Evaluation and Parsing

Announcements

- Chapter on dependency parsing moved to courseworks

Today

- Beam Search
- Evaluation
- Classification
- Arc-eager: a different approach

Beam Search

- In greedy search we always took the best option next
- *Might we get stuck?*

e of sentence could cause deterministic pars stuck?

very long sentence

a sentence containing a word that has two or more possible POS tags

a garden path sentencce

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Exploring other options

- Beam search
  - Instead of exploring only the best choice, explores N best choices
  - Form of best-first search, but with a limited “beam” of best choices
  - If the beam = all possible next choices, then we have breadth-first search

Example: Beam of two

The cotton clothing is made of grows.

- [Root] [The… grows]
  SHIFT
- [Root The] [cotton .. grows]
  SHIFT
- [Root The cotton]
  [clothing… grows]
  LEFT ARC
- [Root cotton] [clothing … grows] (Att 1 2)
  SHIFT
- [Root cotton clothing] [is.. Grows]
  LEFT ARC
- [Root clothing] [is … grows] (Att 1 2) (Att 2 3)
- [Root] [The… grows]

  .
  .

  .
  .

  .

  .

  [Root cotton clothing] [is .. Grows]
  SHIFT
- [Root cotton clothing is] [made .. Grows] (Att 1 2)

Parsing Algorithm

- Input: $\{w_t\}_{t \varepsilon T}$, sentence x of length m, beam width K
- Beam: (c,s)εC x R organized by score (s)
- Output: arcs representing a dependency tree for x

1. B <- Beam ({(c_0,0), K)
2. While c.β ≠ [] for some (c,s)εβ
   1. B′ <- Beam ({},K)
   2. For (c,s) ε β, for t ε LEGAL (c),
      1. B'.push(t(c), s + $\text{score}_{x(t|c))}$
   3. B <- β'
3. Return c^*.A where c^* <- β.pop().

Evaluation (and Training)

- Penn Treebank
  - Early 90's
  - Developed at Univ of Pennsylvania, Linguistics Data Consortium
  - 40,000 training, 2400 test
  - Wall Street Journal
- Treebank-3
  - http://catalog.ldc.upenn.edu/LDC99T42
- Original version
  - http://catalog.ldc.upenn.edu/LDC95T7

Example sentence

- **`WSJ/12/WSJ_1273.MRG, sentence 11`**
- Because the CD had an effective yield of 13.4 % when it was issued in 1984 , and interest rates in general had declined sharply since then , part of the price Dr. Blumenfeld paid was a premium -- an additional amount on top of the CD 's base value plus accrued interest that represented the CD 's increased market value .

Slide from Dragomir Radev

Parsed sentence

```
(S
  (SBAR-PRP
    (IN Because)
    (S
      (S
        (NP-SBJ (DT the) (NNP CD))
        (VP
          (VBD had)
          (NP
            (NP (DT an) (JJ effective) (NN yield))
            (PP (IN of) (NP (CD 13.4) (NN %))))
          (SBAR-TMP
            (WHADVP-4 (WRB when))
            (S
              (NP-SBJ-1 (PRP it))
              (VP
                (VBD was)
                (VP
                  (VBN issued)
                  (NP (-NONE- *-1))
                  (PP-TMP (IN in) (NP (CD 1984)))
                  (ADVP-TMP (-NONE- *T*-4)))))))))
                     ...
```

Slide from Dragomir Radev

```
(S
  (SBAR-PRP
    (IN Because)
    (S
      (S
        (NP-SBJ (DT the) (NNP CD))
        (VP
          (VBD had)
          (NP
            (NP (DT an) (JJ effective) (NN yield))
            (PP (IN of) (NP (CD 13.4) (NN %))))
          (SBAR-TMP
            (WHADVP-4 (WRB when))
            (S
              (NP-SBJ-1 (PRP it))
              (VP
                (VBD was)
                (VP
                  (VBN issued)
                  (NP (-NONE- *-1))
                  (PP-TMP (IN in) (NP (CD 1984)))
                  (ADVP-TMP (-NONE- *T*-4))))))))
      (, ,)
      (CC and)
      (S
        (NP-SBJ
          (NP (NN interest) (NNS rates))
          (PP (IN in) (ADJP (JJ general))))
        (VP
          (VBD had)
          (VP
            (VBN declined)
            (ADVP-MNR (RB sharply))
            (PP-TMP (IN since) (NP (RB
then))))))))
```

```
(, ,)
(NP-SBJ
  (NP (NN part))
  (PP
    (IN of)
    (NP
      (NP (DT the) (NN price))
      (SBAR
        (WHNP-3 (-NONE- 0))
        (S
          (NP-SBJ (NNP Dr.) (NNP Blumenfeld))
          (VP (VBD paid) (NP (-NONE- *T*-3)))))))))
```

```
(VP
  (VBD was)
  (NP-PRD
    (NP (DT a) (NN premium))
    (: --)
    (NP
      (NP
        (NP (DT an) (JJ additional) (NN amount))
        (PP-LOC
          (IN on)
          (NP
            (NP (NN top))
            (PP
              (IN of)
              (NP
                (NP (DT the) (NNP CD) (POS 's))
                (NN base)
                (NN value))))))
      (CC plus)
      (NP (VBN accrued) (NN interest)))
    (SBAR
      (WHNP-2 (WDT that))
      (S
        (NP-SBJ (-NONE- *T*-2))
        (VP
          (VBD represented)
          (NP
            (NP (DT the) (NNP CD) (POS 's))
            (VBN increased)
            (NN market)
            (NN value)))))))
(. .))
```

Slide from Dragomir Radev

```
(S
  (SBAR-PRP
    (IN Because)
    (S
      (S
        (NP-SBJ (DT the) (NNP CD))
        (VP
          (VBD had)
          (NP
            (NP (DT an) (JJ effective) (NN yield))
            (PP (IN of) (NP (CD 13.4) (NN %))))
          (SBAR-TMP
            (WHADVP-4 (WRB when))
            (S
              (NP-SBJ-1 (PRP it))
              (VP
                (VBD was)
                (VP
                  (VBN issued)
                  (NP (-NONE- *-1))
                  (PP-TMP (IN in) (NP (CD 1984)))
                  (ADVP-TMP (-NONE- *T*-4))))))))
      (, ,)
      (CC and)
      (S
        (NP-SBJ
          (NP (NN interest) (NNS rates))
          (PP (IN in) (ADJP (JJ general))))
        (VP
          (VBD had)
          (VP
            (VBN declined)
            (ADVP-MNR (RB sharply))
            (PP-TMP (IN since) (NP (RB
then))))))))
```

```
(, ,)
(NP-SBJ
  (NP (NN part))
  (PP
    (IN of)
    (NP
      (NP (DT the) (NN price))
      (SBAR
        (WHNP-3 (-NONE- 0))
        (S
          (NP-SBJ (NNP Dr.) (NNP Blumenfeld))
          (VP (VBD paid) (NP (-NONE- *T*-3)))))))))
```

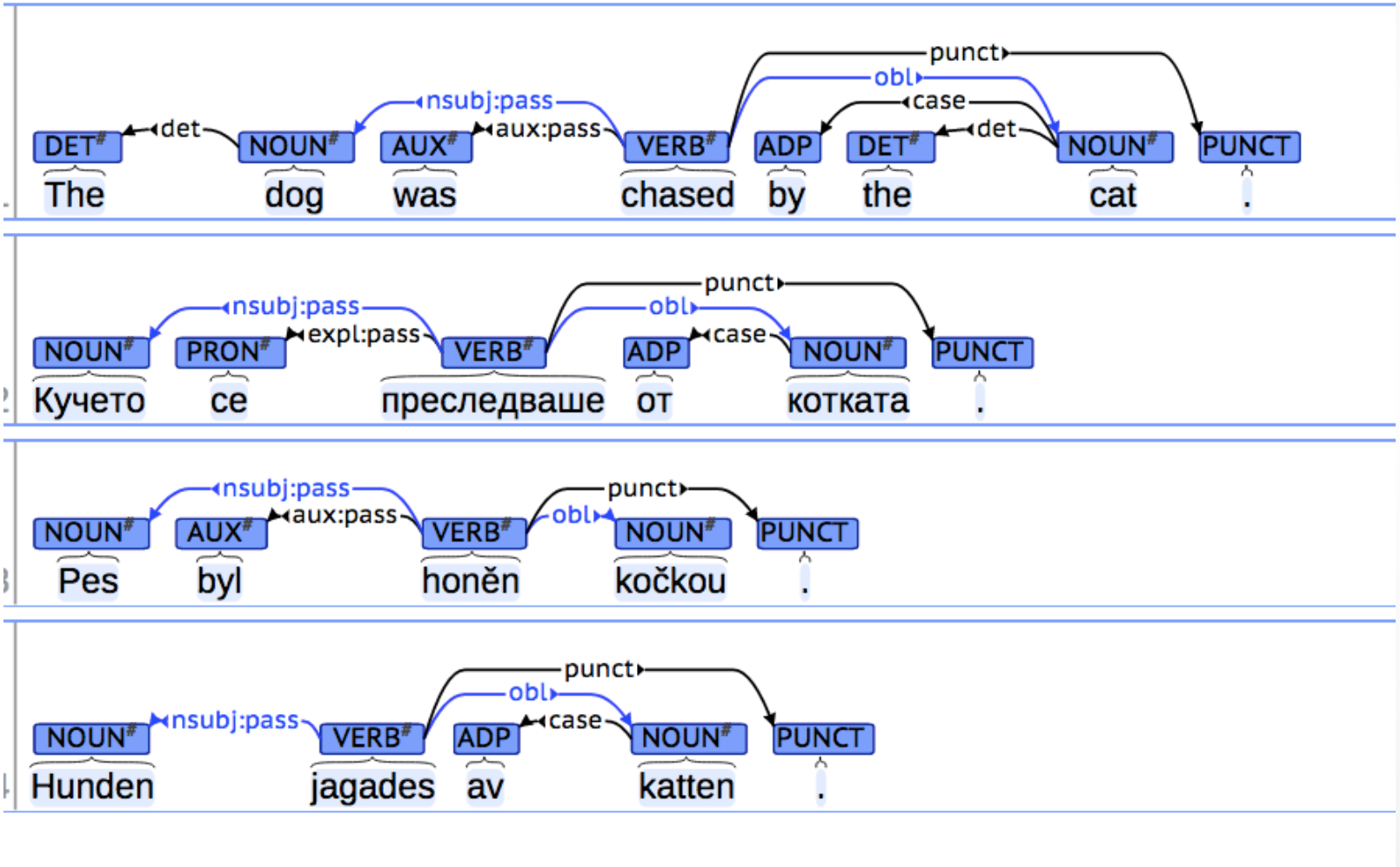
```
(VP
  (VBD was)
  (NP-PRD
    (NP (DT a) (NN premium))
    (: --)
    (NP
      (NP
        (NP (DT an) (JJ additional) (NN amount))
        (PP-LOC
          (IN on)
          (NP
            (NP (NN top))
            (PP
              (IN of)
              (NP
                (NP (DT the) (NNP CD) (POS 's))
                (NN base)
                (NN value))))))
      (CC plus)
      (NP (VBN accrued) (NN interest)))
    (SBAR
      (WHNP-2 (WDT that))
      (S
        (NP-SBJ (-NONE- *T*-2))
        (VP
          (VBD represented)
          (NP
            (NP (DT the) (NNP CD) (POS 's))
            (VBN increased)
            (NN market)
            (NN value)))))))
(. .))
```

Slide from Dragomir Radev

Universal Dependencies

- Cross-linguistically consistent treebank annotation for many languages
- Goal of facilitating multilingual parser development, cross-lingual learning
- Annotation scheme based on (universal) Stanford dependencies (de Marneffe et al., 2006, 2008, 2014), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008)
- http://universaldependencies.org/

Universal Dependencies Example: English, Bulgarian, Czech, Sweedish



Example from universaldependencies.org

Web Treebank – CoNNL-X format

sent_id = email-enronsent28_02-0006

text = He gave no indication on the value of the highest bid.

ID	FORM	LEMMA	CPOS	FPOS	FEATS	HEAD	DEPREL	PHEAD	PDEPREL
1	He	he	PRON	PRP	Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs	2	nsubj	_	_
2	gave	give	VERB	VBD	Mood=Ind|Tense=Past|VerbForm=Fin	0	root	_	_
3	no	no	DET	DT	_	4	det	_	_
4	indication	indication	NOUN	NN	Number=Sing	2	obj	_	_
5	on	on	ADP	IN	_	7	case	_	_
6	the	the	DET	DT	Definite=Def|PronType=Art	7	det	_	_
7	value	value	NOUN	NN	Number=Sing	4	nmod		

Treebanks

- Is it more work to annotate 40K+ sentences than to write a grammar?
- How do we write a grammar?
- Why might it still be advantageous to use a treebank rather than writing a grammar?

ght it be advantageous to use a treebank rath writing a grammar?

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

How do we decide which transition to take?

- He gave no indication on the value of the highest bid

  σ = He gave β = no indication +> SHIFT (ORACLE output)

  What information would help us learn that?

```
2   gave   give   VERB  VBD   Mood=Ind|Tense=Past|VerbForm=Fin   0   root   _   _
3   no     no     DET   DT    _   4   det   _   _
4   indication   indication   NOUN   NN   Number=Sing   2   obj   _   _
```

What information would help us learn this?

How do we decide which transition to take?

- He gave no indication on the value of the highest bid

  σ = He gave no β = indication +> SHIFT (ORACLE output)

  What information would help us learn that?

```
2   gave   give   VERB  VBD  Mood=Ind|Tense=Past|VerbForm=Fin   0   root   _   _
3   no     no     DET   DT   _   4   det   _   _
4   indication   indication   NOUN   NN   Number=Sing   2   obj   _   _
```

How do we decide which transition to take?

- He gave no indication on the value of the highest bid

  σ = He gave no indication β = on +> LEFT ARC (ORACLE output)

  What information would help us learn that?

```
2   gave   give   VERB  VBD   Mood=Ind|Tense=Past|VerbForm=Fin   0   root   _   _
3   no     no     DET   DT    _   4   det   _   _
4   indication   indication   NOUN   NN   Number=Sing   2   obj   _   _
```

How do we decide which transition to take?

- He gave no indication on the value of the highest bid

  σ = He gave indication β = on +> LEFT ARC (ORACLE output) Arcs = ((3 4 ATT))

  What information would help us learn that?

```
2   gave   give   VERB  VBD   Mood=Ind|Tense=Past|VerbForm=Fin   0   root   _   _
3   no     no     DET   DT    _   4   det   _   _
4   indication   indication   NOUN   NN   Number=Sing   2   obj   _   _
```

What information would help us learn this?

Many examples

which gives employees one day a week
gave it that blessed, laudatory lack
give them all a piece of coal
I was given a charge
gives a company an inexpensive way

I bought a car from Fette Ford.
I bought the textbook from Amazon.
Kevin bought the textbook from Book Culture
I bought my mother a gift.

Ebay sells books to students
I sold the Block family my house.

But what about?

- Ate a sandwich
- Painted a picture

Linear Classifier

- Parameters: $w_t \in \mathbb{R}^d$ for each $t \in \mathcal{T}$

- Each $c \in \mathcal{C}$ for sentence x is "featurized" as $\phi^x(c) \in \mathbb{R}^d$.
  - Classical approach: **binary features** providing useful signals
  - Assumes we have access to POS tags of $x_1 \dots x_m$.

$$\phi^x_{20134}(c) := \begin{cases} 1 & \text{if } x_{c.\sigma[0]}.\text{POS} = \texttt{NN} \text{ and } x_{c.\beta[0]}.\text{POS} = \texttt{VBD} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi^x_{1988}(c) := \begin{cases} 1 & \text{if } x_{c.\sigma[0]}.\text{POS} = \texttt{VBD} \text{ with leftmost arc } \texttt{SUBJ} \\ 0 & \text{otherwise} \end{cases}$$

$$\phi^x_{42}(c) := \begin{cases} 1 & \text{if } x_{c.\beta[1]} = \texttt{cat} \\ 0 & \text{otherwise} \end{cases}$$

Our Assignment

- Specify Feature types

- These are converted to binary types with specific words (e.g., if you're looking at a specific word)

Evaluation methodology

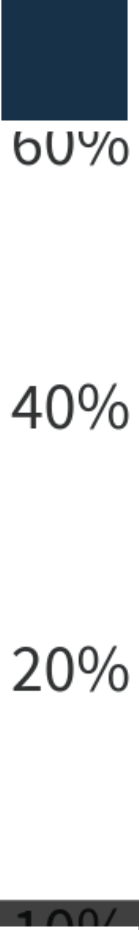
- Train, Dev and Test split
- Baselines
  - Dumb baseline
  - Intelligent baseline
  - Human performance (ceiling)
- New method
- Evaluation methods
  - Accuracy
  - Precision and Recall
- Multiple references
  - Interjudge agreement

Kappa

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

- Agreement vs. expected agreement
  - P(A) is the level of agreement of the judges
  - P(E) is the expected probability of agreement by chance
- When $\kappa > .7$ – agreement is considered high
- Question
  - Judge agreement on a binary classification task is 60%, is this high?

bute Kappa for the case of judge agreement =



Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Scoring: UAS, LAS

- Unlabeled Attachment Score (UAS)

$$\frac{\text{\# words w/correct parent}}{\text{\# words}}$$

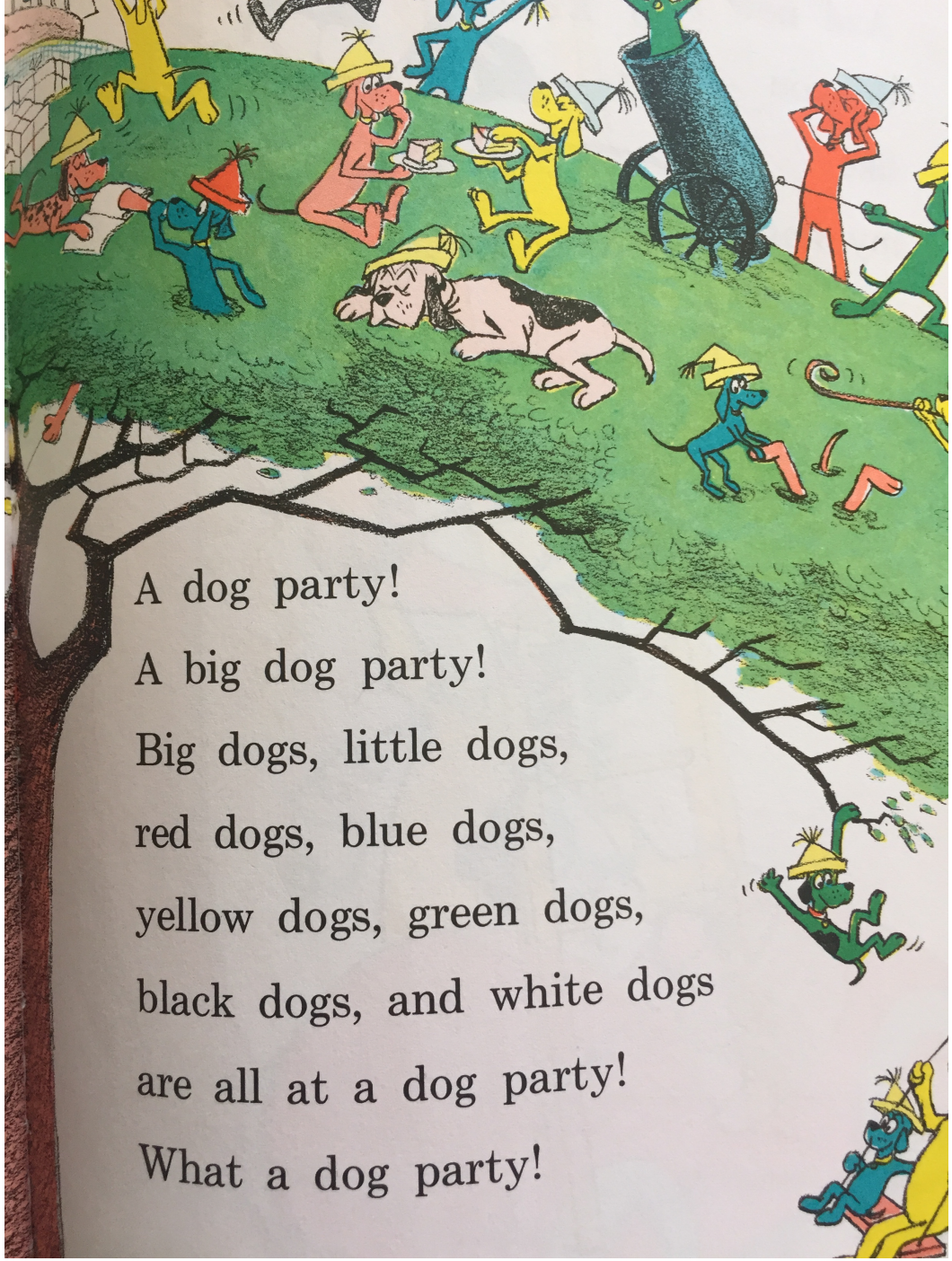
- Labeled Attachment Score (LAS)

$$\frac{\text{\# words w/correct parent and label}}{\text{\# words}}$$

State of the art: 93-95 UAS, 91-93 LAS

Example

- A big dog party



A dog party!
A big dog party!
Big dogs, little dogs,
red dogs, blue dogs,
yellow dogs, green dogs,
black dogs, and white dogs
are all at a dog party!
What a dog party!

Two big dogs

going up.

Labeled Dependency accuracy

- Root A big dog party
  0 1 2 3 4

Gold Labels

1 2 Att

3 4 Att

2 4 Att

4 0 Dobj

Labeled Dependency accuracy

- Root A big dog party
  0 1 2 3 4

Gold Labels	System Output
1 2 Att	1 2 Att
3 4 Att	2 3 Att
2 4 Att	3 4 Att
4 0 Dobj	4 0 Subj

Labeled Dependency accuracy

- Root A big dog party
  0 1 2 3 4

Gold Labels	System Output
1 2 Att	1 2 Att
3 4 Att	2 3 Att
2 4 Att	3 4 Att
4 0 Dobj	4 0 Subj

- Unlabeled Accuracy: #correct/total # = ¾

Labeled Dependency accuracy

- Root A big dog party
  0 1 2 3 4

Gold Labels	System Output
1 2 Att	1 2 Att
3 4 Att	2 3 Att
2 4 Att	3 4 Att
4 0 Dobj	4 0 Subj

- Unlabeled Accuracy: #correct/total # = ¾
  Labeled Accuracy: 2/4 = .5

Representative Performance

- CoNNL-X (2006) shared task provides evaluation numbers for various dependency parsing approaches over 13 languages
  - MALT: LAS from 65-92% depending on language/ treebank

Parser	UAS %
Sagae&Lavie (2005) ensemble of dependency parsers	92.7
Charniak (2000) generative, constituency	92.2
Collins (1999) generative, constituency	91.7
McDonald&Pereira (2005) MST graph-based dependency	91.5
Yamada & Matsumoto (2003) – transition-based dependency	90.4

Slide from Manning

Precision, Recall and F-measure

- F-measure useful when the dataset is unbalanced Informally:
- Precision = $\frac{\text{\# System correct answers}}{\text{\# total system answers}}$

- Recall = $\frac{\text{\# System correct answers}}{\text{\# total correct answers}}$

- F-measure = harmonic mean of precision and recall

$$= 2 \bullet ((p \bullet r)/(p+r))$$

NOTE: Can also weight either P or R more heavily

More formally

- Precision = TP/ (TP + FP)
- Recall = TP/(TP + FN)

  Where TP = true positive, FP = false positive, FN = false negative

- $F_\beta = (1 + \beta^2) \bullet (\text{precision} \bullet \text{recall}) / (\beta^2 \bullet (\text{precision} + \text{recall}))$

- F_1 is the harmonic mean. F_2 weights precision higher. $F_{.5}$ weights recall higher

For the homework

- Precision per sentence = correct_predicated_arcs/predicted_arcs
- Recall per sentence = correct_predicted_arcs/gold_arcs
- F-measure will be F_1
- For the full test set, average of P, R and F

Other alternative approaches

- Deterministic vs beam search
- Arc-standard vs arc-eager
- One pass vs many passes

Arc-standard Problem

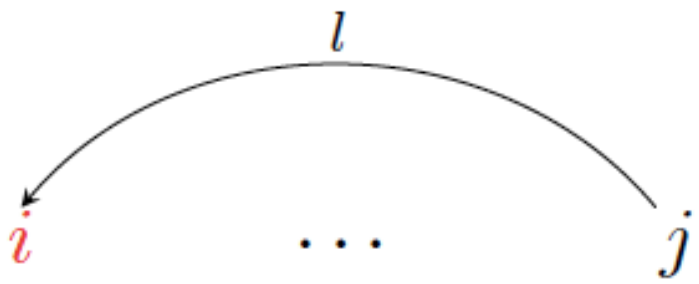
- Right dependents cannot be attached to their head until all their dependents have been attached
- Should a right-arc be taken?

Arc-Eager

- All arcs are added ASAP
- Right arc redefined so that the dependent word is shifted onto the stack
- We add an operation, REDUCE, to pop the dependent word at a later time.

"Eager" Left-Arc

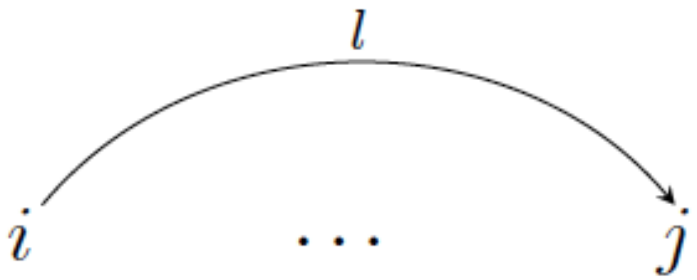
$$\textbf{LEFT}_l^e \quad (\sigma|i, j|\beta, A) \Rightarrow (\sigma, j|\beta, A \cup \{(j, l, i)\})$$



Illegal if either $i = 0$ or i already has a parent.

"Eager" Right-Arc

$$\textbf{RIGHT}_l^e \quad (\sigma|i,j|\beta,A) \Rightarrow (\sigma|i|j,\beta,A \cup \{(i,l,j)\})$$



Illegal if j already has a parent.

Shift and Reduce

SHIFT $(\sigma, i|\beta, A) \Rightarrow (\sigma|i, \beta, A)$

Illegal if β is empty.

REDUCE $(\sigma|i, \beta, A) \Rightarrow (\sigma, \beta, A)$

Illegal if i does not have a parent.

Characteristics of Arc-Eager

- Top-Down
- Complexity: O(n)
- Sound and complete with respect to projective trees

Other resources

- Stanford core NLP
  https://stanfordnlp.github.io/CoreNLP/

- Phrase Structure Parser
- Dependency Parser

- And much more

Next time

- Semantics