  By [$_{A1}$ working [$_{A1}$ hard] , he] ***said*** *, you can achieve a lot.*
  - Pradhan et al. (04) – greedy search for a best set of non-overlapping arguments
  - Toutanova et al. (05) – exact search for the best set of non-overlapping arguments (dynamic programming, linear in the size of the tree)
  - Punyakanok et al. (05) – exact search for best non-overlapping arguments using integer linear programming
- Other constraints ([Punyakanok et al. 04, 05])
  - no repeated core arguments (good heuristic)
  - phrases do not overlap the predicate

53

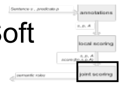
Joint Scoring: Integrating Soft Preferences



- There are many statistical tendencies for the sequence of roles and their syntactic realizations
  - When both are before the verb, AM-TMP is usually before A0
  - Usually, there aren't multiple temporal modifiers
  - Many others which can be learned automatically

54

Joint Scoring: Integrating Soft Preferences



- Gildea and Jurafsky (02) – a smoothed relative frequency estimate of the probability of frame element multi-sets:

  $$P(\{A0, AM_{TMP}, A1, AM_{TMP}\}|hit)$$

  - Gains relative to local model 59.2 → 62.9 FrameNet automatic parses
- Pradhan et al. (04) – a language model on argument label sequences (with the predicate included)

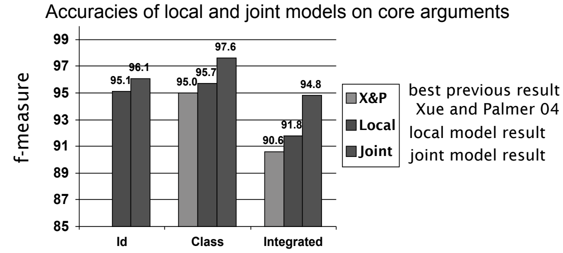
  $$P(A0, AM_{TMP}, hit, A1, AM_{TMP})$$

  - Small gains relative to local model for a baseline system 88.0 → 88.9 on core arguments PropBank correct parses
- Toutanova et al. (05) – a joint model based on CRFs with a rich set of joint features of the sequence of labeled arguments
  - Gains relative to local model on PropBank correct parses 88.4 → 91.2 (24% error reduction); gains on automatic parses 78.2 → 80.0
- Also tree CRFs [Cohn & Brunson] have been used

55

Semantic roles: joint models boost results [Toutanova et al. 2005]

Accuracies of local and joint models on core arguments



Error reduction from best published result:

***44.6%** on Integrated **52%** on Classification*

56

System Properties

- Features
  - Most modern systems use the standard set of Gildea, Pradhan, and Surdeanu features listed above
  - *Lots of features important for building a good system*
- Learning Methods
  - SNoW, MaxEnt, AdaBoost, SVM, CRFs, etc.
  - ➢ *The choice of learning algorithms is less important.*

57

System Properties – Continued

- Syntactic Information
  - Charniak's parser, Collins' parser, clauser, chunker, etc.
  - Top systems use Charniak's parser or some mixture
  - ➢ *Quality of syntactic information is important*
- System/Information Combination
  - Greedy, Re-ranking, Stacking, ILP inference
  - ➢ *Combination of systems or syntactic information is a good strategy to reduce the influence of incorrect syntactic information!*

58

Per Argument Performance

CoNLL-05 Results on WSJ-Test

- Core Arguments (Freq. ~70%)

	Best F_1	Freq.
A0	88.31	25.58%
A1	79.91	35.36%
A2	70.26	8.26%
A3	65.26	1.39%
A4	77.25	1.09%

- Adjuncts (Freq. ~30%)

	Best F_1	Freq.
TMP	78.21	6.86%
ADV	59.73	3.46%
DIS	80.45	2.05%
MNR	59.22	2.67%
LOC	60.99	2.48%
MOD	98.47	3.83%
CAU	64.62	0.50%
NEG	98.91	1.36%

Arguments that need to be improved

Data from Carreras&Màrquez's slides (CoNLL 2005) 59

Summary

- Semantic role labeling
  - An important attempt at a general approach to shallow semantic extraction
- Relatively successful in terms of approximating
  - Human FrameNet labels
  - Human PropBank labels
- Are these good for anything?
  - We don't know for sure yet

60