COMP 790.139 (Fall 2016)
Natural Language Processing
(with some vision, robotics, and deep learning)

Aug 31, 2016

Mohit Bansal

(various slides adapted/borrowed from courses by Dan Klein, Chris Manning)

Announcements

- Must have received system email from me reg. readings
- Tentative schedule and paper list decided
- Paper list on website
- Then start deciding which topic you would like to present
- Send me top choices and I will create an assignment ((I will notify you via email)
- Some topics will be in groups

NLP Basics and Core Tasks 1

- Part-of-Speech Tagging
- Syntactic Parsing: Constituent, Dependency, CCG, others
- Coreference Resolution
- Distributional Semantics: PMI, Neural, CCA
- Compositional Semantics: Logical-form, Semantic Parsing, Vector-form, Neural (RNNs/CNNs)

Note: we will be covering some of these briefly (so as to be able to reach the paper reading weeks quickly), so definitely follow up for more details in the prescribed readings and references, and talk to me in office hours!

Part-of-Speech Tagging

- Tag sequence of words with syntactic categories (noun, verb, preposition, …)

- Useful in itself:
  - Text-to-speech: *read, lead, record*
  - Lemmatization: *saw*[v] → *see*, *saw*[n] → *saw*
  - Shallow Chunking: grep {JJ | NN}* {NN | NNS}

- Useful for downstream tasks (e.g., in parsing, and as features in various word/text classification tasks)

- Demos: http://nlp.stanford.edu:8080/corenlp/

Penn Treebank Tagset

CC	conjunction, coordinating	and both but either or
CD	numeral, cardinal	mid-1890 nine-thirty 0.5 one
DT	determiner	a all an every no that the
EX	existential there	there
FW	foreign word	gemeinschaft hund ich jeux
IN	preposition or conjunction, subordinating	among whether out on by if
JJ	adjective or numeral, ordinal	third ill-mannered regrettable
JJR	adjective, comparative	braver cheaper taller
JJS	adjective, superlative	bravest cheapest tallest
MD	modal auxiliary	can may might will would
NN	noun, common, singular or mass	cabbage thermostat investment subhumanity
NNP	noun, proper, singular	Motown Cougar Yvette Liverpool
NNPS	noun, proper, plural	Americans Materials States
NNS	noun, common, plural	undergraduates bric-a-brac averages
POS	genitive marker	' 's
PRP	pronoun, personal	hers himself it we them
PRP$	pronoun, possessive	her his mine my our ours their thy your
RB	adverb	occasionally maddeningly adventurously
RBR	adverb, comparative	further gloomier heavier less-perfectly
RBS	adverb, superlative	best biggest nearest worst
RP	particle	aboard away back by on open through
TO	"to" as preposition or infinitive marker	to
UH	interjection	huh howdy uh whammo shucks heck
VB	verb, base form	ask bring fire see take
VBD	verb, past tense	pleaded swiped registered saw
VBG	verb, present participle or gerund	stirring focusing approaching erasing
VBN	verb, past participle	dilapidated imitated reunifed unsettled
VBP	verb, present tense, not 3rd person singular	twist appear comprise mold postpone
VBZ	verb, present tense, 3rd person singular	bases reconstructs marks uses
WDT	WH-determiner	that what whatever which whichever
WP	WH-pronoun	that what whatever which who whom
WP$	WH-pronoun, possessive	whose
WRB	Wh-adverb	however whenever where why

Part-of-Speech Ambiguities

- A word can have multiple parts of speech

VBD		VB			
VBN	VBZ	VBP	VBZ		
NNP	NNS	NN	NNS	CD	NN
Fed	raises	interest	rates	0.5	percent

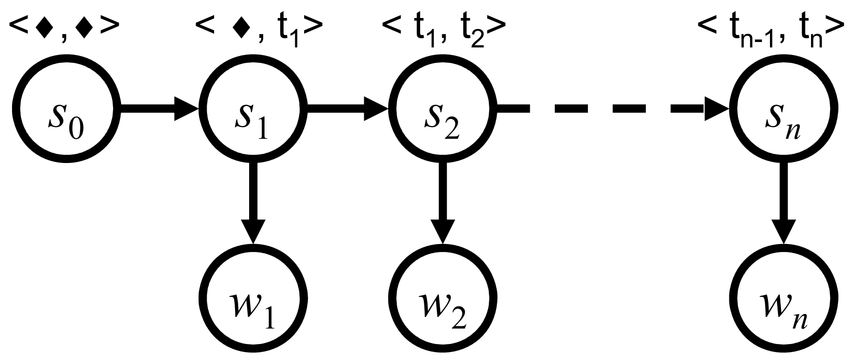
Mrs./NNP Shaefer/NNP never/RB got/VBD **around/RP** to/TO joining/VBG

All/DT we/PRP gotta/VBN do/VB is/VBZ go/VB **around/IN** the/DT corner/NN

Chateau/NNP Petrus/NNP costs/VBZ **around/RB** 250/CD

- Disambiguating features: lexical identity (word), context, morphology (suffixes, prefixes), capitalization, gazetteers (dictionaries), …

Classic Solution: HMMs



$$P(\mathbf{s}, \mathbf{w}) = \prod_i P(s_i | s_{i-1}) P(w_i | s_i)$$

- Trigram HMM: states = tag-pairs
- Estimating Transitions: Standard smoothing w/ backoff
- Estimating Emissions: Use unknown word classes (affixes, shapes) and estimate P(t|w) and invert
- Inference: choose most likely (Viterbi) sequence under model

[Brants, 2000]

POS Tagging: Other Models

- Discriminative sequence models with richer features: MEMMs, CRFs (SoA ~= 97%/90% known/unknown)

- Universal POS tagset for multilingual and cross-lingual tagging and parsing [Petrov et al., 2012]

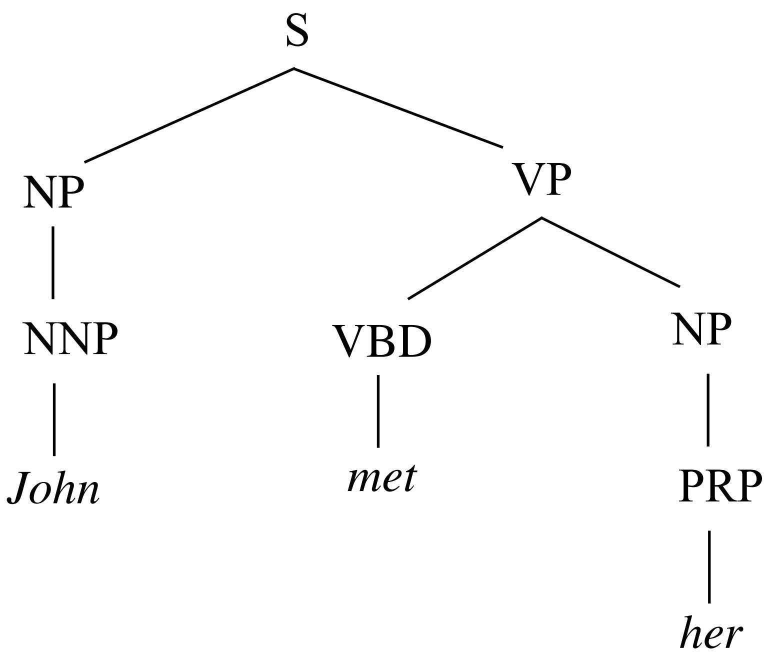
  12 tags: NOUN, VERB, ADJ, ADV, PRON, DET, ADP, NUM, CONJ, PRT, ., X

- Unsupervised tagging also works reasonably well!
  [Yarowsky et al., 2001; Xi and Hwa, 2005; Berg-Kirkpatrick et al., 2010; Christodoulopoulos et al., 2010; Das and Petrov, 2011]

[Brill, 1995; Ratnaparkhi, 1996; Toutanova and Manning, 2000; Toutanova et al., 2003]

Syntactic Parsing -- Constituent

- Phrase-structure parsing or Bracketing



- Demos: http://tomato.banatao.berkeley.edu:8080/parser/parser.html

Probabilistic Context-free Grammars

- A context-free grammar is a tuple <*N, T, S, R*>

  N : the set of non-terminals

  Phrasal categories: S, NP, VP, ADJP, etc.

  Parts-of-speech (pre-terminals): NN, JJ, DT, VB

  T : the set of terminals (the words)

  S : the start symbol

  Often written as ROOT or TOP

  Not usually the sentence non-terminal S

  R : the set of rules

  Of the form $X \rightarrow Y_1\ Y_2 \ldots Y_k$, with $X, Y_i \in N$

  Examples: $S \rightarrow NP\ VP$, $VP \rightarrow VP\ CC\ VP$

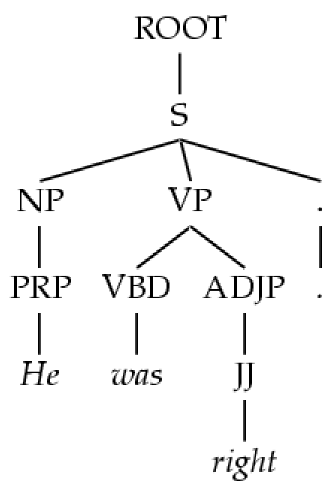
  Also called rewrites, productions, or local trees

Probabilistic Context-free Grammars

- A PCFG:
  - Adds a top-down production probability per rule $P(Y_1 \; Y_2 \; \ldots \; Y_k \mid X)$
  - Allows us to find the ‘most probable parse’ for a sentence
  - The probability of a parse is just the product of the probabilities of the individual rules

Treebank PCFG

- Need a PCFG for broad coverage parsing
- Extracting a grammar right off the trees is not effective:



ROOT → S 1

S → NP VP . 1

NP → PRP 1

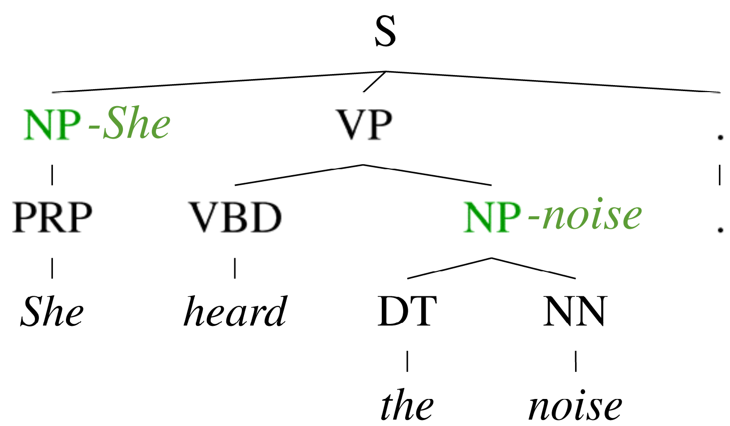
VP → VBD ADJP 1

…..

Model	*F1*
Baseline	72.0

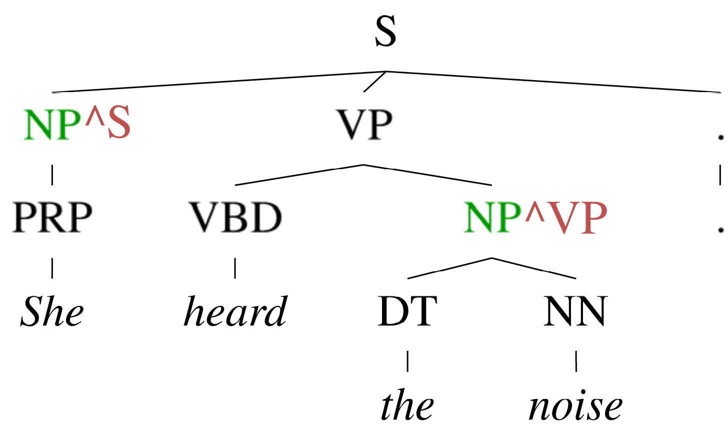
[Charniak, 1996]

Grammar Refinement



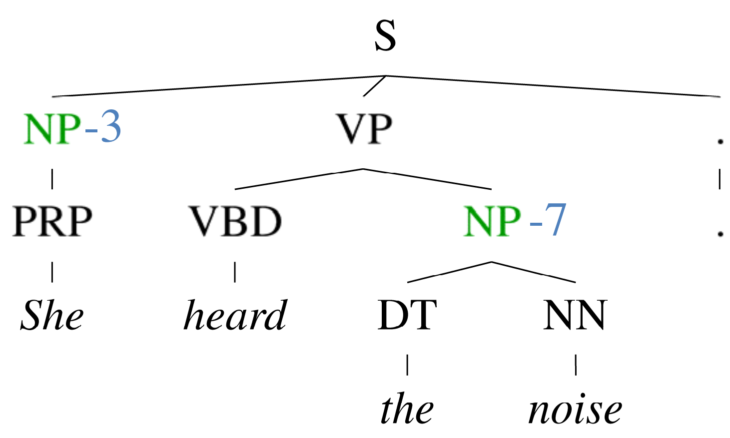
- Conditional independence assumptions often too strong! Not every NP expansion can fill every NP slot
- Better results by enriching the grammar e.g.,
  - Lexicalization [Collins, 1999; Charniak, 2000]

Grammar Refinement



- Conditional independence assumptions often too strong! Not every NP expansion can fill every NP slot
- Better results by enriching the grammar e.g.,
  - Lexicalization [Collins, 1999; Charniak, 2000]
  - Markovization, Manual Tag-splitting [Johnson, 1998; Klein & Manning, 2003]

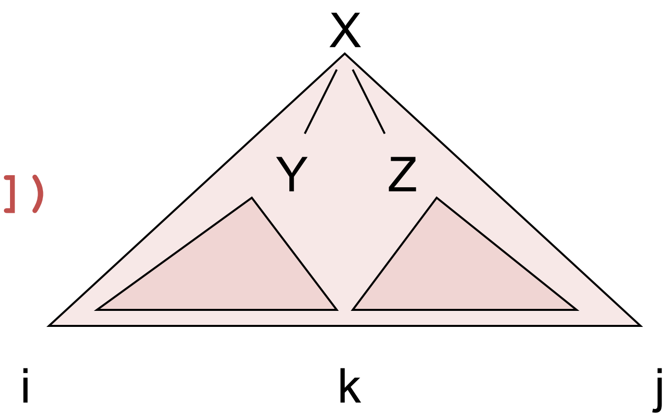
Grammar Refinement



- Conditional independence assumptions often too strong! Not every NP expansion can fill every NP slot
- Better results by enriching the grammar e.g.,
  - Lexicalization [Collins, 1999; Charniak, 2000]
  - Markovization, Manual Tag-splitting [Johnson, 1998; Klein & Manning, 2003]
  - Latent Tag-splitting [Matsuzaki et al., 2005; Petrov et al., 2006]

CKY Parsing Algorithm (Bottom-up)

```
bestScore(s)
 for (i : [0,n-1])
   for (X : tags[s[i]])
     score[X][i][i+1] = tagScore(X,s[i])
 for (diff : [2,n])
   for (i : [0,n-diff])
     j = i + diff
     for (X->YZ : rule)
       for (k : [i+1, j-1])
         score[X][i][j] = max{score[X][i][j], score(X->YZ)
                                              *score[Y][i][k]
                                              *score[Z][k][j]}
```



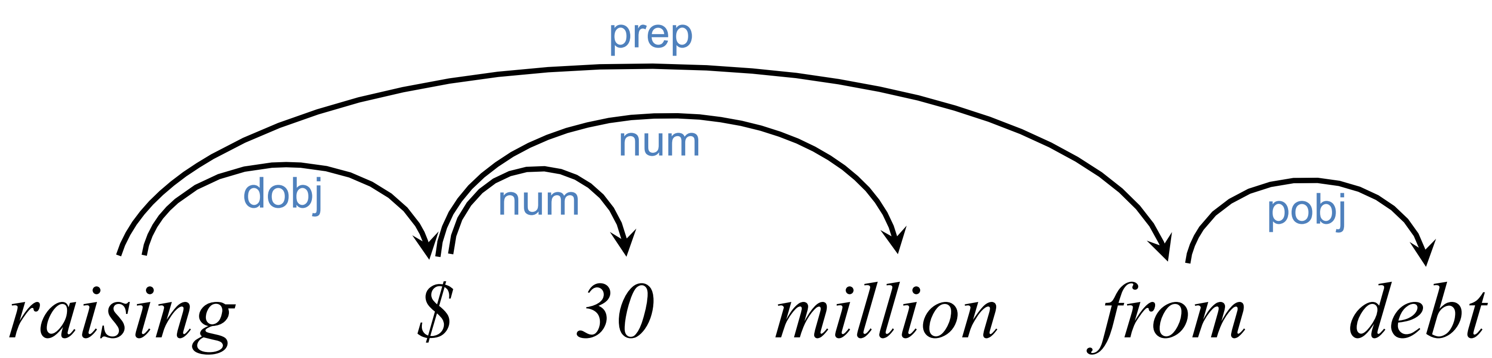
[Cocke, 1970; Kasami, 1965; Younger, 1967]

Some Results

- Collins, 1999 → 88.6 F1 (generative lexical)
- Charniak and Johnson, 2005 → 89.7 / 91.3 F1 (generative lexical / reranking)
- Petrov et al., 2006 → 90.7 F1 (generative unlexical)
- McClosky et al., 2006 – 92.1 F1 (generative + reranking + self-training)

Syntactic Parsing -- Dependency

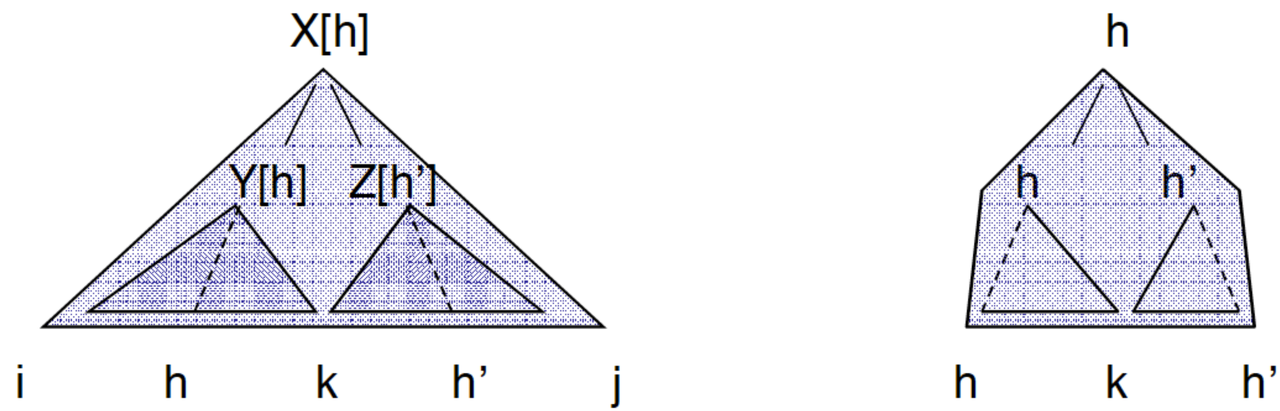
- Predicting directed head-modifier relationship pairs



- Demos: http://nlp.stanford.edu:8080/corenlp/

Syntactic Parsing -- Dependency

- Pure (projective, 1st order) dependency parsing is only cubic [Eisner, 1996]



- Non-projective dependency parsing useful for Czech & other languages – MST algorithms [McDonald et al., 2005]

Parsing: Other Models and Methods

- Combinatory Categorial Grammar [Steedman, 1996, 2000; Clark and Curran, 2004]
- Transition-based Dependency Parsing [Yamada and Matsumoto, 2003; Nivre, 2003]
- Tree-Insertion Grammar, DOP [Schabes and Waters, 1995; Hwa, 1998; Scha, 1990; Bod, 1993; Goodman, 1996; Bansal and Klein, 2010]
- Tree-Adjoining Grammar [Resnik, 1992; Joshi and Schabes, 1998; Chiang, 2000]
- Shift-Reduce Parser [Nivre and Scholz, 2004; Sagae and Lavie, 2005]
- Other: Reranking, A*, K-Best, Self-training, Co-training, System Combination, Cross-lingual Transfer [Sarkar, 2001; Steedman et al., 2003; Charniak and Johnson, 2005; Hwa et al., 2005; Huang and Chiang, 2005; McClosky et al., 2006; Fossum and Knight, 2009; Pauls and Klein, 2009; McDonald et al., 2011]
- Other Demos: http://svn.ask.it.usyd.edu.au/trac/candc/wiki/Demo, http://4.easy-ccg.appspot.com/

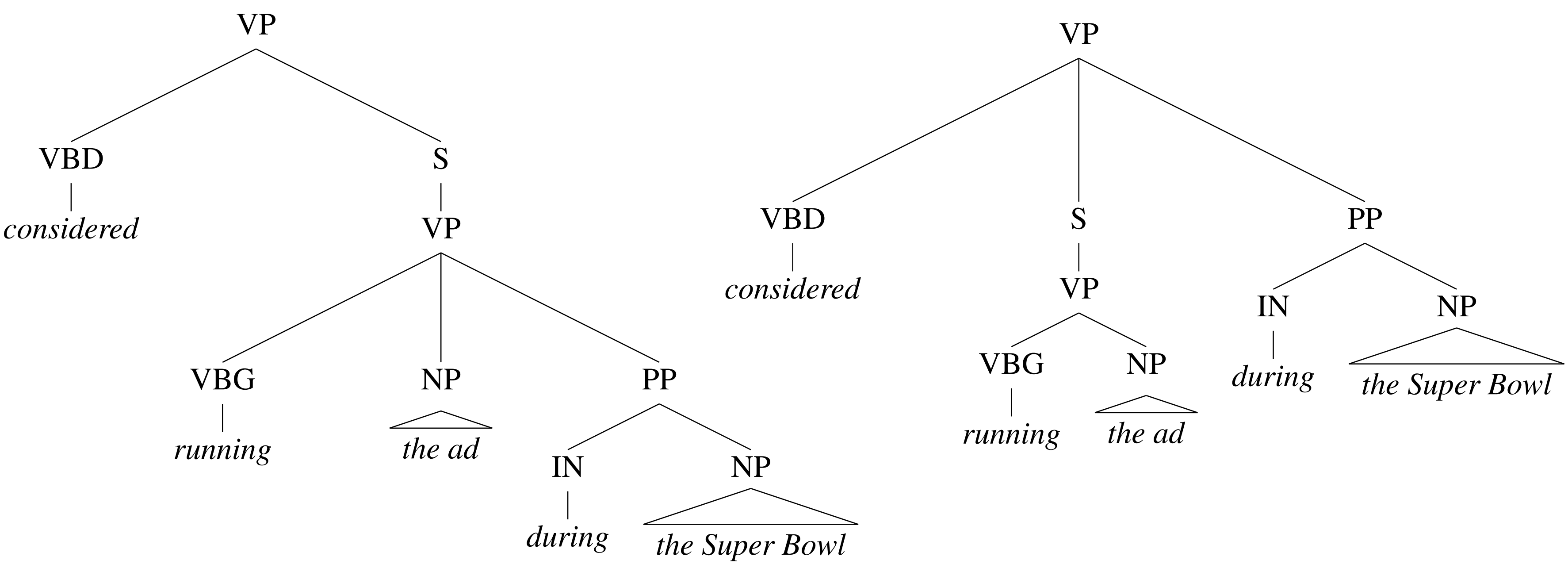
World Knowledge is Important

Web Features for Syntactic Parsing

Dependency:

They considered running the ad during the Super Bowl.

Constituent:



[Nakov and Hearst 2005; Pitler et al., 2010; Bansal and Klein, 2011]

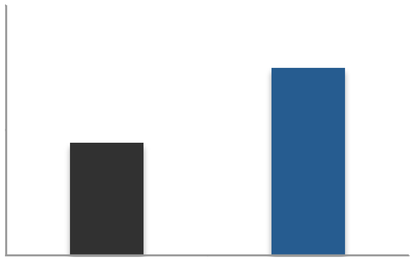
Web Features for Syntactic Parsing

They considered running the ad during the Super Bowl.

Web Ngrams

count(*running it during*) $>$ count(*considered it during*)

- 7-10% relative error reduction over 90-92% parsers

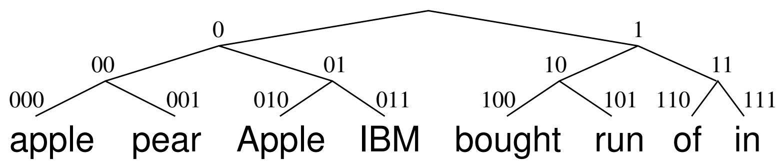


[Bansal and Klein, 2011]

Unsup. Representations for Parsing

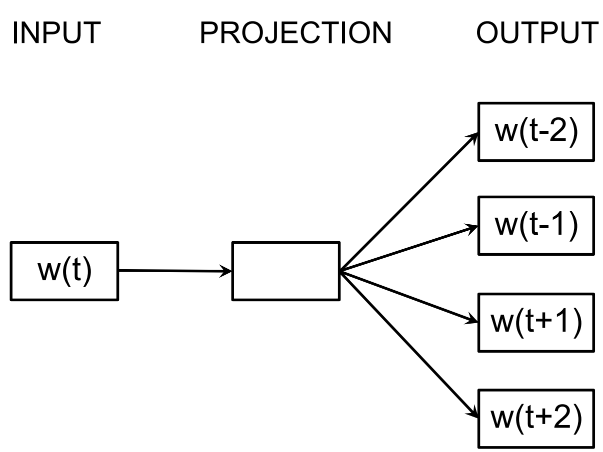
- Discrete or continuous, trained on large amounts of context

- BROWN (Brown et al., 1992):



apple	→	000
pear	→	001
Apple	→	010

- SKIPGRAM (Mikolov et al., 2013):



apple	→	[0.65	0.15	-0.21	0.15	0.70	-0.90]
pear	→	[0.51	0.05	-0.32	0.20	0.80	-0.95]
Apple	→	[0.11	0.33	0.51	-0.05	-0.41	0.50]

[Koo et al., 2008; Bansal et al., 2014]

Unsup. Representations for Parsing

- Condition on dependency context instead of linear, then convert each dependency to a tuple:

$$\begin{array}{ccccc} \text{dep label} & \text{grandparent} & \text{parent} & \text{child} & \text{dep label} \\ \big[\mathit{PMOD}_{<L>} & \mathit{regulation}_{<G>} & \mathit{of} & \mathit{safety} & \mathit{PMOD}_{<L>}\big] \end{array}$$

[*Mr., Mrs., Ms., Prof., III, Jr., Dr.*]

[*Jeffrey, William, Dan, Robert, Stephen, Peter, John, Richard, ...*]

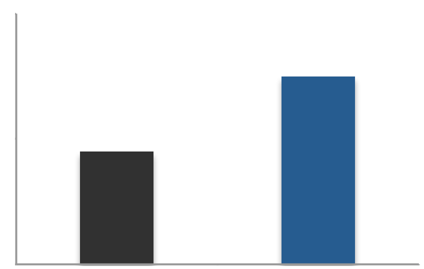
[*Portugal, Iran, Cuba, Ecuador, Greece, Thailand, Indonesia, ...*]

[*his, your, her, its, their, my, our*]

[*Your, Our, Its, My, His, Their, Her*]

[*truly, wildly, politically, financially, completely, potentially, ...*]

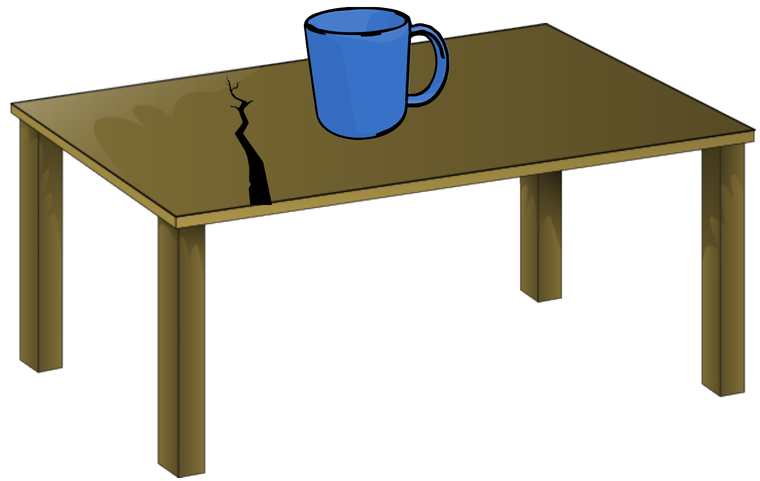
- 10% rel. error reduction over 90-92% parsers



[Bansal et al., 2014]

Visual Recognition Cues

- Joint parsing and image recognition

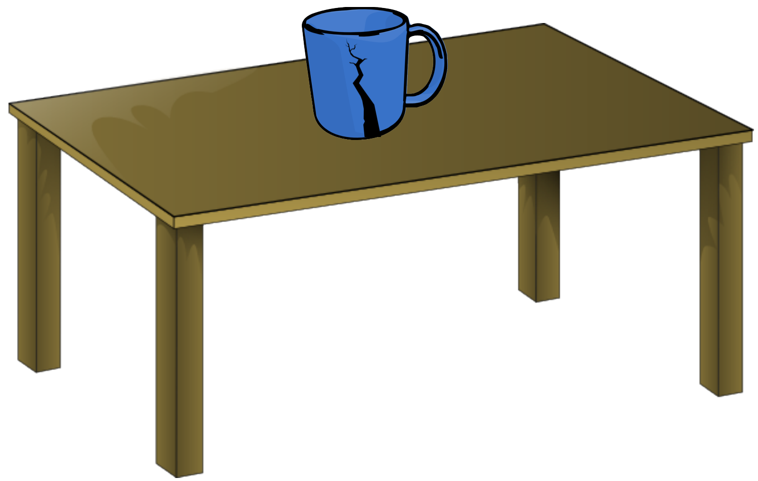


the mug on the table with a crack

[Christie et al., 2016]

Visual Recognition Cues

- Joint parsing and image recognition



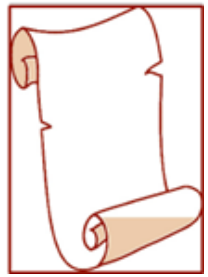
the mug on the table with a crack

red chair and table
light green table

[Christie et al., 2016]

10-min Break?

Coreference Resolution

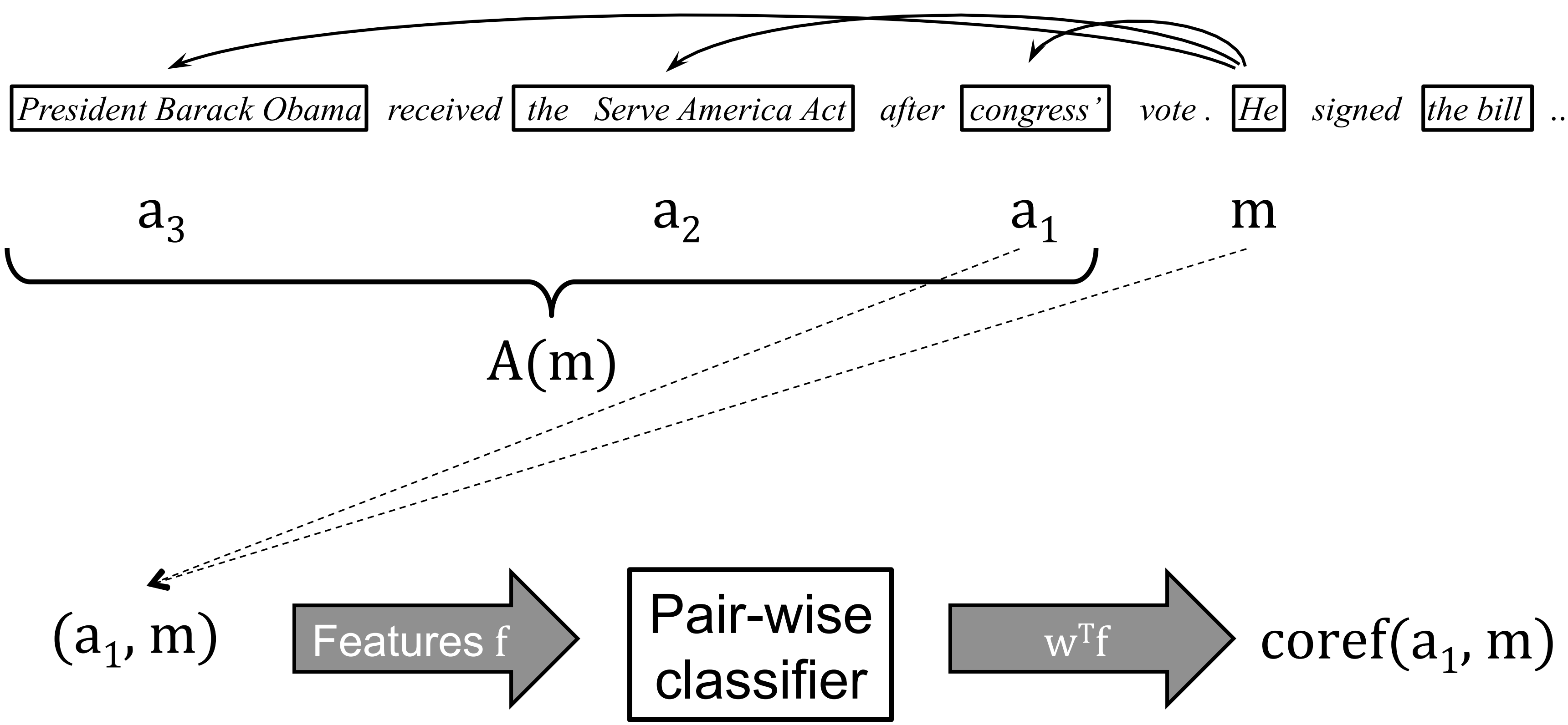



President Barack Obama received the Serve America Act after congress' vote. He signed the bill last Thursday. The president said it would greatly increase service opportunities for the American people.

- Mentions to entity/event clusters
- Demos: http://nlp.stanford.edu:8080/corenlp/process

Mention-pair Models

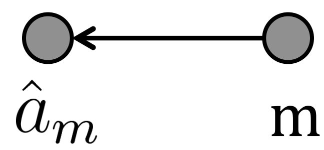
- Pair-wise classification approach:

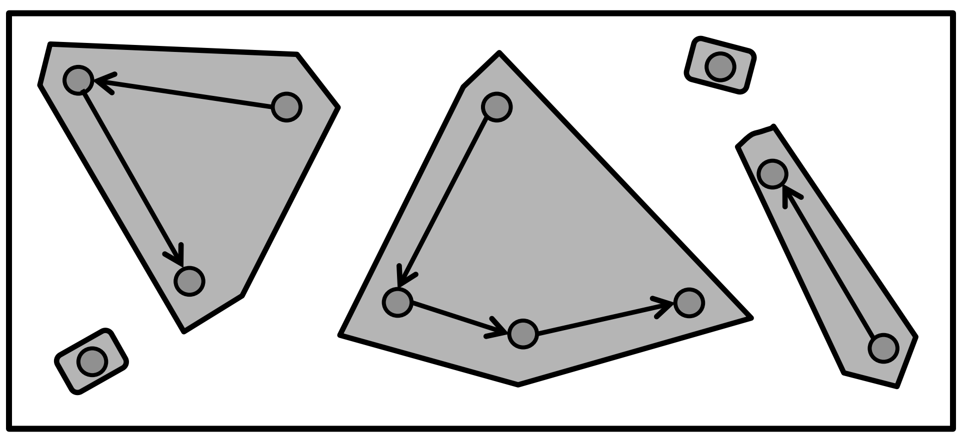


[Soon et al. 2001, Ng and Cardie 2002; Bengtson and Roth, 2008; Stoyanov et al., 2010]

Mention-pair Model

For each mention m, $\hat{a}_m = \underset{a_i \in A(m)}{\operatorname{argmax}}\, coref(a_i, m)$





[Soon et al. 2001, Ng and Cardie 2002; Bengtson and Roth, 2008; Stoyanov et al., 2010]

Standard features

Type	Feature	Description
LEXICAL	SOON_STR	Do the strings match after removing determiners ?
GRAMMATICAL	NUMBER	Do NP_i and NP_j agree in number ?
	GENDER	Do NP_i and NP_j agree in gender ?
	APPOSITIVE	Are the NPs in an appositive relationship ?
SEMANTIC	WORDNET_CLASS	Do NP_i and NP_j have the same WordNet class ?
	ALIAS	Is one NP an alias of the other ?
POSITIONAL	SENTNUM	Distance between the NPs in terms of # of sentences

- Weaknesses: All pairs, Transitivity/Independence errors (He – $Obama$ – She), Insufficient information

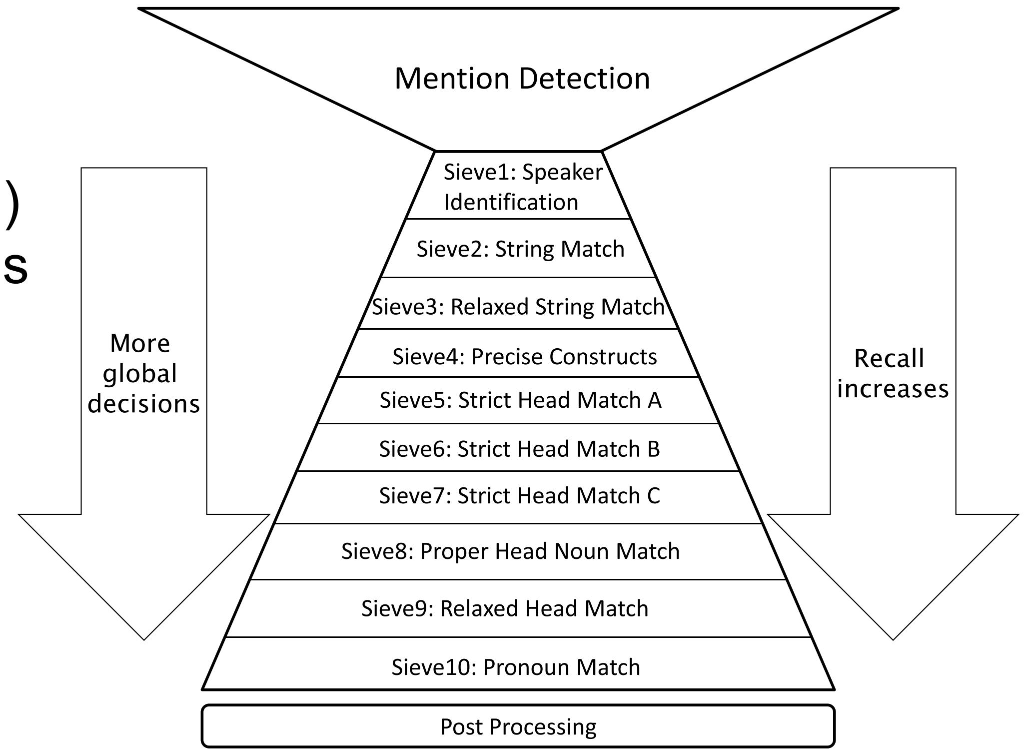
[Soon et al. 2001, Ng and Cardie 2002; Bengtson and Roth, 2008; Stoyanov et al., 2010]

Entity-centric Models

- Each coreference decision is globally informed by previously clustered mentions and their shared attributes

- Lee et al., 2013's deterministic (rule-based) system: multiple, cautious sieves from high to low precision

- Durrett et al., 2013's entity-level model is discriminative, probabilistic using factor graphs and BP

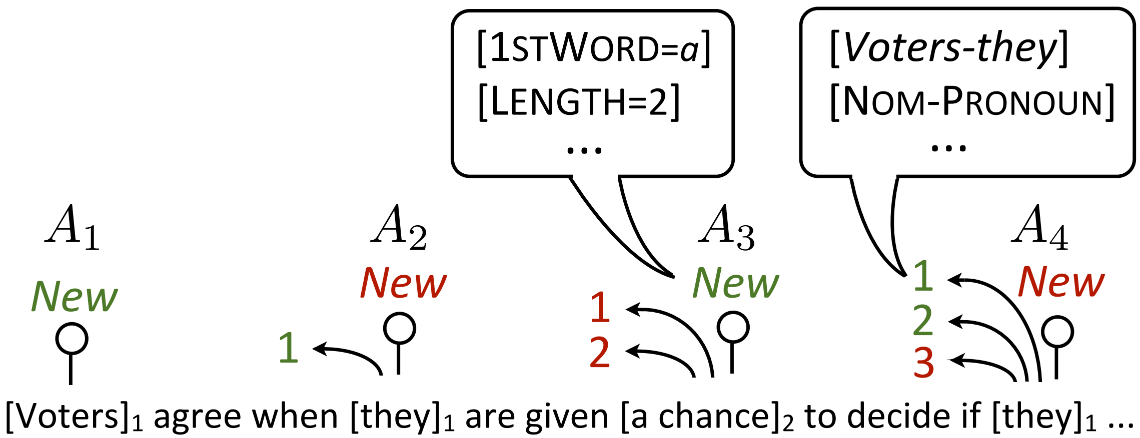


[Haghighi and Klein, 2009; Lee et al., 2013; Durrett et al., 2013]

Mention-Ranking Models (Learned)

- Log-linear model to select at most 1 antecedent for each mention or determine that it begins a new cluster

$$Pr(A_i = a|x) \propto \exp(w^\top f(i, a, x))$$



[Denis and Baldridge, 2008; Durrett and Klein, 2013]

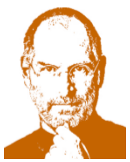
- Recent work (Wiseman et al., 2016, Clark & Manning, 2016) has used NNs for non-linear and vector-space coreference features to achieve SoA!

Adding Knowledge to Coref

- External corpora: Web, Wikipedia, YAGO, FrameNet, Gender/Number/Person lists/classifiers, 3D Images, Videos
- Methods:
  - Self-training, Bootstrapping
  - Co-occurrence, Distributional, and Pattern-based Features
  - Entity Linking
  - Visual Cues from 3D Images and Videos
- Daumé III and Marcu, 2005; Markert and Nissim, 2005; Bergsma and Lin, 2006; Ponzetto and Strube, 2006; Haghighi and Klein, 2009; Kobdani et al., 2011; Rahman and Ng, 2011; Bansal and Klein, 2012; Durrett and Klein, 2014; Kong et al., 2014; Ramanathan et al., 2014

Web Features for Coreference

count(*Obama * president*) vs count(*Jobs * president*)

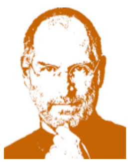


When Obama met Jobs , the president discussed the …

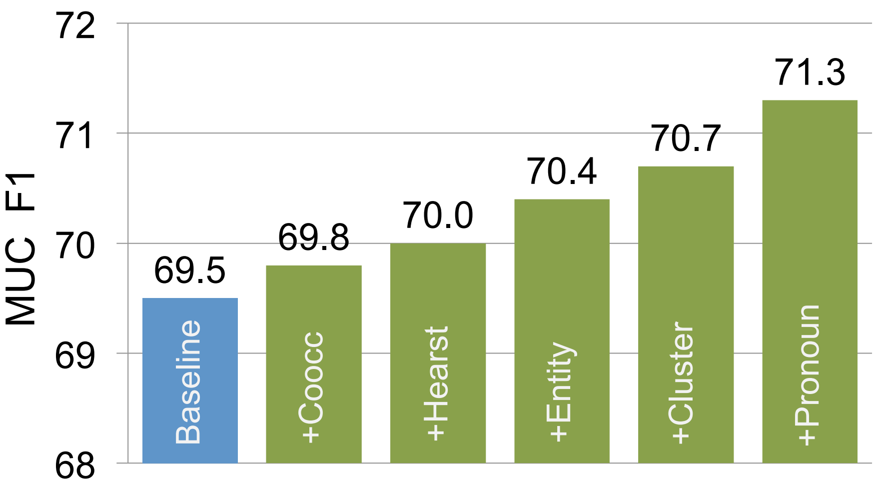
[Bansal and Klein, 2012]

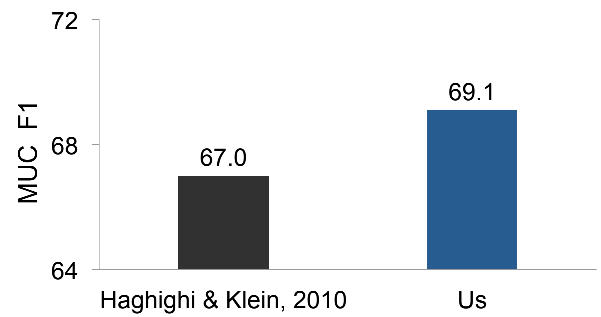
Web Features for Coreference

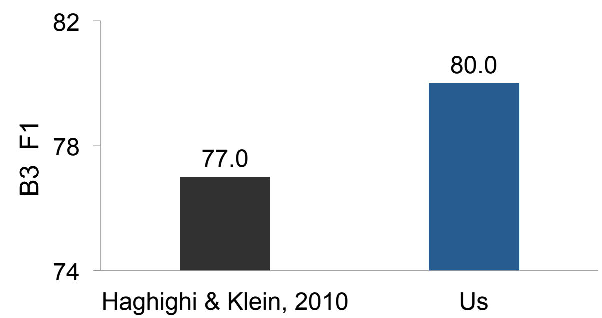
count(*Obama signed bills*) vs count(*Jobs signed bills*)



When Obama met Jobs , the … He signed bills that …



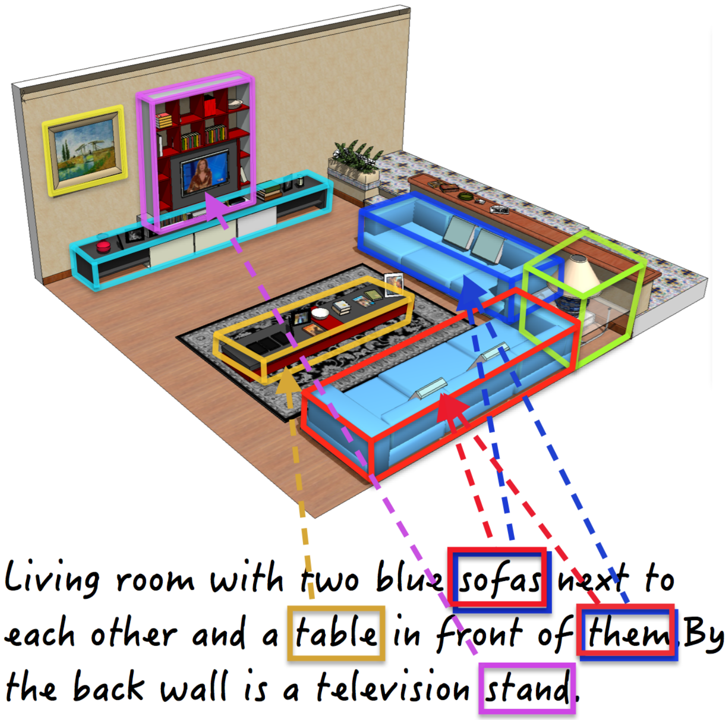




[Bansal and Klein, 2012]

Visual Cues for Coreference

- Joint coreference and 3D image recognition



	MUC			B^3		
Method	precision	recall	F1	precision	recall	F1
Stanford	61.56	62.59	62.07	75.05	76.15	75.59
Ours	83.69	51.08	63.44	88.42	70.02	78.15

[Kong, Lin, Bansal, Urtasun, and Fidler, 2014]

Distributional Semantics

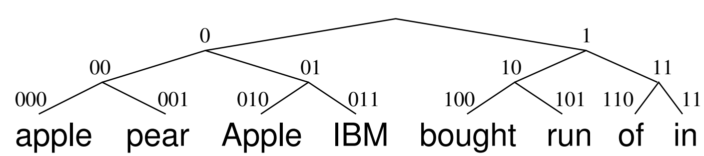
- Words occurring in similar context have similar linguistic behavior (meaning) [Harris, 1954; Firth, 1957]

$$\textit{food} \rightarrow \begin{bmatrix} 0.6 \\ -0.2 \\ 0.9 \\ 0.3 \\ -0.4 \\ 0.5 \end{bmatrix}$$

- Traditional approach: context-counting vectors
  - Count left and right context in window
  - Reweight with PMI or LLR
  - Reduce dimensionality with SVD or NNMF

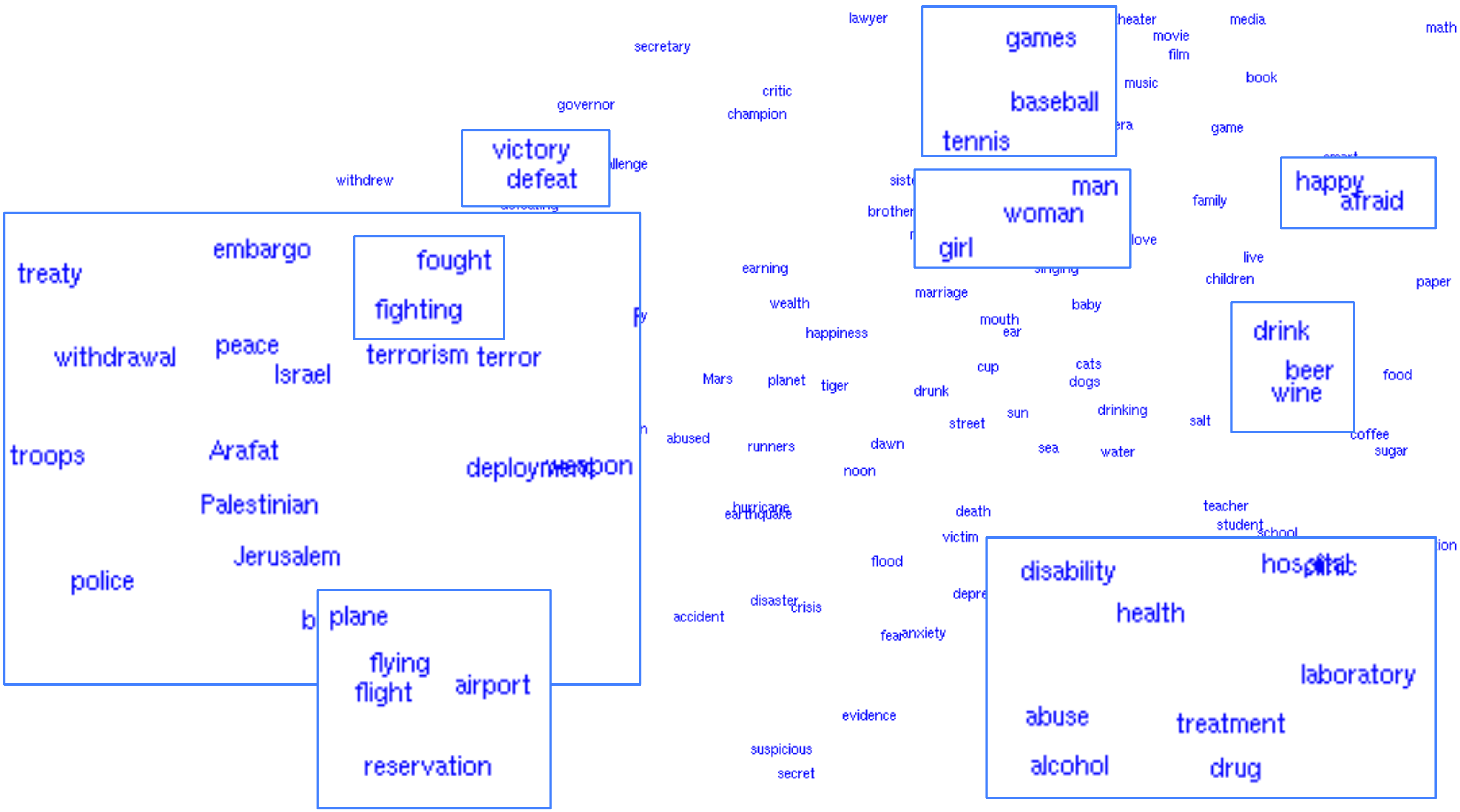
  [Pereira et al., 1993; Lund & Burgess, 1996; Lin, 1998; Lin and Pantel, 2001; Sahlgren, 2006; Pado & Lapata, 2007; Turney and Pantel, 2010; Baroni and Lenci, 2010]

- More word representations: hierarchical clustering based on bigram LM LL
  [Brown et al., 1992]



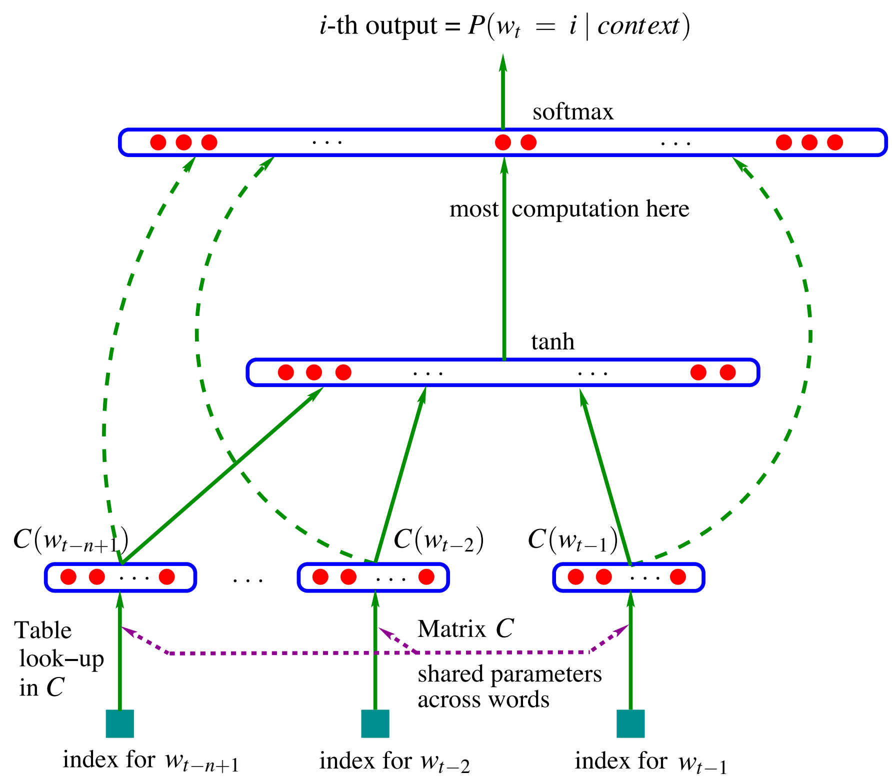
Unsupervised Embeddings

- Vector space representations learned on unlabeled linear context (i.e., left/right words): distributional semantics (Harris, 1954; Firth, 1957)



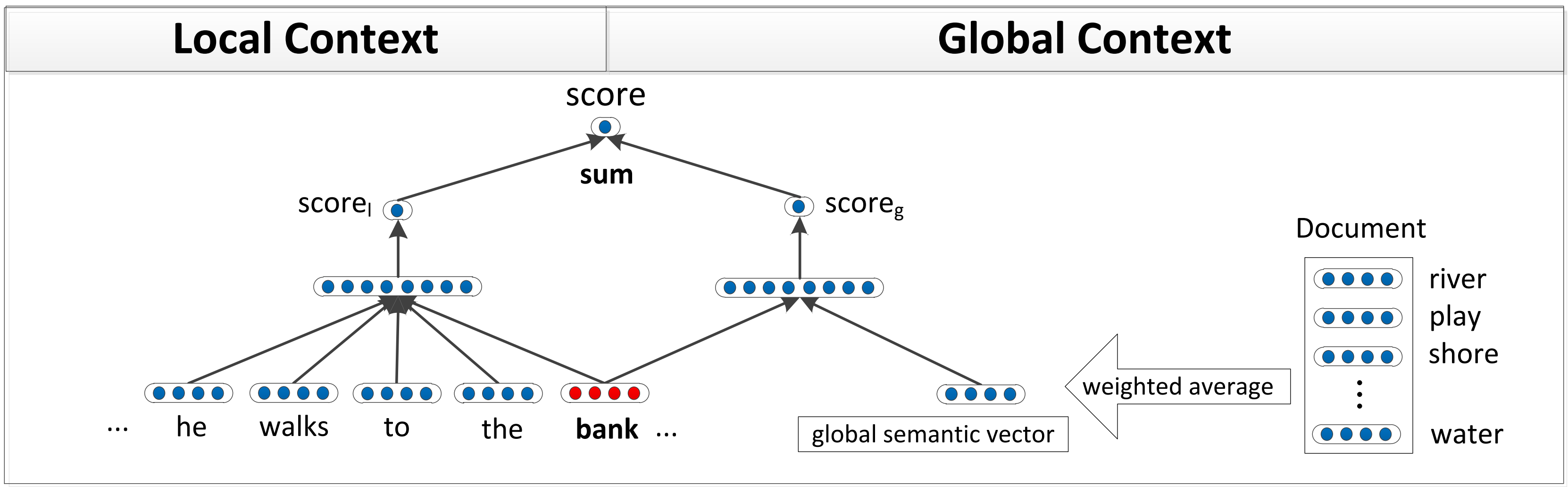
Distributional Semantics -- NNs

- Newer approach: context-predicting vectors (NNs)
  - SENNA [Collobert and Weston, 2008; Collobert et al., 2011]: Multi-layer DNN w/ ranking-loss objective; BoW and sentence-level feature layers, followed by std. NN layers. Similar to [Bengio et al., 2003].



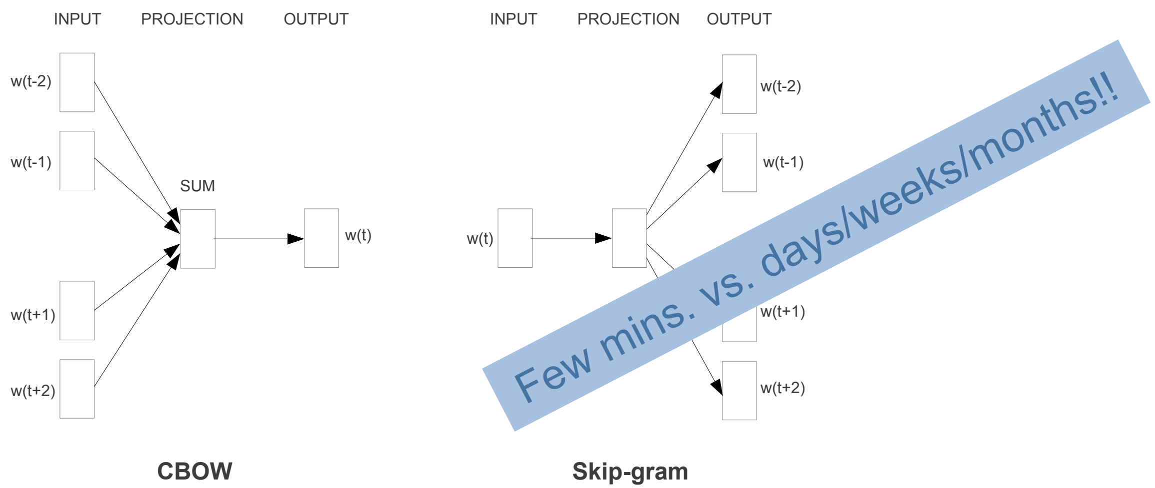
Distributional Semantics -- NNs

- HUANG [Huang et al., 2012]: Add global, document-level context



Distributional Semantics -- NNs

- CBOW, SKIP, word2vec [Mikolov et al., 2013]: Simple, super-fast NN w/ no hidden layer. Continuous BoW model predicts word given context, skip-gram model predicts surrounding context words given current word



  - Other: [Mnih and Hinton, 2007; Turian et al., 2010]

- Demos: https://code.google.com/p/word2vec, http://metaoptimize.com/projects/wordreprs/, http://ml.nec-labs.com/senna/

Distributional Semantics

- Other approaches: spectral methods, e.g., CCA
  - Word-context correlation [Dhillon et al., 2011, 2012]
  - Multilingual correlation [Faruqui and Dyer, 2014; Lu et al., 2015]
- Some recent directions: Train task-tailored embeddings to capture specific types of similarity/semantics, e.g.,
  - Dependency context [Bansal et al., 2014, Levy and Goldberg, 2014]
  - Predicate-argument structures [Hashimoto et al., 2014; Madhyastha et al., 2014]
  - Lexicon evidence (PPDB, WordNet, FrameNet) [Xu et al., 2014; Yu and Dredze, 2014; Faruqui et al., 2014; Wieting et al., 2015]