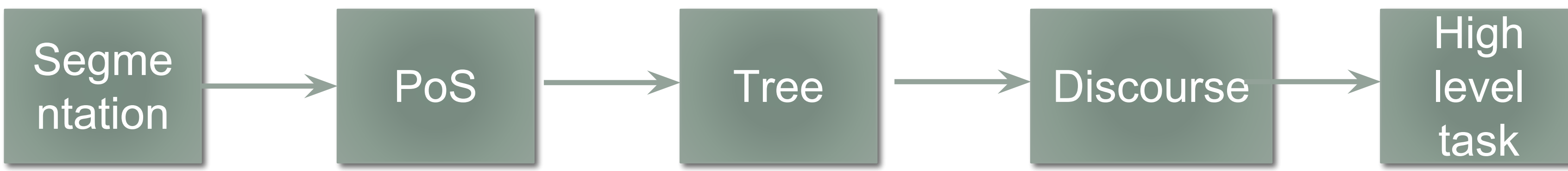
Midterm stuffs

Error propagation and NLP systems

- Typical flow of NLP systems

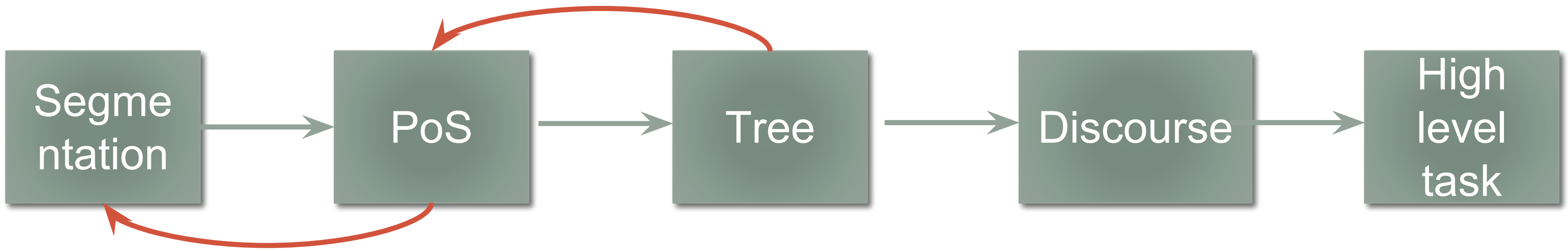
เขาต้องมีหมอนกอด vs เขาต้องมีหมอรกอด



Pipeline analysis

Error propagation and NLP systems

- Use higher level info to help lower level
  - Iterative process



Joint analysis

Error propagation and NLP systems

- Do the task together

Joint analysis

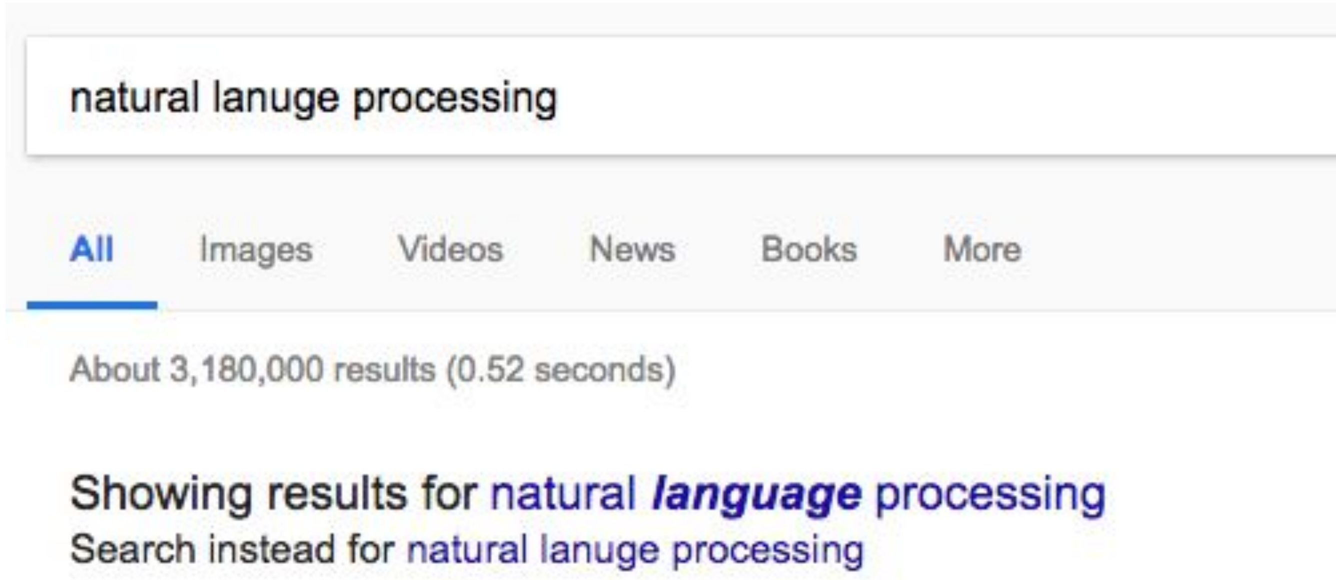
Error propagation and NLP systems

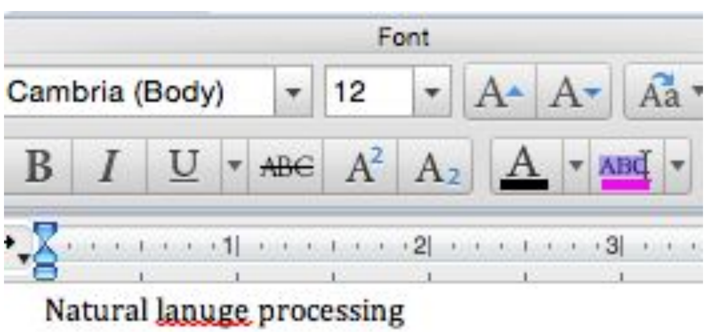
- One single task
  - Input closest to the input form – characters, words

High level task with low level structures as input

End-to-end analysis

Spell correction



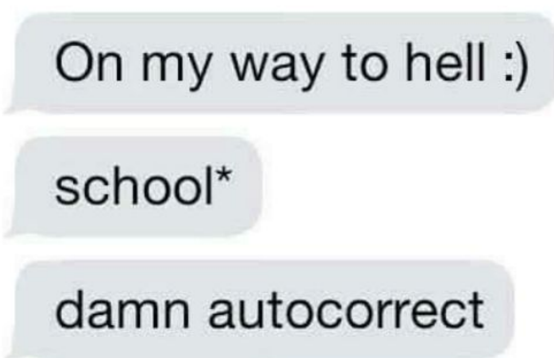


Rates of spelling errors ~1-20% depending on application

Tasks

- Spell error detection
- Spell error correction
  - Suggestion list
  - Suggest a correction
  - Autocorrect

Confidence of the system



Type of spelling errors

- Non-word errors
  - Languaeg -> Language
- Real-word errors
  - Typographical errors
    - Three -> there
    - เขาต้องมีหมอรกอด
  - Cognitive errors (homophones)
    - piece->peace
    - too->two
    - ค่ะ -> คะ

- Dictionary is required.

Corpora

- What kind of text?
  - Pantip, twitter, wikipedia, textbook, …
- Extra: How to actually construct the training corpora?
  - Look at standard sentences and add misspellings
    - By keyboard distance, etc.

Possible solutions

Detecting errors:

- Segmentation failure (word not in dictionary)
- Using prediction model

Correcting errors:

- Simple N-grams
- CBOW (neural based)
- Edit distance

End-to-end

- Seq2seq model (machine transation-based)

A note on dictionary

- Require a large dictionary (millions) to handle real world internet usage
- Automatically generated from unigrams list
  - This list includes misspellings
  - Misspellings occur less frequently than correct words in similar context
- Or let users add new words when not accepting corrections
- Cute solution using NER to ignore NE words if OOV

Whitelaw, C., Hutchinson, B., Chung, G. Y., and Ellis,G. (2009). Using the web for language independent spellchecking and autocorrection

Assumptions

- This only handle one misspelled word per sentence?
  - How to handle multiple missing words in a row?
- Word level correction:
  - Cannot correct OOV words
- Character level correction:
  - Detecting misspellings in the character level might not be able to handle deletion errors

    ฉ ◌ั น ไ ป เ ท ◌ี ◌้ ย ว

    0 0 0 0 0 0 0 0 1 0 0

    ฉ ◌ั น ไ ป เ ท ◌ี ย ว

    0 0 0 0 0 0 0 0 0 0

Assumptions

- Discourse?
  - Only one student actually talks about discourse

A: เดี๋ยวพรุ่งนี้เราต้องไปที่น่าน

B: จะไปที่นะนทำไมละ

นนท์ vs น่าน vs นั่น

Spell correction in literature

1. Non-word
2. Real-word

Non-word spelling error

- Detection:
  - Any word not in dictionary is an error
- Correction:
  1. Generate candidates
  2. Choose the one that has the best
     1. Smallest edit distance
     2. Best noisy channel probability

Real word spelling error

- Detection:
  - ???
- Correction:
  1. Generate candidates
     1. From similar pronunciations
     2. From similar spellings
  2. Choose the one that has the best
     1. Smallest edit distance
     2. Best noisy channel probability

  The criterions has to be context sensitive!

  ไปไหนดีค่ะ -> ไปไหนดีคะ

Candidate generation

- Words with similar spelling
  - Edit distance
  - Typically at most 2 edit distance
  - Can be weighted based on the task, e.g. keyboard
  - Allow merging and splitting of words
    - Ladygaga -> Lady Gaga
    - |การะ|เกด| -> การะเกด
- Words with similar pronunciation
  - Example: Aspell, Soundex – a hash based on sounds
    - Jurafsky, Jarofsky, Jarovsky, and Jarovski all map to J612
  - Modern methods use Letter-2-sound (L2S) rules (also called Grapheme-to-Phoneme G2P)
- How to generate this in an efficient manner?

Candidate generation

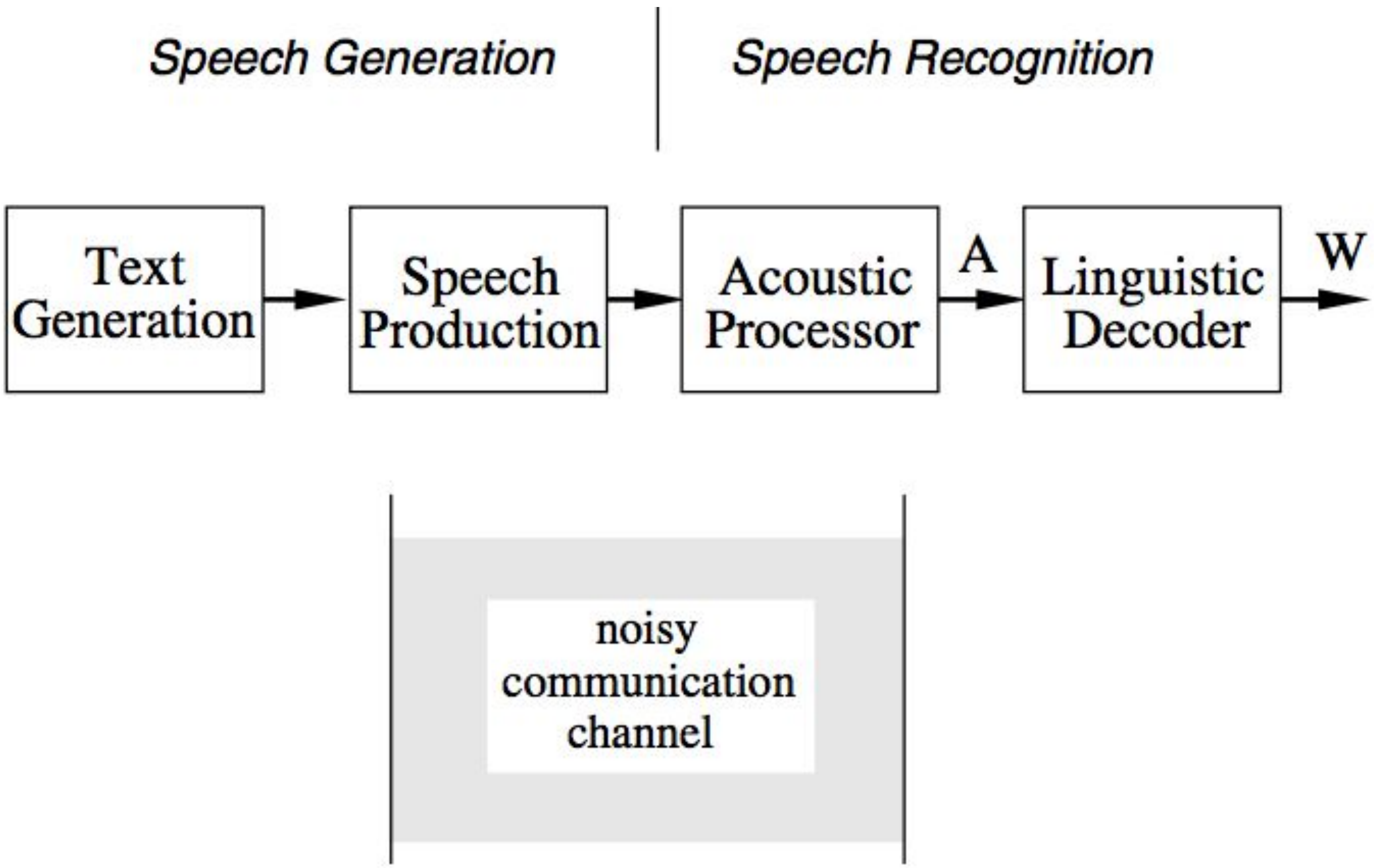
1. Run through dictionary, check edit distance (slow)
2. Generate all words with desired edit distance, then intersect with dictionary
3. Character n-gram inverted index. Find words which shared the most n-gram (fast)
   - Google -> goo oog ogl gle
   - Inverted index: goo -> google good goon good-bye
4. Compute with Finite State Transducer (FST) (fast)
5. Precomputed map of words
6. Mixture (https://github.com/wolfgarbe/SymSpell) (fast)

Tommi A Pirinen. Weighted Finite-State Methods for Spell-Checking and Correction 2014

The noisy channel model

- Used to rank the best candidates

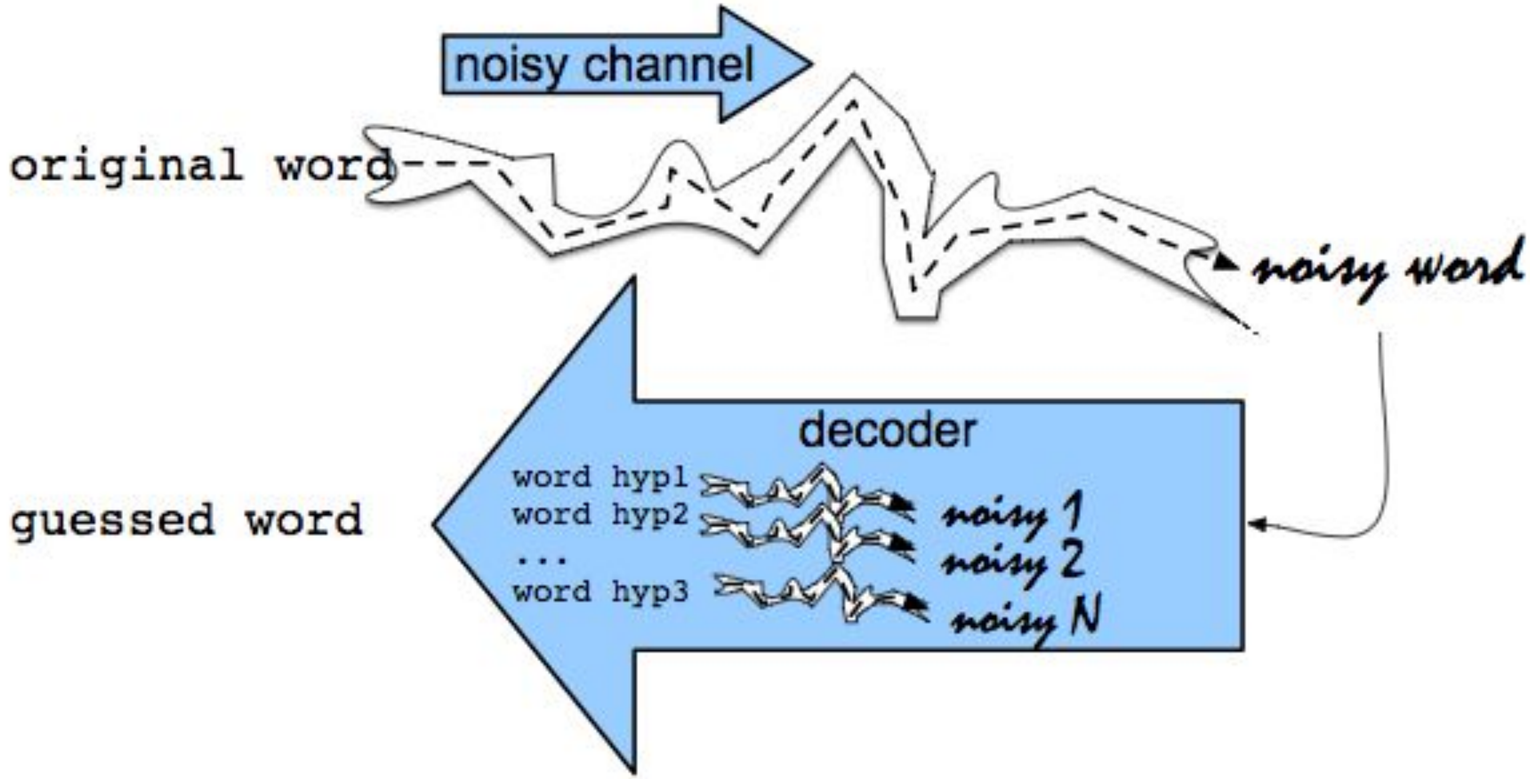
Information theoretic formulation



$$W^* = \arg\max_W P(W \mid A)$$

$$P(W \mid A) = \frac{P(A \mid W)P(W)}{P(A)}$$

The noisy channel intuition



Jurafsky, chapter 5

Bayes rule

- An observation x of the misspelled word
- Find the correct word w* from vocab list V

$$= argmax_{w \in V} P(w|x)$$

$$= argmax_{w \in V} \frac{P(x|w)P(w)}{P(x)}$$

$$= argmax_{w \in V} P(x|w)P(w)$$

Language model

Channel/error model

Noisy channel example

- acress
- Let’s generate candidates

Noisy channel example

Edit distance 1 (Damerau-Levenshtein distance)

Levenshtein edit distance – insertion, deletion, substitution
Damerau-Levenshtein edit distance - insertion, deletion, substitution, transposition

		Transformation			
Error	**Correction**	**Correct Letter**	**Error Letter**	**Position (Letter #)**	**Type**
acress	actress	t	—	2	deletion
acress	cress	—	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	—	s	5	insertion
acress	acres	—	s	4	insertion

Jurafsky, chapter 5

Language model (unigram)

w	count(w)	p(w)
actress	9,321	.0000231
cress	220	.000000544
caress	686	.00000170
access	37,038	.0000916
across	120,844	.000299
acres	12,874	.0000318

Channel/error model

- Easiest to use a confusion matrix conditioned on the previous character.
- Based on counts

$$P(x|w) = \begin{cases} \dfrac{\mathrm{del}[x_{i-1}, w_i]}{\mathrm{count}[x_{i-1} w_i]}, \text{ if deletion} \\ \dfrac{\mathrm{ins}[x_{i-1}, w_i]}{\mathrm{count}[w_{i-1}]}, \text{ if insertion} \\ \dfrac{\mathrm{sub}[x_i, w_i]}{\mathrm{count}[w_i]}, \text{ if substitution} \\ \dfrac{\mathrm{trans}[w_i, w_{i+1}]}{\mathrm{count}[w_i w_{i+1}]}, \text{ if transposition} \end{cases}$$

w_i is the i-th character

Learning the confusion matrix

1. Have a data of misspellings-corrections pair
2. Have a data of misspellings and use EM
    1. Initialize the confusion matrix to be equally likely
    2. E-step: run the spell checker and correct words
    3. M-step: re-estimate the confusion matrix based on corrections

Or use nearby keys
(Each phone need to have different model)

Channel/error model

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)
actress	t	-	c|ct	.000117
cress	-	a	a|#	.00000144
caress	ca	ac	ac|ca	.00000164
access	c	r	r|c	.000000209
across	o	e	e|o	.0000093
acres	-	s	es|e	.0000321
acres	-	s	ss|s	.0000342

Noisy channel model

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)	P(w)	10^9*P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	0.00078
caress	ca	ac	ac|ca	.00000164	.00000170	0.0028
access	c	r	r|c	.000000209	.0000916	0.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Note1: It is typical to use more than unigrams for the LM (very important for real word errors)
Note2: typically added a weight parameter tuned on dev set

$$P(x|w)P(w)^{\alpha}$$

Real word case for noisy channel

- Need to compare candidates against no correction
  - Error model needs a score for no correction
- Set to some parameter, c, representing how errorful our input is

- $P(w|x) = c$ for $w = x$

- The rest of the probability 1-c is weighted by the confusion matrix

Classifier-based methods

- Have a pair of correction and features
  - whether/weather
  - Features:
    1. cloudy in +- 5 words
    2. Followed by “to VERB”
    3. Followed by “or not”

- End-to-end model using neural networks
  - Seq2Seq model (more on this in later lectures)

WISESIGHT Challenge

- Predict sentiment from social media text

Leaderboard

1	NongLAYyy	0.75316	51	1d
2	Nattasit M	0.74556	17	14h
3	poringz	0.74133	32	1d
4	Patiphan W	0.73710	16	14h
5	cstorm3000	0.73372	16	19h
6	sunboy	0.73203	14	13h
7	earthy	0.73119	3	10d
8	Chawan FC	0.73119	17	1d
9	PanTA	0.73119	19	1d
10	nattichai	0.73119	55	13h

1	▲ 4	cstorm3000	0.75968	16	18h
2	▲ 5	earthy	0.75099	3	10d
3	▲ 9	Phakhin Deesee	0.75063	45	6d
4	▲ 4	Chawan FC	0.75063	17	1d
5	▼ 4	NongLAYyy	0.74954	51	1d
6	▲ 13	AneZreT	0.74809	57	13h
7	▲ 16	jumpwmk	0.74629	21	14h
8	▼ 2	sunboy	0.74592	14	13h
9	—	PanTA	0.74448	19	1d
10	▲ 7	Piriyapong Laopongsit	0.74230	5	1d

4th-7th

BoW + classifier

- 3 Dense layers (7th)
- xgboost (6th)
- 3 Dense layer (4th)
- SVC (4th)

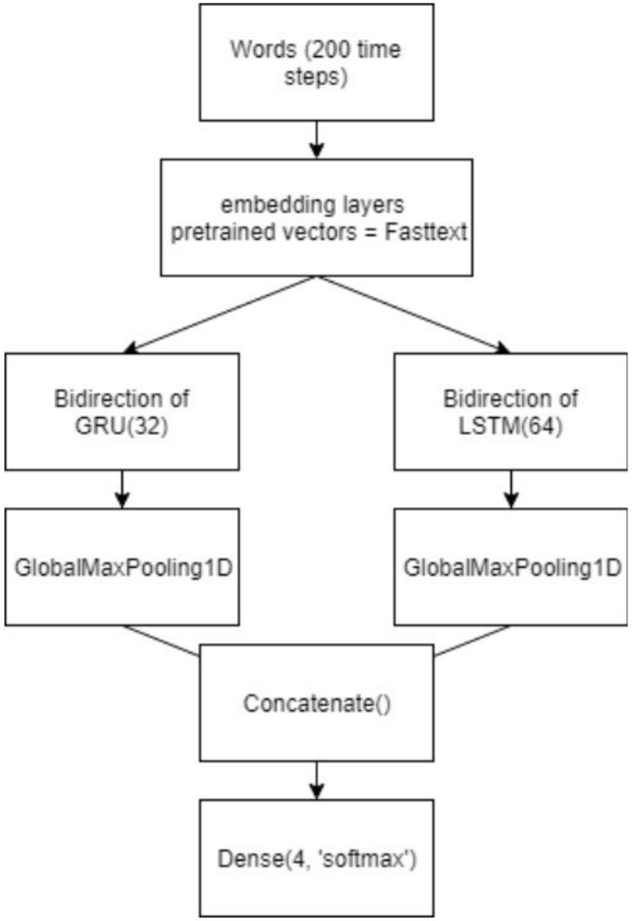
BoW

- Char n-gram+TF-IDF (4th)
- Binarized BoW (7th)

3rd

System combination of

1. Deep model (thai2fit)
2. Deep model (fasttext,
   char embedding for OOV)
3. NBSVM 1
4. NBSVM 2
   word+char n-gram with
   binarized counts and smoothing



1st

Semi-supervised learning

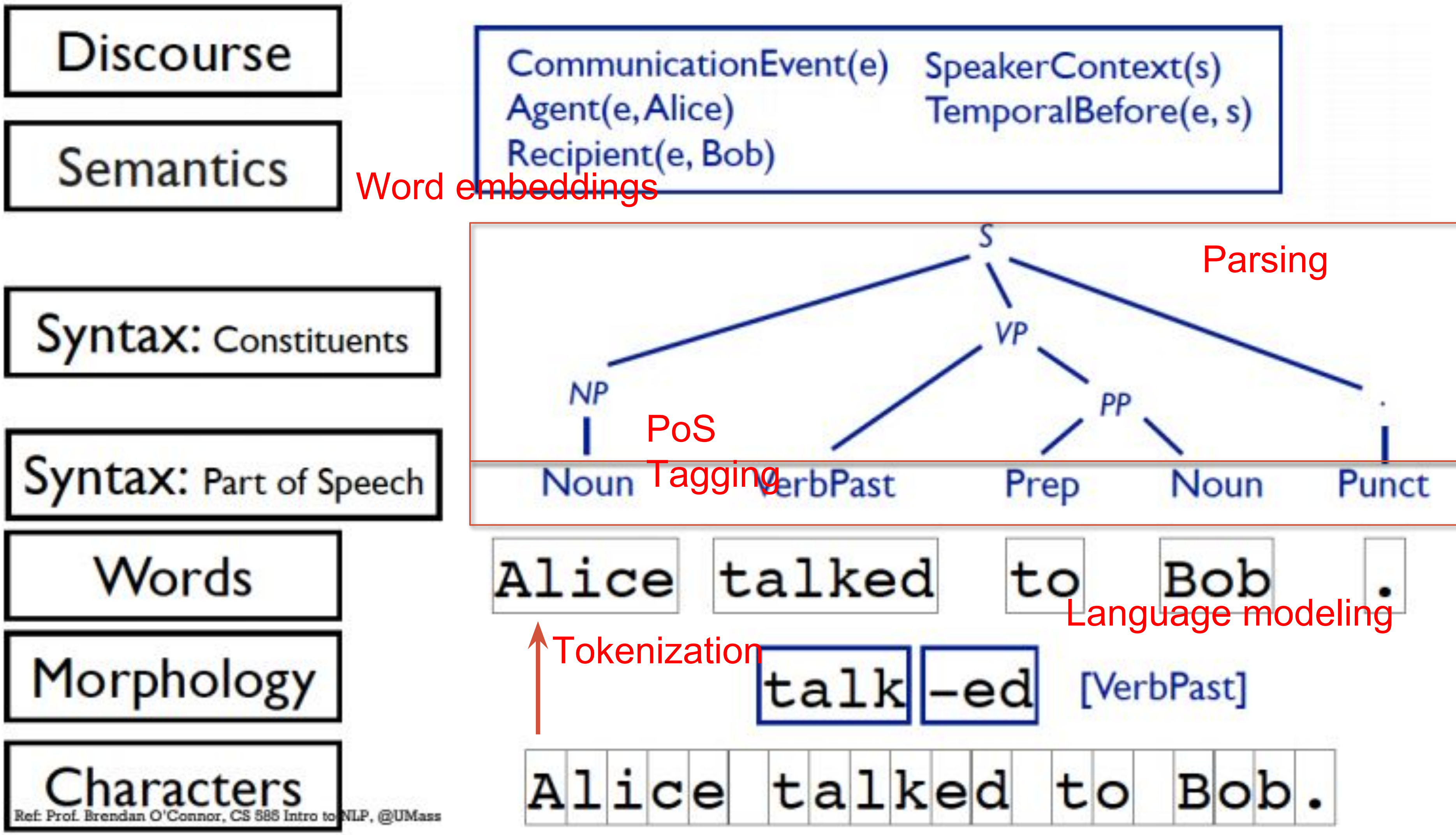
1. logistic regression + tfidf to predict the test set
2. Augment training data with test set and train thai2fit
3. System combination of 1+2

3% improvement on validation set

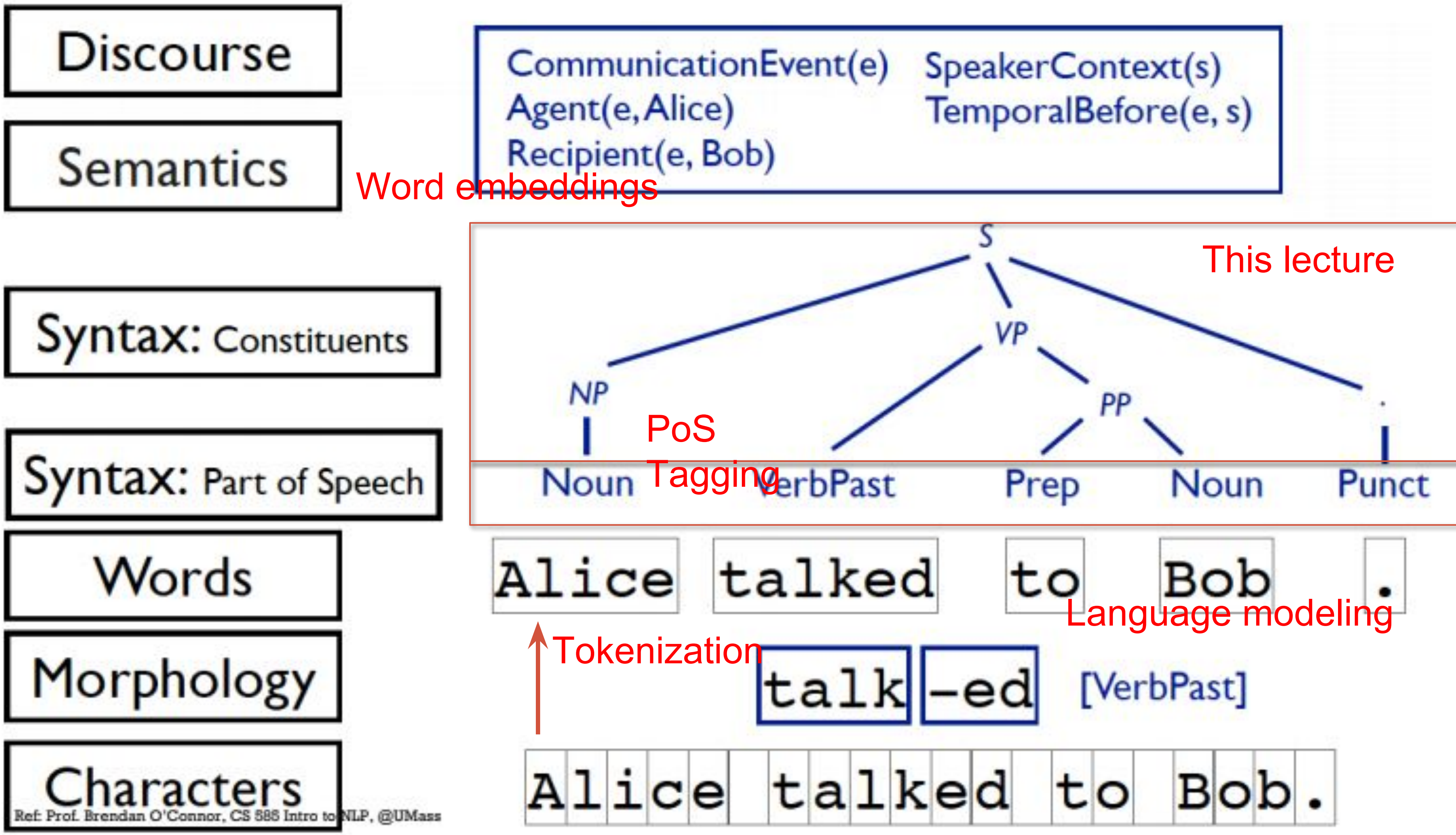
PARSING

PCFG, Recursive Neural Network

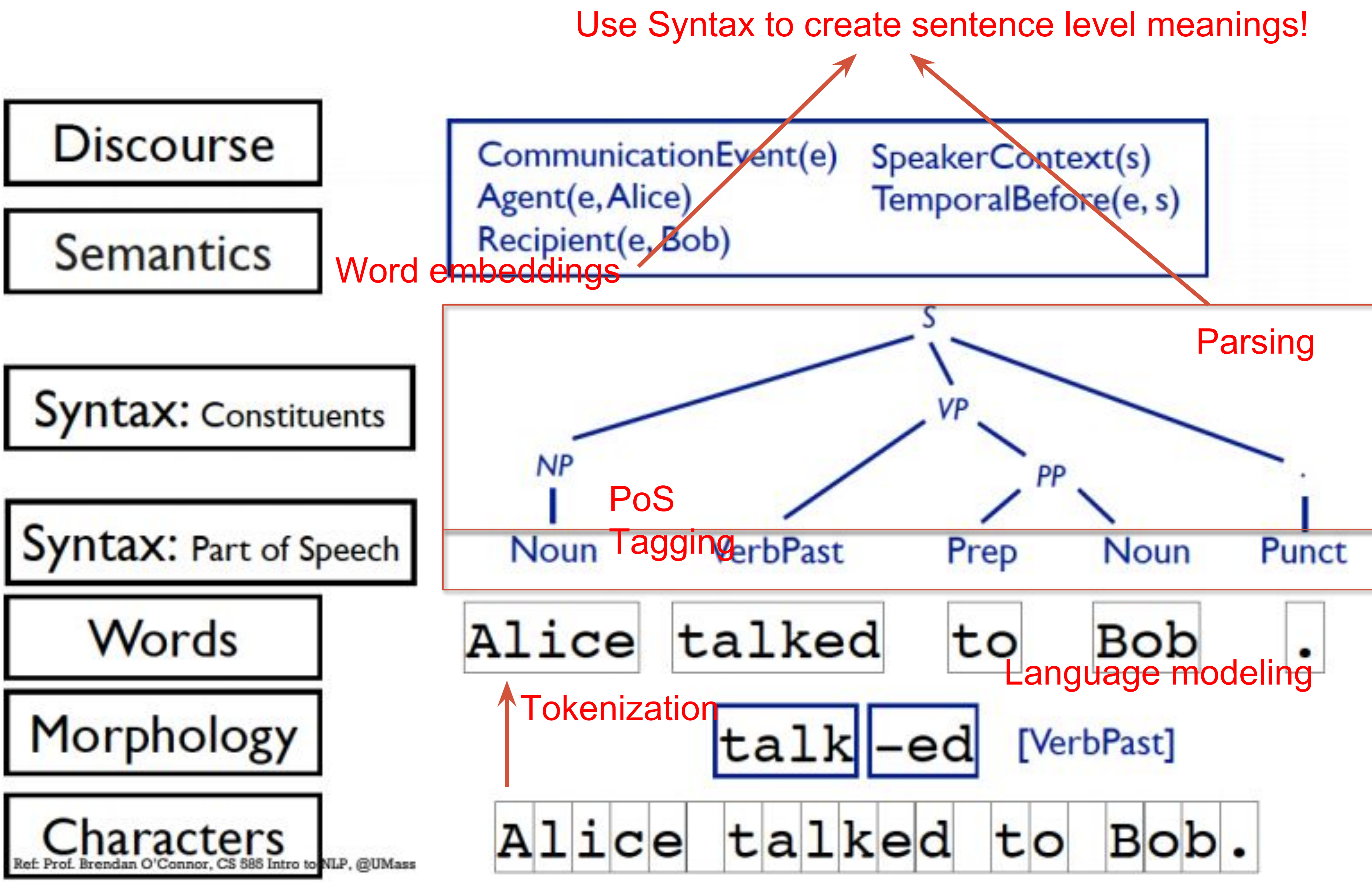
Document classification, sentiment analysis, QA, conversation agents, summarization, translation



Ref: Prof. Brendan O'Connor, CS 585 Intro to NLP, @UMass



Ref: Prof. Brendan O'Connor, CS 585 Intro to NLP, @UMass



Use Syntax to create sentence level meanings!
Discourse
Semantics
Syntax: Constituents
Syntax: Part of Speech
Words
Morphology
Characters
CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)
SpeakerContext(s)
TemporalBefore(e, s)
Word embeddings
S
VP
NP
PP
.
Parsing
PoS
Tagging
Noun
VerbPast
Prep
Noun
Punct
Alice
talked
to
Bob
.
Language modeling
Tokenization
talk
-ed
[VerbPast]
A l i c e t a l k e d t o B o b .
Ref: Prof. Brendan O'Connor, CS 588 Intro to NLP, @UMass

Semantic embeddings of several words

- Compositionality
- We know how to create a dense vector representation for a word
  - What about larger linguistic units? (e.g. phrase, sentence)
- We can combine smaller units into a larger unit

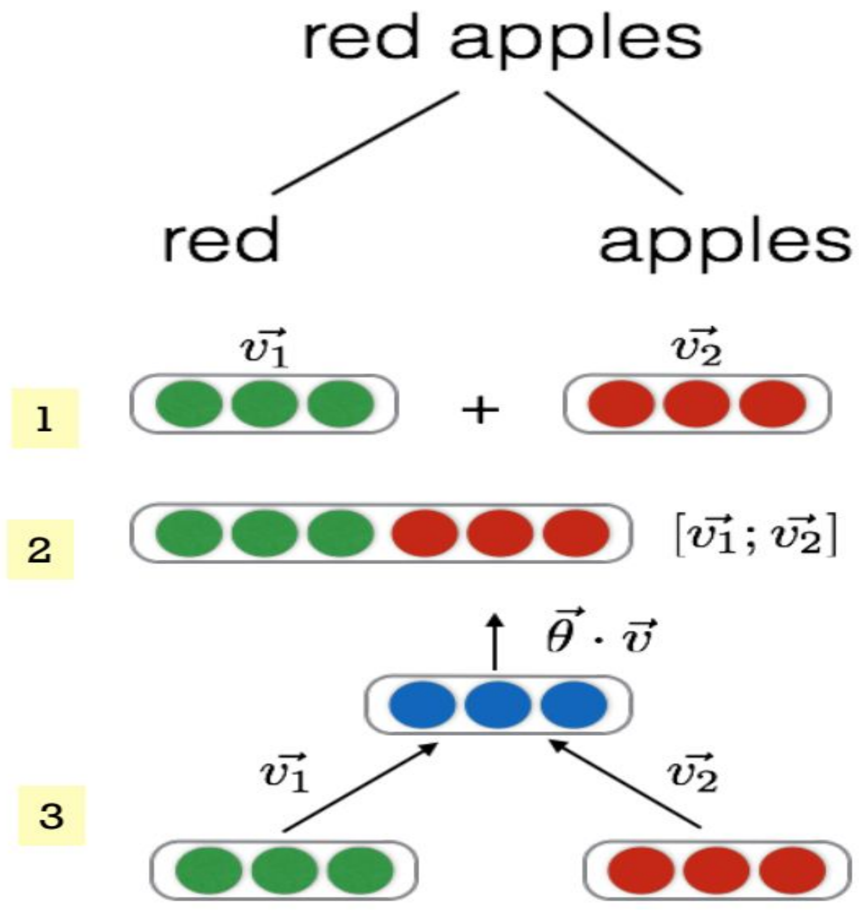


Image ref: Prof. Regina Barzilay , NLP@MIT

Semantic vectors overview

Discourse level

word phrase sentence paragraph document

- Word vectors
  - Co-occurrence
  - PPMI
  - TFIDF
  - Word2vec
    - CBoW
    - Skip-gram

- Doc vectors
  - Term-document
    - Bag of words model
  - Recurrent networks (HW3 ToDo#6)

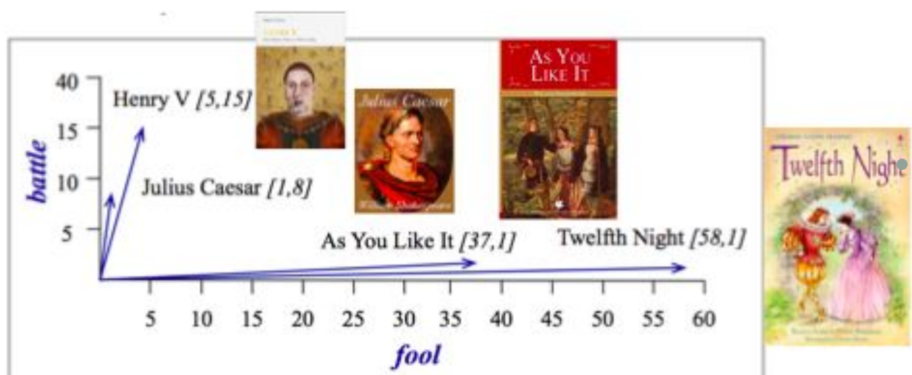
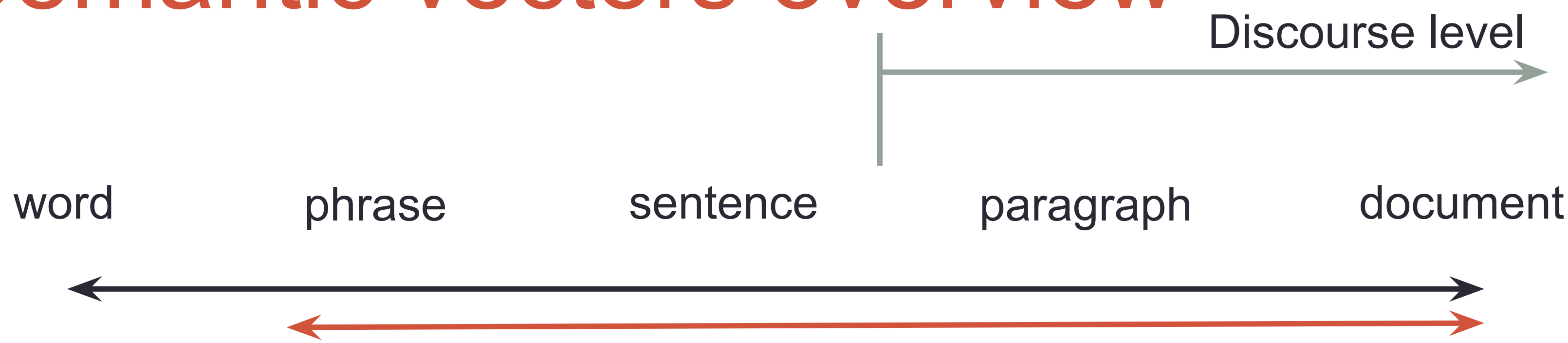


Figure 15.3 A spatial visualization of the document vectors for the four Shakespeare play documents, showing just two of the dimensions, corresponding to the words *battle* and *fool*. The comedies have high values for the *fool* dimension and low values for the *battle* dimension.

Reference: Jurafsky, Dan, and James H. Martin. Speech and language processing. 3rd edition draft, https://web.stanford.edu/~jurafsky/slp3/, August 2017

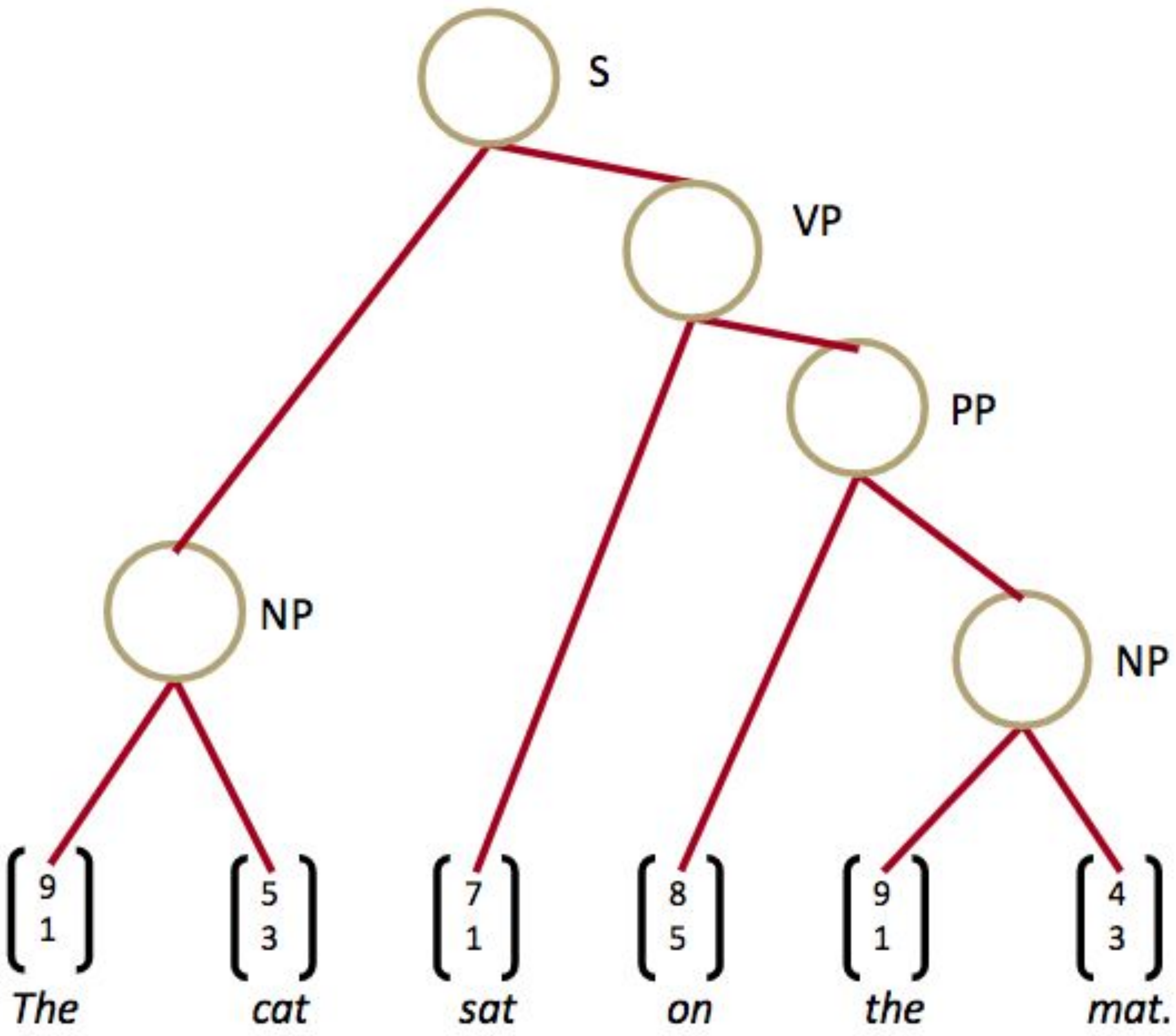
Semantic vectors overview



How do we represent things in this level? Without ignoring word order (bag of words)

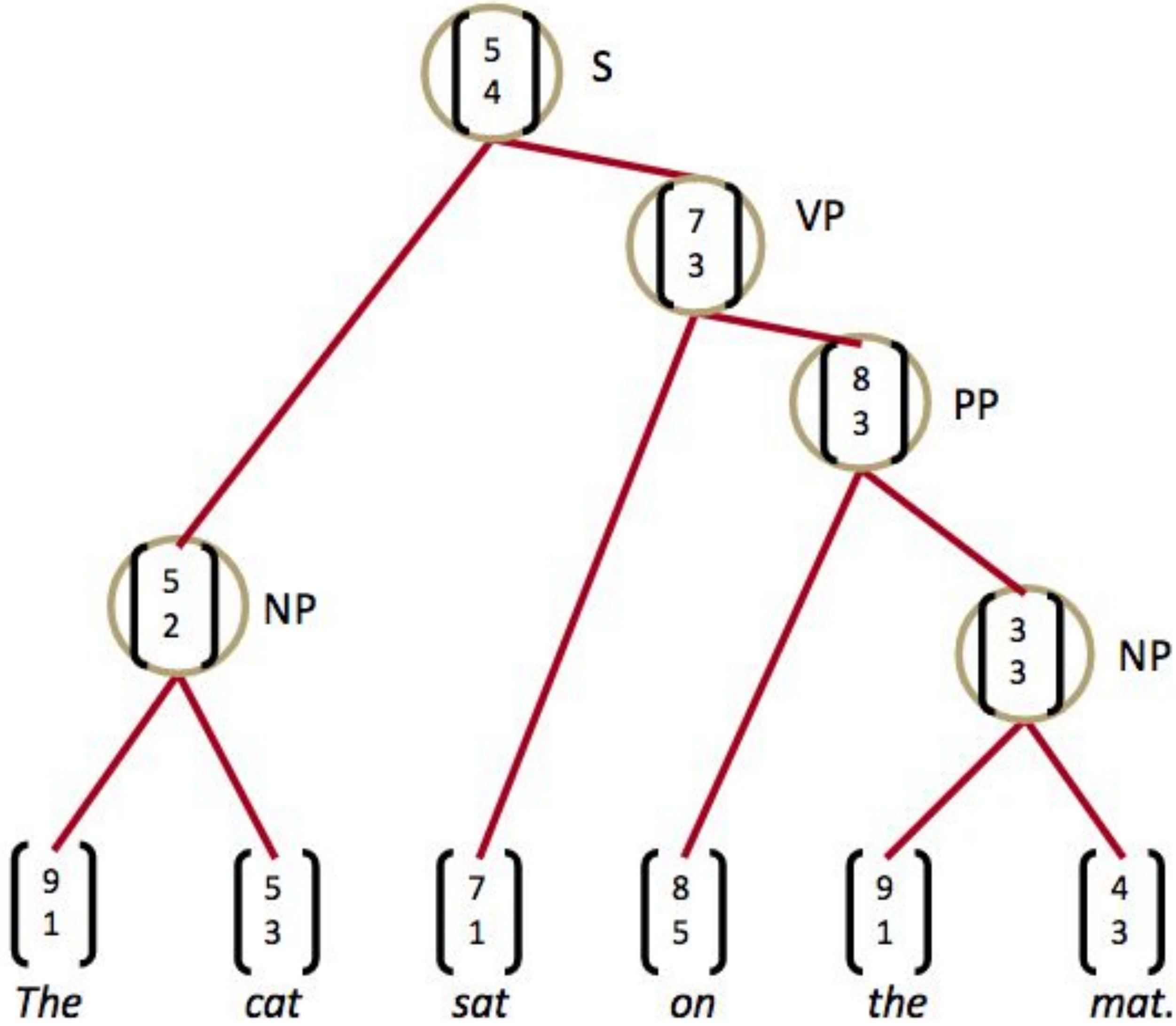
- Word vectors
  - Co-occurrence
  - PPMI
  - TFIDF
  - Word2vec
    - CBoW
    - Skip-gram
- Doc vectors
  - Term-document
    - Bag of words model
    - Doc2Vec
    - LDA2Vec
  - Recurrent networks

Parsing and representation big picture



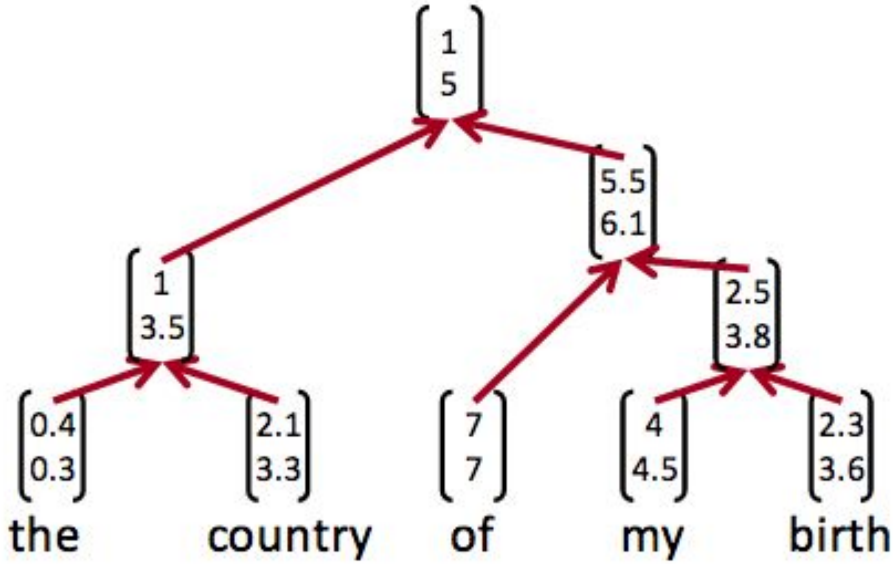
http://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

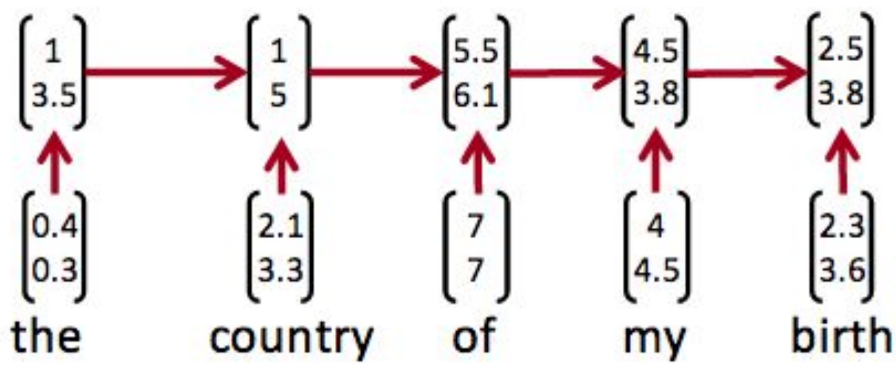
Parsing and representation big picture



http://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

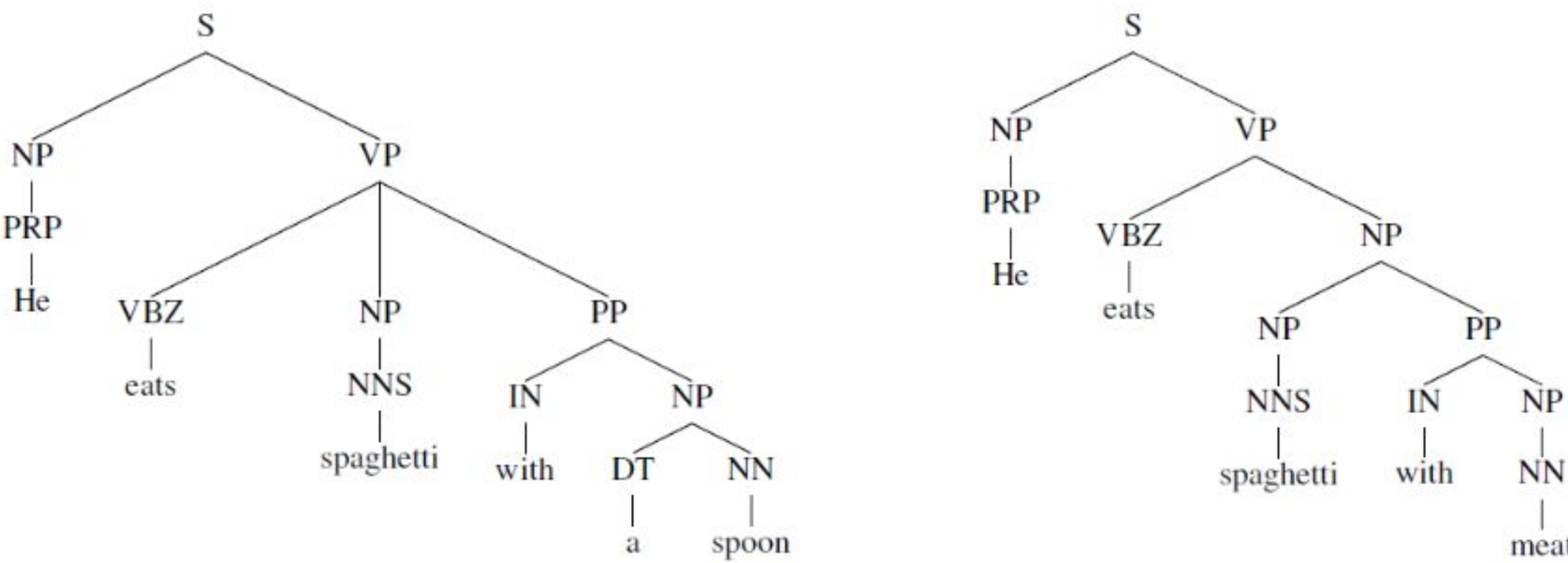
Recursive vs Recurrent representation





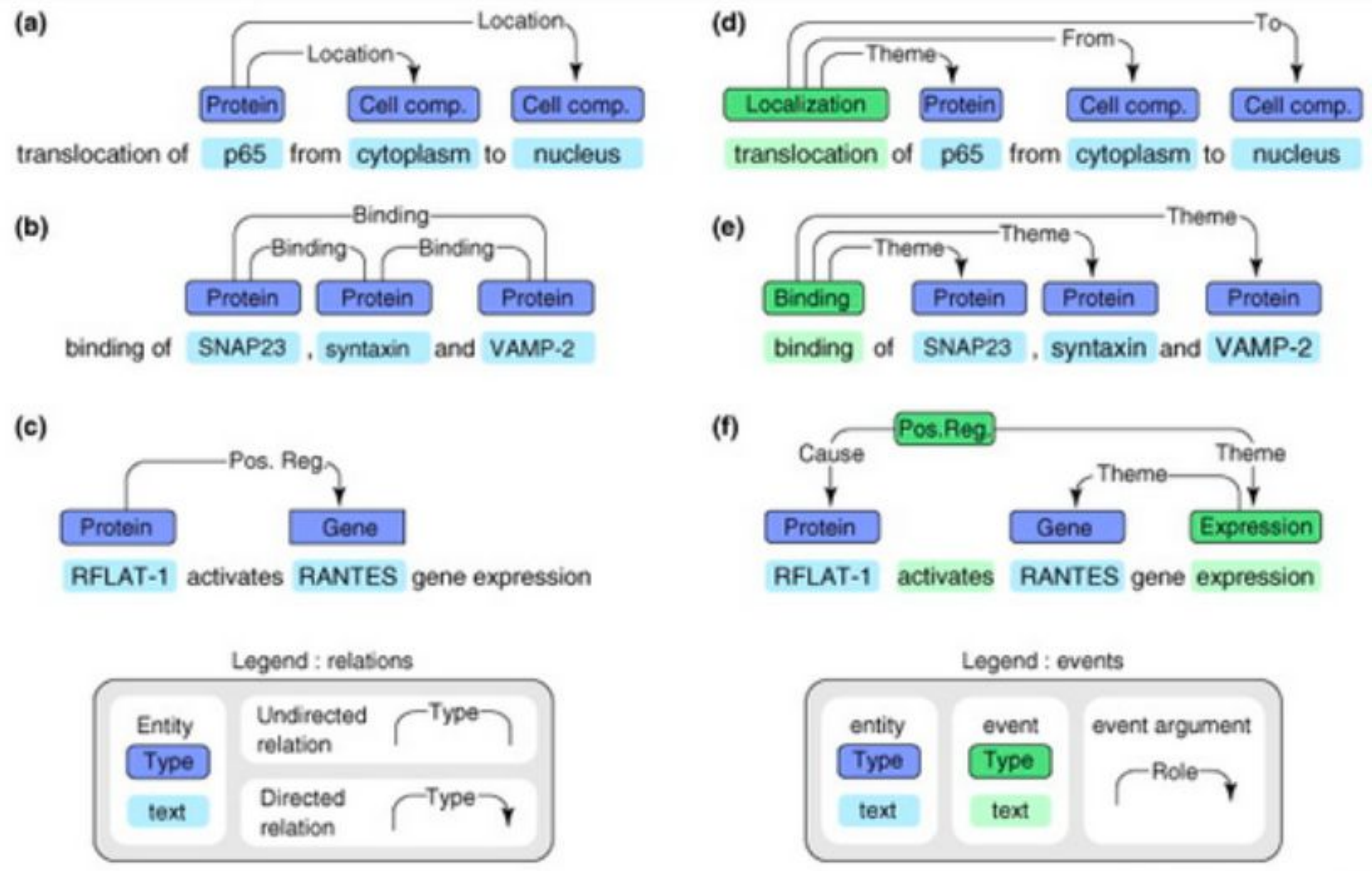
http://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

Cases where recursive might be better?

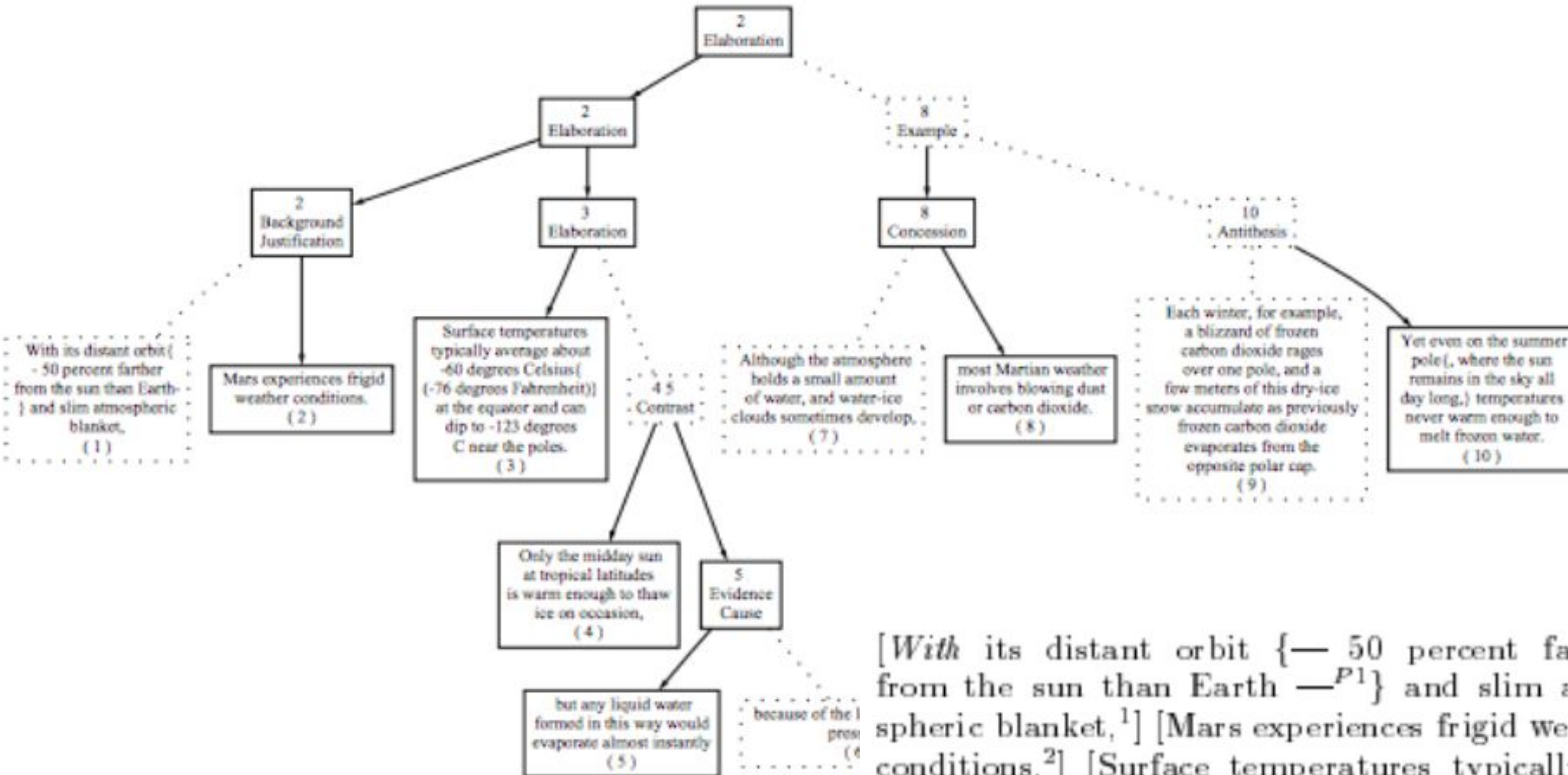


But recurrent structures might be able to learn this too…

Application of parse trees: info extraction



A side note: discourse parsing



[*With* its distant orbit {— 50 percent farther from the sun than Earth —^{P1}} and slim atmospheric blanket,^{1}] [Mars experiences frigid weather conditions.^{2}] [Surface temperatures typically average about −60 degrees Celsius (−76 degrees Fahrenheit) at the equator and can dip to −123 degrees C near the poles.^{3}] [Only the midday sun at tropical latitudes is warm enough to thaw ice

Image ref: Prof. Regina Barzilay , NLP@MIT

Overview

- Types of grammars
  - Context Free Grammar
  - Probabilistic Context Free Grammar
    - CYK parser
  - Dependency Grammar
    - Transition-based parsing
    - Recursive neural networks

Constituents

- Groups of words behaving as a single units
  - Ex: Noun phrase

  Harry the Horse

  The reason he comes into the house

  They

  A high-class spot such as Mindy's
- Checking for constituents
  - See if they can appear in similar syntactic environments

  They sit...

  The reason he comes into the house is…

http://web.stanford.edu/~jurafsky/slp3/11.pdf

Context-Free Grammar (CFG)

- A grammar specifies what kind of parse tree can be generated.
- CFG or Phrase-Structure Grammars assumes the grammar is context-free
  - Most forms of natural language are context-free
    - Thai and English are typically CFG languages
  - Used in many programming languages
- CFG is based on constituents structures

A context-free grammar example

A CFG is defined by
G = (N,S,Σ,R)

- N = {S, NP, VP, PP, DT, Vi, Vt, NN, IN} Nonterminals
- S = S Starting symbol
- Σ = {sleeps, saw, man, woman, telescope, the, with, in}

Terminals

R =

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Production rules
X -> $Y_1 \ldots Y_n$
Y_i can be terminal or nonterminal
The rule only relies on X (no context)
(vs context-sensitive)

Note: S=sentence, VP=verb phrase, NP=noun phrase, PP=prepositional phrase, DT=determiner, Vi=intransitive verb, Vt=transitive verb, NN=noun, IN=preposition

Example ref: Prof. Regina Barzilay, NLP@MIT

Generation using left-most derivation strategy

- Recursive strategy

$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

```
left_most_derivation(A):
    if terminal(A):
        return A          Return if terminal
    rule = choose (rules(A))
    rhs = right_hand_side(rule)
    return concatenate([left_most_derivation(A') for A' in rhs])
```

- To generate we call left_most_derivation(S) (top down)
- A string belongs to the language of a CFG if there exist a sequence of left-most derivation that can generate the string

$$L = \{s \in \Sigma^* | s = left\_most\_derivation(S)\}$$

Example

- The woman saw the man with the telescope

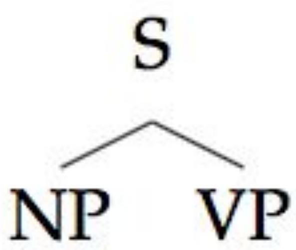
$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Example

- The woman saw the man with the telescope



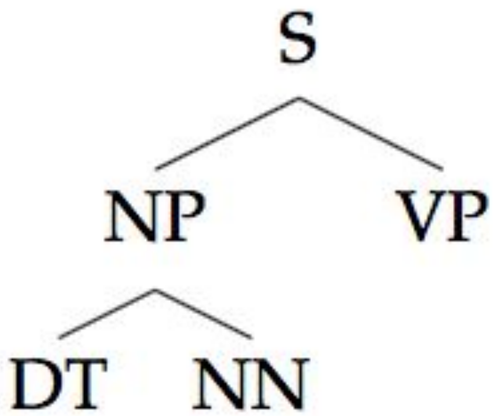
$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Example

- The woman saw the man with the telescope



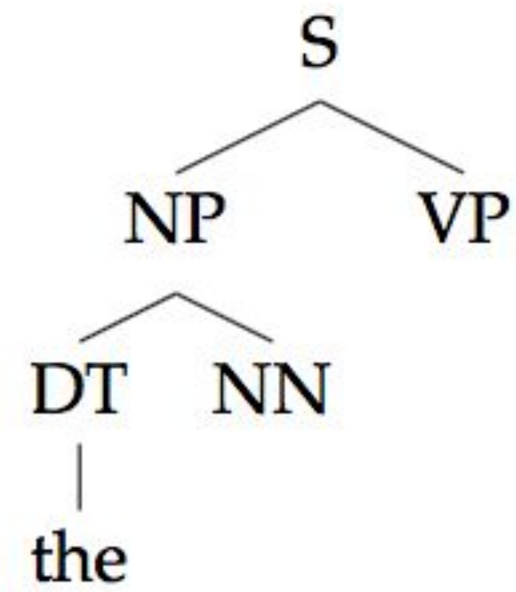
$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Example

- The woman saw the man with the telescope



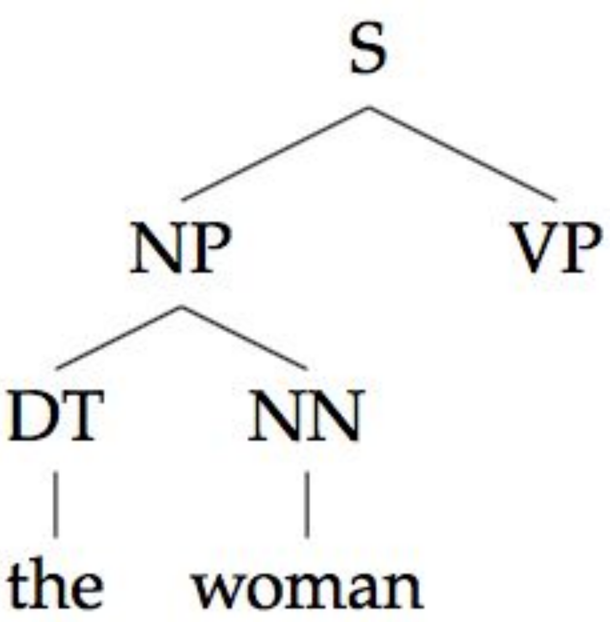
$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Example

- The woman saw the man with the telescope



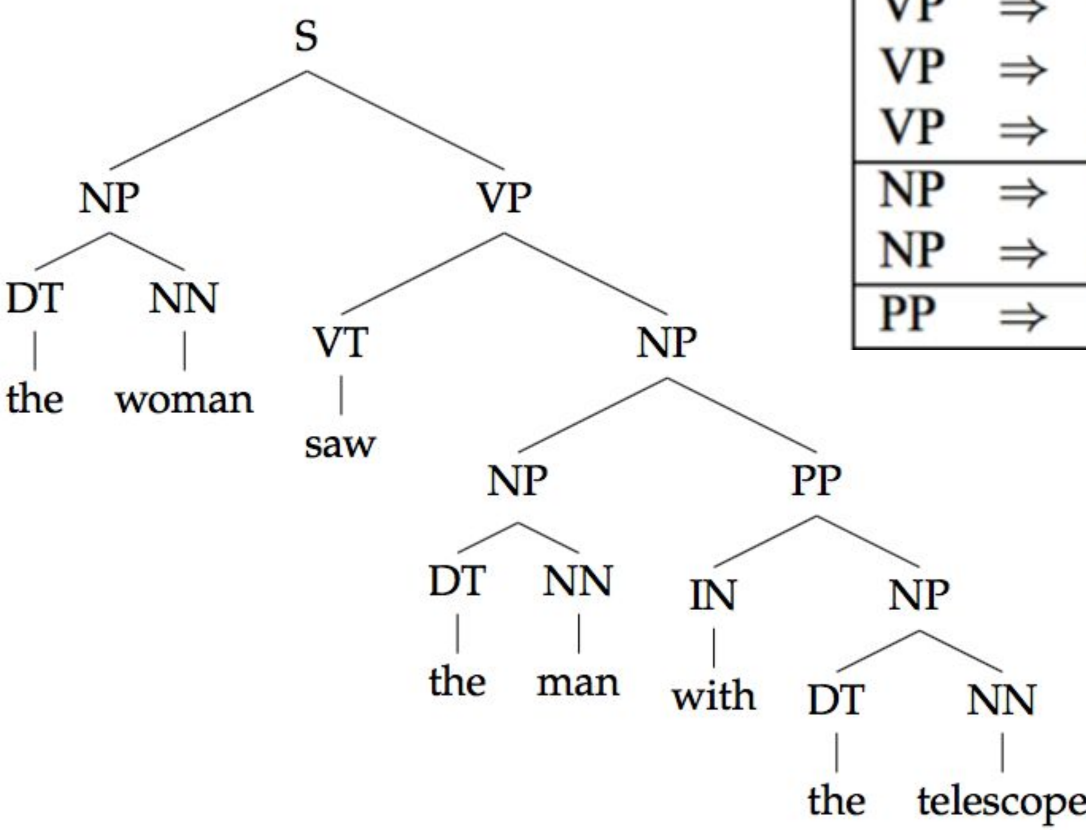
$R =$

S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Example

- The woman saw the man with the telescope



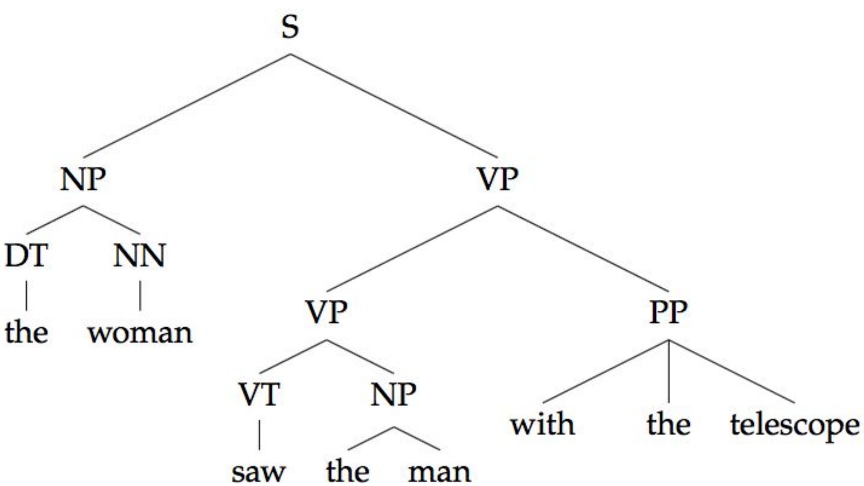
$R =$

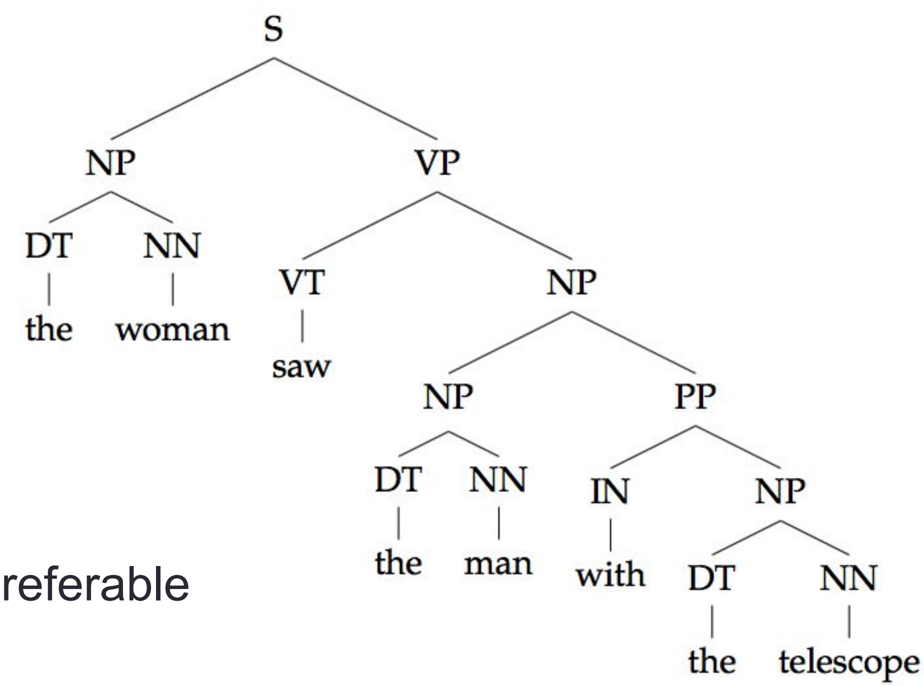
S	$\Rightarrow$	NP	VP
VP	$\Rightarrow$	Vi	
VP	$\Rightarrow$	Vt	NP
VP	$\Rightarrow$	VP	PP
NP	$\Rightarrow$	DT	NN
NP	$\Rightarrow$	NP	PP
PP	$\Rightarrow$	IN	NP

Vi	$\Rightarrow$	sleeps
Vt	$\Rightarrow$	saw
NN	$\Rightarrow$	man
NN	$\Rightarrow$	woman
NN	$\Rightarrow$	telescope
DT	$\Rightarrow$	the
IN	$\Rightarrow$	with
IN	$\Rightarrow$	in

Ambiguities

- There can be multiple derivations for the same string
- These sentences are ambiguous as each parse represents a different meaning





Need a way to say which one is more preferable

Add probabilities!

Probabilistic Context-Free Grammar

- Production rules now have probabilities

S	⇒	NP	VP	1.0
VP	⇒	Vi		0.4
VP	⇒	Vt	NP	0.4
VP	⇒	VP	PP	0.2
NP	⇒	DT	NN	0.3
NP	⇒	NP	PP	0.7
PP	⇒	P	NP	1.0

Vi	⇒	sleeps	1.0
Vt	⇒	saw	1.0
NN	⇒	man	0.7
NN	⇒	woman	0.2
NN	⇒	telescope	0.1
DT	⇒	the	1.0
IN	⇒	with	0.5
IN	⇒	in	0.5

The probability of a (sentence, parse tree) pair is

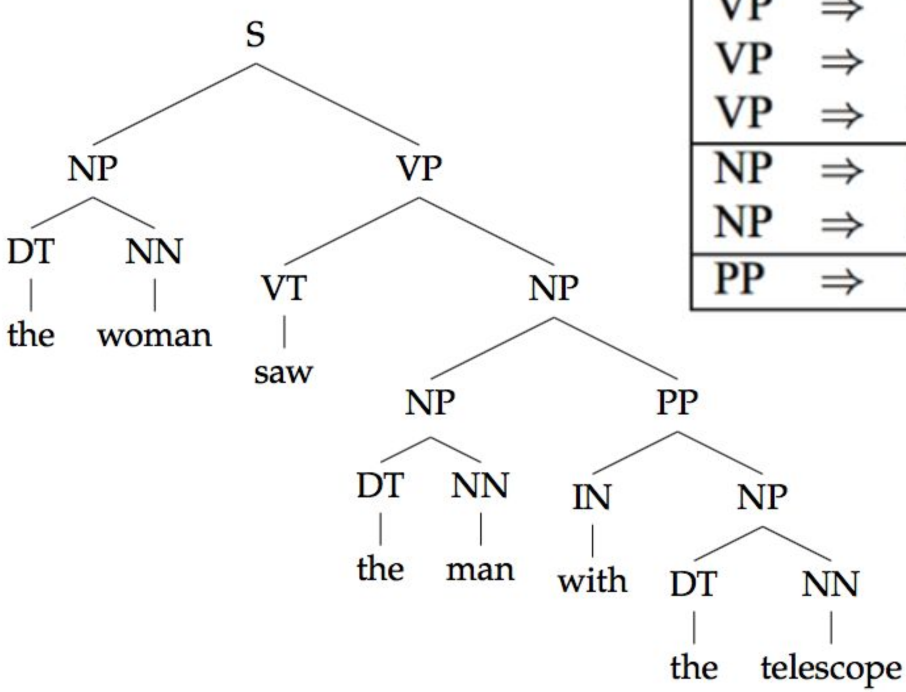
$$p(S,T) = \prod_{i=1}^{m} p(\alpha_i \rightarrow \beta_i | \alpha_i)$$

m is the number of transitions

P(NN -> man | NN) = 0.7

P(S,T)

- P(S,T) = [1.0 * 0.3 * 1.0 * 0.2] * [0.4 * 1.0 * 0.7 * (0.3 * 1.0 * 0.7) * (1.0 * 0.5 * 0.3 * 1.0 * 0.1)]



S	⇒	NP	VP	1.0
VP	⇒	Vi		0.4
VP	⇒	Vt	NP	0.4
VP	⇒	VP	PP	0.2
NP	⇒	DT	NN	0.3
NP	⇒	NP	PP	0.7
PP	⇒	P	NP	1.0

Vi	⇒	sleeps	1.0
Vt	⇒	saw	1.0
NN	⇒	man	0.7
NN	⇒	woman	0.2
NN	⇒	telescope	0.1
DT	⇒	the	1.0
IN	⇒	with	0.5
IN	⇒	in	0.5

Estimating transition probabilities

- Counts from training set!

$$P(S \rightarrow NP\ VP|S) = \frac{count(S \rightarrow NP\ VP)}{count(S)}$$

PCFG related questions

- What is the most likely parse? (parsing task)
  - $\text{argmax}_T\ P(T,S)$
- What is the probability of the sentence, P(S)? (Language modeling task)
  - $P(S) = \Sigma_T\ P(T,S)$

How?
Dynamic programming
(CYK algorithm –
Cocke–Younger–Kasami algorithm)

Chomsky Normal Form

- CYK can be used if the CFG is in Chomsky Normal Form
- A CFG is in Chomsky Normal Form if each rule either converts to two nonterminals or a single terminal.

$$X \rightarrow Y_1 Y_2$$

$$X \rightarrow y$$

Uppercase letters mean nonterminal
Lowercsae letters mean terminal

- Any CFG can be converted to CNF
  - For example NP -> DT, ADJ, NN
    - Turns into two rules
    - NP -> DT, ADJP
    - ADJP -> ADJ, NN

- N = {S, NP, VP, PP, DT, Vi, Vt, NN, IN} Nonterminals
- S = S Starting symbol
- Σ = {sleeps, saw, man, woman, telescope, the, with, in} Terminals

R =

S	⇒	NP	VP
VP	⇒	Vi	
VP	⇒	Vt	NP
VP	⇒	VP	PP
NP	⇒	DT	NN
NP	⇒	NP	PP
PP	⇒	IN	NP

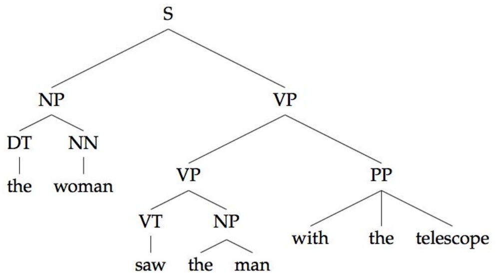
Vi	⇒	sleeps
Vt	⇒	saw
NN	⇒	man
NN	⇒	woman
NN	⇒	telescope
DT	⇒	the
IN	⇒	with
IN	⇒	in

Production rules
$X \rightarrow Y_1 \ldots Y_n$
Y_i can be terminal or nonterminal
The rule only relies on X (no context)
(vs context-sensitive)

CYK algorithm for parsing

3 dimensional

- $\pi(i,j,N)$ probability that words i to j can be generated by nonterminal N
- Base case: $\pi(i,i,N) = P(N \rightarrow w_i|N)$



- Inductive case:

1 2 3 4 5 6 7 8

k is the index that splits the subsentence

$$\pi(i,j,N) = \max_{k,P,Q} Pr(N \rightarrow P\ Q|N) \cdot \pi(\underline{i,k},P) \cdot \pi(\underline{k+1,j},Q)$$

where $k \in \{i,\ldots,j-1\}$, $P \in \mathcal{N}$, and $Q \in \mathcal{N}$.

Saves the rule that gave max probability to backtrack

CYK algorithm for language modeling

- $\pi(i,j,N)$ probability that words i to j can be generated by nonterminal N

- Base case: $\pi(i,i,N) = P(N \rightarrow w_i|N)$

- Inductive case:

$$\pi(i,j,N) = \max_{k,P,Q} Pr(N \rightarrow P\ Q|N) \cdot \pi(i,k,P) \cdot \pi(k+1,j,Q)$$

argmax_T P(T,S) vs P(S) = Σ_T P(T,S)

$$\pi(i,j,N) = \sum_{k,P,Q} P(N \rightarrow P\ Q|N) \cdot \pi(\underline{i,k},P) \cdot \pi(\underline{k+1,j},Q)$$

where $k \in \{i,\ldots,j-1\}$, $P \in \mathcal{N}$, and $Q \in \mathcal{N}$.

Language modeling example CYK with PCFG

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P(“abc”)

$A \rightarrow AB$	0.8
$A \rightarrow a$	0.2
$B \rightarrow BB$	0.7
$B \rightarrow b$	0.1
$B \rightarrow c$	0.2

$$\pi(i, j, N)$$

A

a			
b			
c			

B

a			
b			
c			

Base case

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P("abc")

$A \rightarrow AB$	0.8
$A \rightarrow a$	0.2
$B \rightarrow BB$	0.7
$B \rightarrow b$	0.1
$B \rightarrow c$	0.2

$$\pi(i, i, N) = P(N \rightarrow w_i | N)$$

A

a	0.2		
b		0	
c			0

B

a	0		
b		0.1	
c			0.2

1^{st} step

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P(“abc”)

$A \to AB$	0.8
$A \to a$	0.2
$B \to BB$	0.7
$B \to b$	0.1
$B \to c$	0.2

$$\pi(i,j,N) = \sum_{k,P,Q} P(N \to P\,Q|N) \cdot \pi(i,k,P) \cdot \pi(k+1,j,Q)$$

A

a	0.2		
b		0	
c			0

B

a	0		
b		0.1	
c			0.2

Only consider i, j where i < j

1^{st} step

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P("abc")

$A \to AB$	0.8
$A \to a$	0.2
$B \to BB$	0.7
$B \to b$	0.1
$B \to c$	0.2

$$\pi(i,j,N) = \sum_{k,P,Q} P(N \to P\ Q|N) \cdot \pi(i,k,P) \cdot \pi(k+1,j,Q)$$

A

a	0.2	0.016	
b		0	
c			0

$$\begin{aligned}\pi[1,2,A] =& P(A \to AA) \cdot P(A \to a|A) \cdot P(A \to b|A) \\ &+ P(A \to AB) \cdot P(A \to a|A) \cdot P(B \to b|B) \\ &+ P(A \to BA) \cdot P(B \to a|B) \cdot P(A \to b|A) \\ &+ P(A \to BB) \cdot P(B \to a|B) \cdot P(B \to b|B) \\ =& 0 + P(A \to AB) \cdot \pi[1,1,A] \cdot \pi[2,2,B] + 0 + 0 \\ =& 0.8 \cdot 0.2 \cdot 0.1 = 0.016\end{aligned}$$

1^{st} step

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P("abc")

$A \rightarrow AB$	0.8
$A \rightarrow a$	0.2
$B \rightarrow BB$	0.7
$B \rightarrow b$	0.1
$B \rightarrow c$	0.2

$$\pi(i,j,N) = \sum_{k,P,Q} P(N \rightarrow P\ Q|N) \cdot \pi(i,k,P) \cdot \pi(k+1,j,Q)$$

$\boldsymbol{\pi}$(1,2,B) = P(B -> BB)P(B->a|B)P(B->b|B)
= P(B -> B B)$\boldsymbol{\pi}$(1,1,B)$\boldsymbol{\pi}$(2,2,B)
= 0.7*0*0.1 = 0

B

a	0	0	
b		0.1	
c			0.2

2^{nd} step

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P("abc")

$A \to AB$	0.8
$A \to a$	0.2
$B \to BB$	0.7
$B \to b$	0.1
$B \to c$	0.2

$$\pi(i,j,N) = \sum_{k,P,Q} P(N \to P\ Q|N) \cdot \pi(i,k,P) \cdot \pi(k+1,j,Q)$$

A

a	0.2	0.016	
b		0	0
c			0

B

a	0	0	
b		0.1	0.014
c			0.2

$\boldsymbol{\pi}(2,3,A) = P(A \text{ -> } A\ B)\boldsymbol{\pi}(2,2,A)\boldsymbol{\pi}(3,3,B) = 0.8*0*0.2 = 0$

2^{nd} step

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P("abc")

$A \to AB$	0.8
$A \to a$	0.2
$B \to BB$	0.7
$B \to b$	0.1
$B \to c$	0.2

$$\pi(i,j,N) = \sum_{k,P,Q} P(N \to P\ Q|N) \cdot \pi(i,k,P) \cdot \pi(k+1,j,Q)$$

A

a	0.2	0.016	
b		0	0
c			0

B

a	0	0	
b		0.1	0.014
c			0.2

$\boldsymbol{\pi}(2,3,B) = P(B \text{ -> } B\ B)\boldsymbol{\pi}(2,2,B)\boldsymbol{\pi}(3,3,B) + 0 = 0.7*0.1*0.2 = 0.014$

Finish

- N = {A, B}
- Σ = {a, b, c}
- S = {A}

Find P("abc")

$A \rightarrow AB$	0.8
$A \rightarrow a$	0.2
$B \rightarrow BB$	0.7
$B \rightarrow b$	0.1
$B \rightarrow c$	0.2

$$\pi(i,j,N) = \sum_{k,P,Q} P(N \rightarrow P\, Q|N) \cdot \pi(i,k,P) \cdot \pi(k+1,j,Q)$$

A

a	0.2	0.016	0.0048
b		0	0
c			0

B

a	0	0	0
b		0.1	0.014
c			0.2

$\boldsymbol{\pi}(1,3,A) = P(A \text{ -> } A\ B)\boldsymbol{\pi}(1,2,A)\boldsymbol{\pi}(3,3,B) + P(A \text{ -> } A\ B)\boldsymbol{\pi}(1,1,A)\boldsymbol{\pi}(2,3,B)$

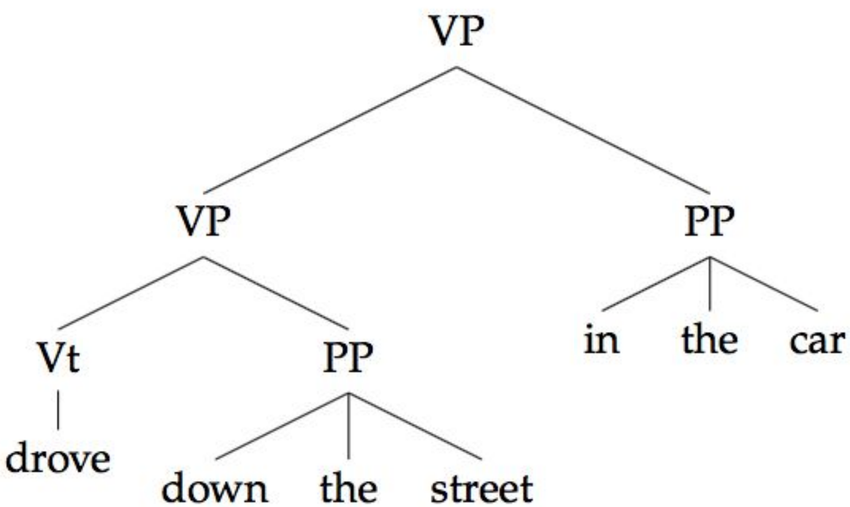
$= 0.8 * 0.016 * 0.2 \quad + 0.8 * 0.2 * 0.014 \quad = 0.0048$

PCFG weakness

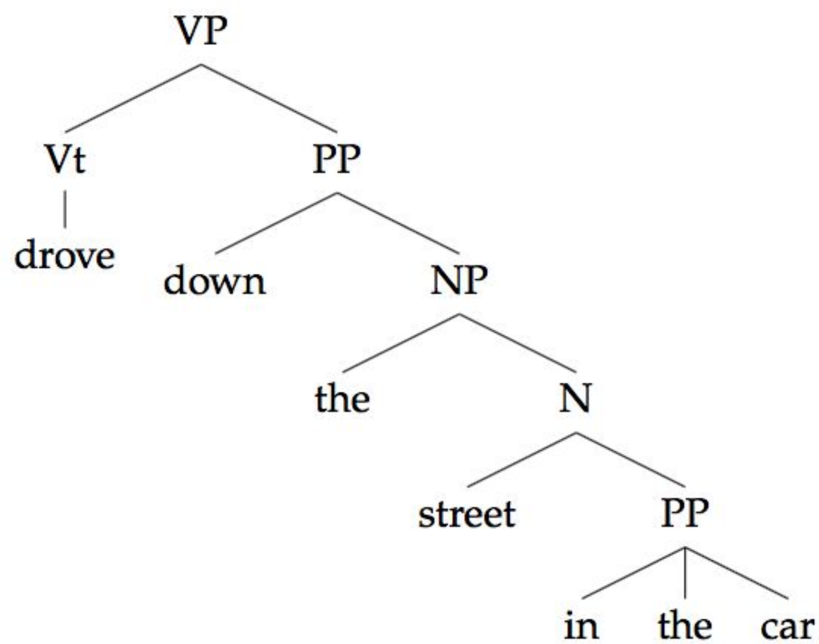
- Lack of sensitivity to lexical info (does not consider semantics)
- Lack of sensitivity to structural frequency

Lack of lexical info

The probabilities only see terminals and expansions

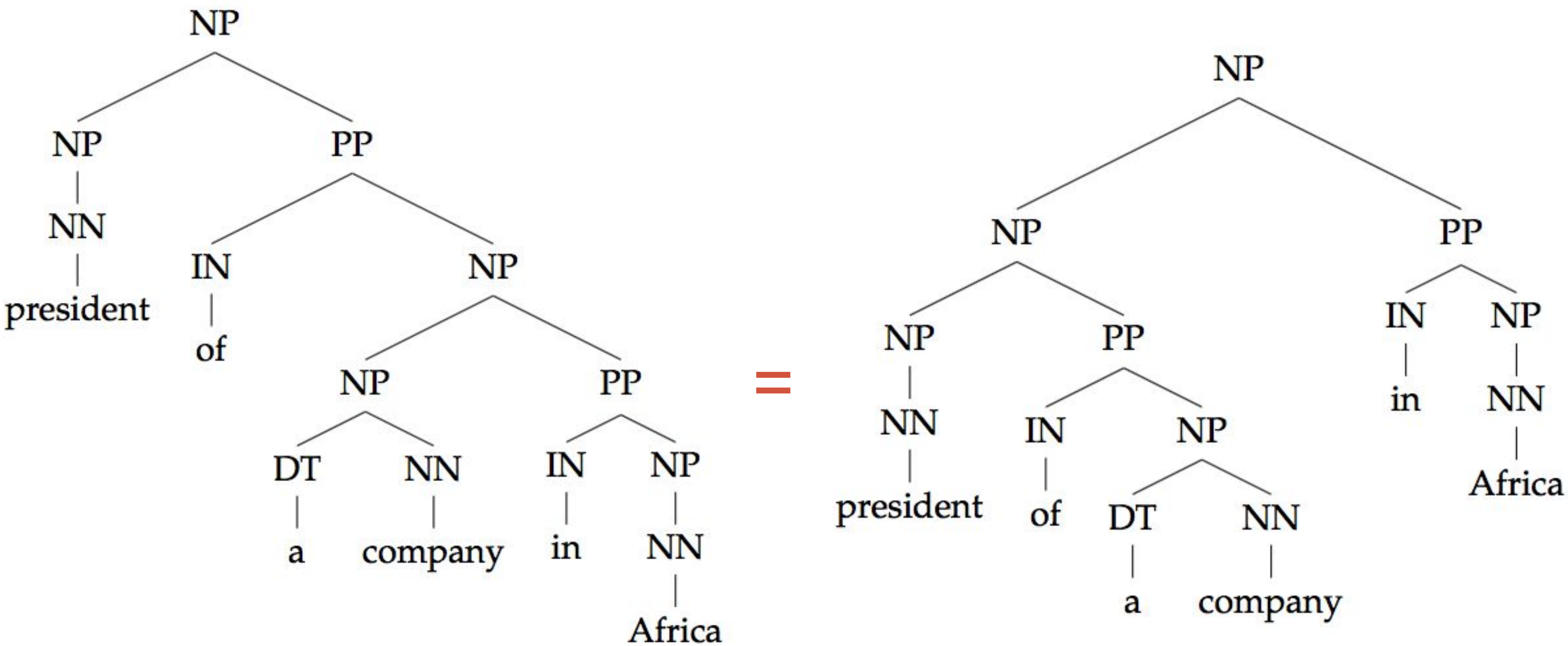


<



The street in the car

Lack of sensitivity to structural frequency



Both trees have same expansion and so same probability

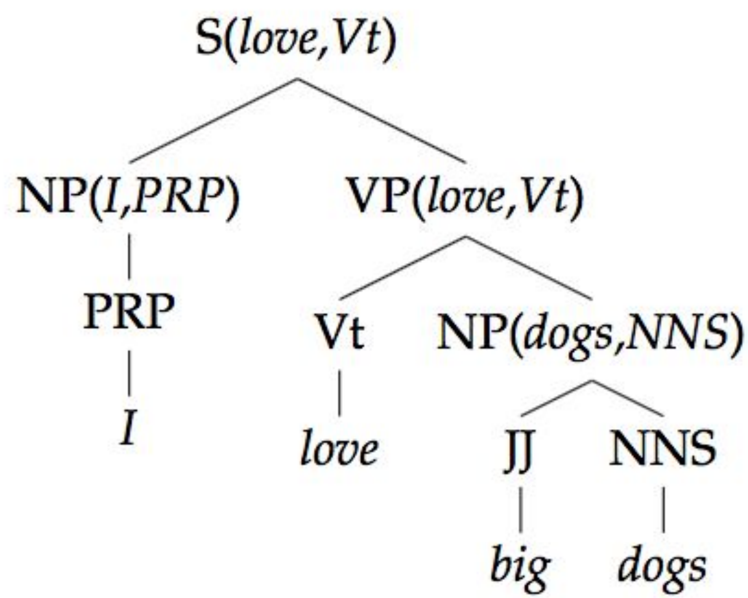
Turns out, left hand side structure appears more frequently in the training set

Methods to improve PCFG

- Add lexical information to the tree
  - Label each node with associate words
    - Lexicalized trees
    - Latent variable grammars
- Look at bigger portions of the tree at a time

S	⇒	NP	VP	1.0
VP	⇒	Vi		0.4
VP	⇒	Vt	NP	0.4
VP	⇒	VP	PP	0.2
NP	⇒	DT	NN	0.3
NP	⇒	NP	PP	0.7
PP	⇒	P	NP	1.0

Vi	⇒	sleeps	1.0
Vt	⇒	saw	1.0
NN	⇒	man	0.7
NN	⇒	woman	0.2
NN	⇒	telescope	0.1
DT	⇒	the	1.0
IN	⇒	with	0.5
IN	⇒	in	0.5



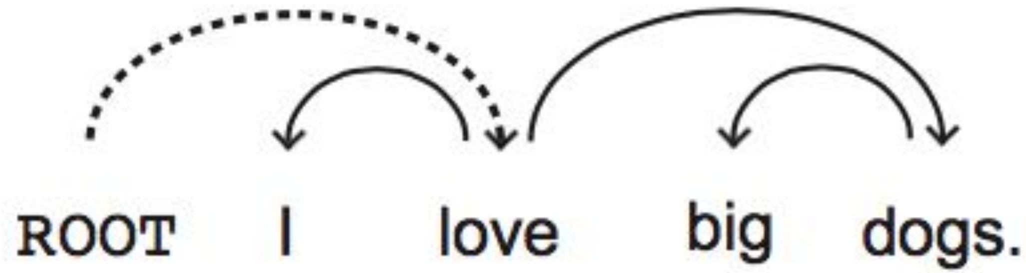
Lexicalized tree

Dependency grammar

- CFG is based on constituency relation
- In dependency grammar the structure is composed of lexical items (words) linked by edges to form a tree
- Assumptions
  - Each words in a sentence is related or modifies another word
  - All words have a direct or indirect relation to the main verb

Example

- Add ROOT node as the root of the tree
- The main verb always point to ROOT



A -> B means
A governs B or
B depends on A

A is the head of B

- Each arc can have a category for the relationship.
- Each word can have a PoS label

Constituency structures vs dependency structures

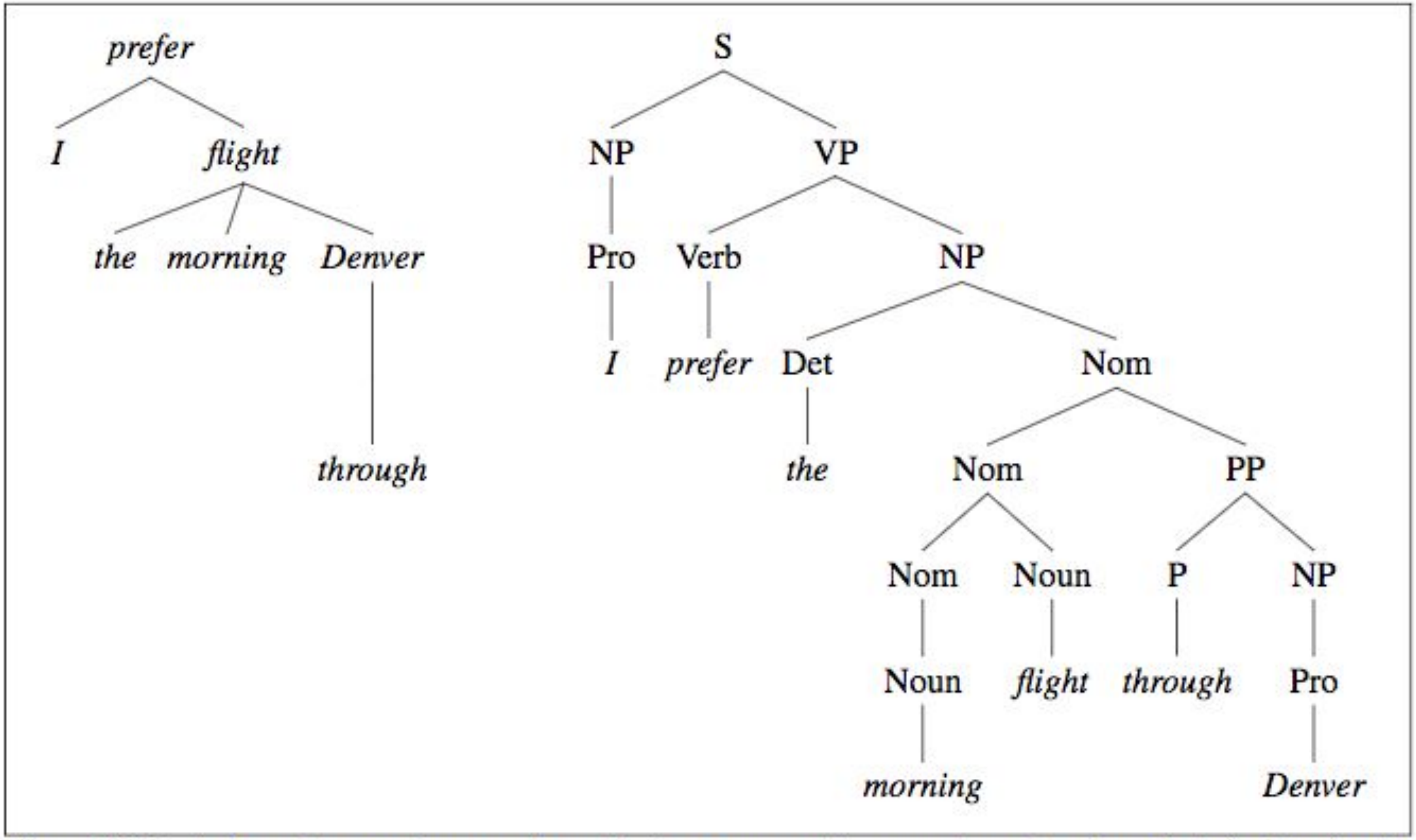


Figure 14.1 A dependency-style parse alongside the corresponding constituent-based analysis for *I prefer the morning flight through Denver.*

http://web.stanford.edu/~jurafsky/slp3/14.pdf

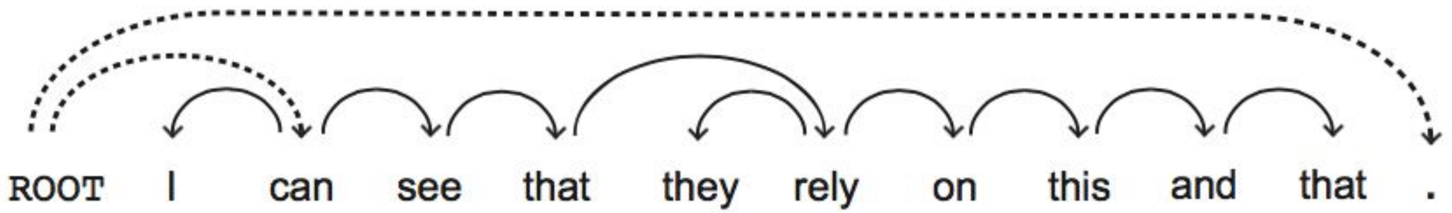
Constituency structures vs dependency structures

- Constituency structures use more nodes to represent sentences at different levels.
- Constituency structures explicitly label non-terminal nodes (NP vs VP)
- Constituency structures encode more info than dependency structures
- You can convert constituency structures to dependency structures
  - Dependency parsers trained on this is usually better than dependency parsers trained on original dependency structures

Criteria for heads (basics)

- Head, H. Dependent D
- D modifies H
  - Big (D) dogs (H), willow (D) tree (H)
- H can often replace D
  - I love big (D) dogs (H) -> I love dogs
- H is obligatory while D sometimes is optional
- H determines whether D is obligatory
  - Sarah sneezed (H) vs George kicks (H) the chair (D)
- More criterias! Mostly depends on corpus

Ex

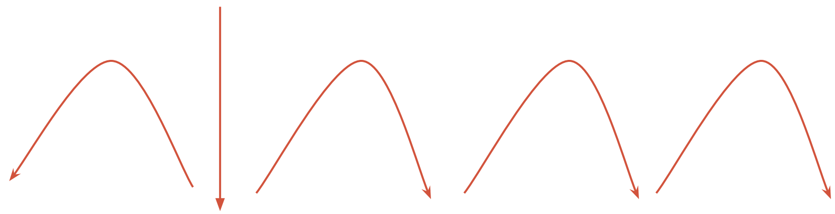


Dependency and meaning

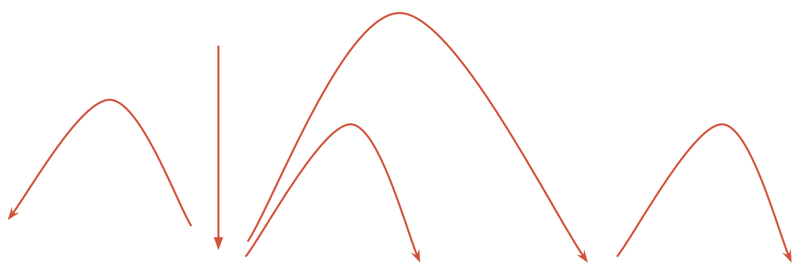
- Scientists study whales from space

- Scientists study whales from space

Dependency and meaning



- Scientists study whales from space



- Scientists study whales from space

Sample Thai sentences

- วัน หนึ่ง เกิด เหตุ ที่ ไม่ คาดคิด ขึ้น มา
- การ ที่ เรา จะ รู้ ถึง สิ่ง เหล่านั้น ได้ ก็ ต้อง อาศัย ปัจจัย หลาย ๆ อย่าง ประกอบ เข้า ด้วยกัน

Note these criteria depends on the corpus convention

Sample Thai sentences

- วัน หนึ่ง เกิด เหตุ ที่ ไม่ คาดคิด ขึ้น มา

Note these criteria depends on the corpus convention

Sample Thai sentences

- การ ที่ เรา จะ รู้ ถึง สิ่ง เหล่านั้น ได้ ก็ ต้อง อาศัย

ปัจจัย หลาย ๆ อย่าง ประกอบ เข้า ด้วยกัน

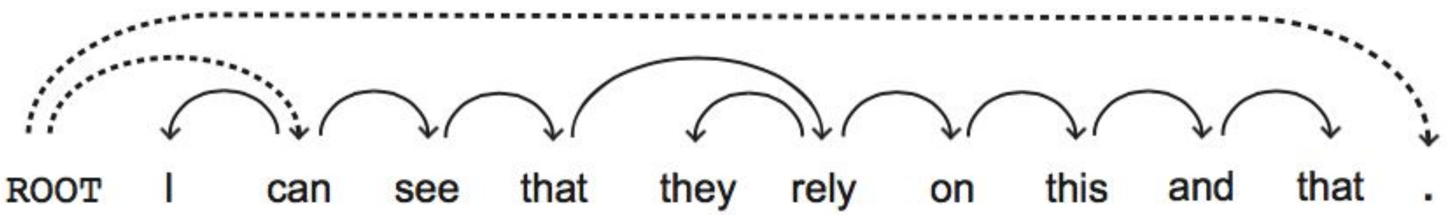
Note these criteria depends on the corpus convention

Dependency graph requirements

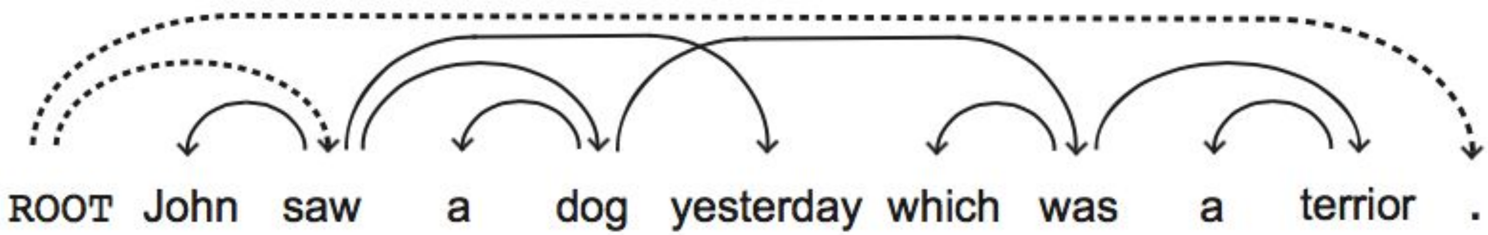
- Syntactic structure is complete (connectedness, spanning)
- Hierarchical (acyclic)
- Every word has a single head

Projectivity

- A dependency graph is projective if the arcs do not cross



projective



Non-projective

English and Thai are mostly projective.
Some languages are more non-projective than others,
for example German, Dutch, Czech.

When picking a parser algo, check whether it assumes projectivity

Transition-based parsing (Nivre 2007)

- Use a stack and buffer data structure and sequentially add edges
- Characteristics
  - Greedy algo. Only goes left to right. No backtracking
    - Fast O(n)
  - Requires projectivity
  - The algo is closely related to how human parse sentences (left to right one word at a time instead of looking at the sentence as a whole)

Arc-standard Transition-based parsing

- A stack σ, written with top of the stack to the right
  - Starts with the ROOT symbol
- A buffer β, written with top to the left
  - Starts with the input sentence
- A set of dependency arcs A
  - Starts of empty
- A set of actions

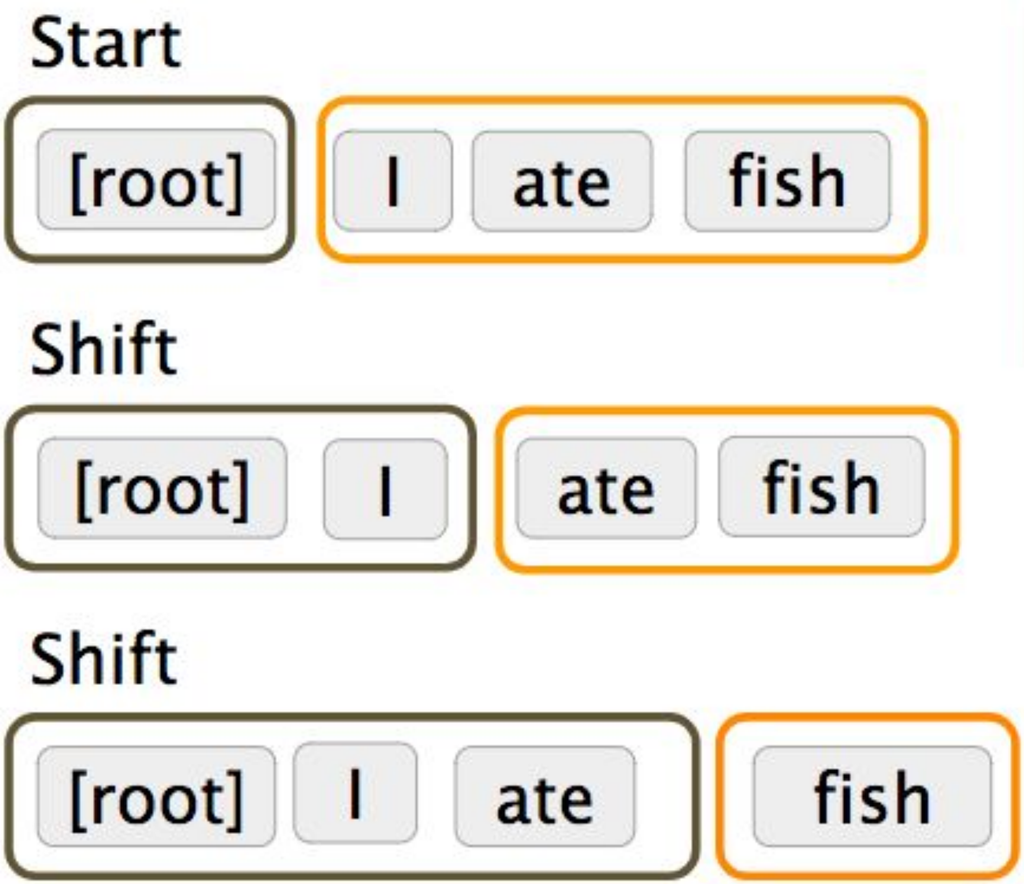
$$\sigma = [\text{ROOT}],\ \beta = w_1, \ldots, w_n,\ A = \varnothing$$

1. Shift $\quad \sigma, w_i|\beta, A \rightarrow \sigma|w_i, \beta, A$ Move buffer to stack
2. Left-Arc$_r$ $\quad \sigma|w_i|w_j, \beta, A \rightarrow \sigma|w_j, \beta, A \cup \{r(w_j,w_i)\}$
3. Right-Arc$_r$ $\quad \sigma|w_i|w_j, \beta, A \rightarrow \sigma|w_i, \beta, A \cup \{r(w_i,w_j)\}$

- Finishes when β becomes empty

http://web.stanford.edu/class/cs224n/lectures/lecture7.pdf

I ate fish



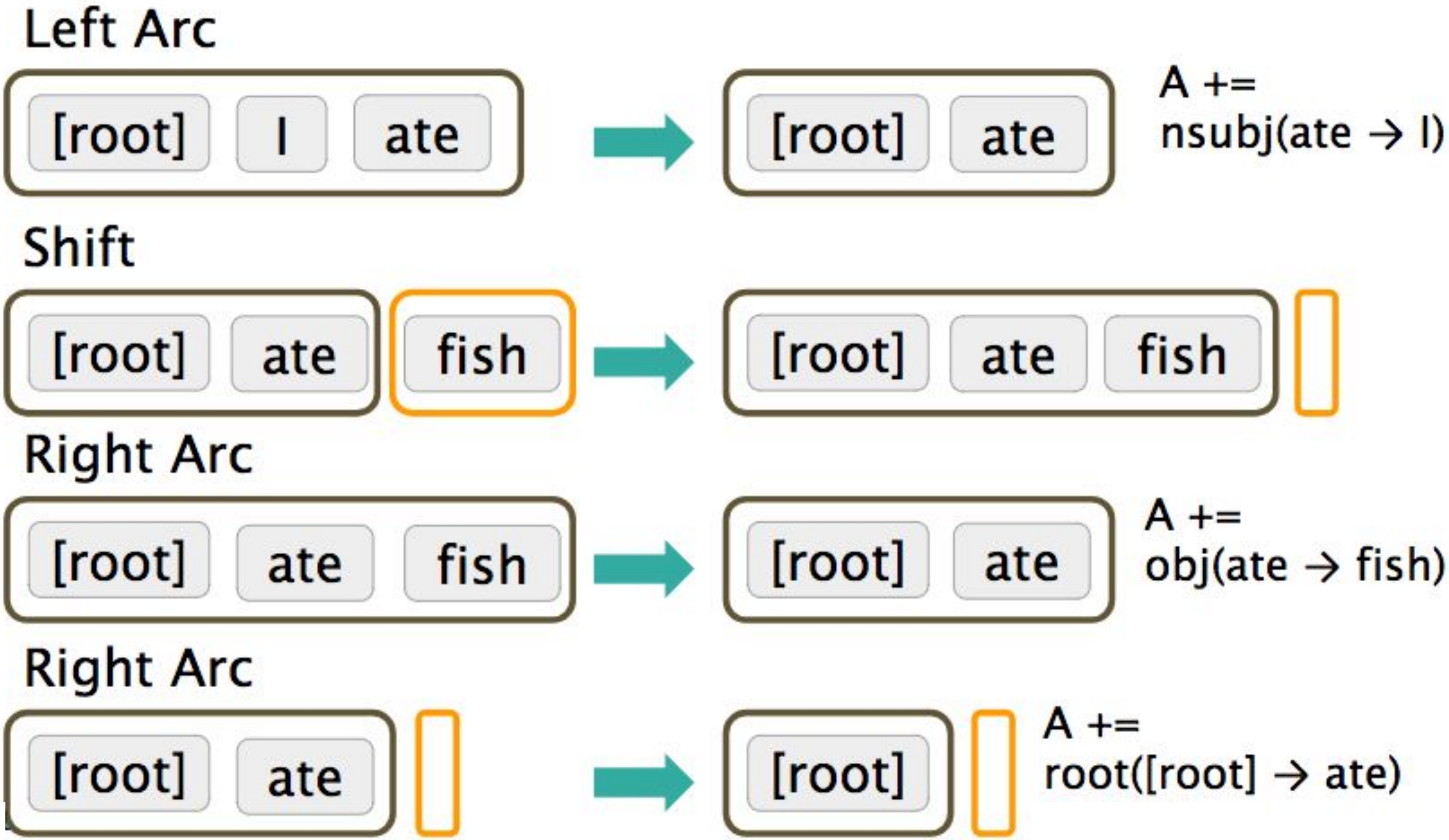
Start: $\sigma = [\text{ROOT}], \beta = w_1, \ldots, w_n, A = \varnothing$

1. Shift $\sigma, w_i|\beta, A \rightarrow \sigma|w_i, \beta, A$
2. Left-Arc$_r$ $\sigma|w_i|w_j, \beta, A \rightarrow \sigma|w_j, \beta, A \cup \{r(w_j, w_i)\}$
3. Right-Arc$_r$ $\sigma|w_i|w_j, \beta, A \rightarrow \sigma|w_i, \beta, A \cup \{r(w_i, w_j)\}$

Finish: $\beta = \varnothing$

http://web.stanford.edu/class/cs224n/lectures/lecture7.pdf

I ate fish



Left Arc
[root]
I
ate
[root]
ate
A +=
nsubj(ate → I)
Shift
[root]
ate
fish
[root]
ate
fish
Right Arc
[root]
ate
fish
[root]
ate
A +=
obj(ate → fish)
Right Arc
[root]
ate
[root]
A +=
root([root] → ate)

Discriminative parsing

- How to choose an action?
  - Shift, left-arc, right-arc
- Each action is predicted by a discriminative classifier (SVM, logistic regression, Neural networks) over legal moves
  - Features: top two word from stack, POS, children info; first word in buffer, POS, children info; etc.
- Greedy and no beamsearch
  - But you can include beamsearch (modern parsers do)

https://cs.stanford.edu/people/danqi/papers/emnlp2014.pdf

Discriminative parsing with neural networks

Shift, left, right (in actual, left/right + type of dependency)
So 2N+1, where N = dependency types

Softmax layer:

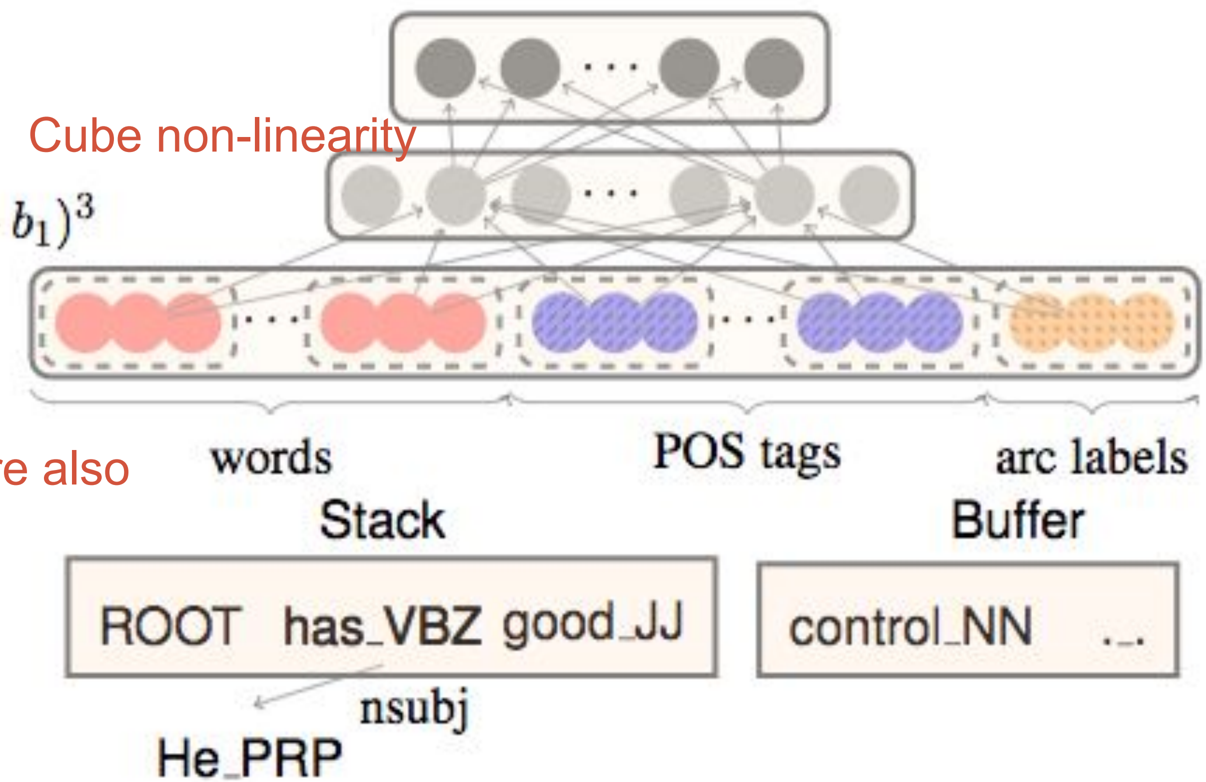
$$p = \texttt{softmax}(W_2 h)$$

Hidden layer:

$$h = (W_1^w x^w + W_1^t x^t + W_1^l x^l + b_1)^3$$

Input layer: $[x^w, x^t, x^l]$

Cube non-linearity



POS and arc labels are also embeddings

Configuration

https://cs.stanford.edu/people/danqi/papers/emnlp2014.pdf

Improvements

- Bigger networks
- Beam search
- Global inference over the sequence (CRF style)

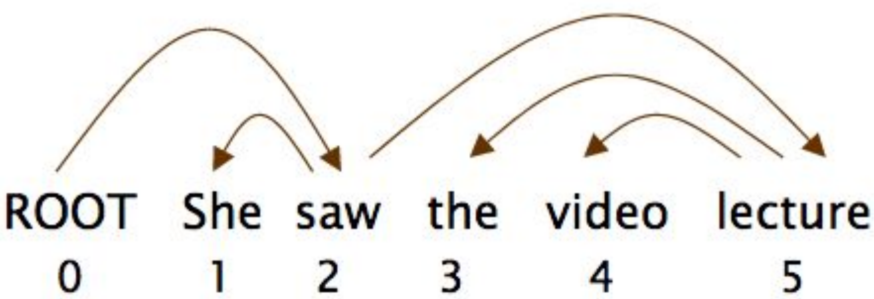
- Lead to SyntaxNet and Parsey McParseFace model

https://github.com/tensorflow/models/tree/master/research/syntaxnet

https://research.googleblog.com/2016/05/announcing-syntaxnet-worlds-most.html

Parsing evaluation

- Labeled parsing accuracy (LAS)
- Unlabeled parsing accuracy (UAS)



$$\text{Acc} = \frac{\#\text{ correct deps}}{\#\text{ of deps}}$$

UAS = 4 / 5 = 80%

LAS = 2 / 5 = 40%

Gold

1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	obj

Parsed

1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

http://web.stanford.edu/class/cs224n/lectures/lecture7.pdf

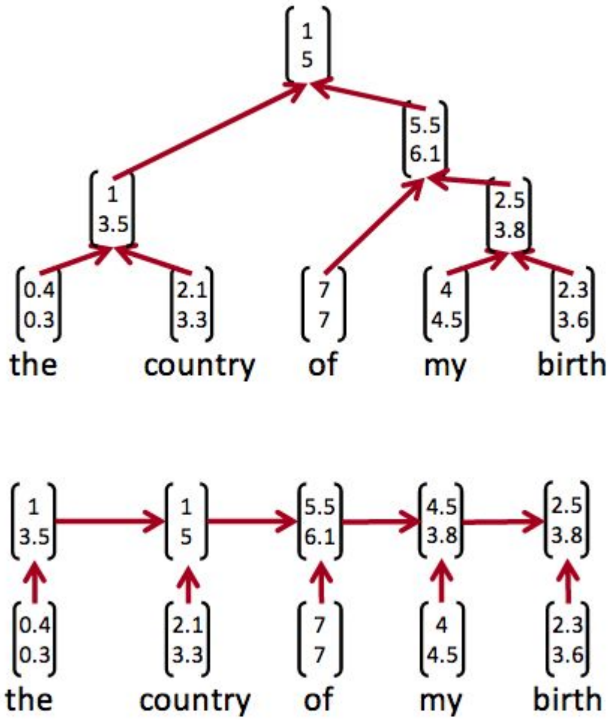
UAS of Parsey McParseFace

Model	News	Web	Questions
Martins et al. (2013)	93.10	88.23	94.21
Zhang and McDonald (2014)	93.32	88.65	93.37
Weiss et al. (2015)	93.91	89.29	94.17
Andor et al. (2016)*	94.44	90.17	95.40
Parsey McParseface	94.15	89.08	94.77

https://github.com/tensorflow/models/tree/master/research/syntaxnet

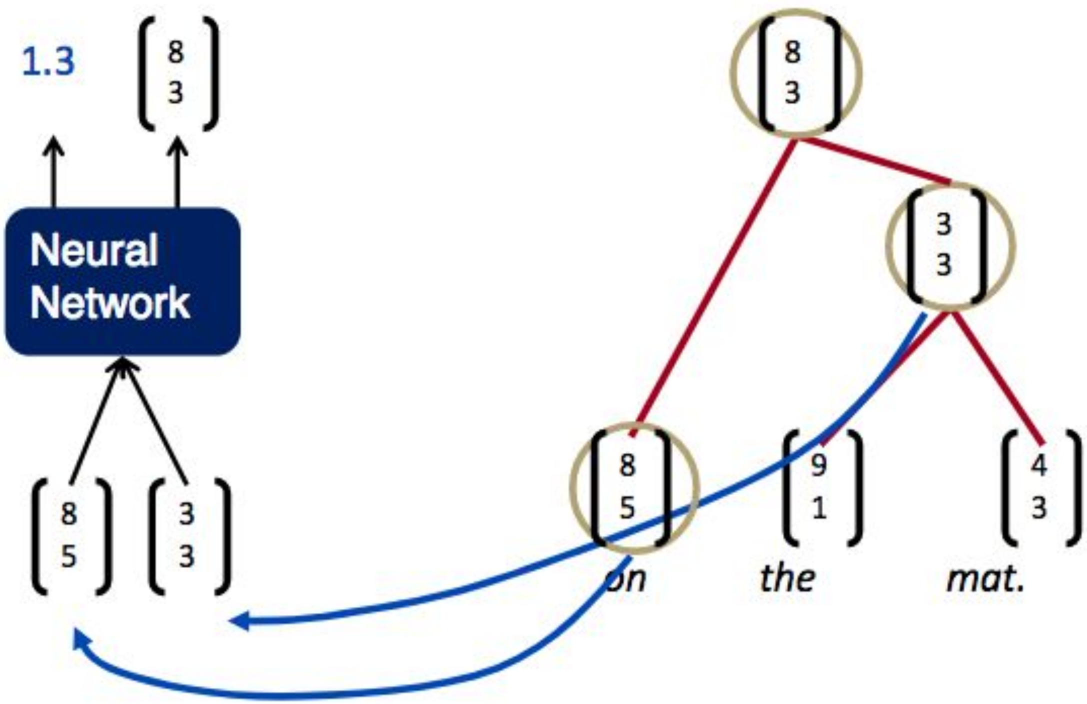
Recursive neural networks

- Not really used in parsing anymore but interesting concept
- Recursive vs Recurrent



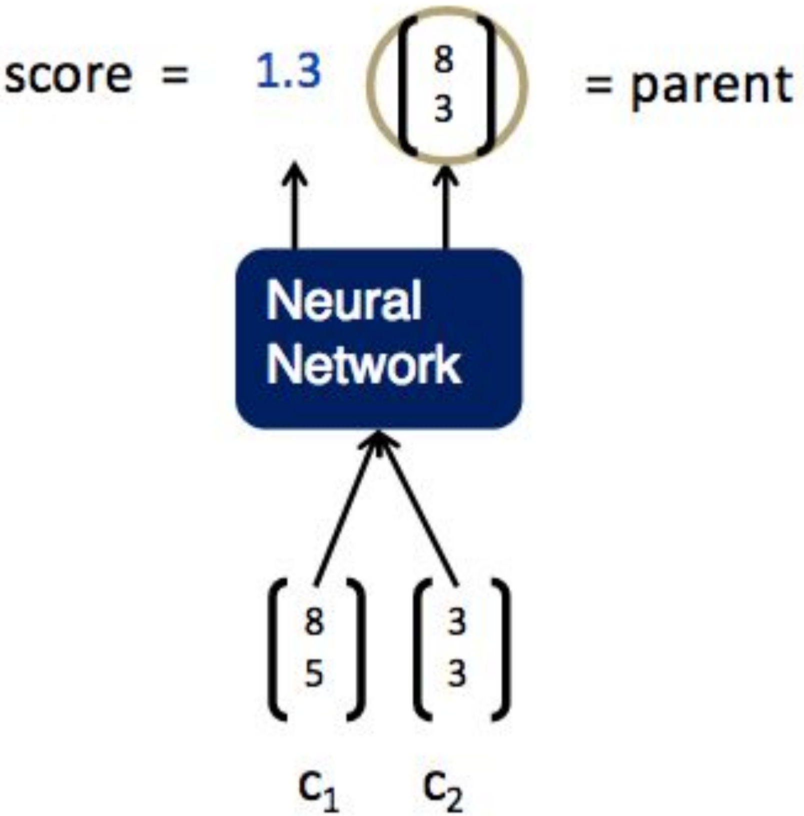
Recursive neural networks

- Concept: try different connections, see if which one gives the highest score (graph-based dependency parsers)
- Inputs: two candidate children representations
- Output:
  - Semantic representation of the parent
  - Score of new node



http://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

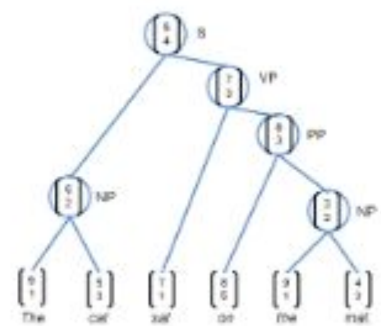
Shared recursive structure



$$\text{score} = U^{T}p$$

$$\boxed{p = \tanh\left(W\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + b\right),}$$

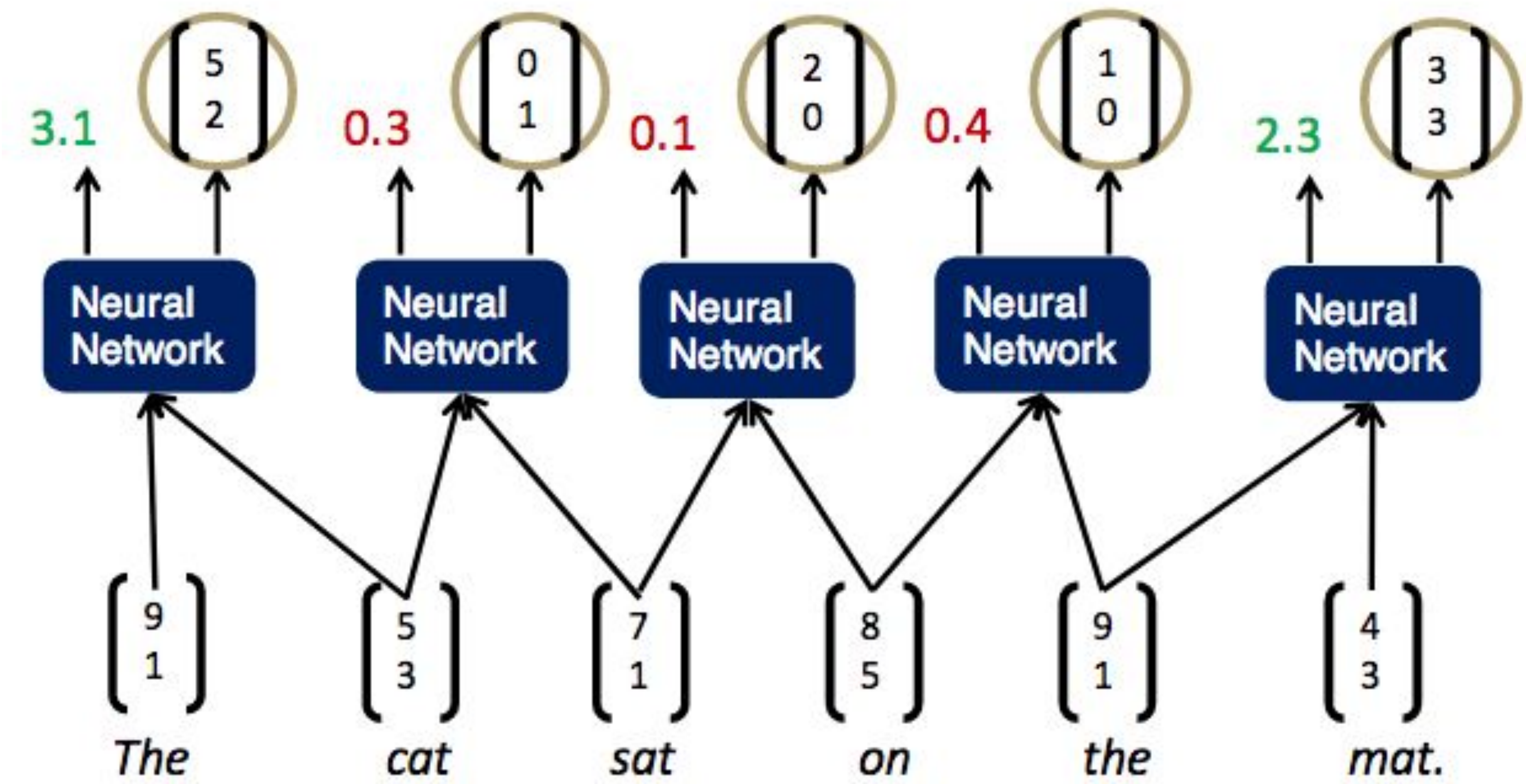
Same *W* parameters at all nodes of the tree

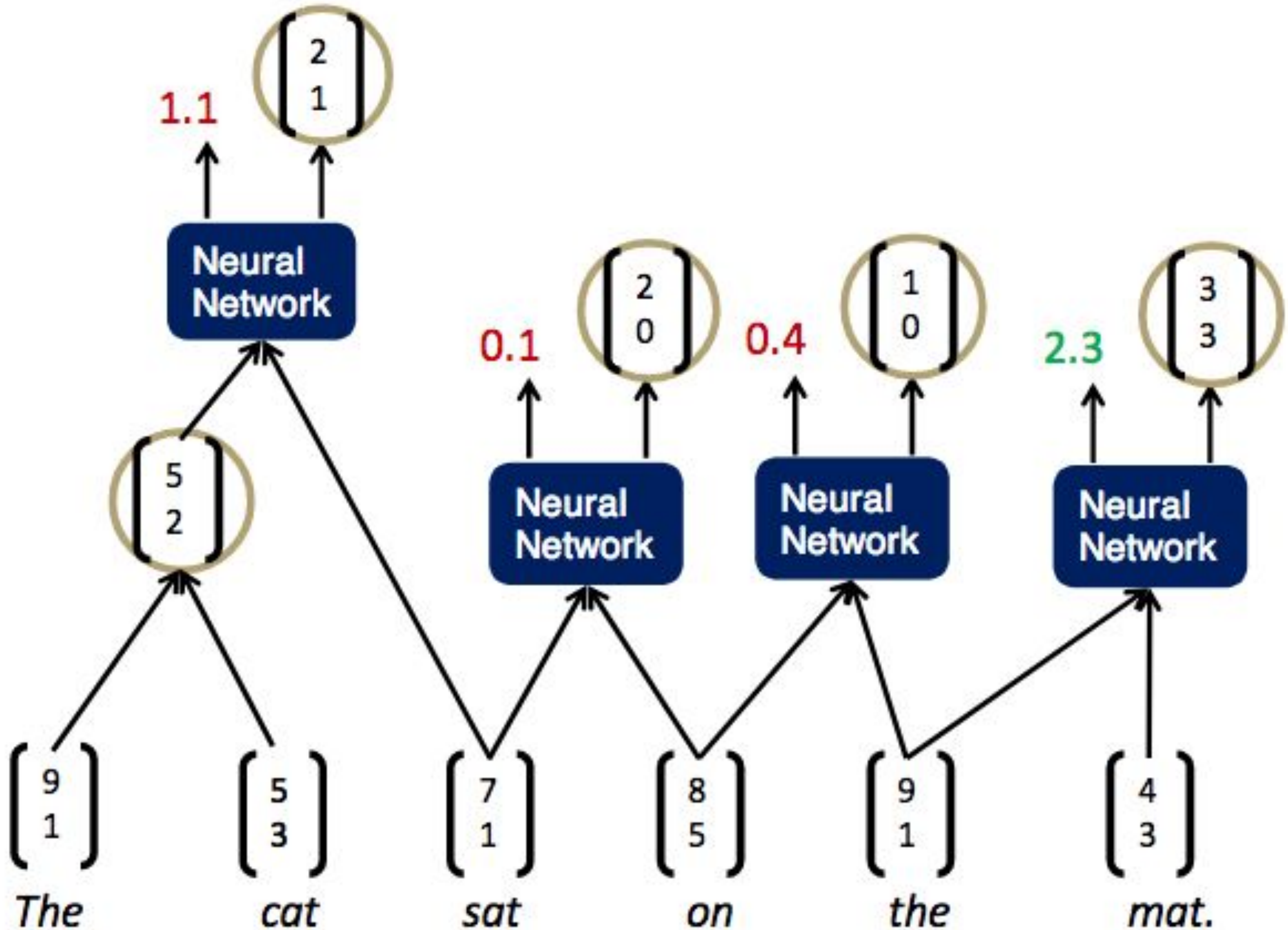


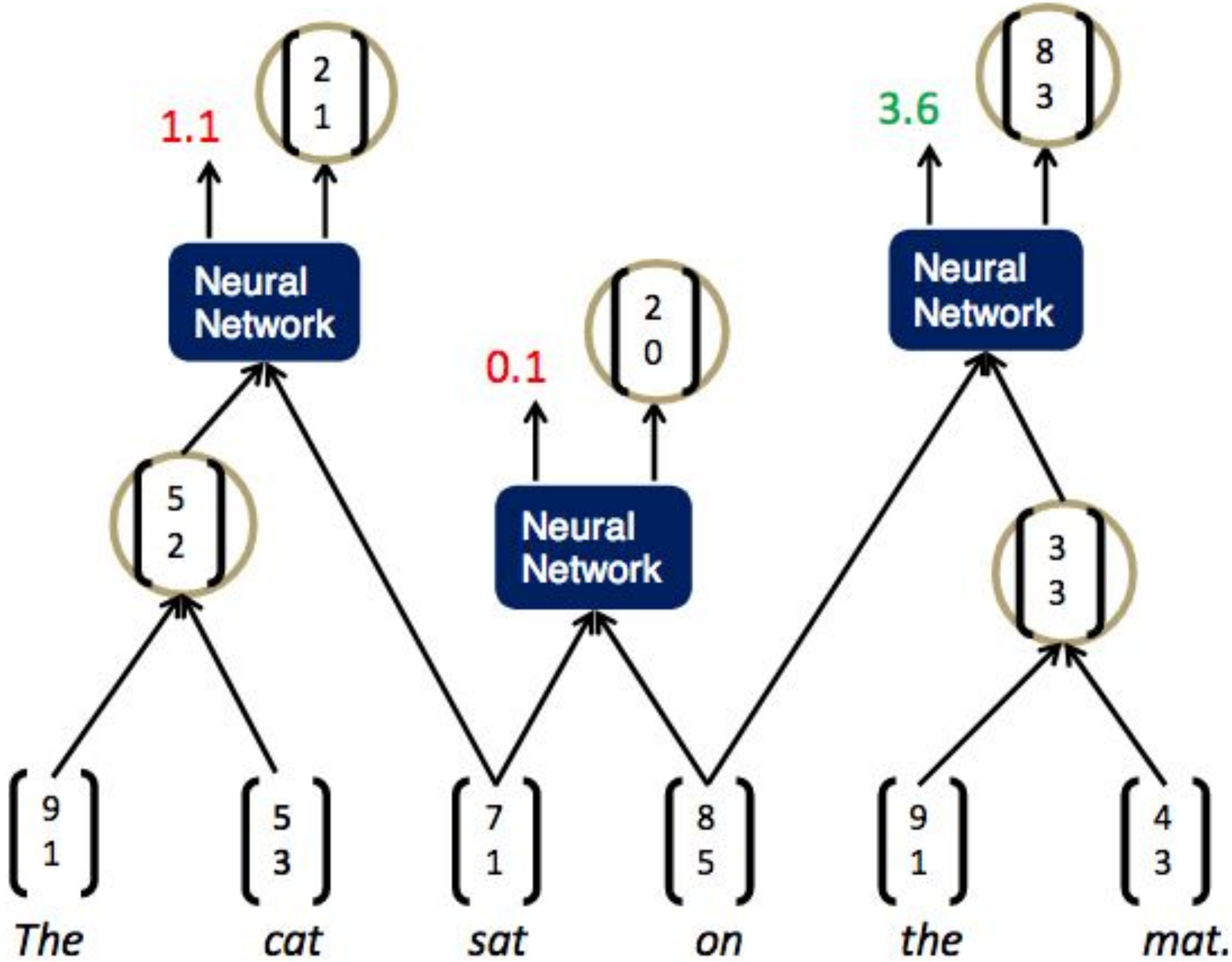
http://cs224d.stanford.edu/lectures/CS224d-Lecture10.pdf

Parsing with recursive networks

Follow the highest scoring pair



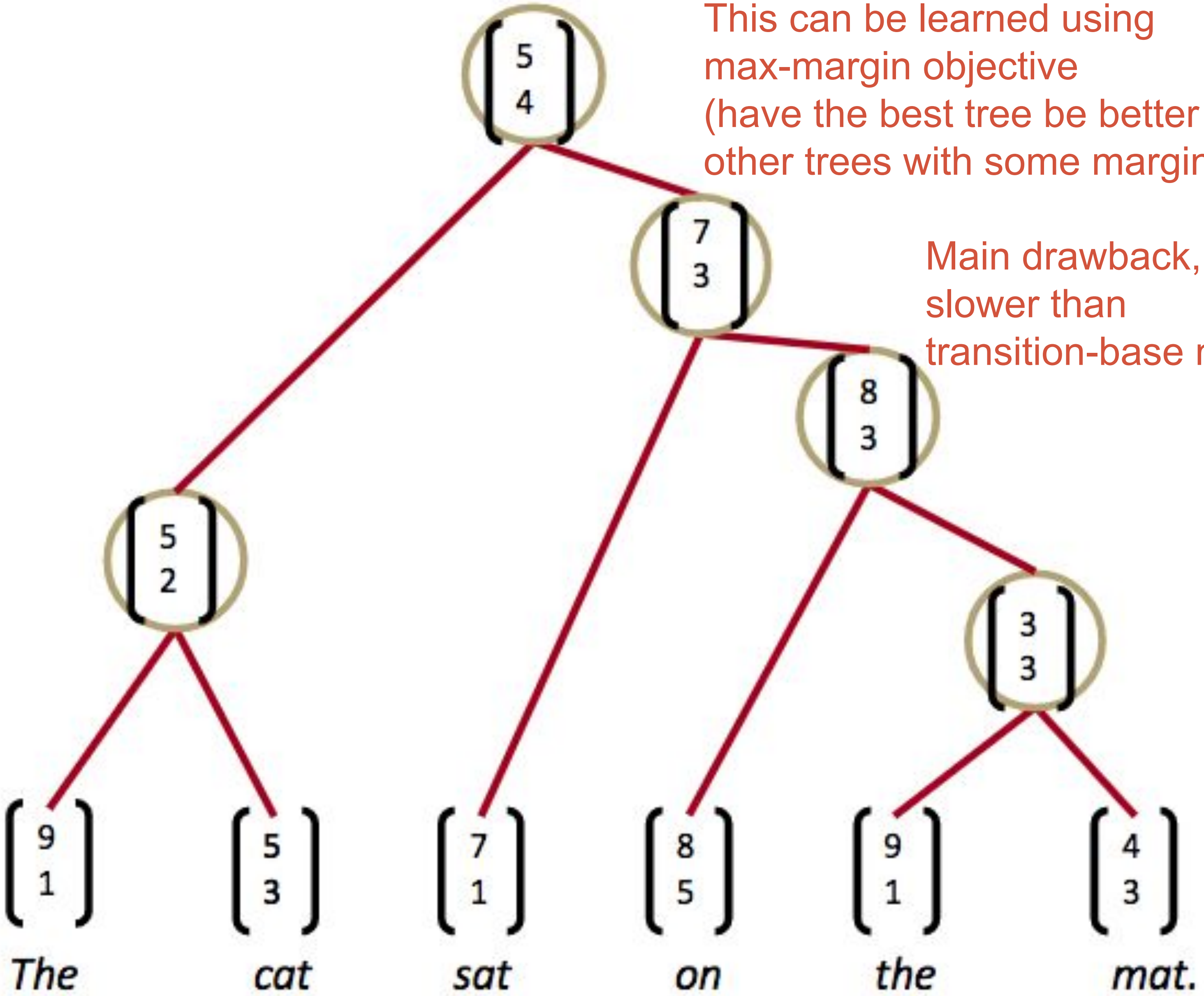

1.1
2
1
Neural Network
5
2
0.1
2
0
Neural Network
0.4
1
0
Neural Network
2.3
3
3
Neural Network
9
1
5
3
7
1
8
5
9
1
4
3
The
cat
sat
on
the
mat.



1.1
2
1
Neural Network
0.1
2
0
Neural Network
3.6
8
3
Neural Network
5
2
3
3
9
1
5
3
7
1
8
5
9
1
4
3
The
cat
sat
on
the
mat.

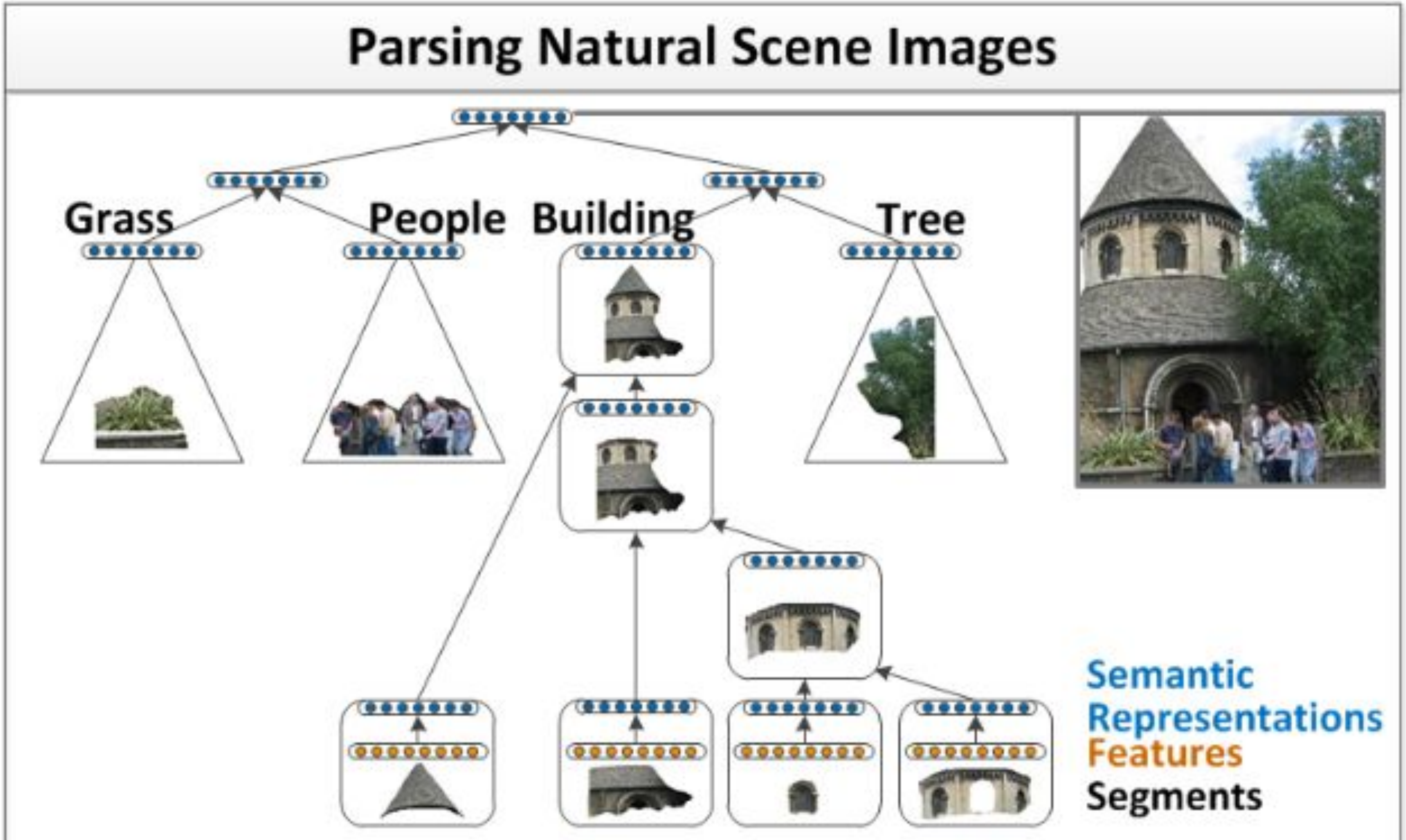
This can be learned using max-margin objective (have the best tree be better than other trees with some margin)

Main drawback, slower than transition-base methods



http://ai.stanford.edu/~ang/papers/icml11-ParsingWithRecursiveNeuralNetworks.pdf

Recursive neural networks for scene parsing



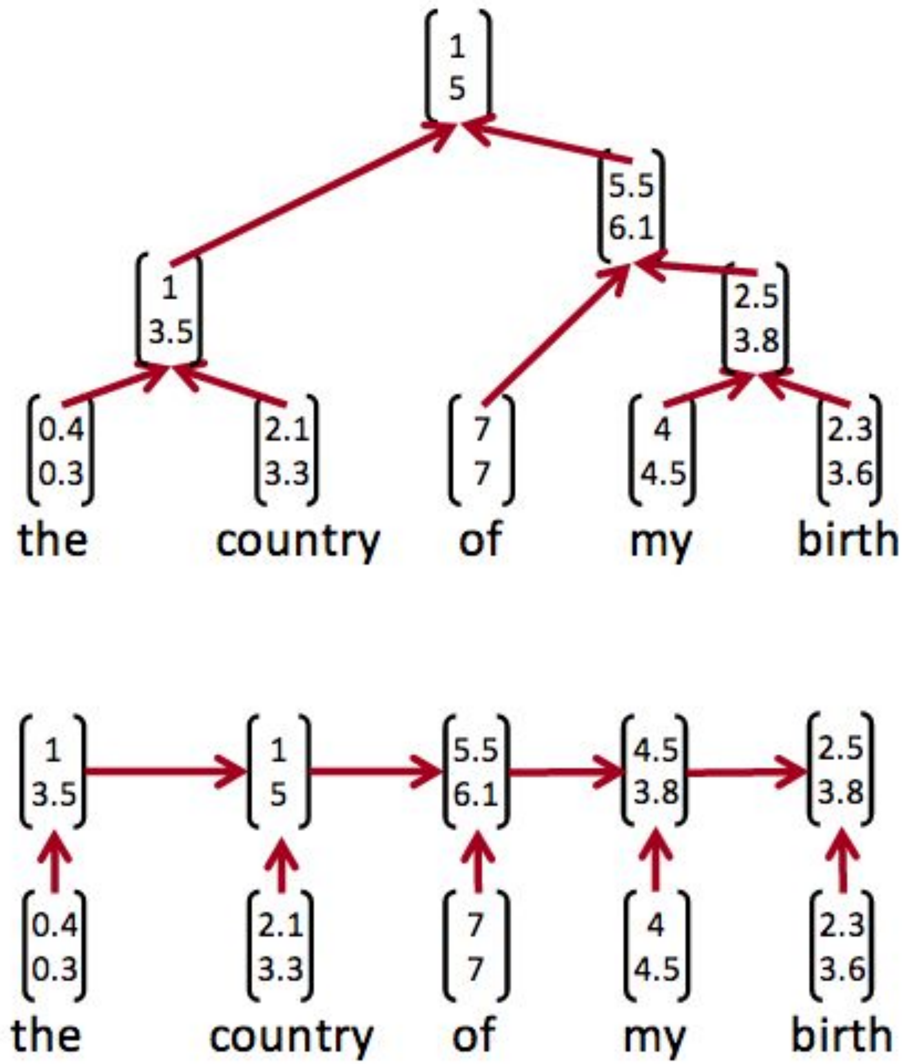
http://ai.stanford.edu/~ang/papers/icml11-ParsingWithRecursiveNeuralNetworks.pdf

Resources for parsing

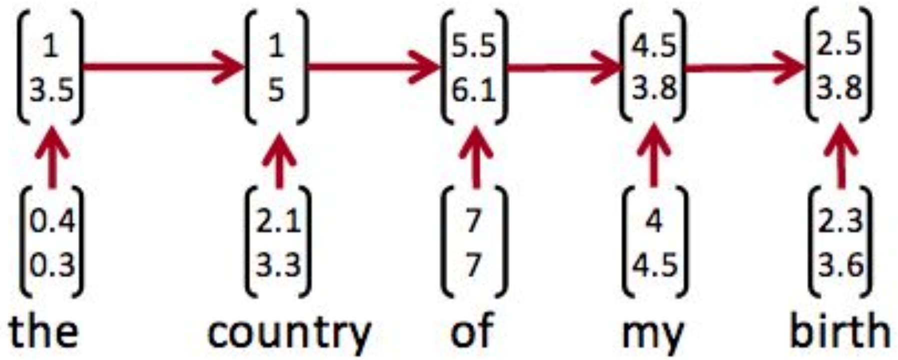
- Orchid & some corpus in development by NECTEC
- http://universaldependencies.org/

Russian	3	1,226K	IE, Slavic
Sanskrit	1	1K	IE, Indic
Serbian	1	86K	IE, Slavic
Slovak	1	106K	IE, Slavic
Slovenian	2	170K	IE, Slavic
Spanish	3	1,004K	IE, Romance
Swedish	3	195K	IE, Germanic
Swedish Sign Language	1	1K	Sign Language
Tamil	1	9K	Dravidian, Southern
Telugu	1	6K	Dravidian, South Central
Thai	1	23K	Tai-Kadai
Turkish	2	74K	Turkic, Southwestern
Ukrainian	1	100K	IE, Slavic
Upper Sorbian	1	10K	IE, Slavic
Urdu	1	138K	IE, Indic
Uyghur	1	15K	Turkic, Southeastern
Vietnamese	1	43K	Austro-Asiatic

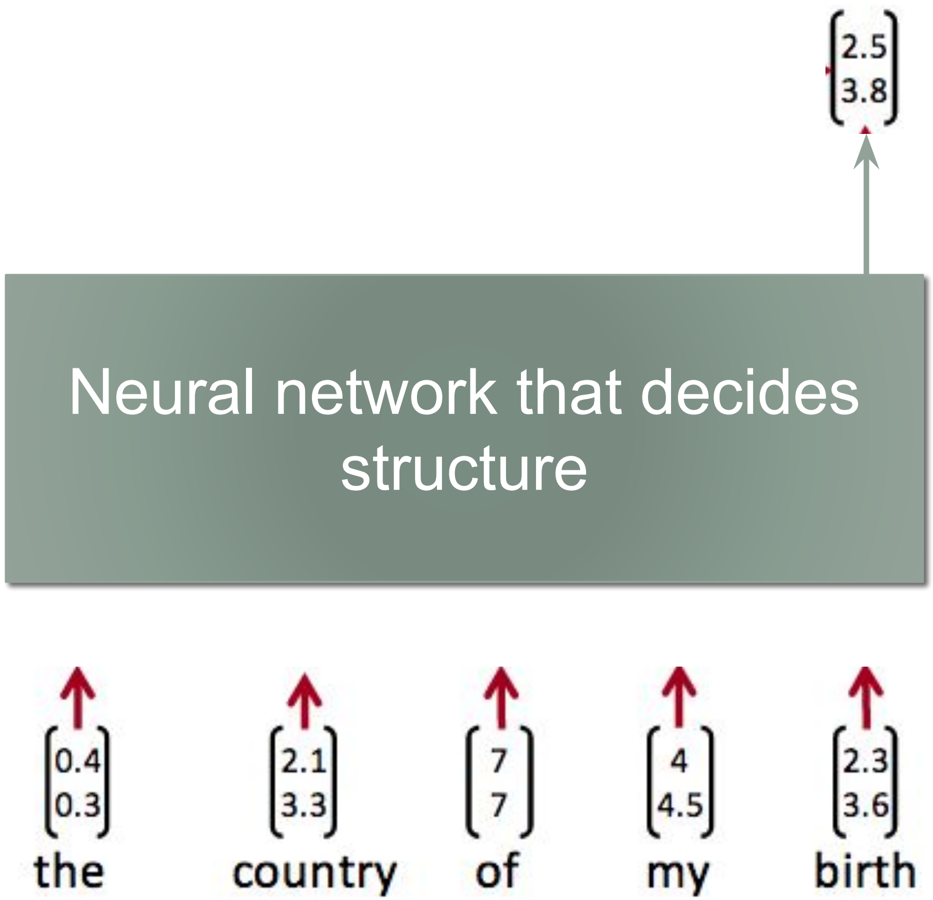
Recursive vs Recurrent representation



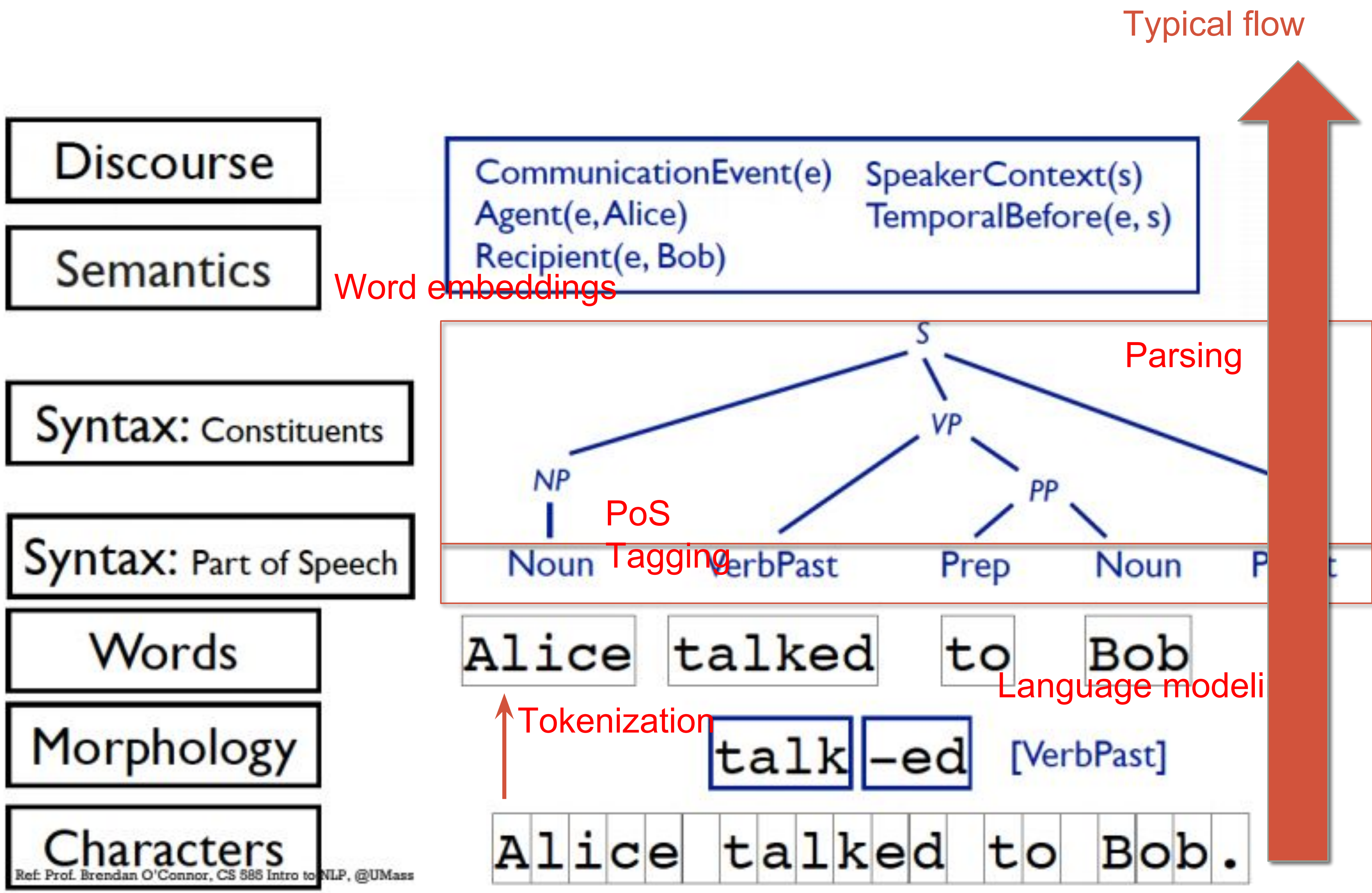
Towards unsupervised dependency parsing



Towards unsupervised dependency parsing



What is this piece? Next lecture! (attention modeling)

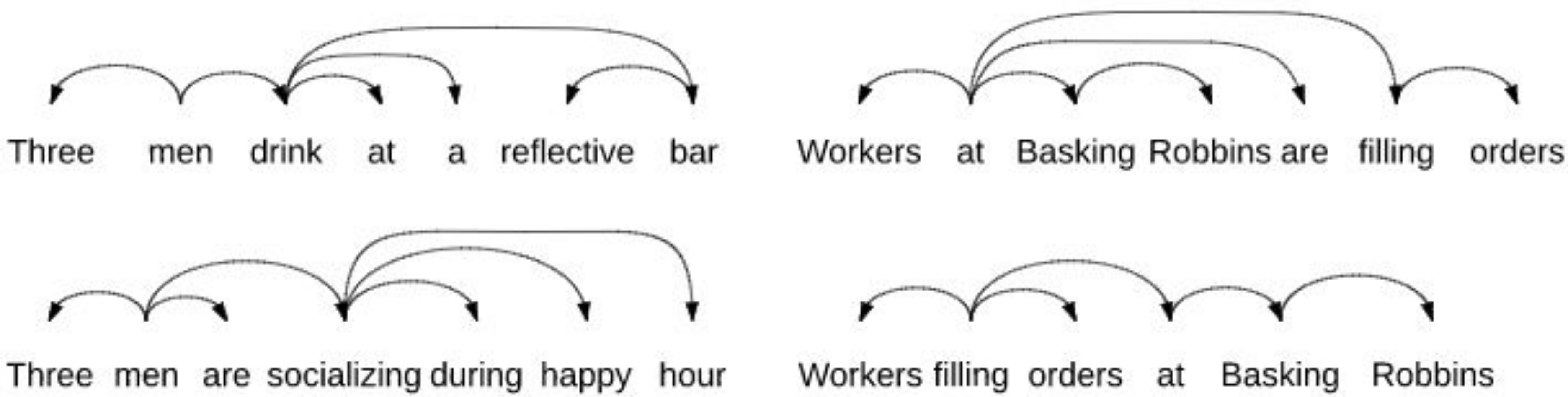


Typical flow
Discourse
Semantics
Syntax: Constituents
Syntax: Part of Speech
Words
Morphology
Characters
CommunicationEvent(e)
Agent(e, Alice)
Recipient(e, Bob)
SpeakerContext(s)
TemporalBefore(e, s)
Word embeddings
S
VP
NP
PP
Parsing
PoS Tagging
Noun
VerbPast
Prep
Noun
Alice
talked
to
Bob
Language modeli
Tokenization
talk
-ed
[VerbPast]
Alice talked to Bob.
Ref: Prof. Brendan O'Connor, CS 588 Intro to NLP, @UMass

An top down end-to-end approach

- Input: text
- Output: some task, sentiment analysis score
- Automatically gets parse tree (without any treebank corpus)

Example of generated parse tree



https://arxiv.org/pdf/1705.09207.pdf

Can also use recursive neural networks to learn unsupervised parse trees
Example https://arxiv.org/pdf/1707.02786.pdf

Summary

- Parsing
- Types of grammars
  - Context Free Grammar
  - Probabilistic Context Free Grammar (Constituency parsers)
    - CYK parser
  - Dependency Grammar (Dependency parsers)
    - Transition-based parsing
    - Recursive Neural networks