Large-Scale Language Learning

Research at Google

Slav Petrov

Thanks to
Ryan McDonald, Keith Hall, Alexander Rush, Dipanjan Das,
Hao Zhang, Michael Ringgaard, Terry Koo and Kuzman Ganchev
(a.k.a. the Natural Language Parsing Team at Google)

NLP: Where do we stand?

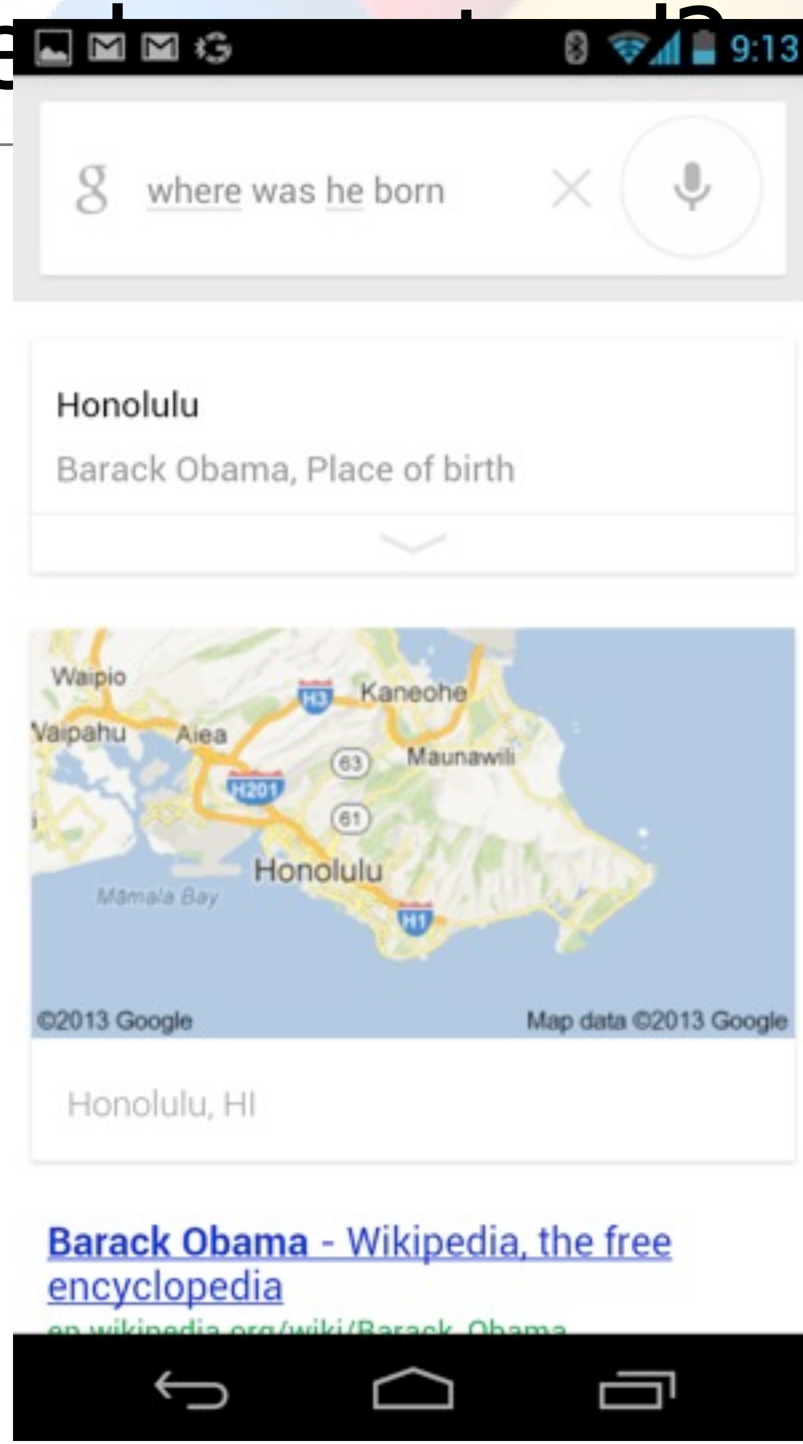
NLP: Where

9:13
who is the president of the US
Barack Obama
United States of America, President
President of the United States - Wikipedia, the free encyclopedia
en.wikipedia.org/.../President_of_the_ ...
On January 20, 2009, Barack Obama became the 44th and current president. On November 6, 2012, he was re-elected ...
List of Presidents of the United ... -
List of Presidents of the United States - Wikipedia, the free ...
en.wikipedia.org/.../List_of_Presidents_o...

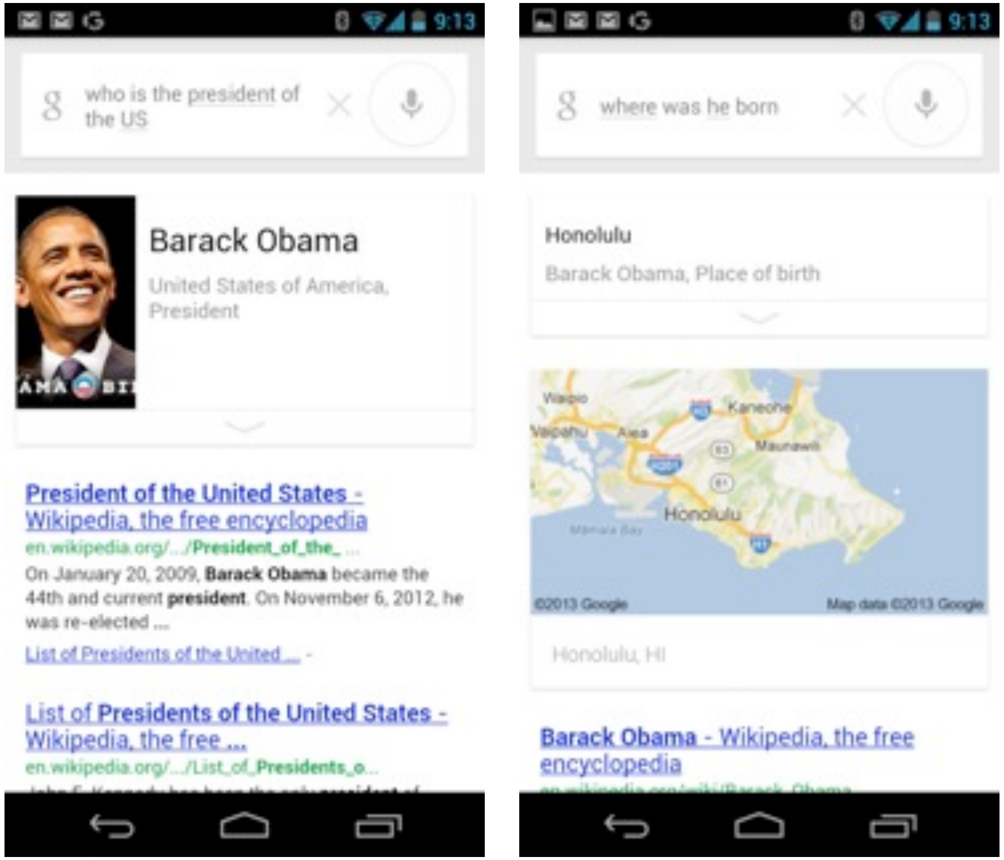
NLP: Where do we stand?

NLP: Where

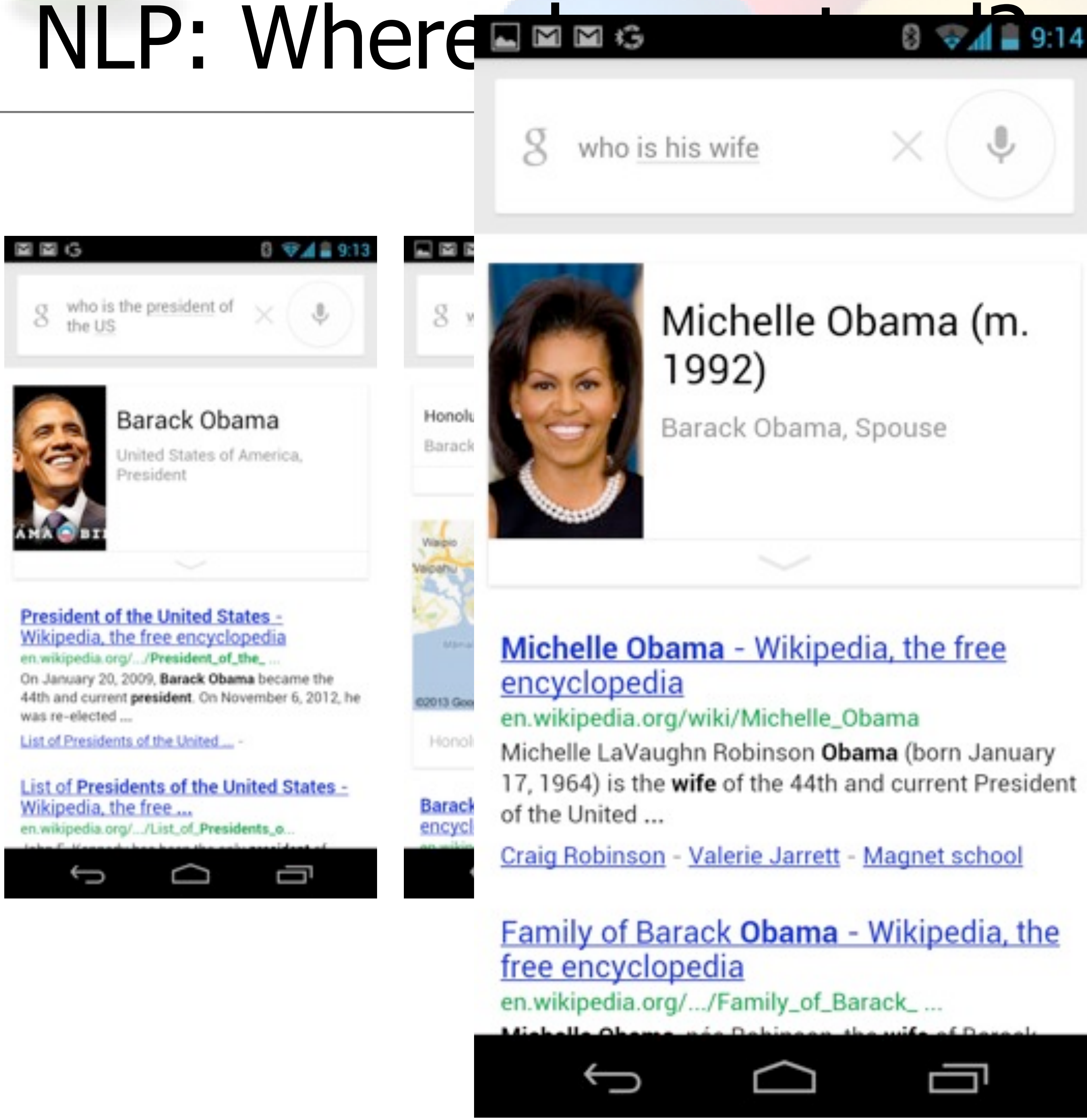

9:13
who is the president of the US
Barack Obama
United States of America, President
President of the United States - Wikipedia, the free encyclopedia
en.wikipedia.org/.../President_of_the_...
On January 20, 2009, Barack Obama became the 44th and current president. On November 6, 2012, he was re-elected ...
List of Presidents of the United ... -
List of Presidents of the United States - Wikipedia, the free ...
en.wikipedia.org/.../List_of_Presidents_o...


9:13
where was he born
Honolulu
Barack Obama, Place of birth
Waipio
Kaneohe
Waipahu
Aiea
Maunawili
H3
H201
63
61
Honolulu
Mamala Bay
H1
©2013 Google
Map data ©2013 Google
Honolulu, HI
Barack Obama - Wikipedia, the free encyclopedia

NLP: Where do we stand?



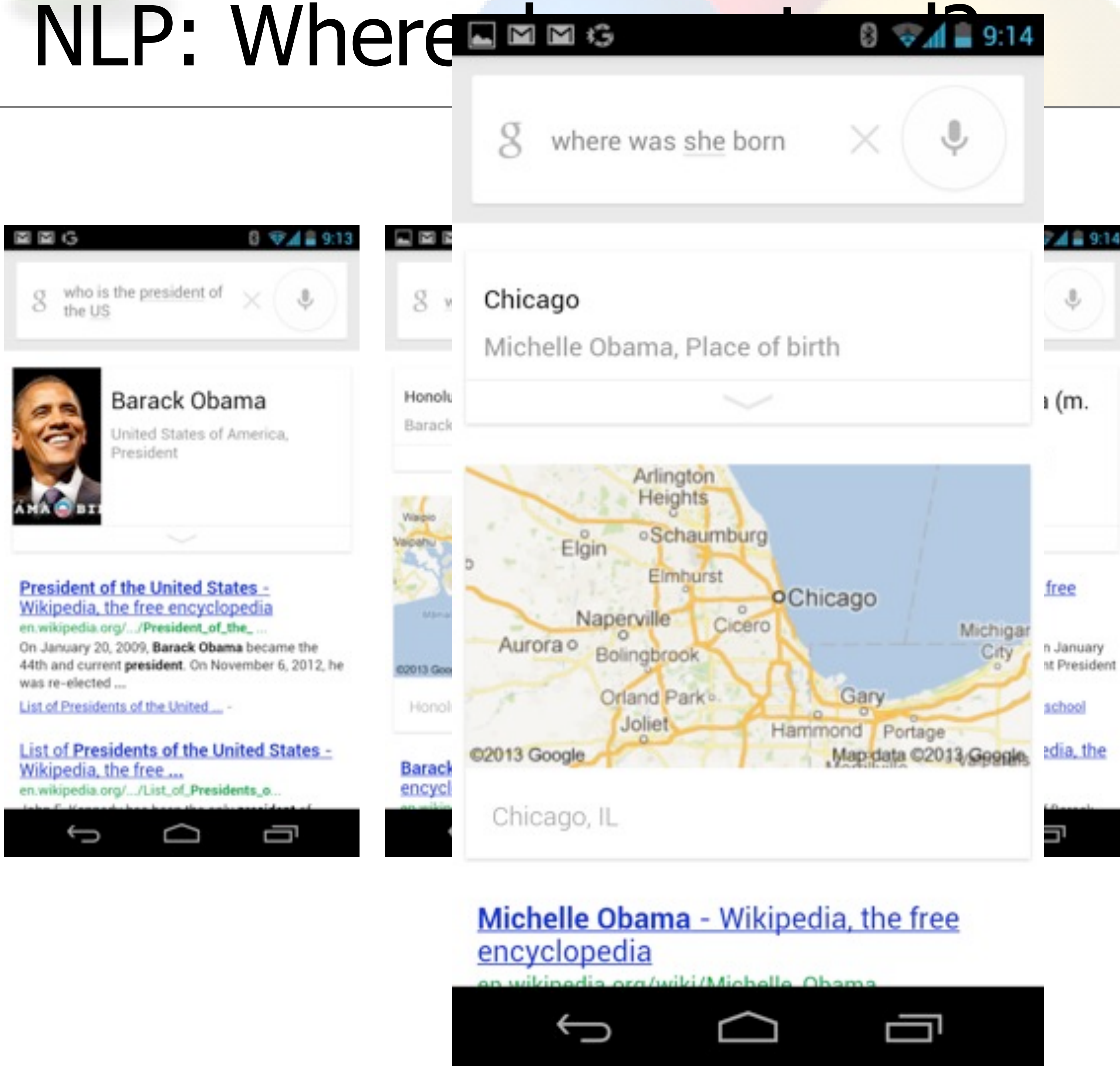
NLP: Where



9:13
who is the president of the US
Barack Obama
United States of America, President
President of the United States - Wikipedia, the free encyclopedia
en.wikipedia.org/.../President_of_the_...
On January 20, 2009, Barack Obama became the 44th and current president. On November 6, 2012, he was re-elected ...
List of Presidents of the United ...
List of Presidents of the United States - Wikipedia, the free ...
en.wikipedia.org/.../List_of_Presidents_o...
9:14
who is his wife
Michelle Obama (m. 1992)
Barack Obama, Spouse
Michelle Obama - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Michelle_Obama
Michelle LaVaughn Robinson Obama (born January 17, 1964) is the wife of the 44th and current President of the United ...
Craig Robinson - Valerie Jarrett - Magnet school
Family of Barack Obama - Wikipedia, the free encyclopedia
en.wikipedia.org/.../Family_of_Barack_ ...

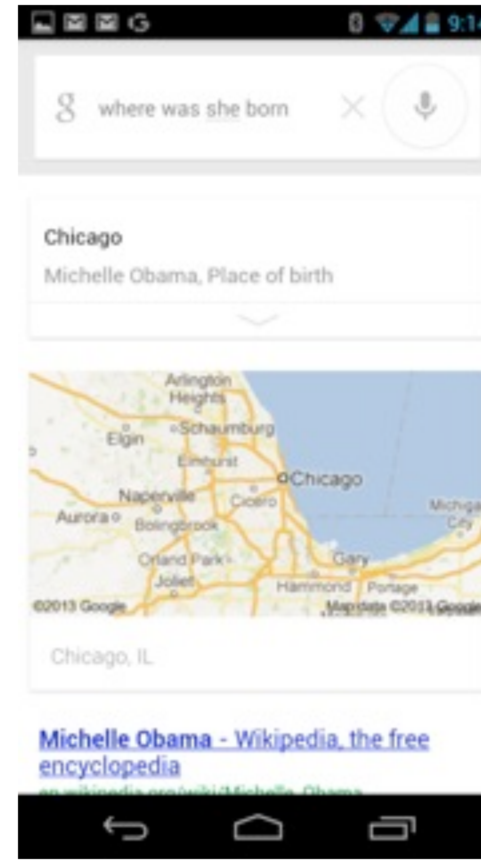
NLP: Where do we stand?

NLP: Where

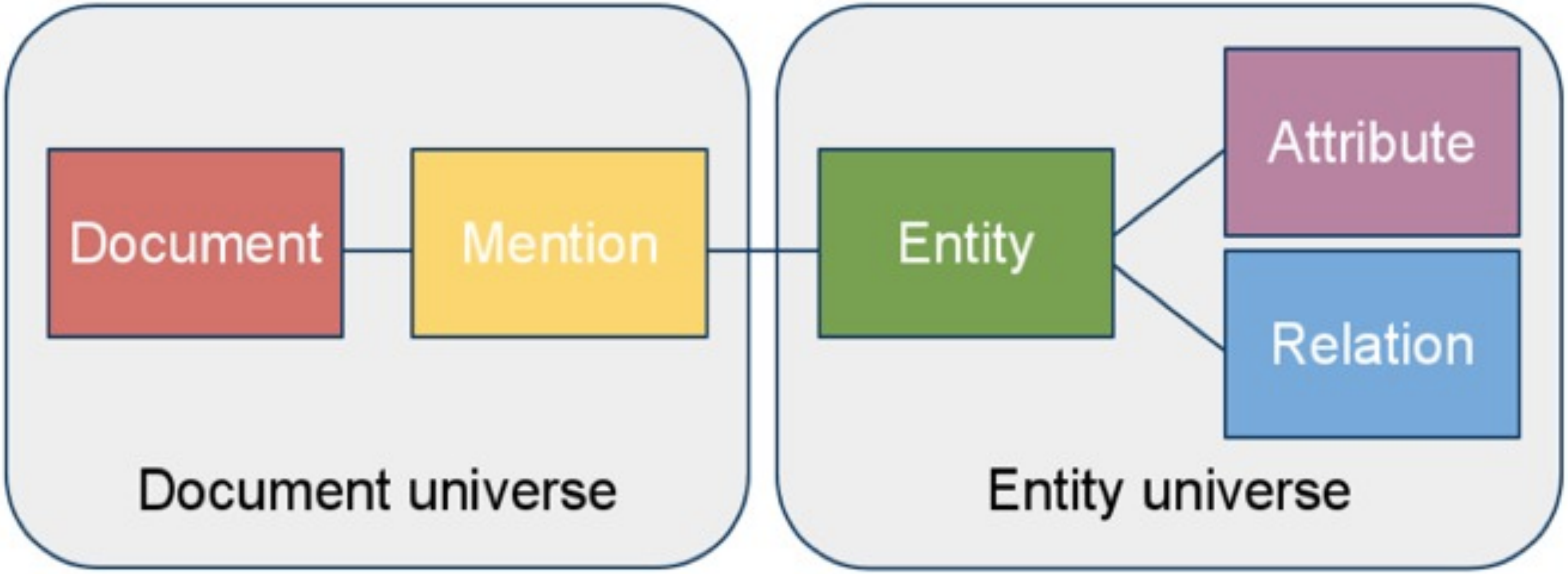


9:14
where was she born
Chicago
Michelle Obama, Place of birth
Arlington Heights
Schaumburg
Elgin
Elmhurst
Chicago
Naperville
Cicero
Aurora
Bolingbrook
Orland Park
Joliet
Gary
Hammond
Portage
Michigan City
©2013 Google
Map data ©2013 Google
Chicago, IL
Michelle Obama - Wikipedia, the free encyclopedia
9:13
who is the president of the US
Barack Obama
United States of America, President
President of the United States - Wikipedia, the free encyclopedia
en.wikipedia.org/.../President_of_the_...
On January 20, 2009, Barack Obama became the 44th and current president. On November 6, 2012, he was re-elected ...
List of Presidents of the United ...
List of Presidents of the United States - Wikipedia, the free ...
en.wikipedia.org/.../List_of_Presidents_o...

NLP: Where do we stand?



Building a Repository



Find *all* *Mentions* of Entities in Documents

- Resolve Mentions to the *Identities* of the Entities
- Extract typed *Relations* between Entities
- Extract typed *Attribute* values for Entities

Semantic Annotation of Documents

President Barack Obama has been re-elected to a second term, defeating Republican challenger Mitt Romney. America's first black president secured more than the 270 votes in the electoral college needed to win. In his victory speech before supporters in Chicago, Mr Obama said he would talk to Mr Romney about "where we can work together to move this country forward". Mr Obama prevailed despite lingering dissatisfaction with the economy and a hard-fought challenge by Mr Romney.

His Democrats also retained their majority in the Senate, which they have held since 2007. The Republicans kept control of the House of Representatives, which analysts say will likely result in more of the gridlock that characterized Mr Obama's first term, with the House and the president at loggerheads on most legislation.

In his address, the president challenged his opponents, asking them to work with him. With only Florida's 29 electoral votes still undecided, Mr Obama won 303 electoral votes to Mr Romney's 206. The popular vote, which is symbolically and politically important but not decisive in the race, remains very close.

Mr Obama congratulated Mr Romney and Republican vice-presidential candidate Paul Ryan on their hard-fought campaign. "We have picked ourselves up, we have fought our way back and we know in our hearts that for the United States of America the best is yet to come," he said. Mr Obama said he was returning to the White House "more determined, and more inspired than ever about the work there is to do, and the future that lies ahead". He pledged to work with Republican leaders in Congress to reduce the government's budget deficit, fix the tax code and reform the immigration system.

Semantic Annotation of Documents

E_1 **Barack Obama**	PER
Mentions	Barack Obama, Obama (6), Mr (6), President, He, His, he (3), his
Gender	Male ♂
Profile	Barack Obama
Freebase MID	/m/02mjmr
Wikipedia	Barack Obama

E_2 **term**	NON
Mentions	term

E_3 **Mitt Romney**	PER
Mentions	Mitt Romney, Romney (4), Mr (4), challenger
Gender	Male ♂
Profile	Mitt Romney
Freebase MID	/m/0271_s
Wikipedia	Mitt Romney

E_4 **Republican**	ORG
Mentions	Republican (3)
Profile	Republican Party (United States)
Freebase MID	/m/07wbk
Wikipedia	Republican Party (United States)

E_5 **president**	PER
Mentions	president

E_6 **United States of America**	LOC
Mentions	America, United States of America
Profile	United States
Freebase MID	/m/09c7w0
Wikipedia	United States

E_7 **votes**	NON

⟨President⟩$_1$ [Barack Obama]$_1$ has been re-elected to a second ⟨term⟩$_2$, defeating [Republican]$_4$ ⟨challenger⟩$_3$ [Mitt Romney]$_3$. [America]$_6$'s first black ⟨president⟩$_5$ secured more than the 270 ⟨votes⟩$_7$ in the ⟨electoral college⟩$_8$ needed to win. In [his]$_1$ ⟨victory speech⟩$_9$ before ⟨supporters⟩$_{10}$ in [Chicago]$_{11}$, ⟨Mr⟩$_1$ [Obama]$_1$ said [he]$_1$ would talk to ⟨Mr⟩$_3$ [Romney]$_3$ about "[where]$_{11}$ [we]$_{10}$ can work together to move this ⟨country⟩$_{12}$ forward". ⟨Mr⟩$_1$ [Obama]$_1$ prevailed despite lingering ⟨dissatisfaction⟩$_{13}$ with the ⟨economy⟩$_{14}$ and a hard-fought ⟨challenge⟩$_{15}$ by ⟨Mr⟩$_3$ [Romney]$_3$.

[His]$_1$ [Democrats]$_{16}$ also retained [their]$_{16}$ ⟨majority⟩$_{17}$ in the [Senate]$_{18}$, [which]$_{18}$ [they]$_{16}$ have held since 2007. The [Republicans]$_{19}$ kept ⟨control⟩$_{20}$ of the [House of Representatives]$_{21}$, [which]$_{21}$ ⟨analysts⟩$_{22}$ say will likely result in more of the ⟨gridlock⟩$_{23}$ [that]$_{23}$ characterized ⟨Mr⟩$_1$ [Obama]$_1$'s first ⟨term⟩$_{24}$, with the [House]$_{21}$ and the ⟨president⟩$_{25}$ at ⟨loggerheads⟩$_{26}$ on most ⟨legislation⟩$_{27}$.

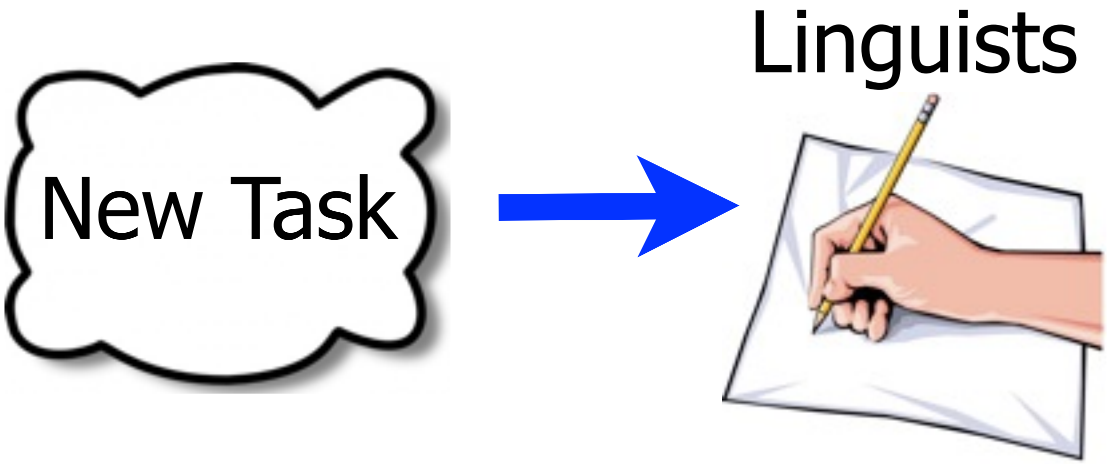
In [his]$_{28}$ ⟨address⟩$_{29}$, the ⟨president⟩$_{28}$ challenged [his]$_{28}$ ⟨opponents⟩$_{30}$, asking [them]$_{30}$ to work with [him]$_{28}$. With only [Florida]$_{32}$'s 29 ⟨electoral votes⟩$_{31}$ still undecided, ⟨Mr⟩$_1$ [Obama]$_1$ won 303 ⟨electoral votes⟩$_{33}$ to ⟨Mr⟩$_3$ [Romney]$_3$'s 206. The popular ⟨vote⟩$_{34}$, [which]$_{34}$ is symbolically and politically important but not decisive in the ⟨race⟩$_{35}$, remains very close.

⟨Mr⟩$_1$ [Obama]$_1$ congratulated ⟨Mr⟩$_3$ [Romney]$_3$ and [Republican]$_4$ ⟨vice-presidential candidate⟩$_{36}$ [Paul Ryan]$_{36}$ on [their]$_{33}$ hard-fought ⟨campaign⟩$_{37}$. "[We]$_{33}$ have picked [ourselves]$_{33}$ up, [we]$_{33}$ have fought [our]$_{33}$ ⟨way⟩$_{38}$ back and [we]$_{33}$ know in [our]$_{33}$ ⟨hearts⟩$_{39}$ that for the [United States of America]$_6$ the best is yet to come," [he]$_1$ said. ⟨Mr⟩$_1$ [Obama]$_1$ said [he]$_1$ was returning to the [White House]$_{40}$ "more determined, and more inspired than ever about the ⟨work⟩$_{42}$ [there]$_{43}$ is to do, and the ⟨future⟩$_{41}$ [that]$_{41}$ lies ahead". [He]$_1$ pledged to work with [Republican]$_4$ ⟨leaders⟩$_{44}$ in [Congress]$_{45}$ to reduce the

Traditional Supervised NLP

New Task

Traditional Supervised NLP

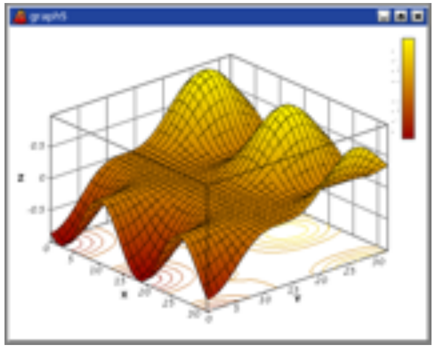


Traditional Supervised NLP

New Task

Linguists

Machine Learning

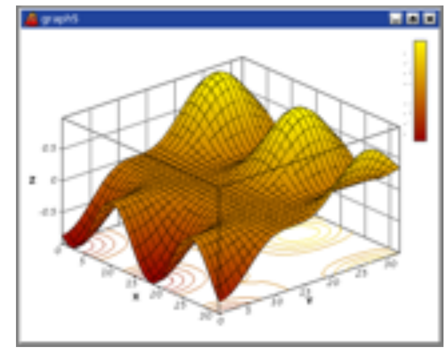


Traditional Supervised NLP

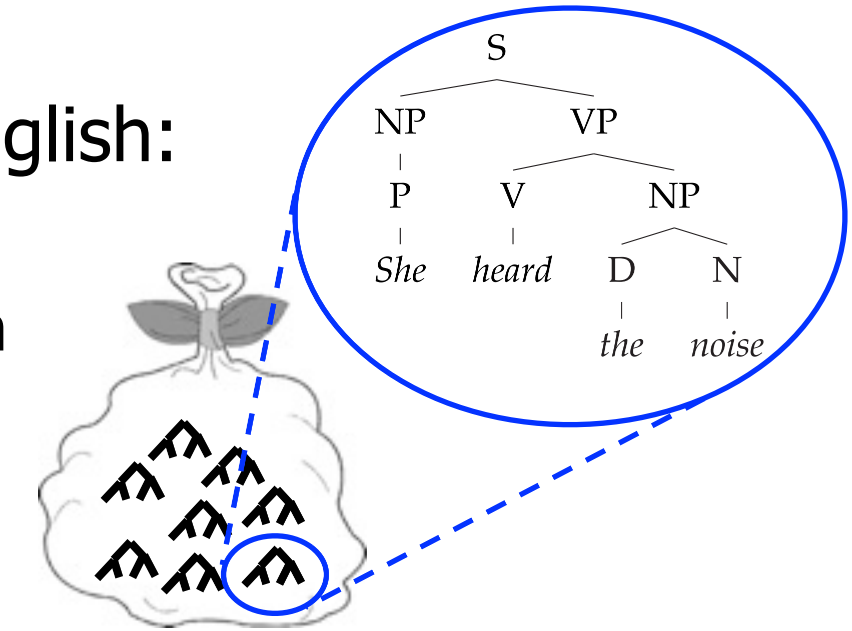
New Task

Linguists

Machine Learning



- Example: Syntactic analysis for English:
  - Not solved, but accuracies are high
  - 97% **(or 90%?)** for parts-of-speech
  - 93% **(or 83%?)** for parse trees



Treebank:
2 million words

A Note on Parallel Computing

A Note on Parallel Computing

- Most NLP models these days operate at sentence level
  - Trivially parallelizable
  - Focus on efficient single-core models

A Note on Parallel Computing

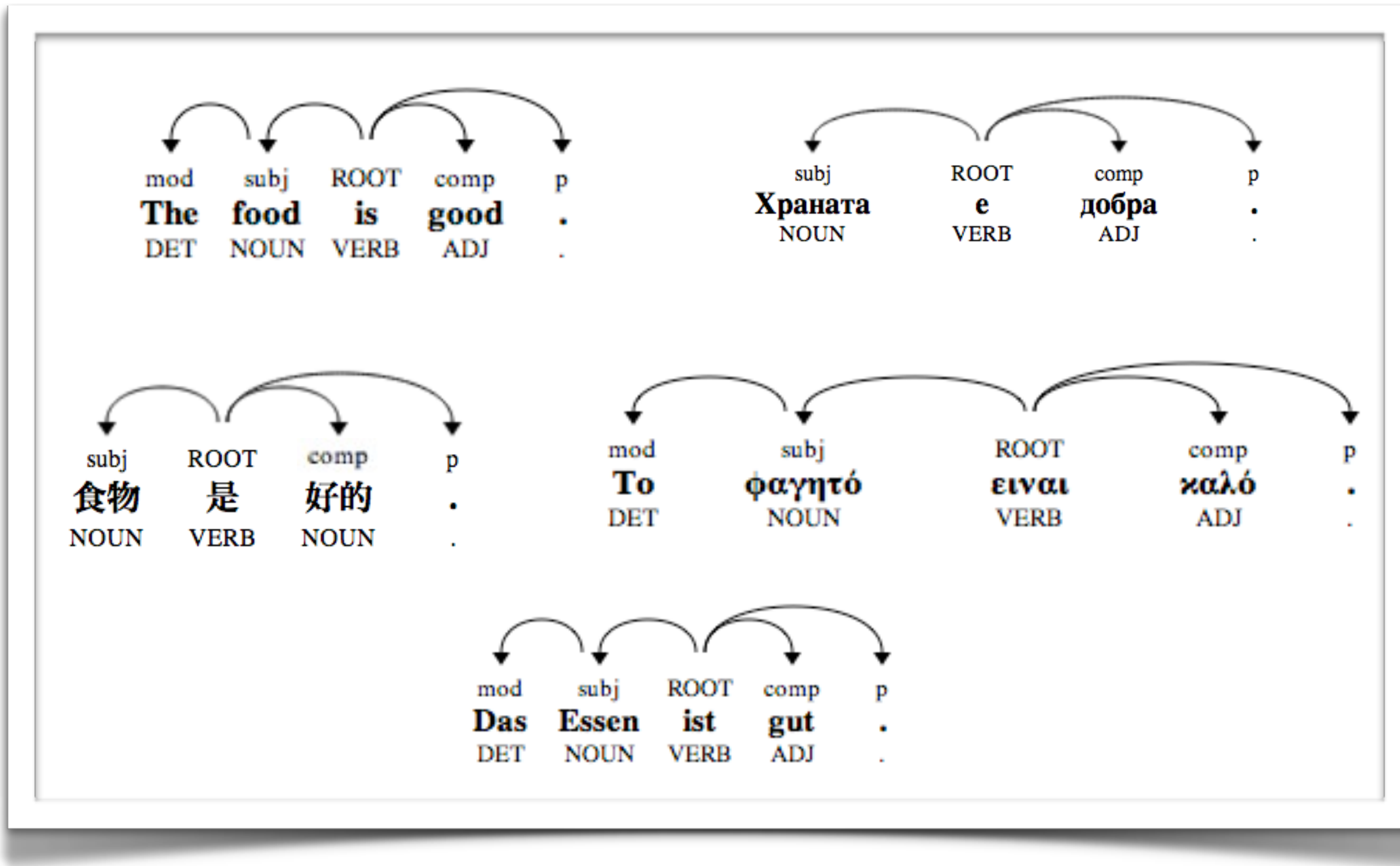
- Most NLP models these days operate at sentence level
  - Trivially parallelizable
  - Focus on efficient single-core models
- Examples:
  - Web-Scale Information Extraction
  - Machine Translation

A Note on Parallel Computing

- Most NLP models these days operate at sentence level
  - Trivially parallelizable
  - Focus on efficient single-core models
- Examples:
  - Web-Scale Information Extraction
  - Machine Translation
- Distributed systems might be needed for storage:
  - Distributed Language Models, Phrase Tables, ...
    in Machine Translation, Speech Recognition, ...

Universal Tagging/Parsing

- Goal: high accuracy parsing in all languages with a single universal representation of syntax



Syntactic Processing for the Web

Syntactic Processing for the Web

Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov '12]

Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov ’12]
- Weakly supervised domain adaptation
  - Training data is not representative
  - Learn from weak signals [Hall et al. ’11, Ganchev et al. ’12]

Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov ’12]
- Weakly supervised domain adaptation
  - Training data is not representative
  - Learn from weak signals [Hall et al. ’11, Ganchev et al. ’12]
- Multilingual tagger and parser projection
  - Will never have labeled resources for all languages
  - Use parallel data to project information
    [Das & Petrov ’11, McDonald et al. ’11]

Syntactic Processing for the Web

- Fast and accurate supervised parsing
  - Many labeled resources are English only
  - Nonetheless: Use them! [Rush & Petrov ’12]
- Weakly supervised domain adaptation
  - Training data is not representative
  - Learn from weak signals [Hall et al. ’11, Ganchev et al. ’12]
- Multilingual tagger and parser projection
  - Will never have labeled resources for all languages
  - Use parallel data to project information [Das & Petrov ’11, McDonald et al. ’11]

Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

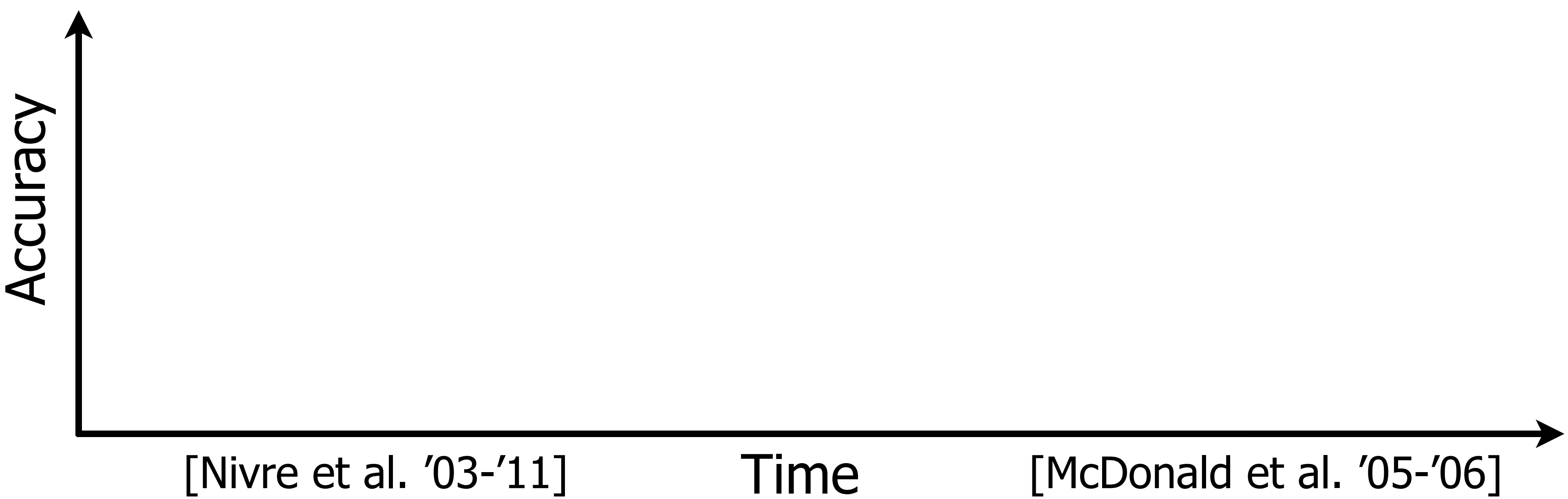
[Nivre et al. ’03-’11]

- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

[McDonald et al. ’05-’06]

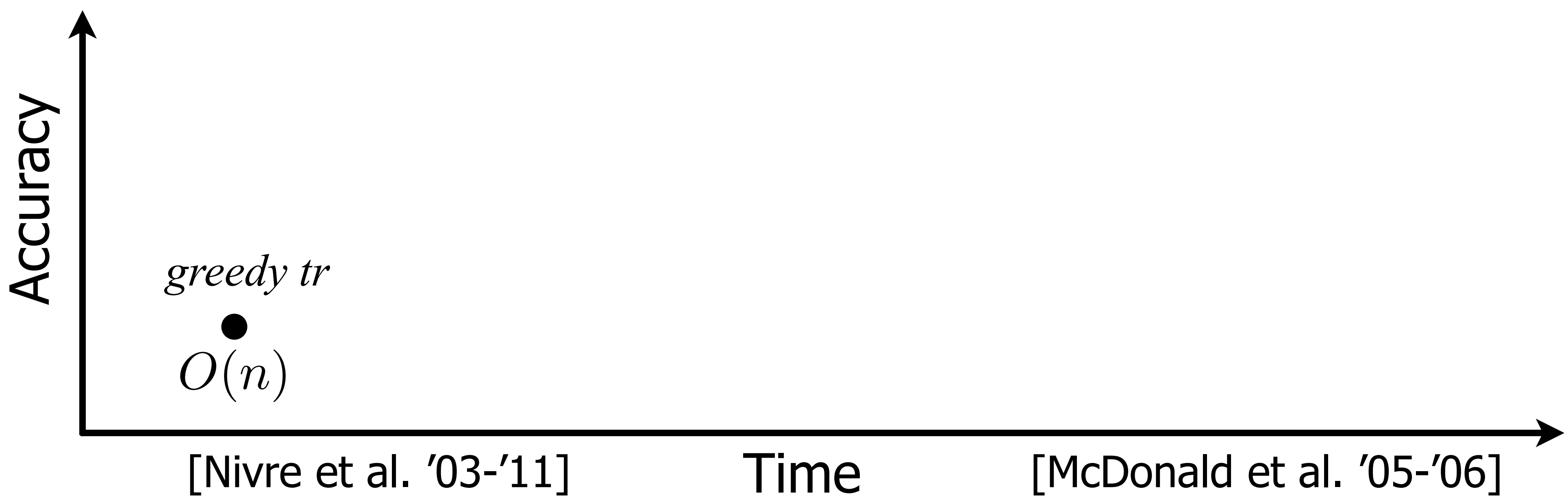
Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search
- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations



Dependency Parsing

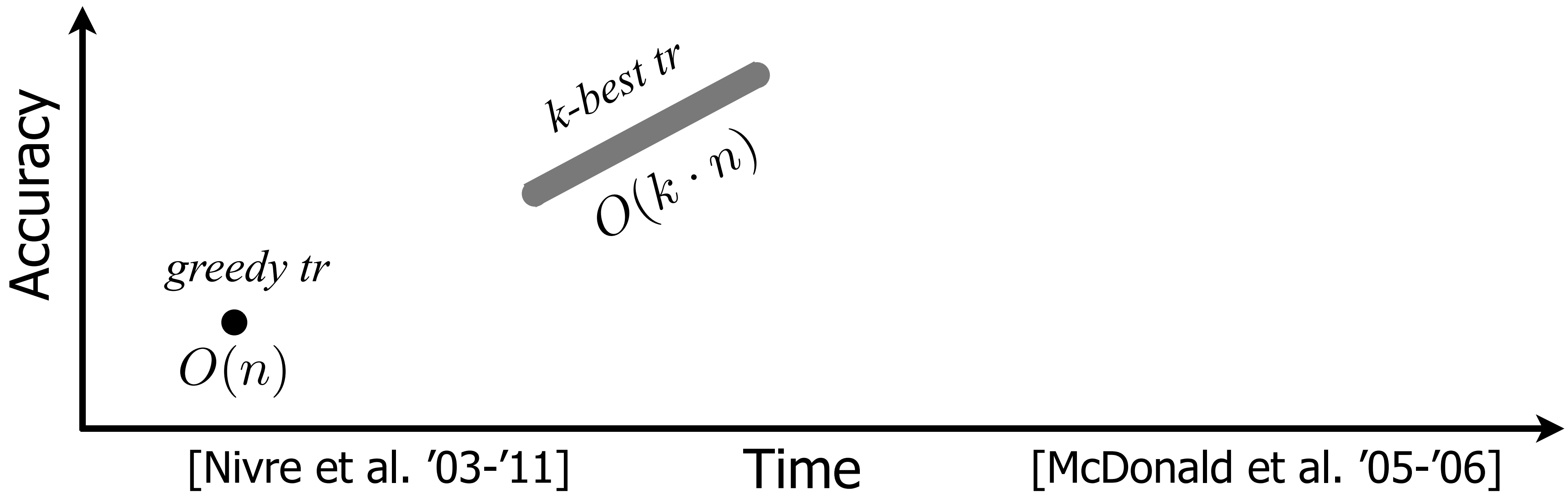
- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search
- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations



Dependency Parsing

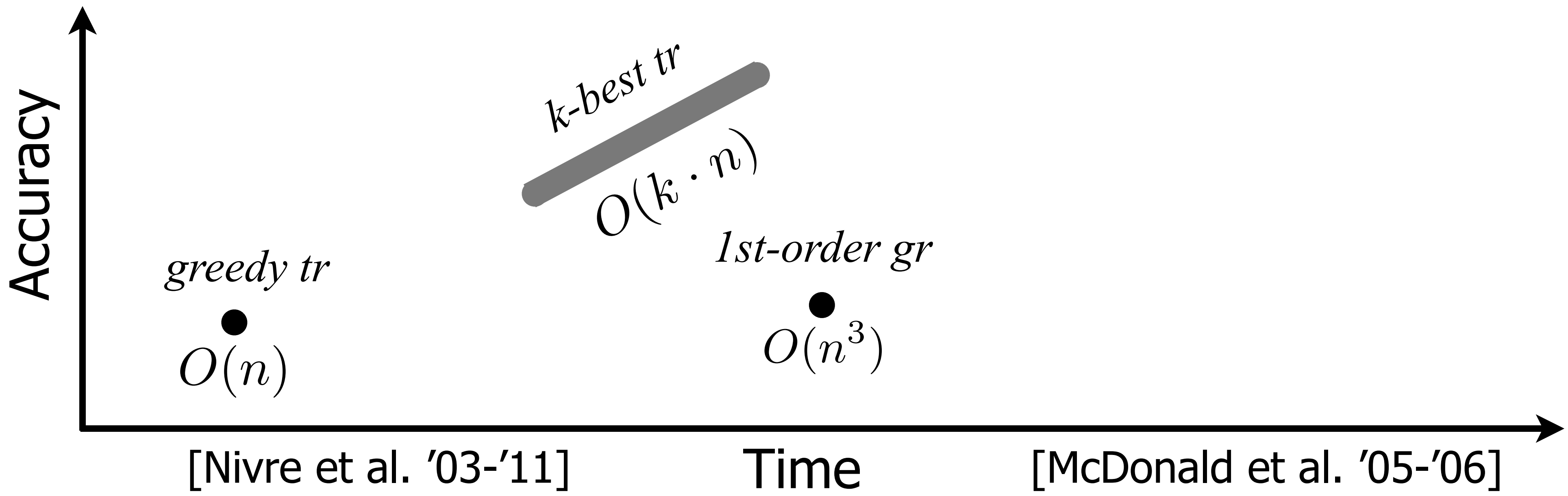
- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search

- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations



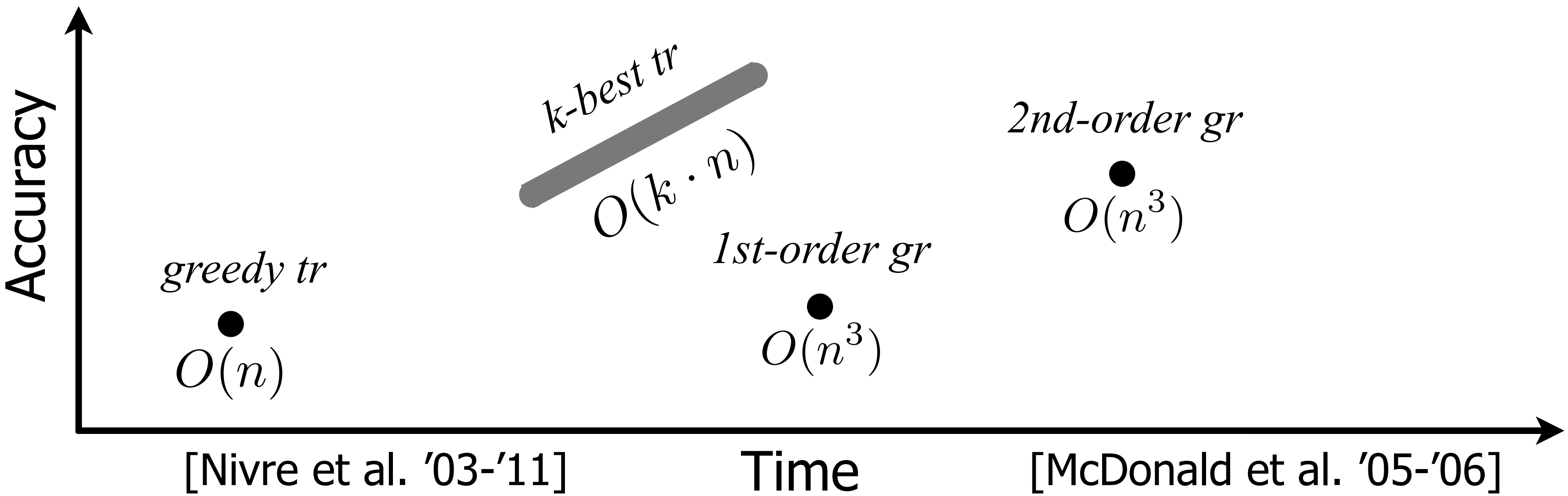
Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search
- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations



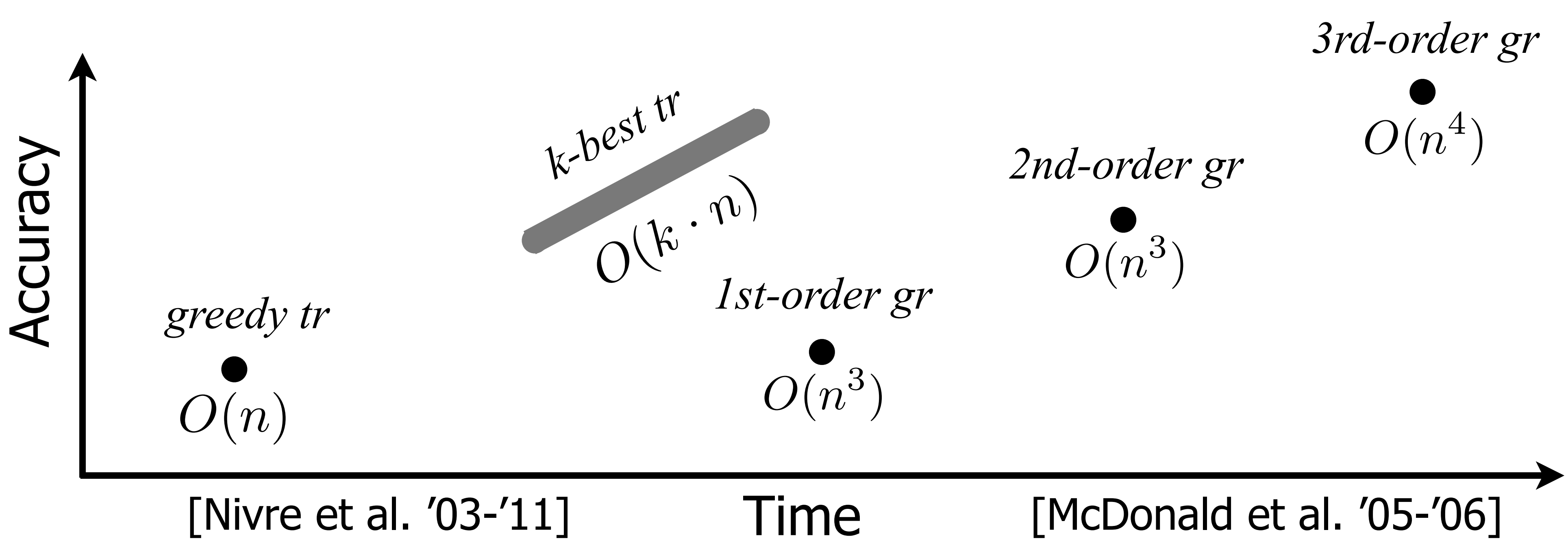
Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search
- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations



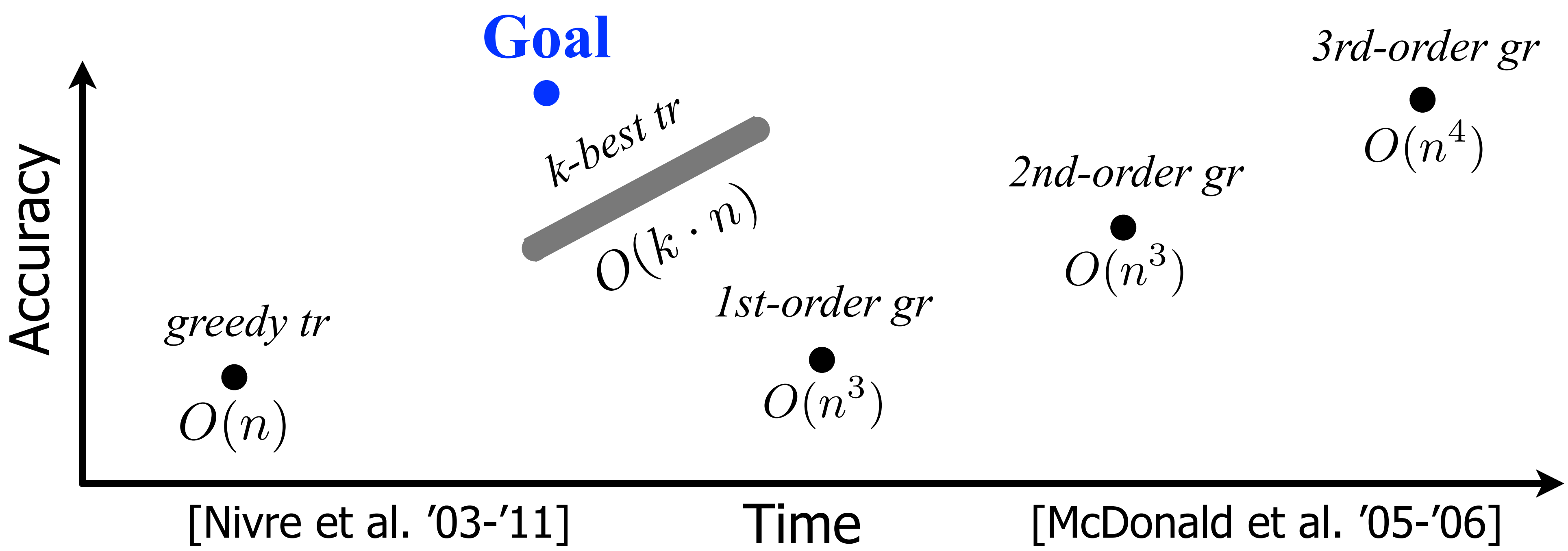
Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search
- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations



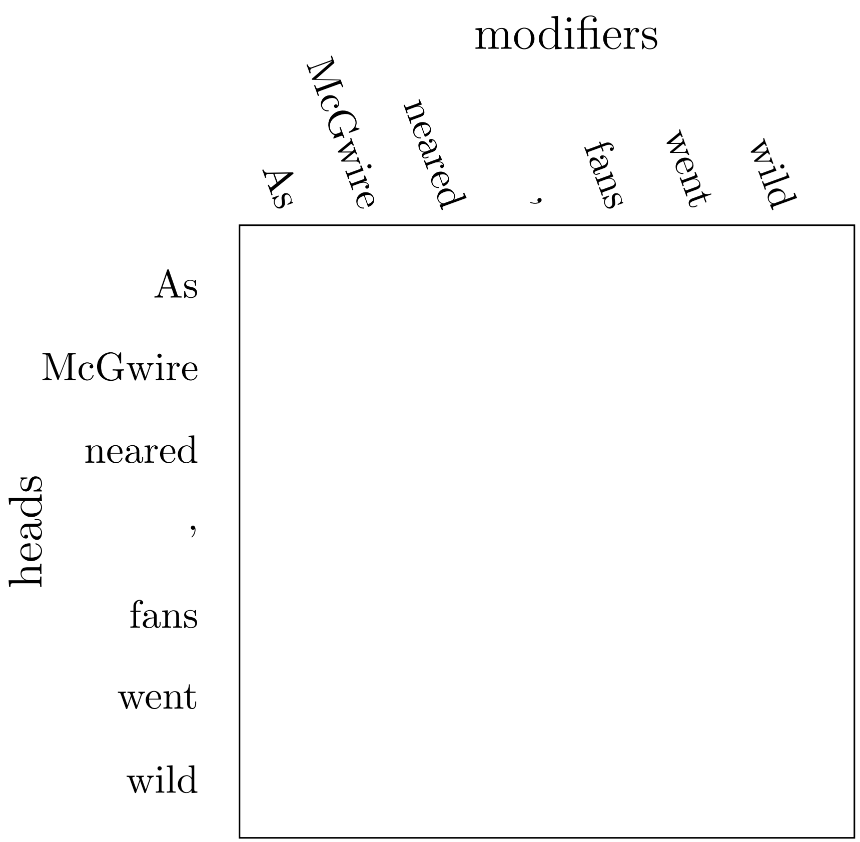
Dependency Parsing

- Transition-Based (tr)
  - Fast, greedy, linear time inference algorithms
  - Trained for greedy search
  - Beam search
- Graph-Based (gr)
  - Slower, exhaustive, dynamic programming inference algorithms
  - Higher-order factorizations

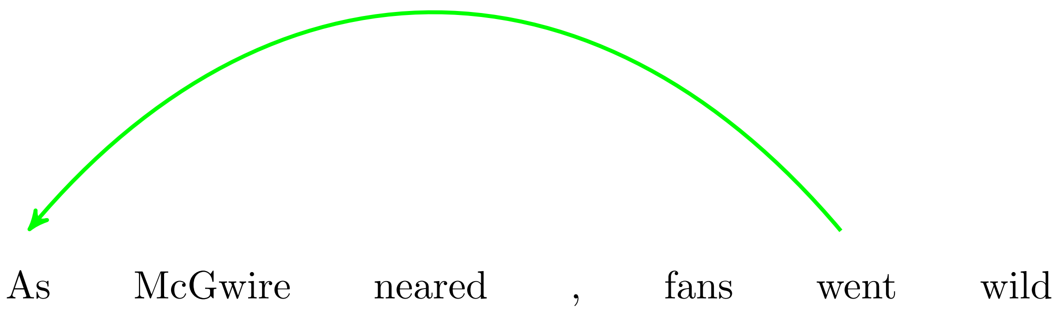


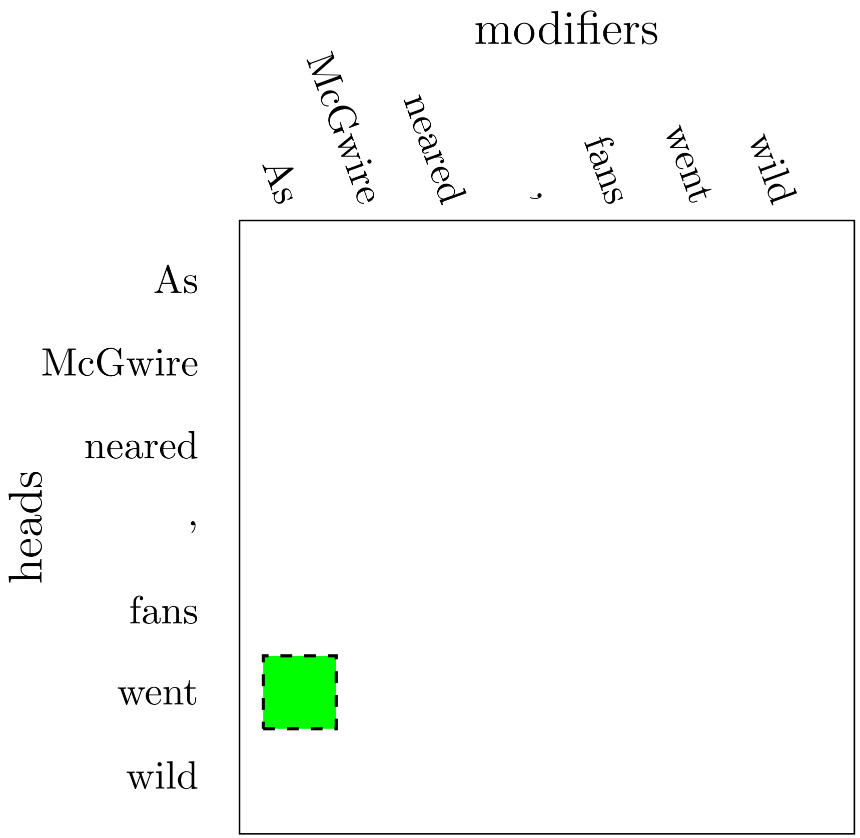
Dependency Representation

As McGwire neared , fans went wild

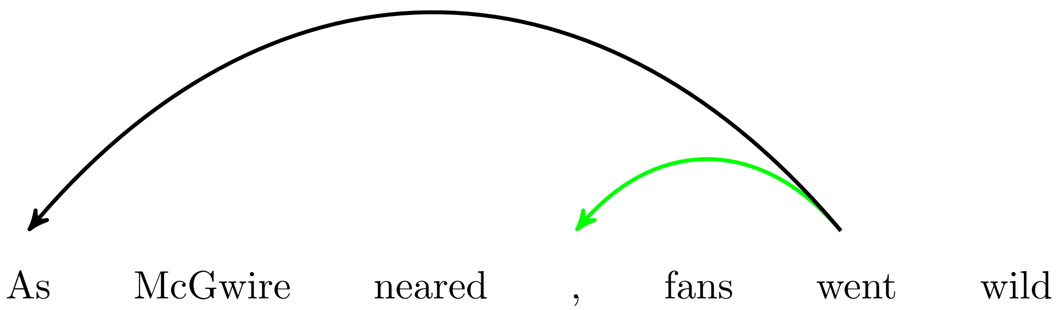


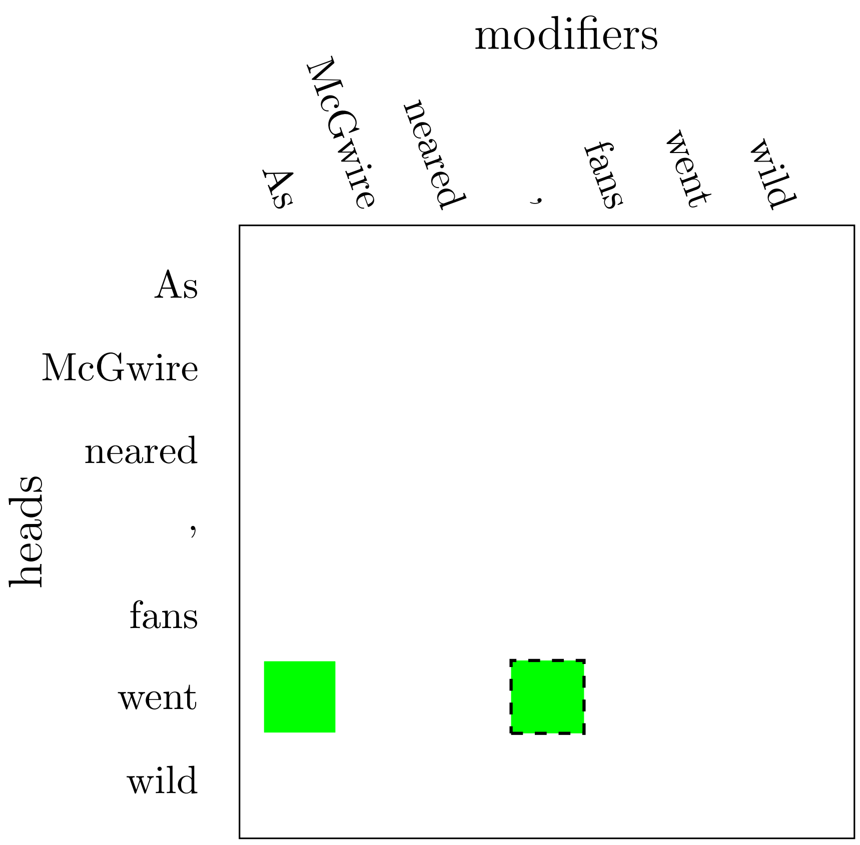
Dependency Representation



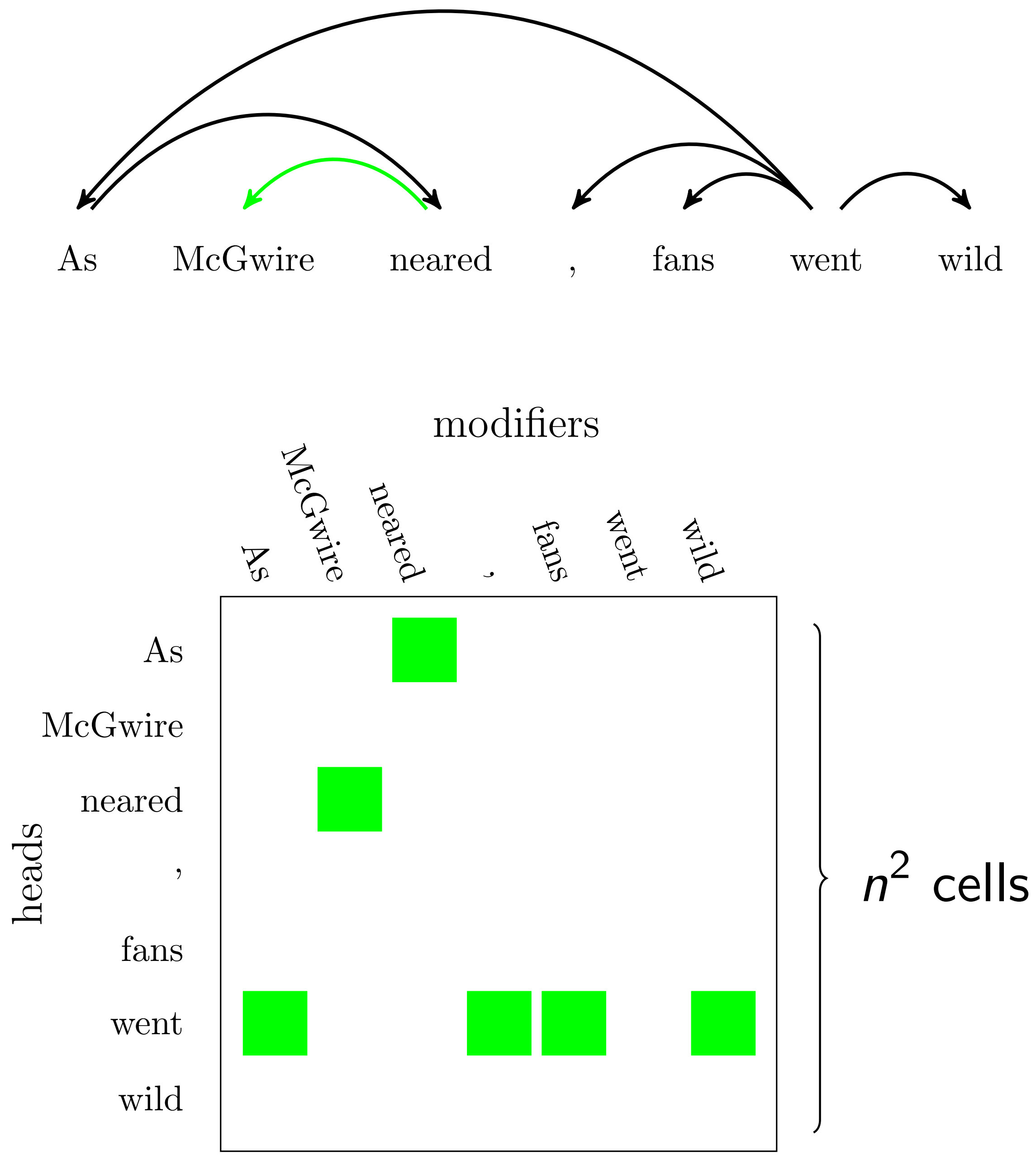


Dependency Representation

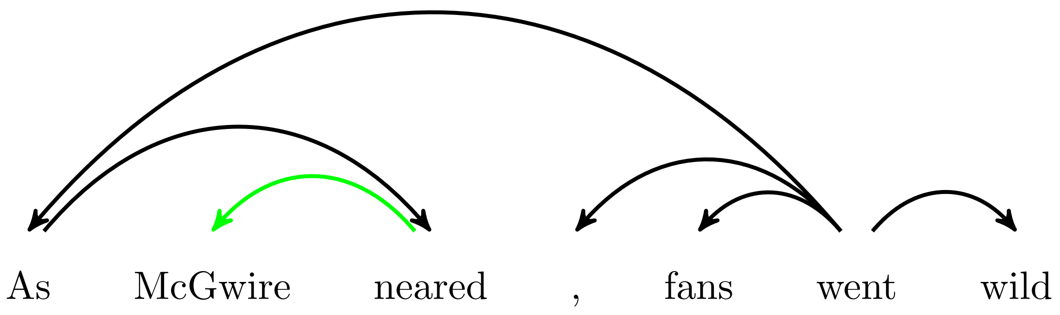


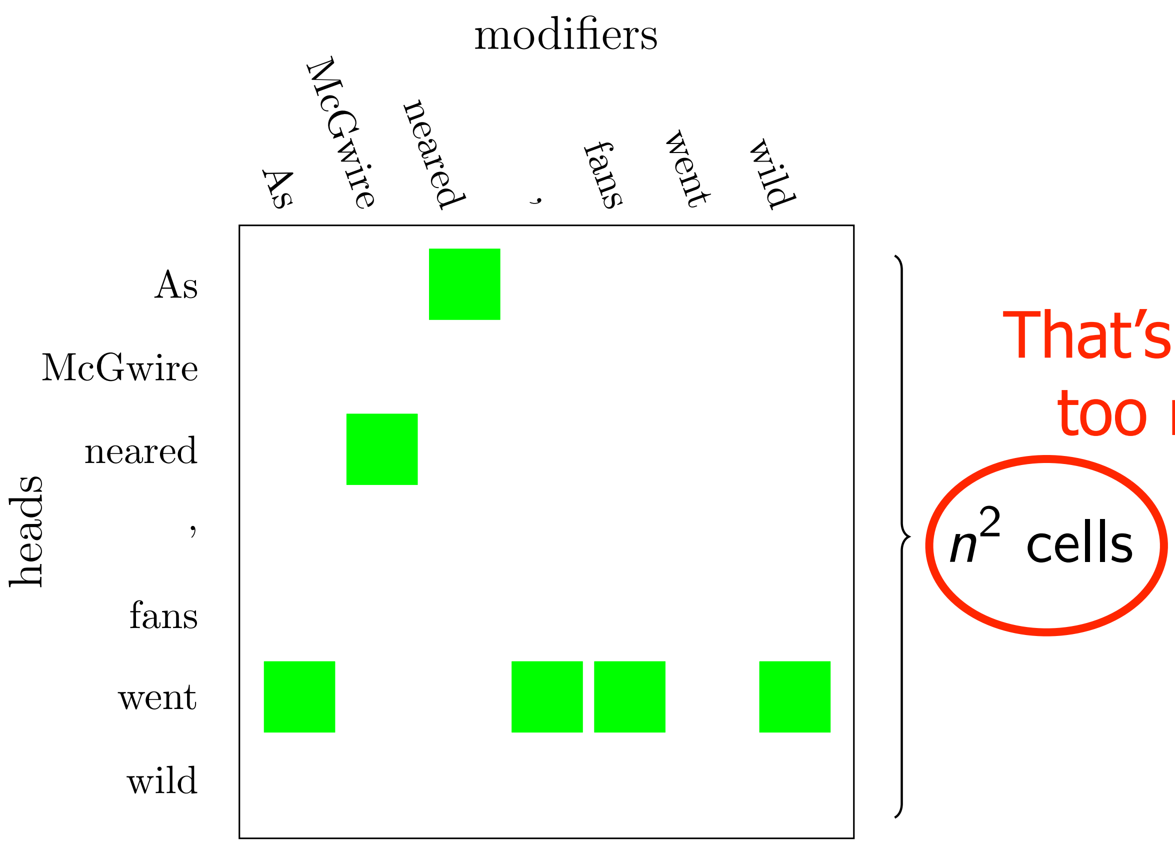


Dependency Representation



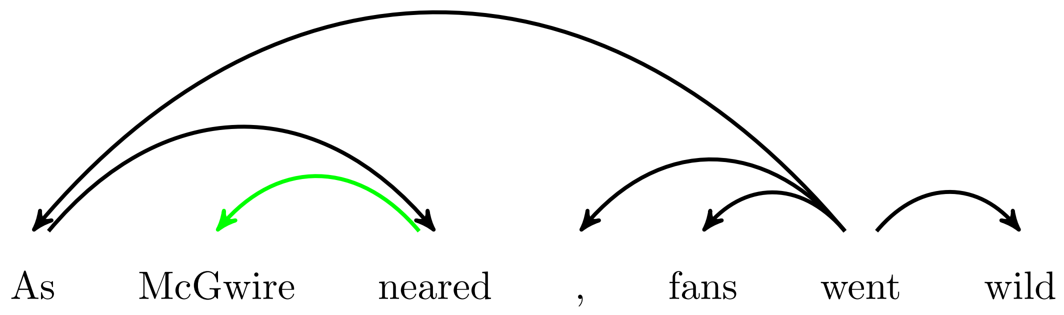
Dependency Representation

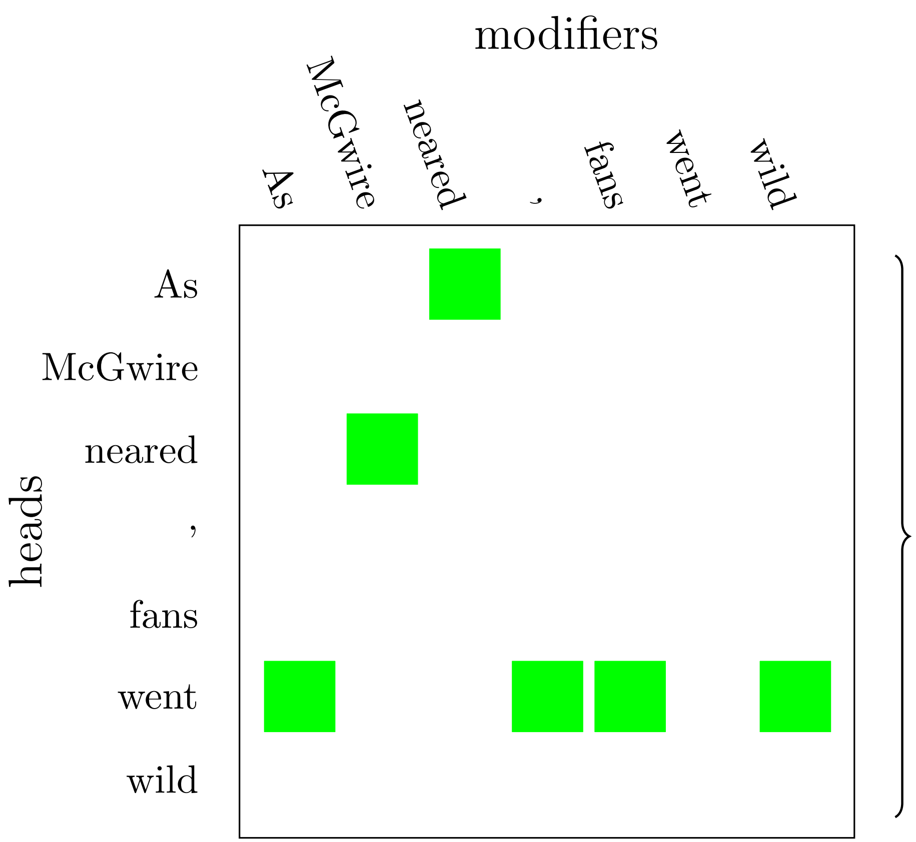




That's already too much!

Dependency Representation





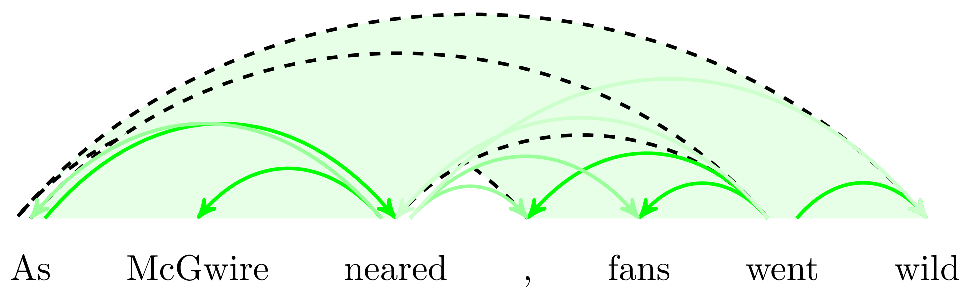
That's already too much!

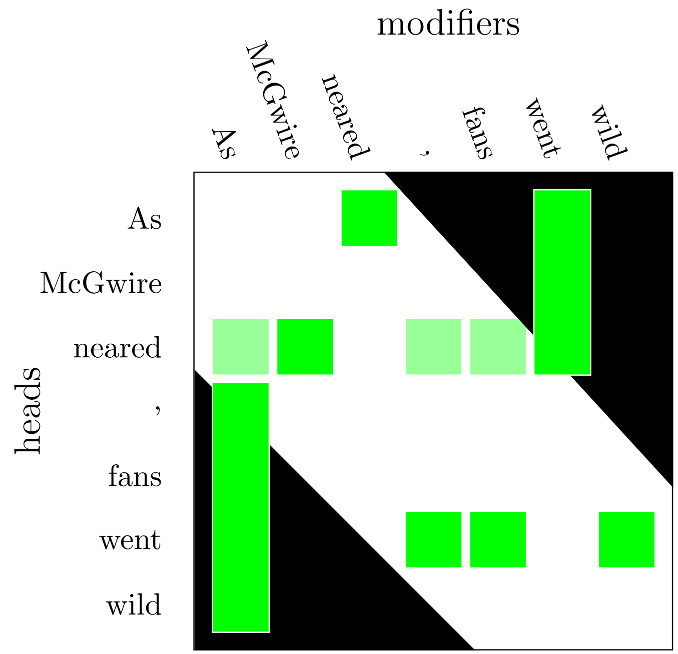
n^2 cells

Exploit problem structure!

Vine Pruning

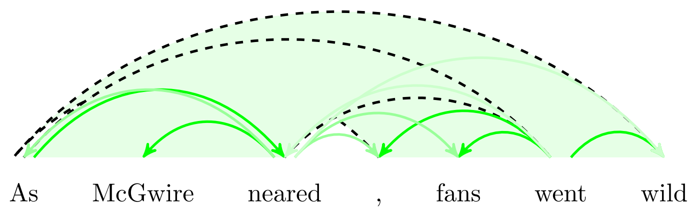
[Rush & Petrov '12]

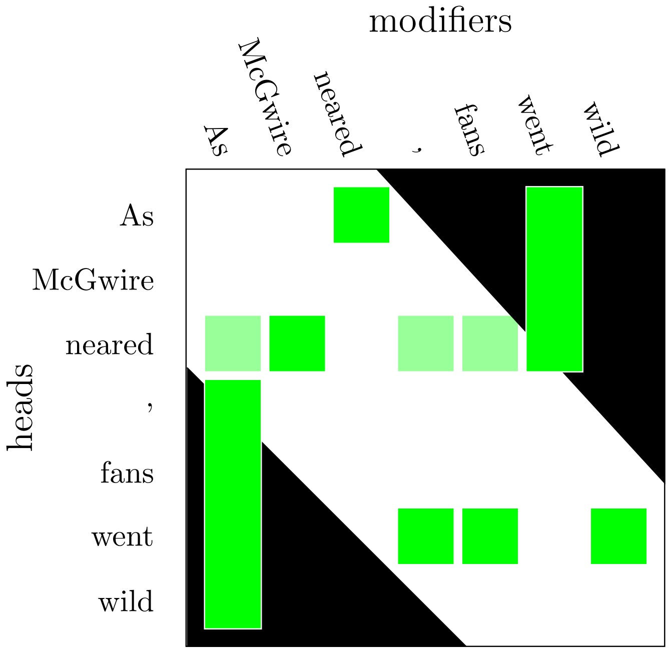




Coarse-to-Fine Cascades

[Rush & Petrov '12]

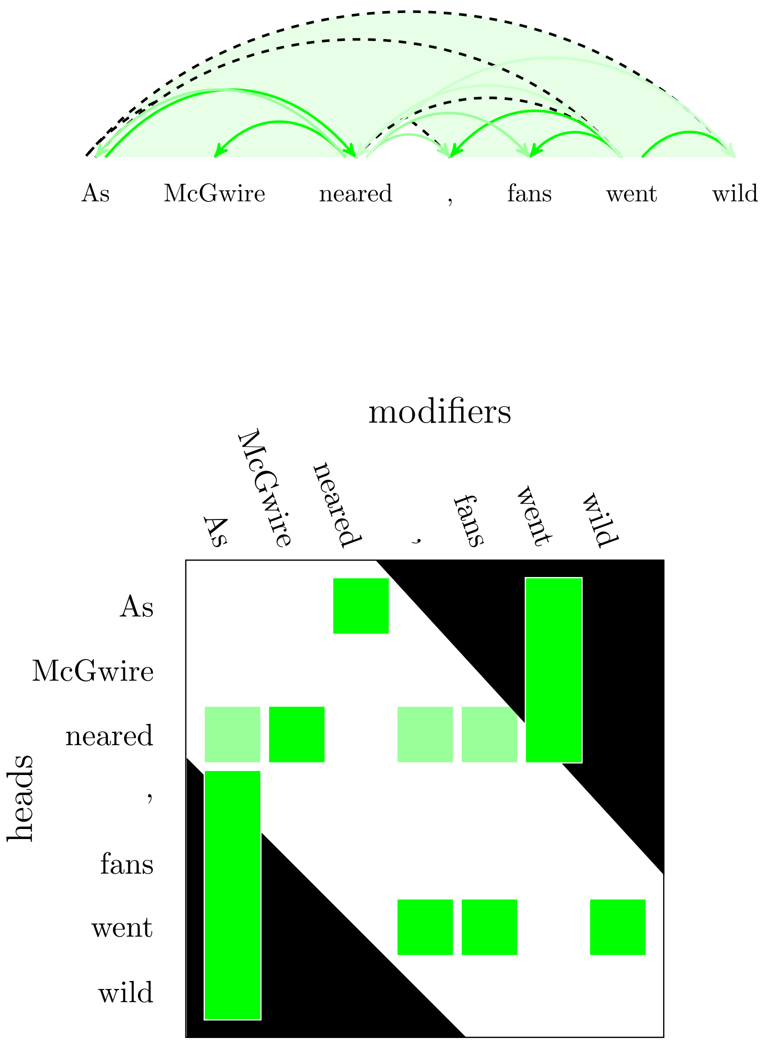




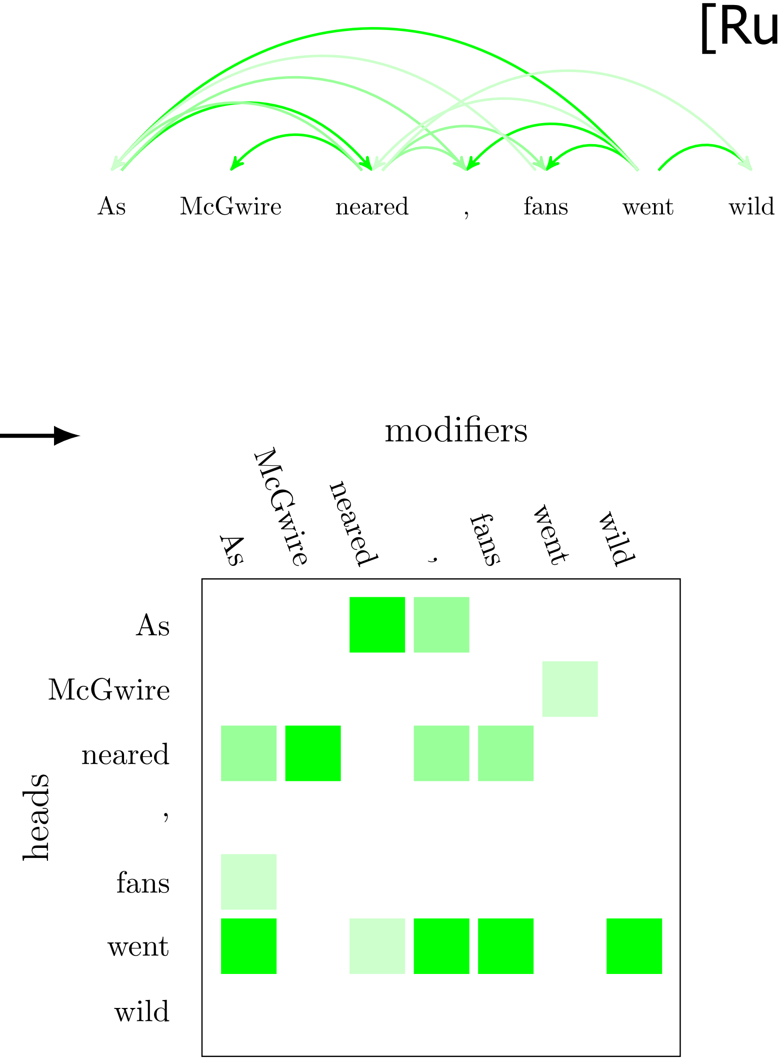
Vine

Coarse-to-Fine Cascades

[Rush & Petrov '12]



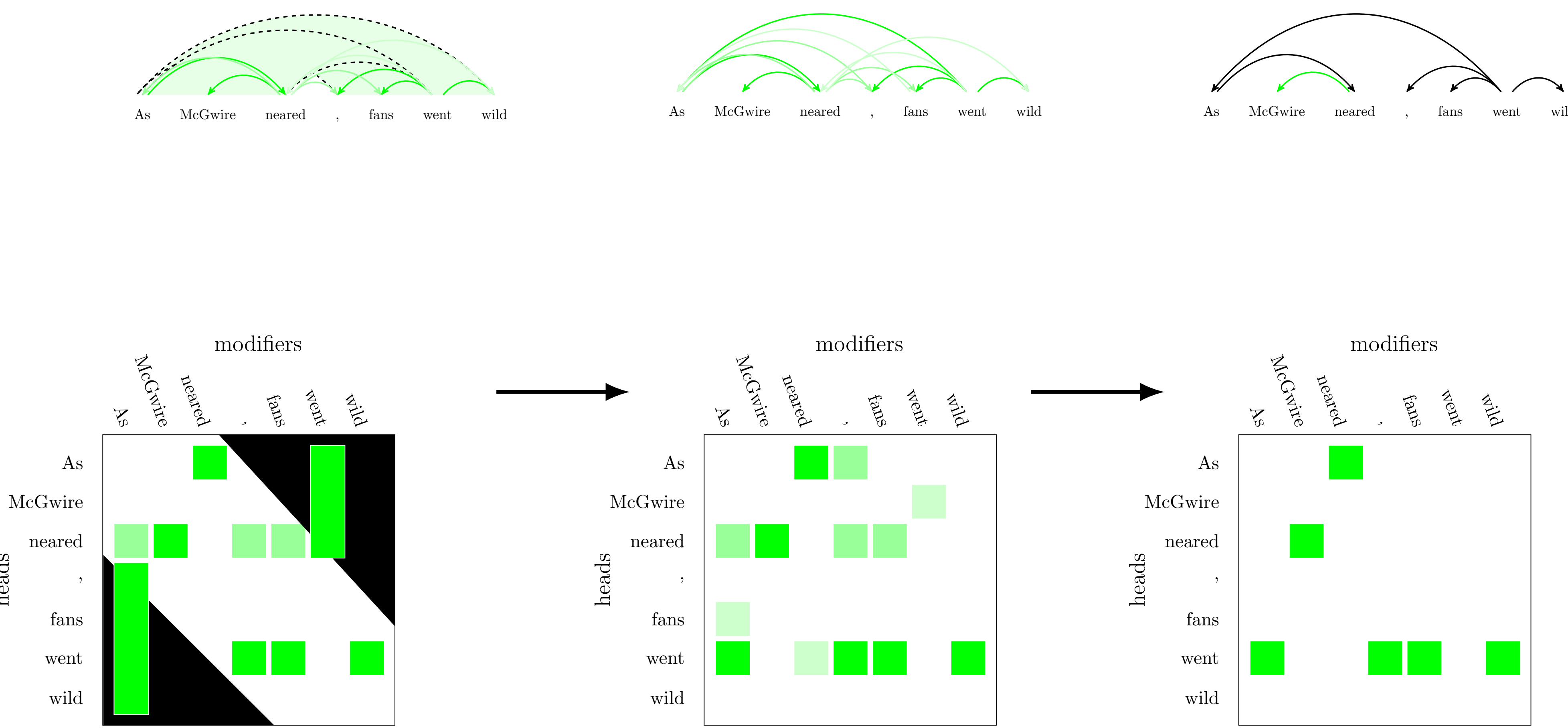
Vine



First-Order

Coarse-to-Fine Cascades

[Rush & Petrov '12]



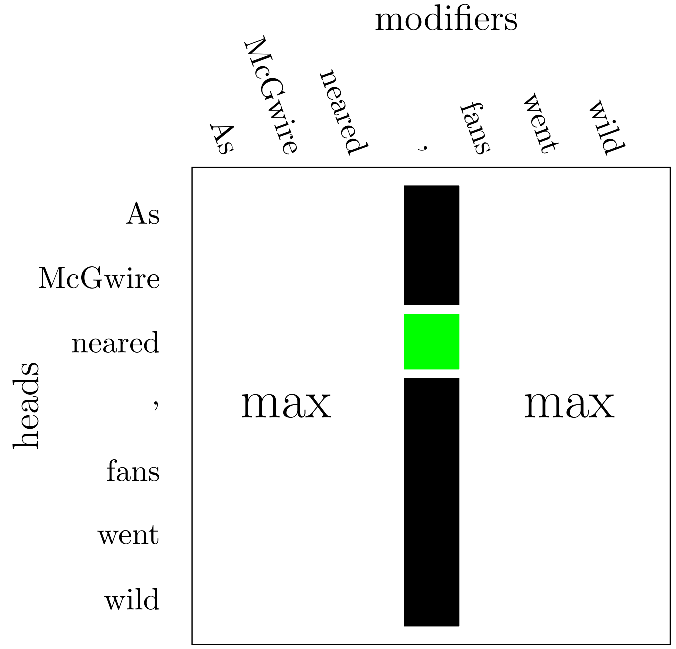
Vine

First-Order

Second-Order

Max-Marginals

As McGwire neared , fans went wild



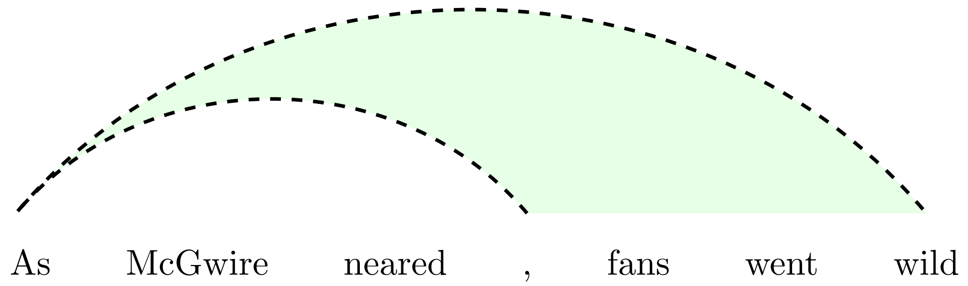
MAP Parse

$$y^* = \arg\max_{y \in \mathcal{Y}} y \cdot w$$

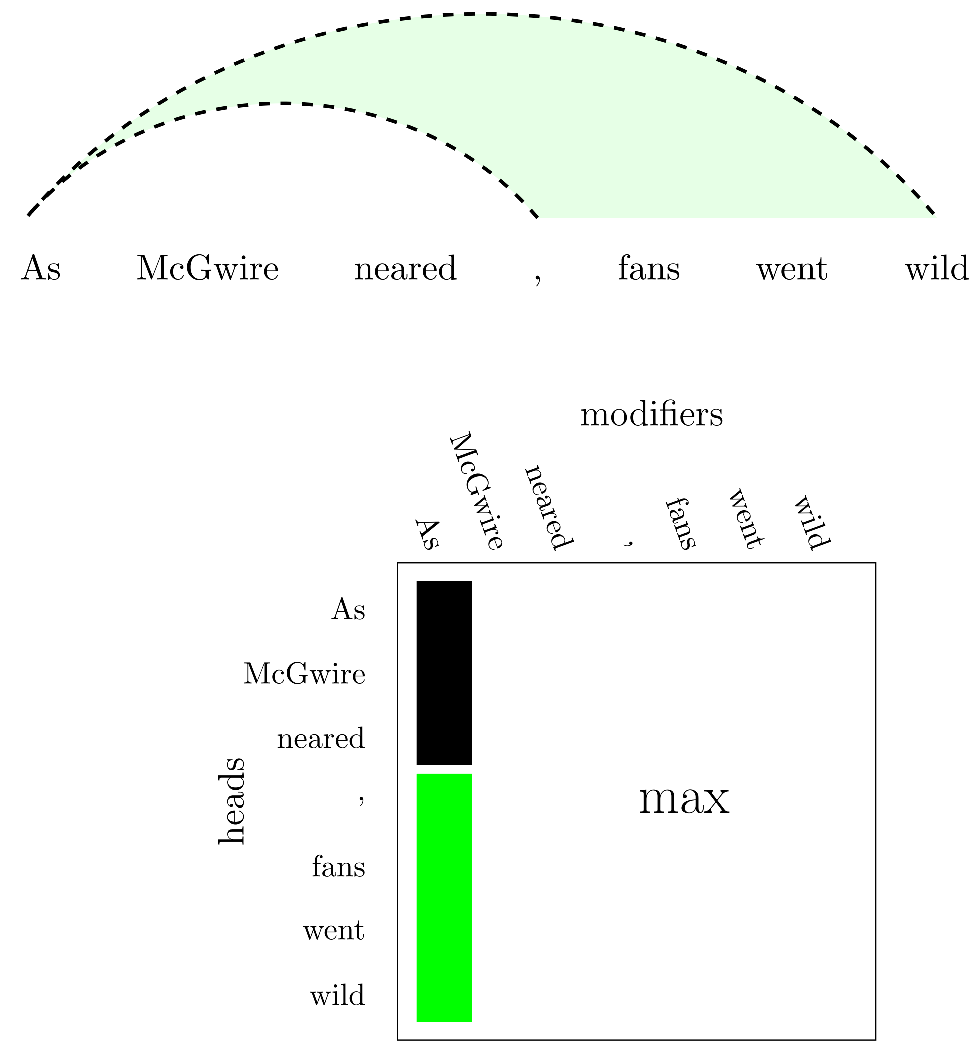
Max-Marginal

$$m(a) = \arg\max_{y \in \mathcal{Y} : a \in A} y \cdot w$$

Max-Marginals



Max-Marginals



Structured Prediction Cascades Training

[Weiss & Taskar ’10]

Structured Prediction Cascades Training

[Weiss & Taskar '10]

- Train to minimize pruning error (rather than 1-best)

Structured Prediction Cascades Training

[Weiss & Taskar '10]

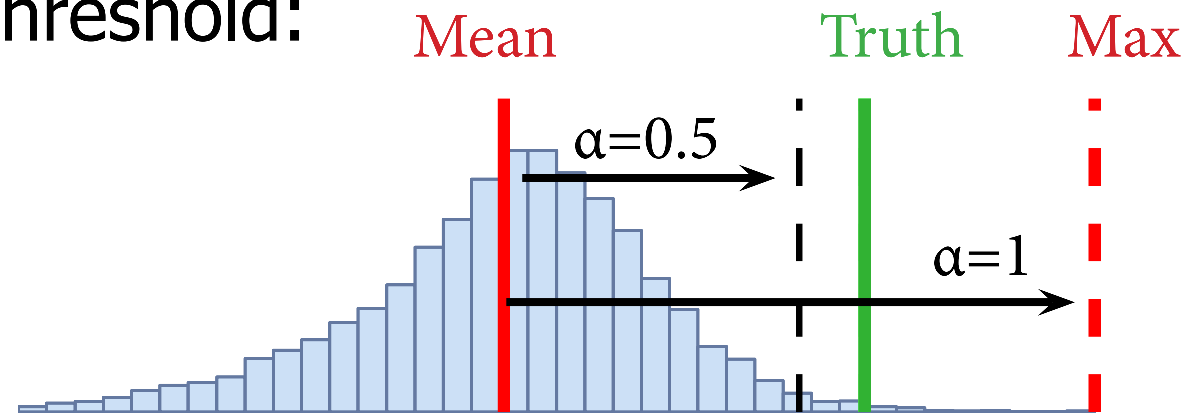
- Train to minimize pruning error (rather than 1-best)
- Pruning threshold:

$$t_\alpha(w) = \alpha y^* \cdot w + (1-\alpha)\frac{1}{|A|}\sum_{a \in A} m(a) \cdot w$$

Structured Prediction Cascades Training

[Weiss & Taskar '10]

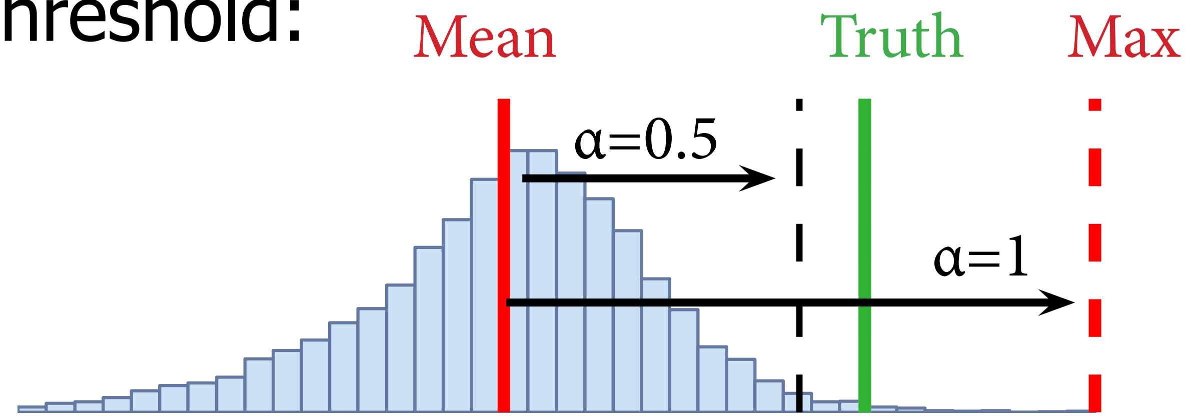
- Train to minimize pruning error (rather than 1-best)
- Pruning threshold:



Structured Prediction Cascades Training

[Weiss & Taskar '10]

- Train to minimize pruning error (rather than 1-best)
- Pruning threshold:



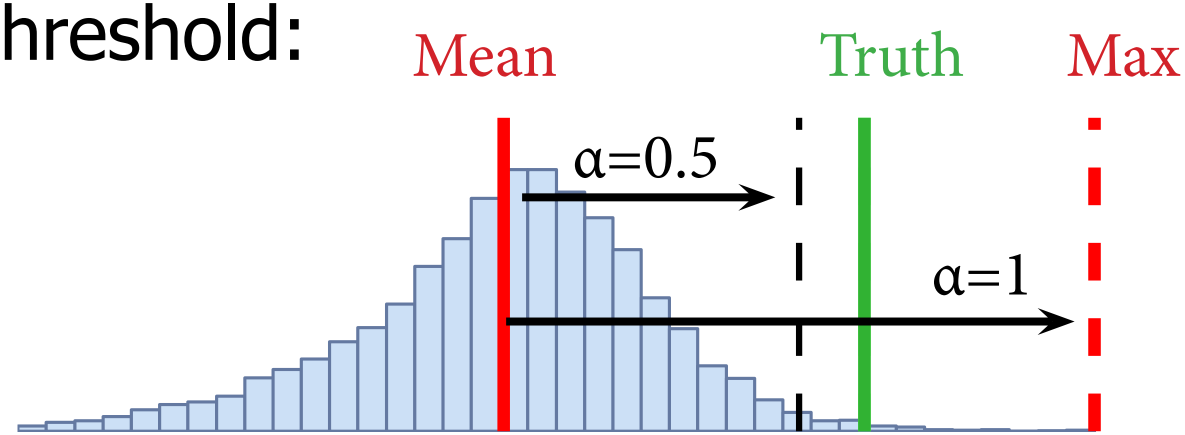
- Training objective:

$$\min_w \lambda \|w\|^2 + \frac{1}{M} \sum_{m=1}^{M} \max\{0, 1 + t_\alpha(w) - y^m \cdot w\}$$

Structured Prediction Cascades Training

[Weiss & Taskar '10]

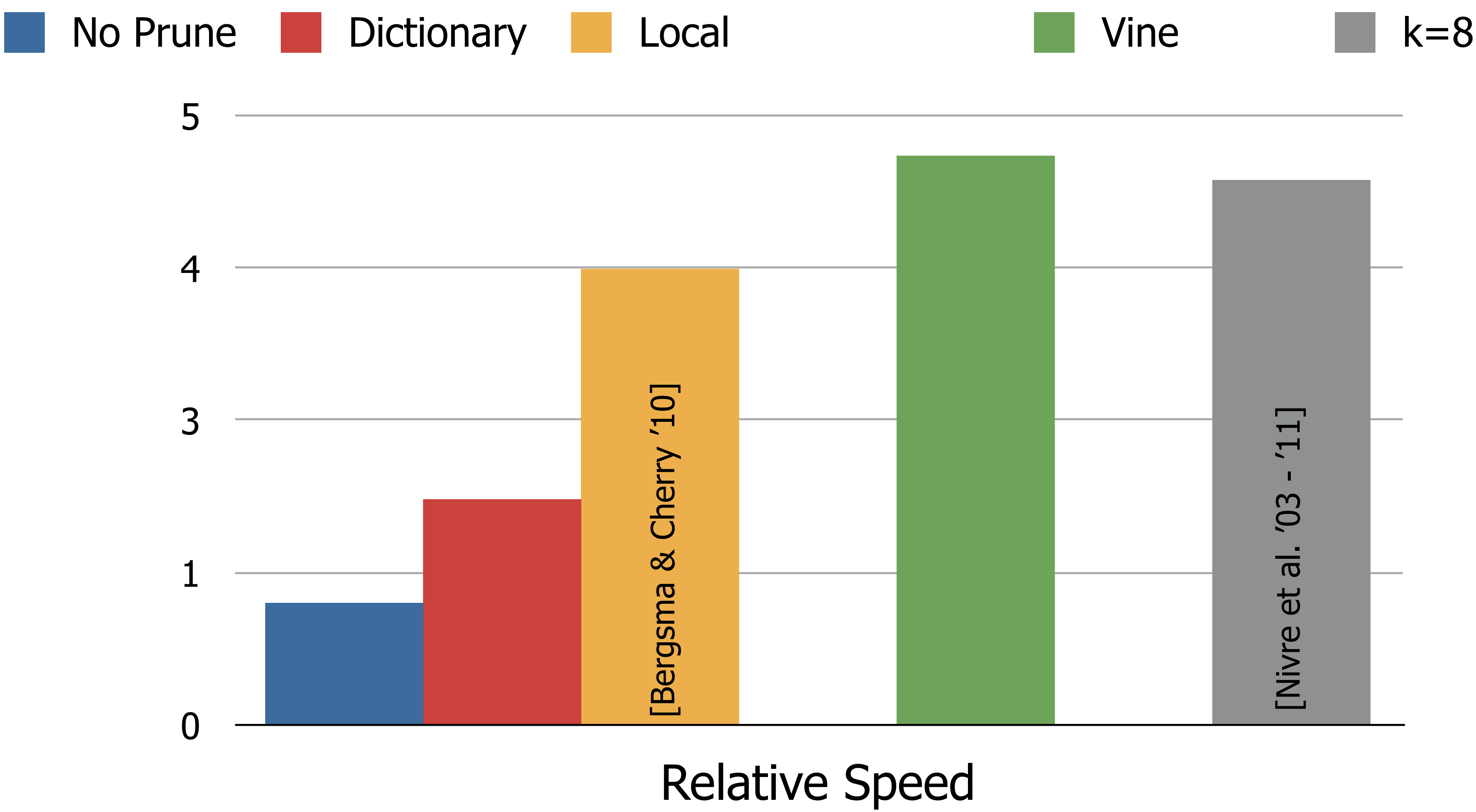
- Train to minimize pruning error (rather than 1-best)
- Pruning threshold:



- Training objective:
  - Filter as many arcs as possible
  - While preserving gold arc
  - Optimize with stochastic gradient decent (not so different from perceptron updates)

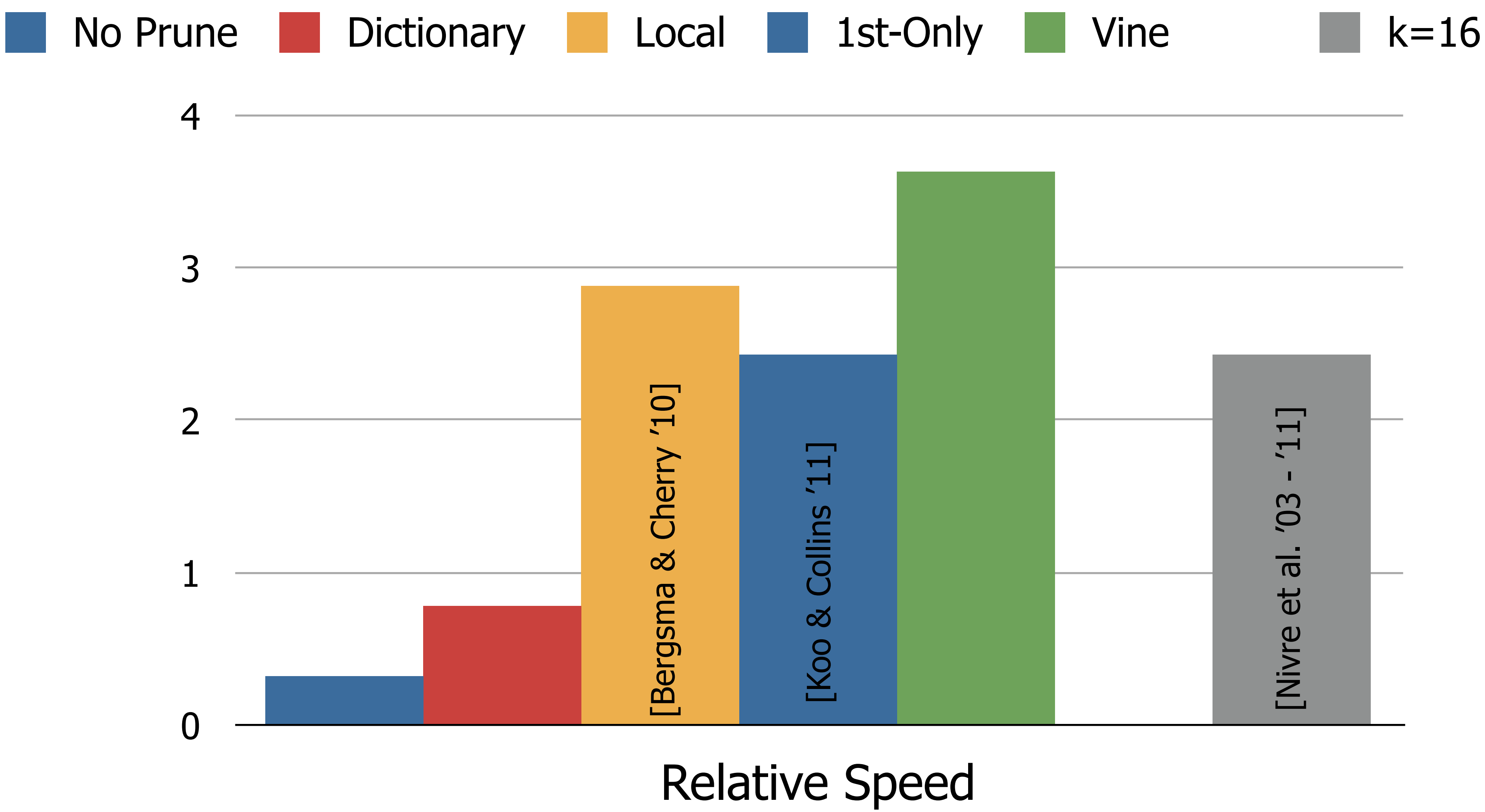
First-Order Parsing

UAS 91.0, Set pruning thresholds for no loss in accuracy



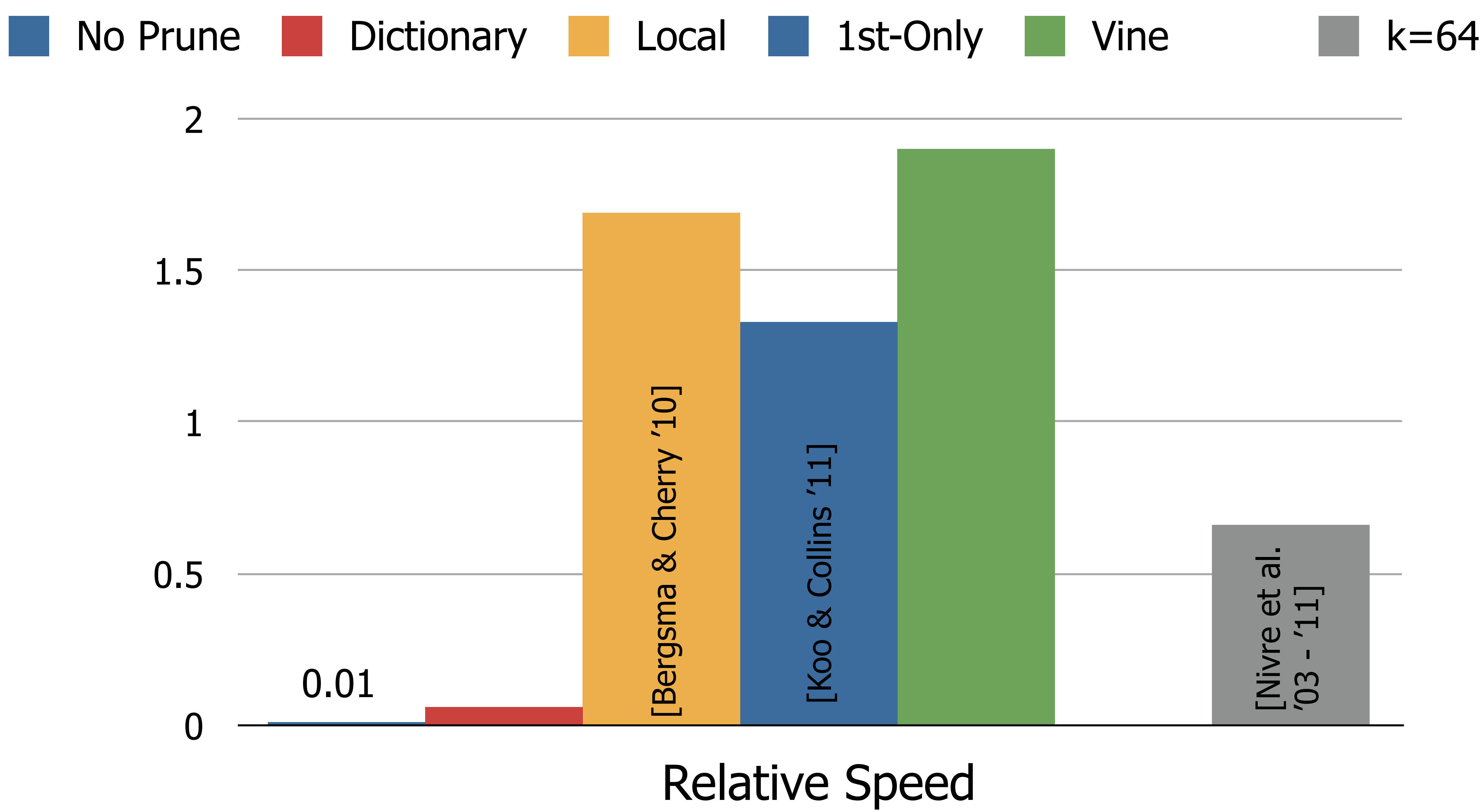
Second-Order Parsing

UAS 92.1, Set pruning thresholds for no loss in accuracy



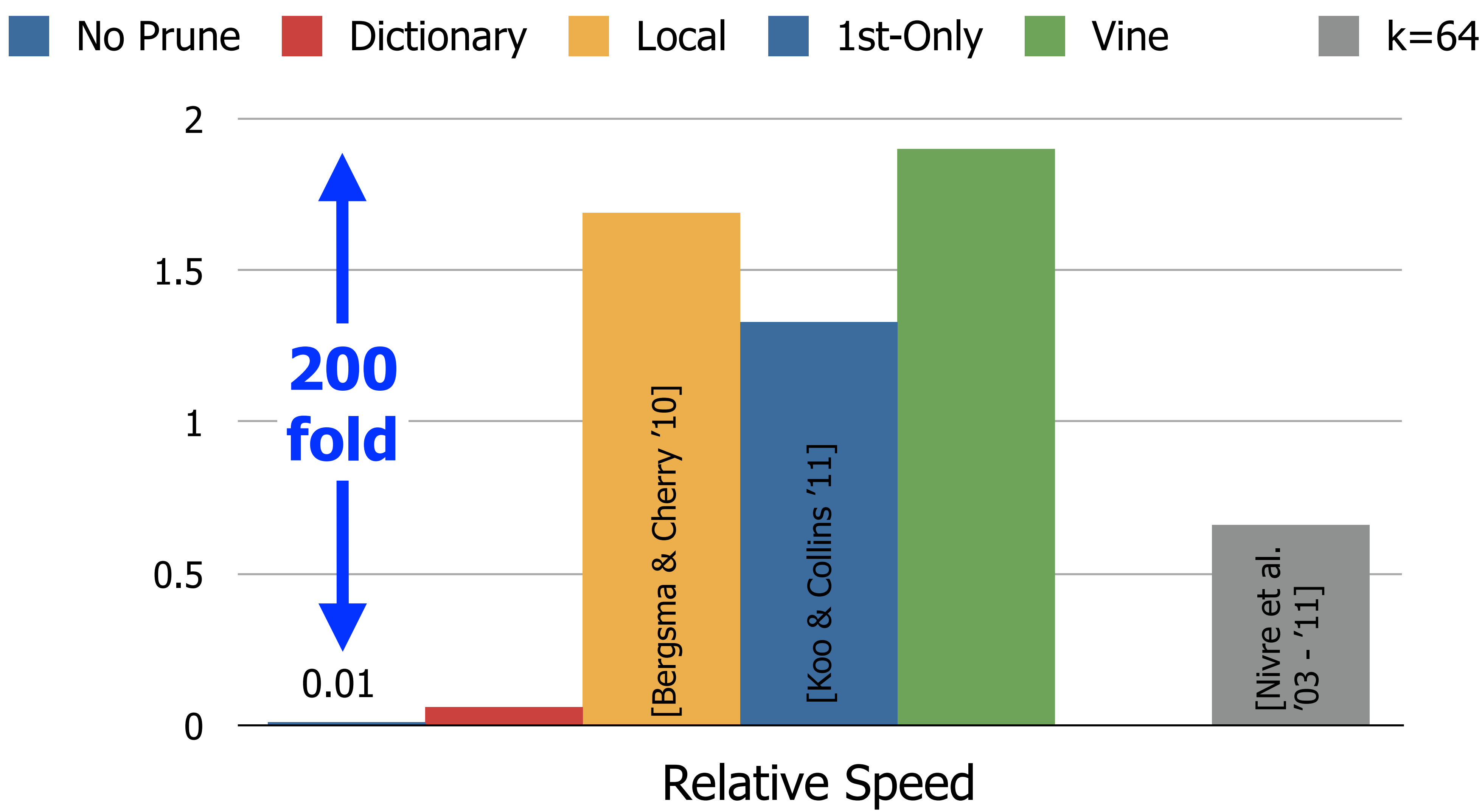
Third-Order Parsing

UAS 92.9, Set pruning thresholds for no loss in accuracy



Third-Order Parsing

UAS 92.9, Set pruning thresholds for no loss in accuracy



Google Web Treebanks

[Petrov & McDonald '12]

- Google Web Treebank
  - Funded by Google, annotated and released by LDC
  - 5 domains: Blogs, Newsgroups, Reviews, Emails, Q&A
  - ~2,000 manually annotated sentences (PTB-style)
  - >100,000 unlabeled sentences

- Shared Task at NAACL '12 Workshop
  - Constituency Trees or Stanford Dependencies
  - Train on WSJ + unlabeled data
  - 2 domains released for development
  - Test on remaining 3 domains

- New: Google Multilingual Treebank (6 languages)

POS Accuracy (SANCL Shared Task)

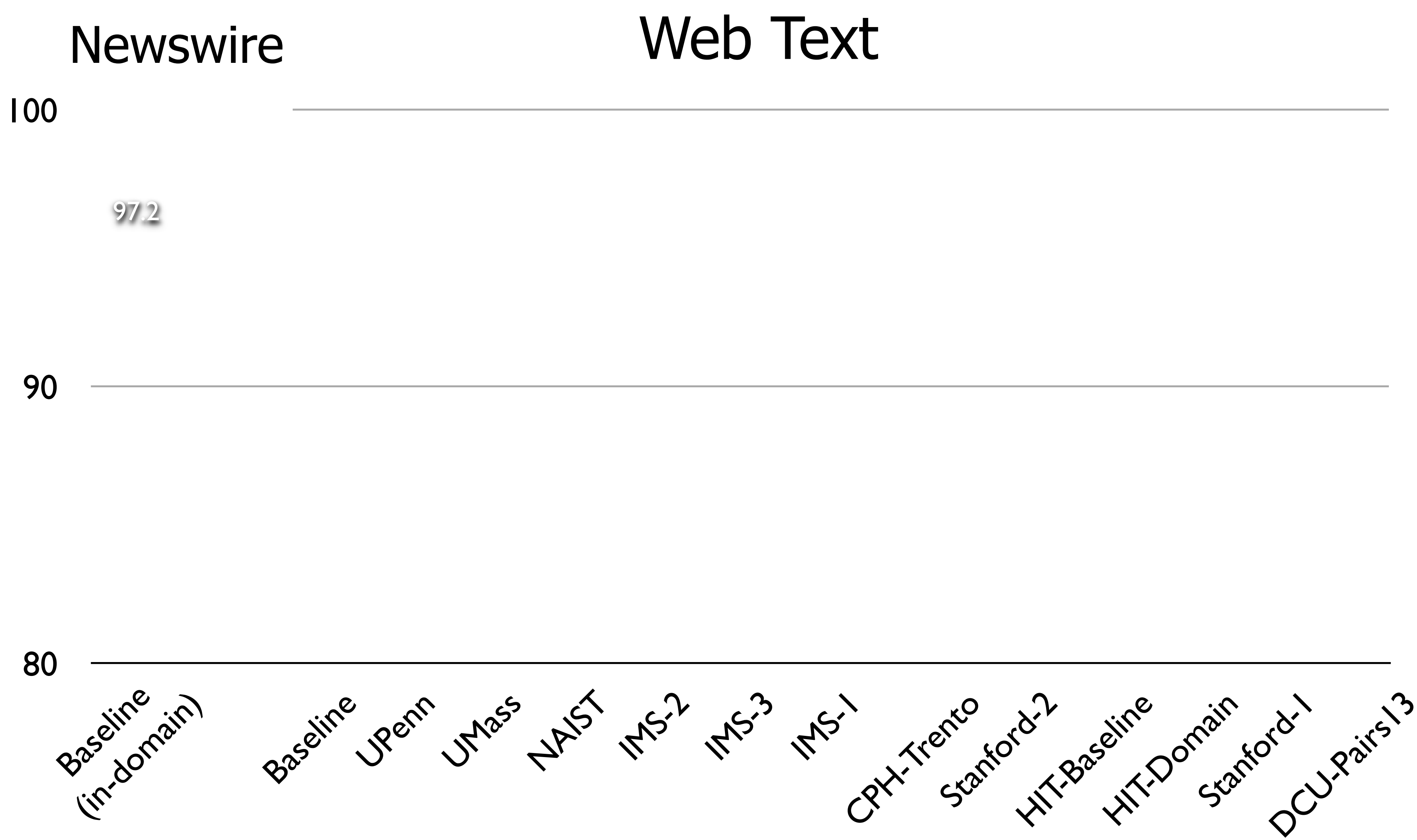
Newswire

Web Text

Baseline

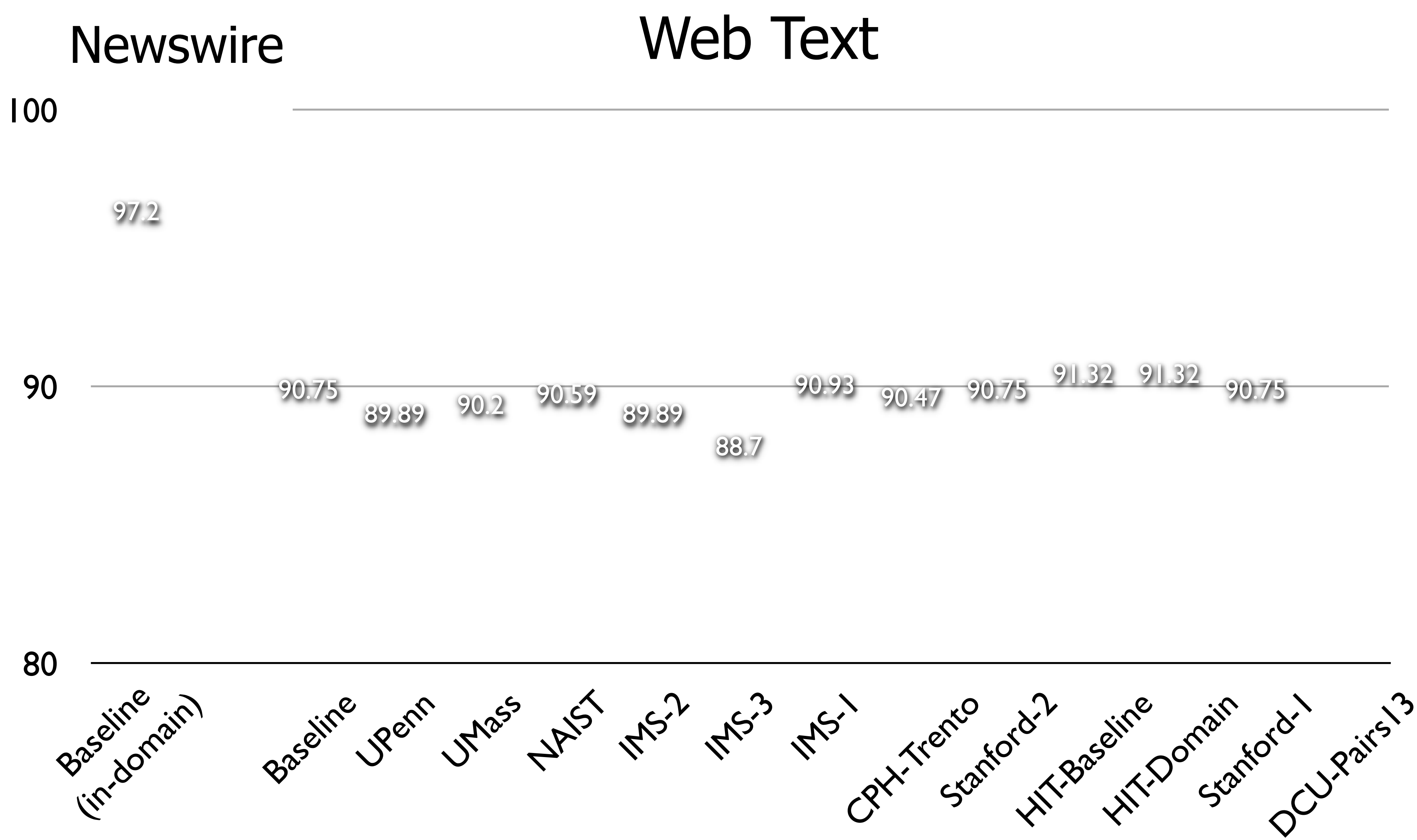
Baseline = StanfordTagger v2.0 [Manning ’11]

POS Accuracy (SANCL Shared Task)



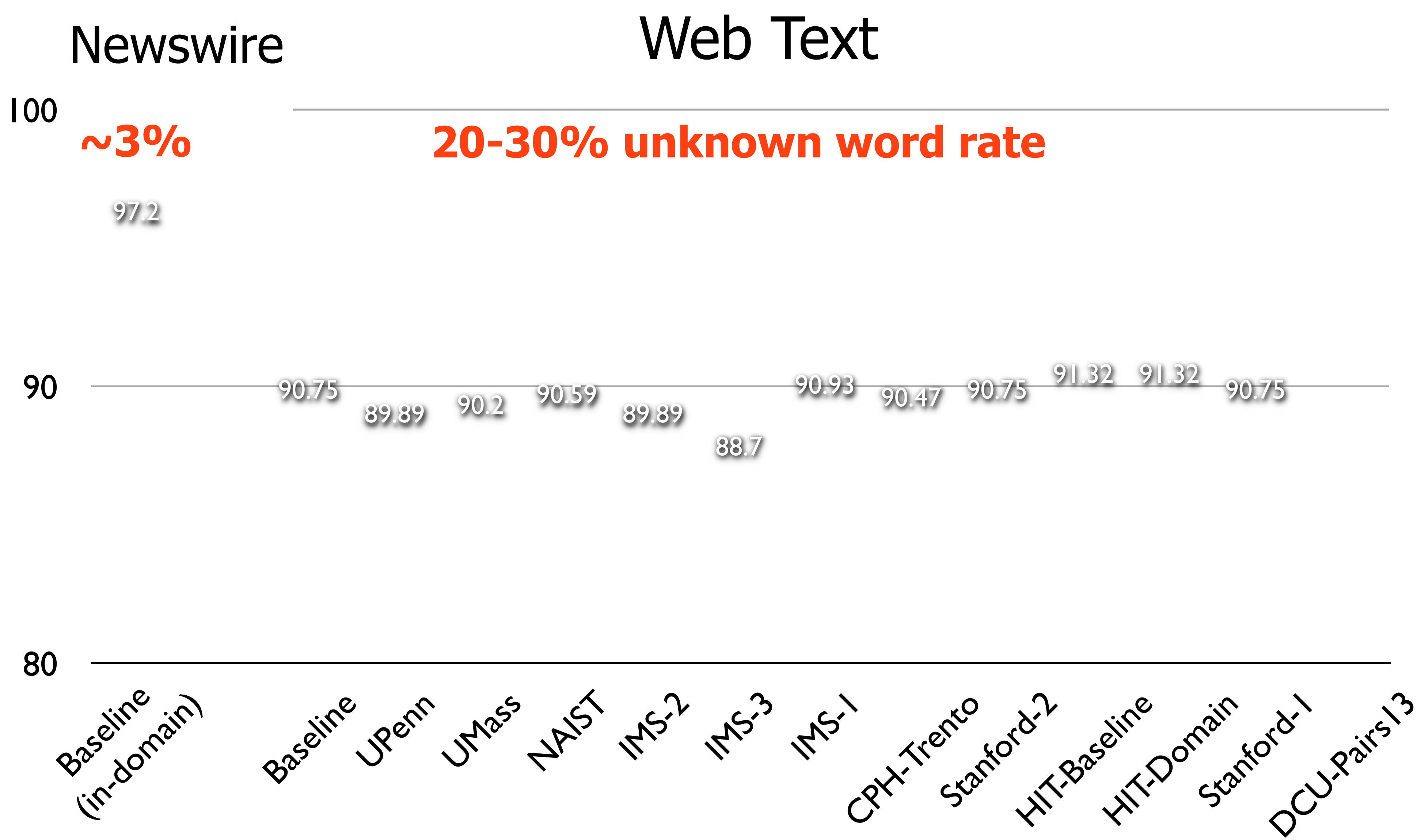
Baseline = StanfordTagger v2.0 [Manning '11]

POS Accuracy (SANCL Shared Task)



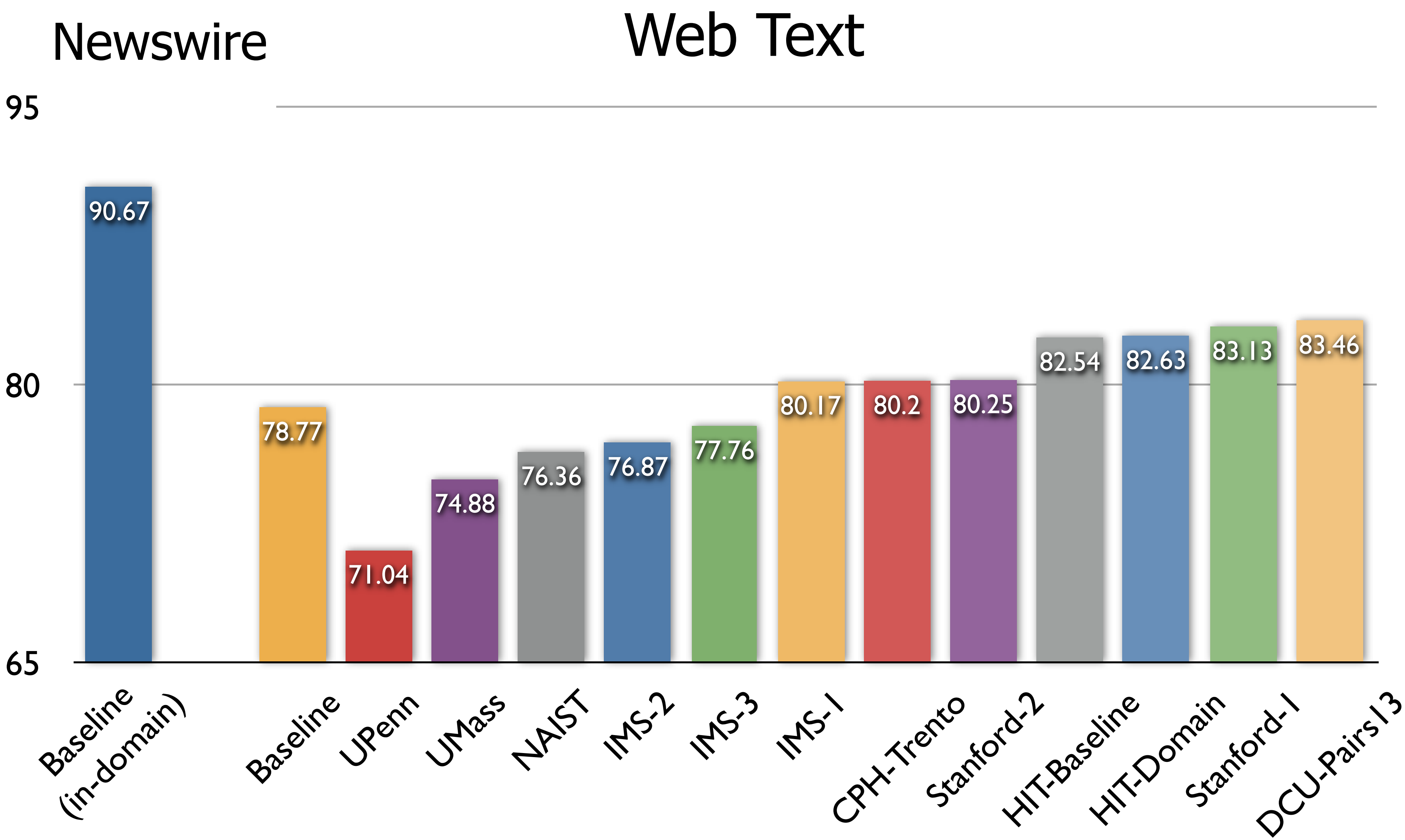
Baseline = StanfordTagger v2.0 [Manning '11]

POS Accuracy (SANCL Shared Task)



Baseline = StanfordTagger v2.0 [Manning '11]

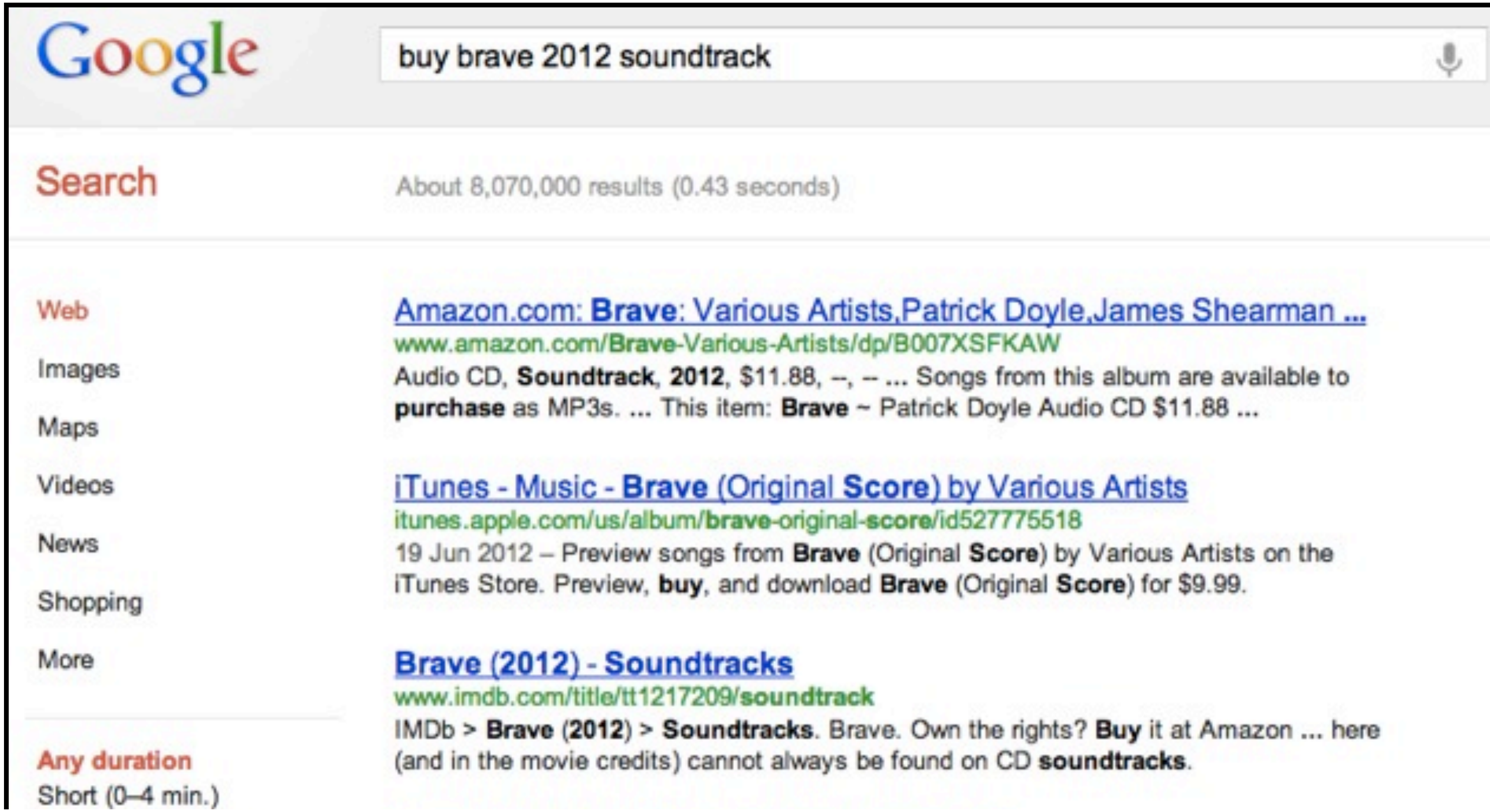
Parsing Accuracy (SANCL Shared Task)



Baseline = MaltParser [Zhang & Nivre '11]

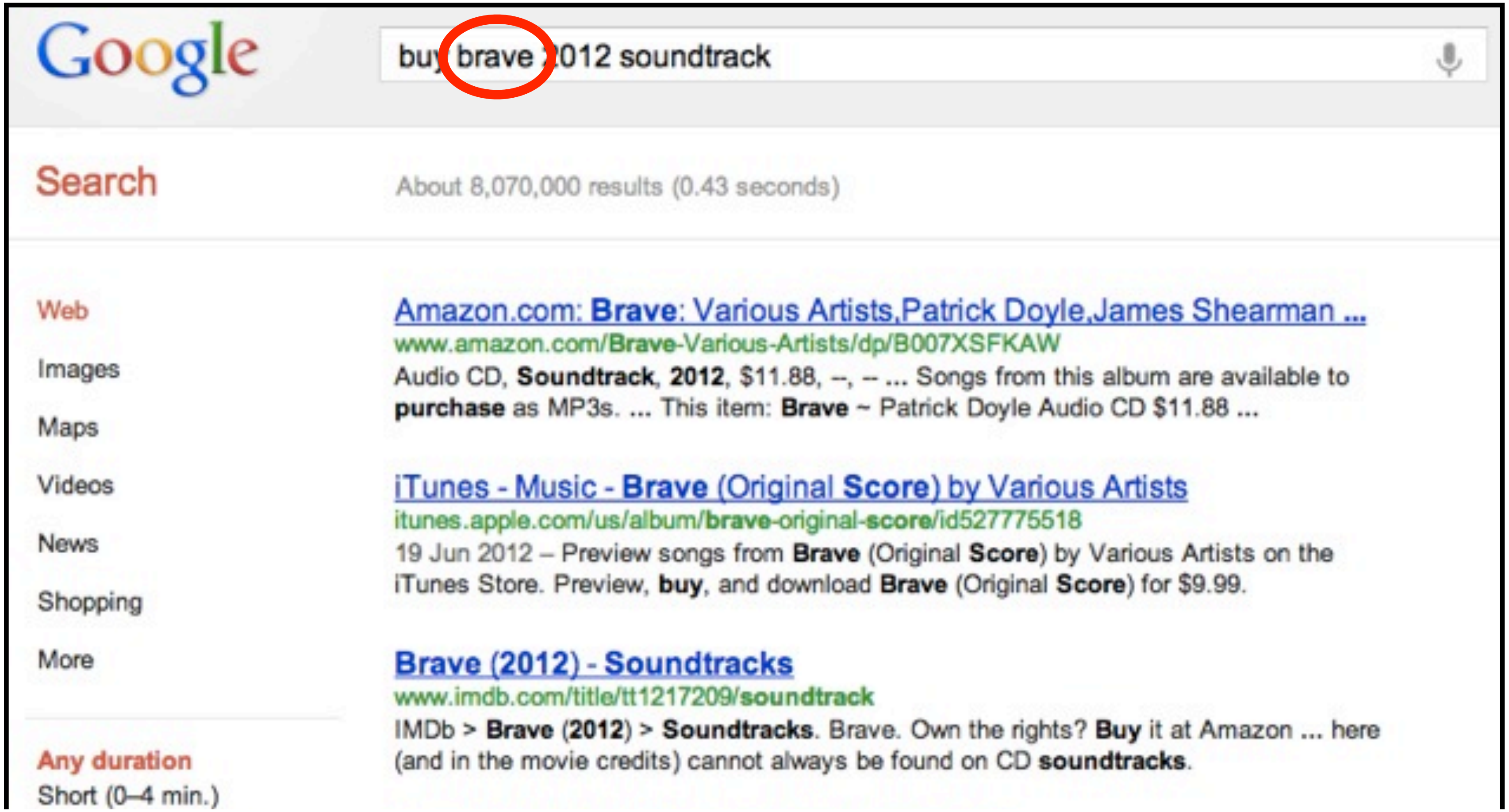
Tagging Search Queries

[Ganchev et al. '12]



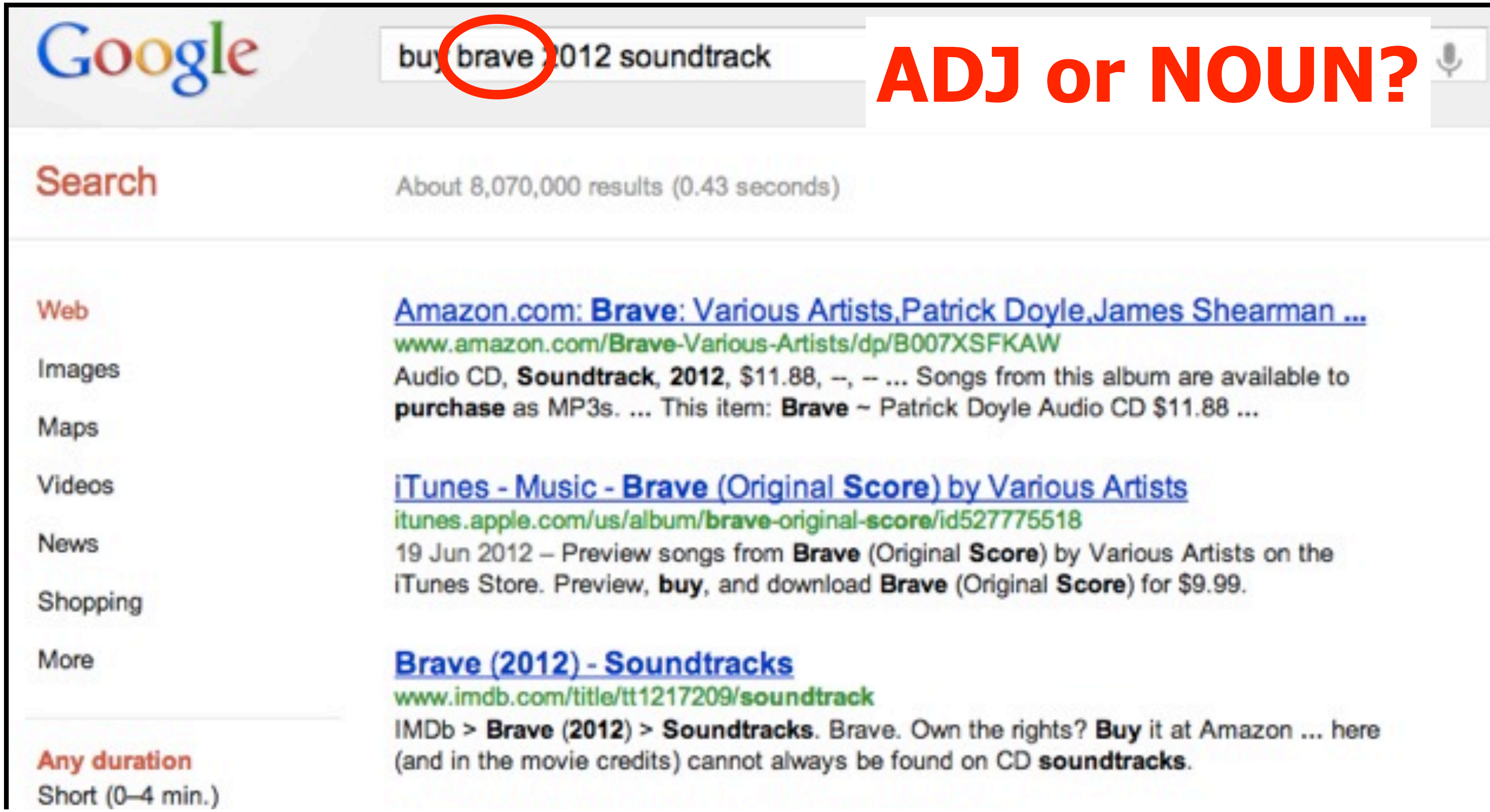
Tagging Search Queries

[Ganchev et al. '12]



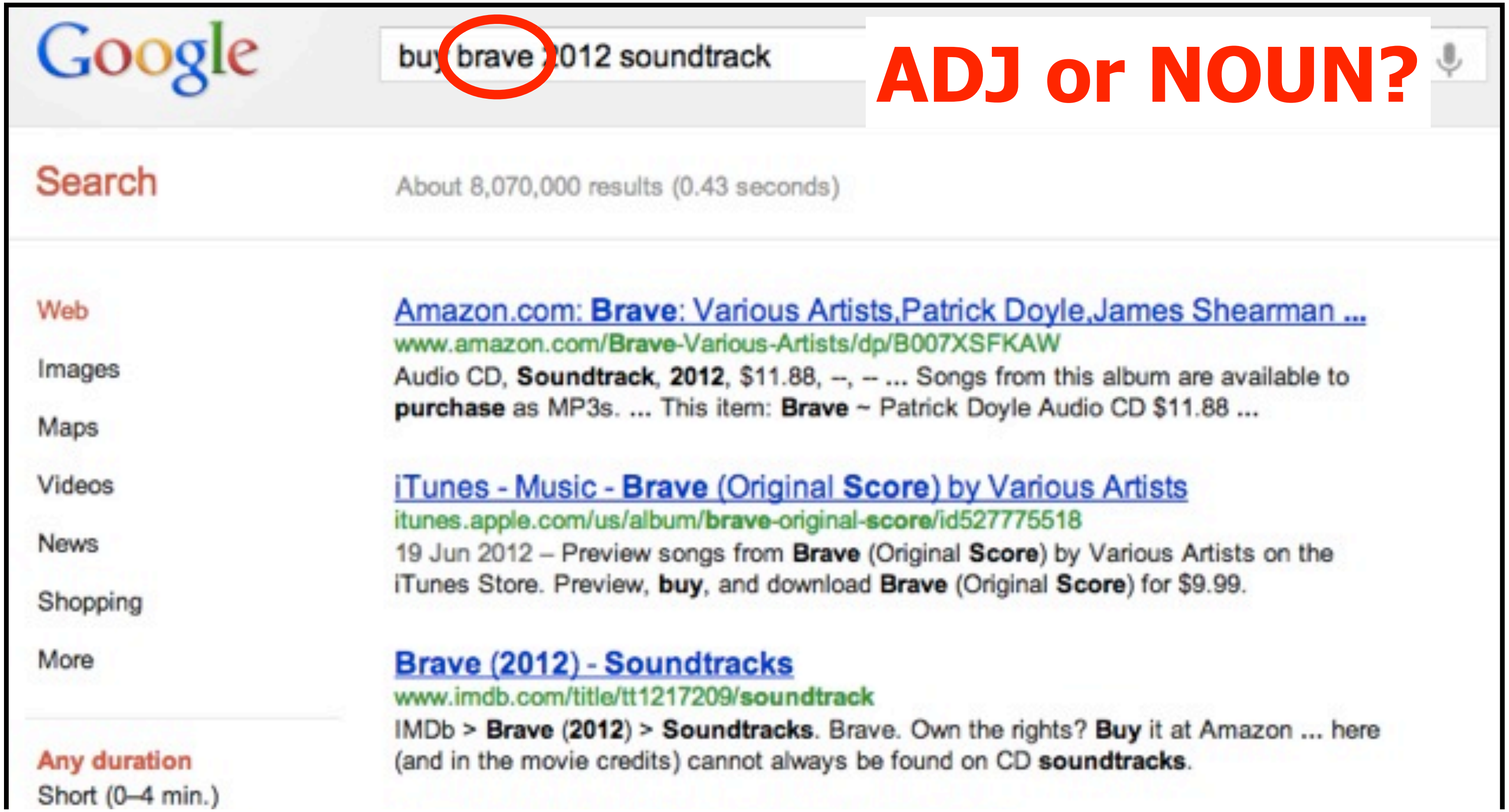
Tagging Search Queries

[Ganchev et al. '12]



Tagging Search Queries

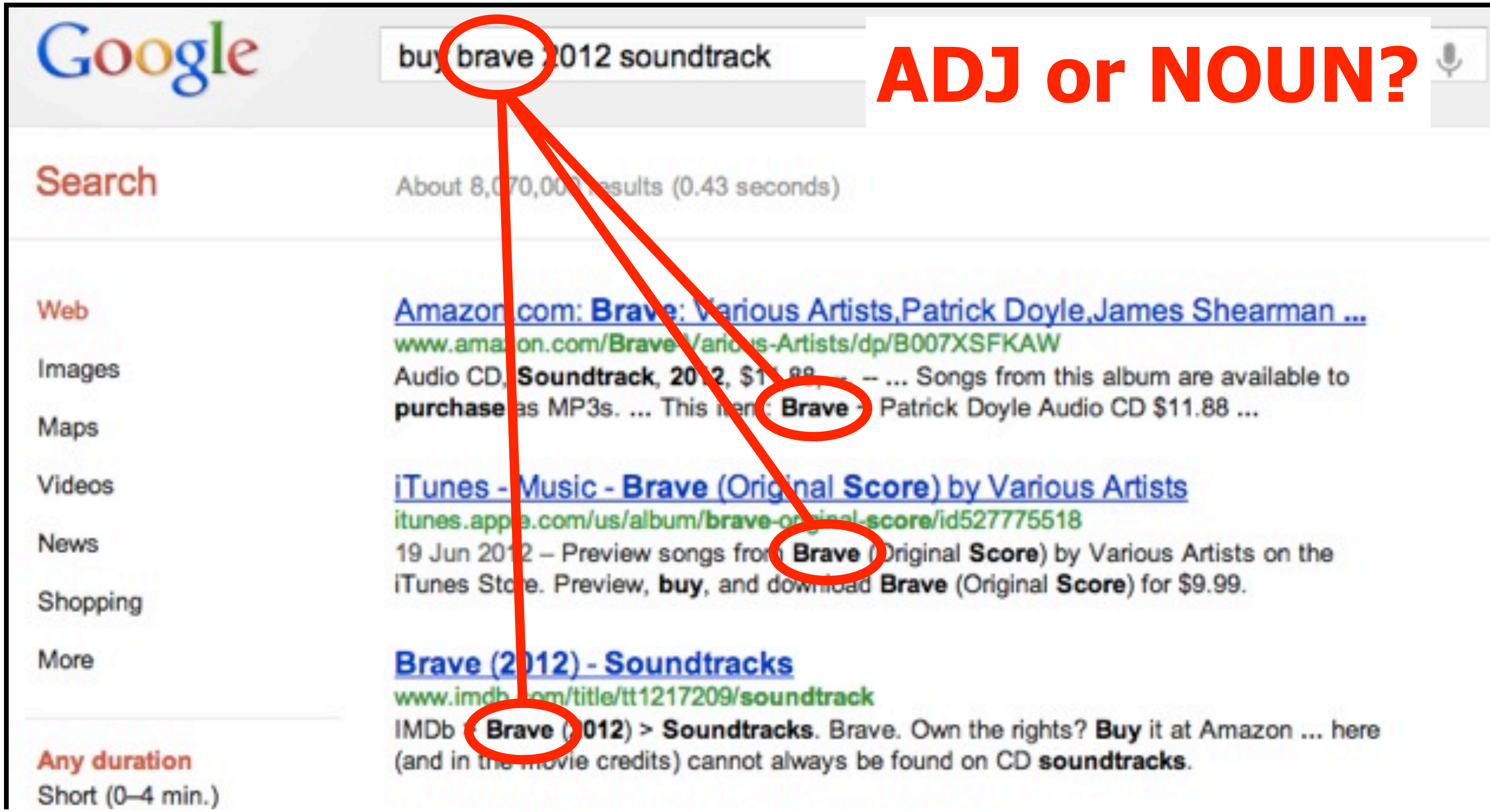
[Ganchev et al. ’12]



buy	**brave**	**2012**	**soundtrack**
VB	JJ	CD	NN
VERB	ADJ	NUM	NOUN

Tagging Search Queries

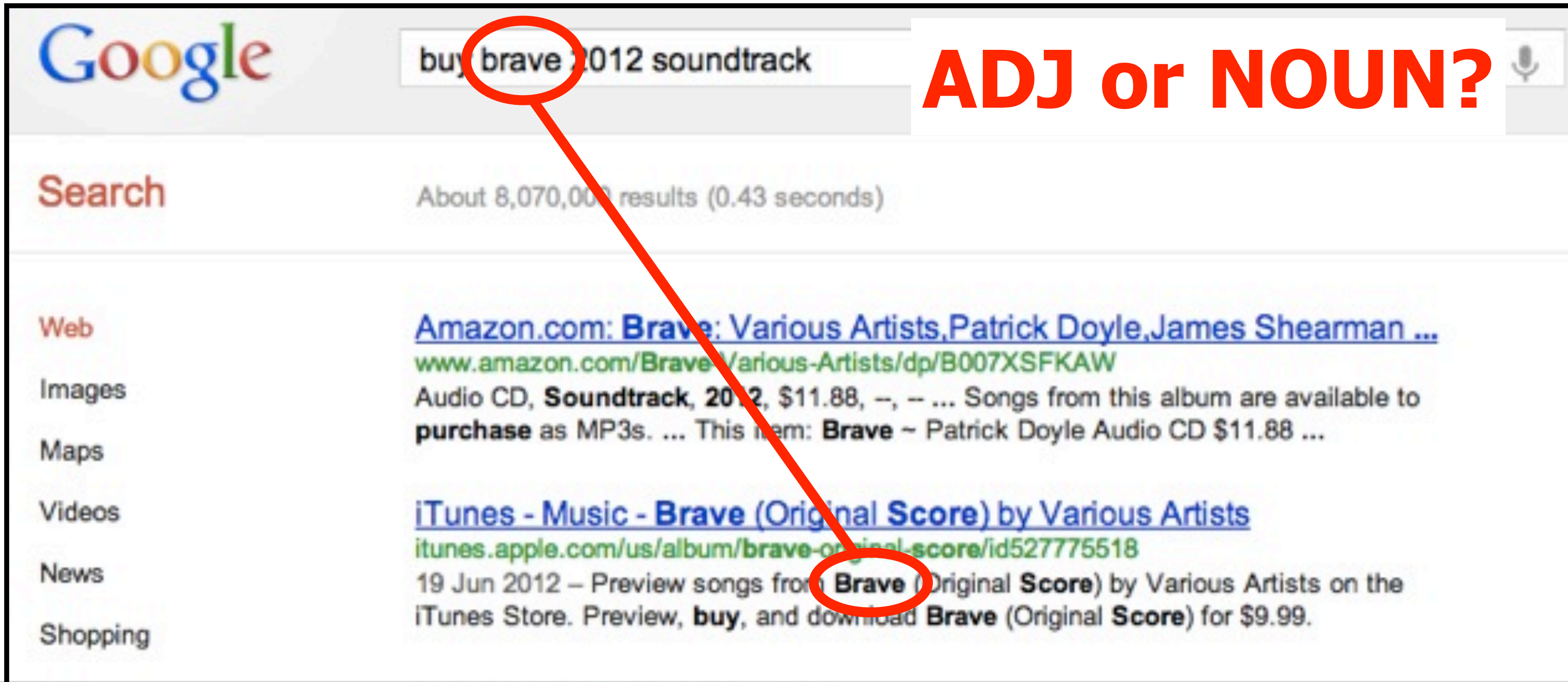
[Ganchev et al. '12]



buy	brave	2012	soundtrack
VB	JJ	CD	NN
VERB	ADJ	NUM	NOUN

Tagging Search Queries

[Ganchev et al. '12]

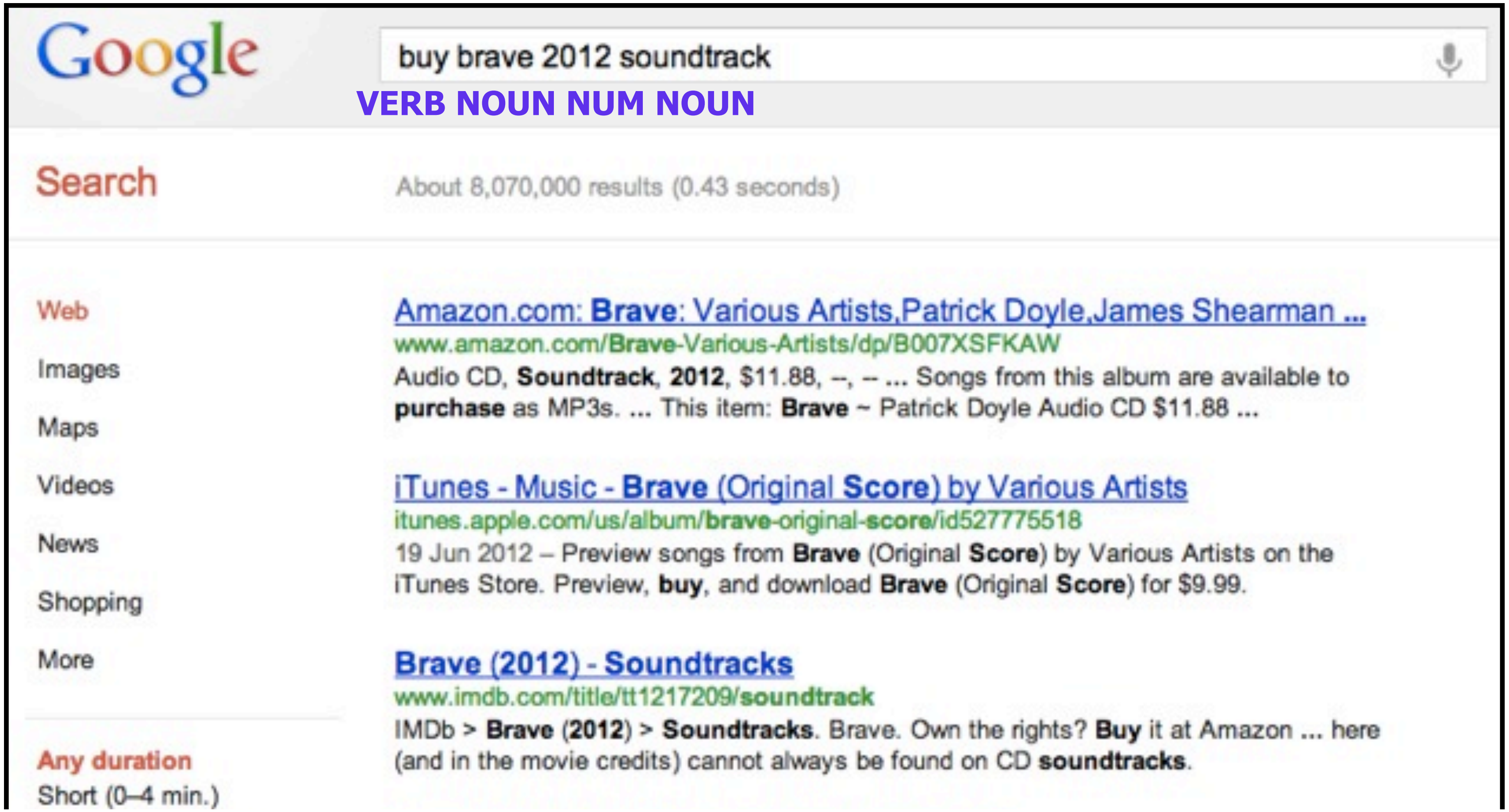


Preview	songs	from	Brave	(	Original	Score	)	by	Various	Artists	.
NN	NNS	IN	NNP	(	JJ	NN	)	IN	JJ	NNS	.
NOUN	NOUN	ADP	NOUN	.	ADJ	NOUN	.	ADP	ADJ	NOUN	.

buy	brave	2012	soundtrack
VB	JJ	CD	NN
VERB	ADJ	NUM	NOUN

Tagging Search Queries

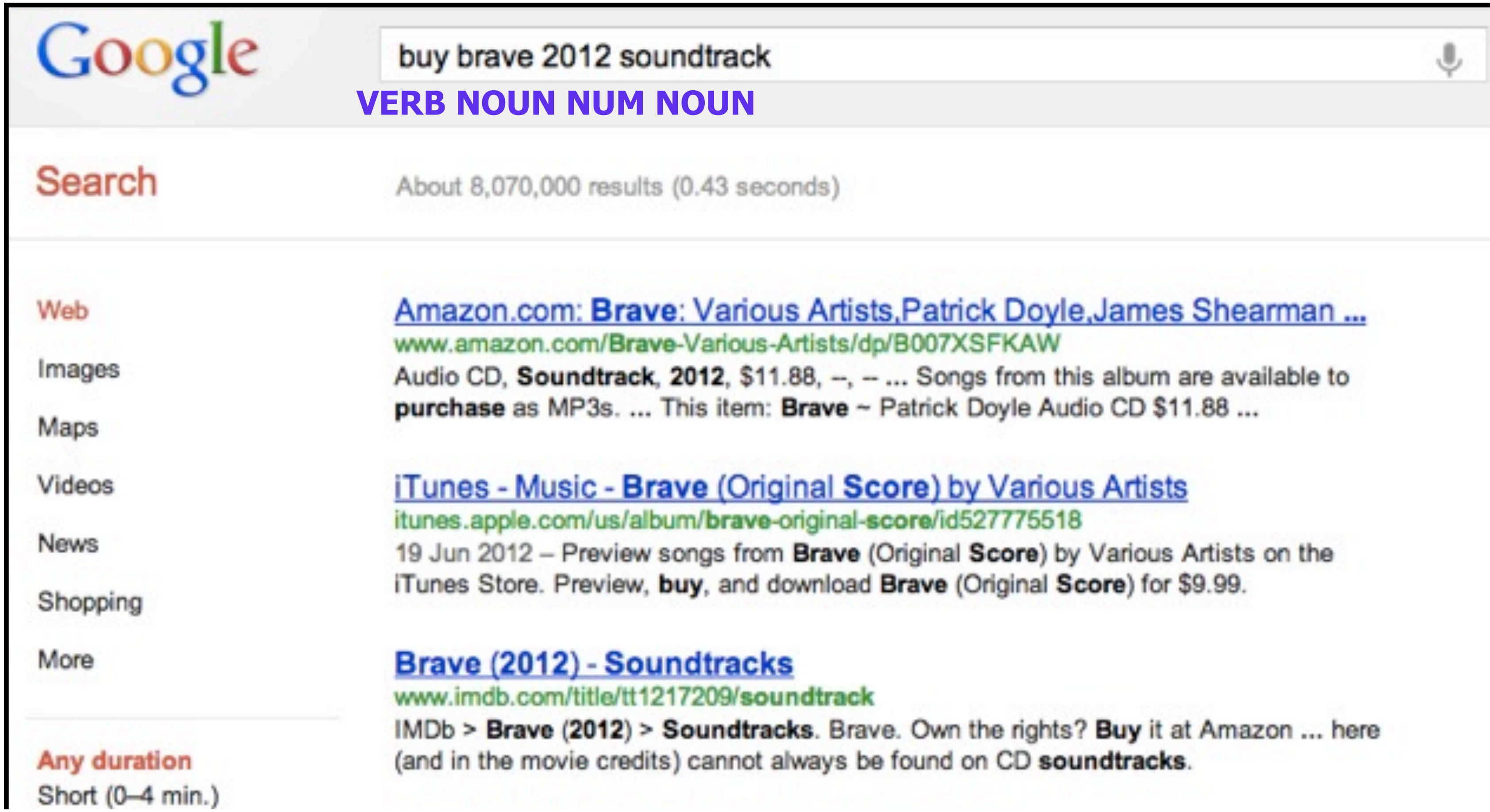
[Ganchev et al. '12]



buy	brave	2012	soundtrack
VB	JJ	CD	NN
VERB	ADJ	NUM	NOUN

Tagging Search Queries

[Ganchev et al. '12]



Baseline

buy	brave	2012	soundtrack
VB	JJ	CD	NN
VERB	ADJ	NUM	NOUN

Retrained

buy	brave	2012	soundtrack
VB	NNP	CD	NN
VERB	NOUN	NUM	NOUN

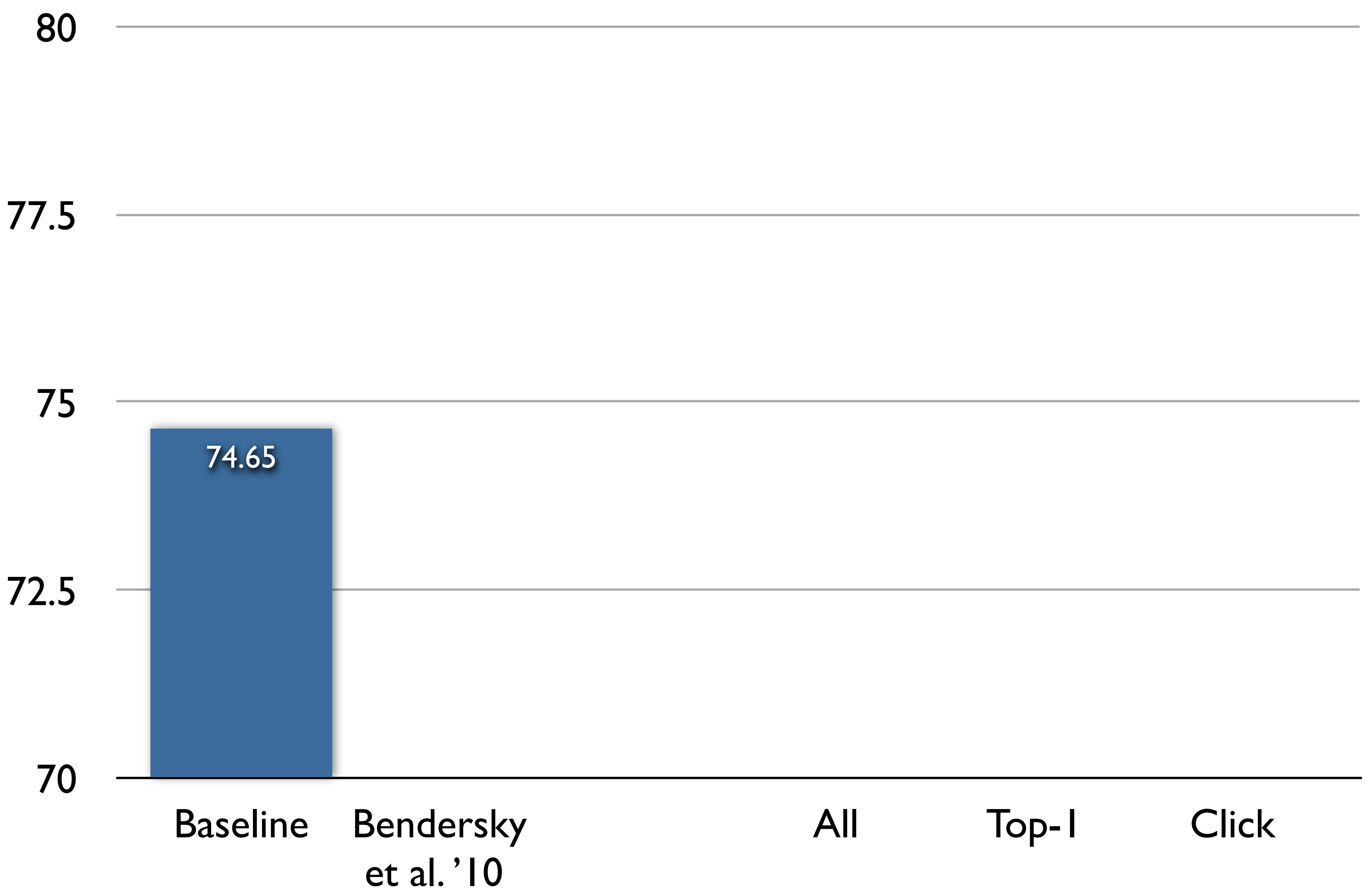
Query Tagging Results

[Ganchev et al. '12]

et al. '10

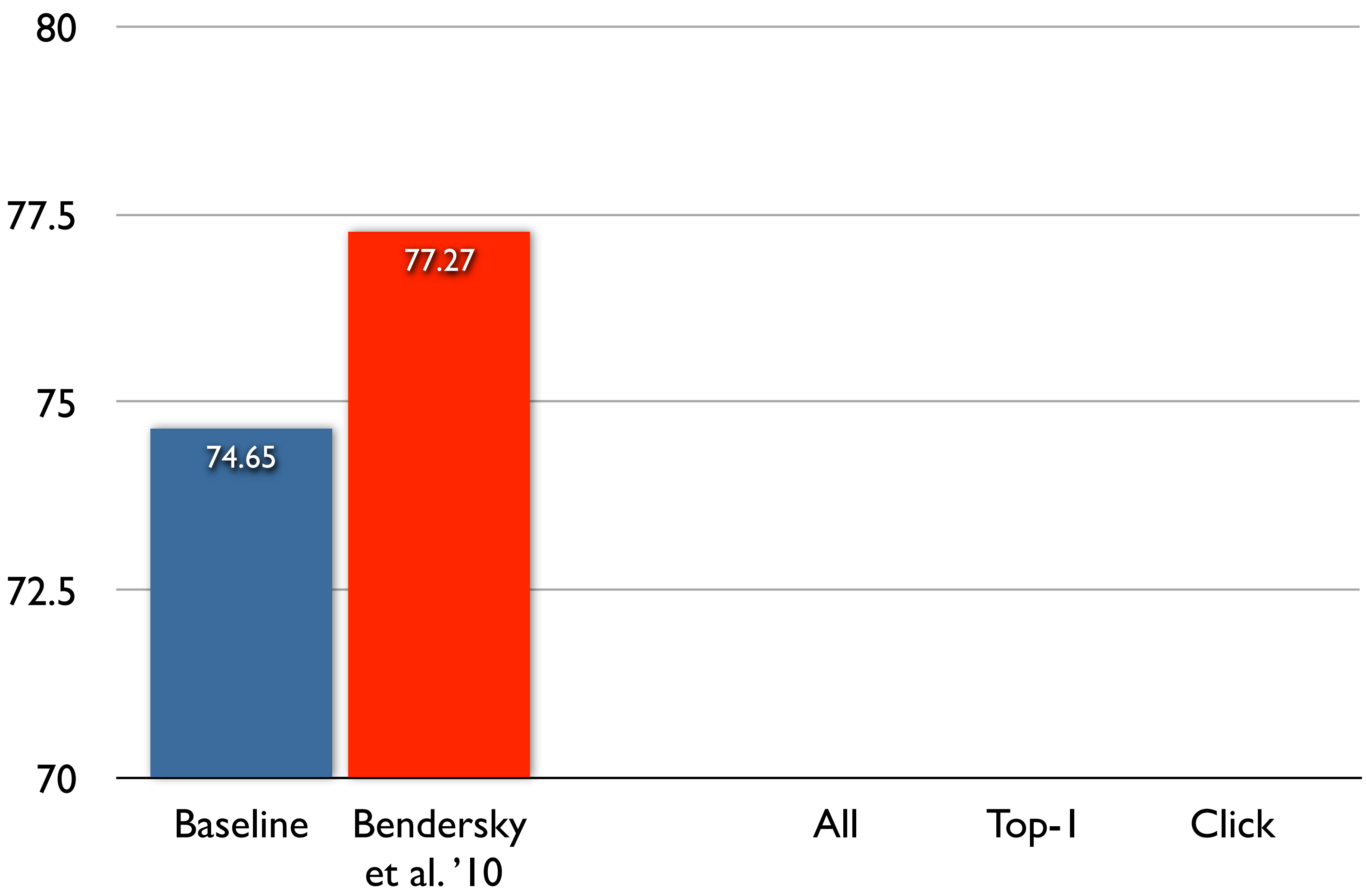
Query Tagging Results

[Ganchev et al. '12]



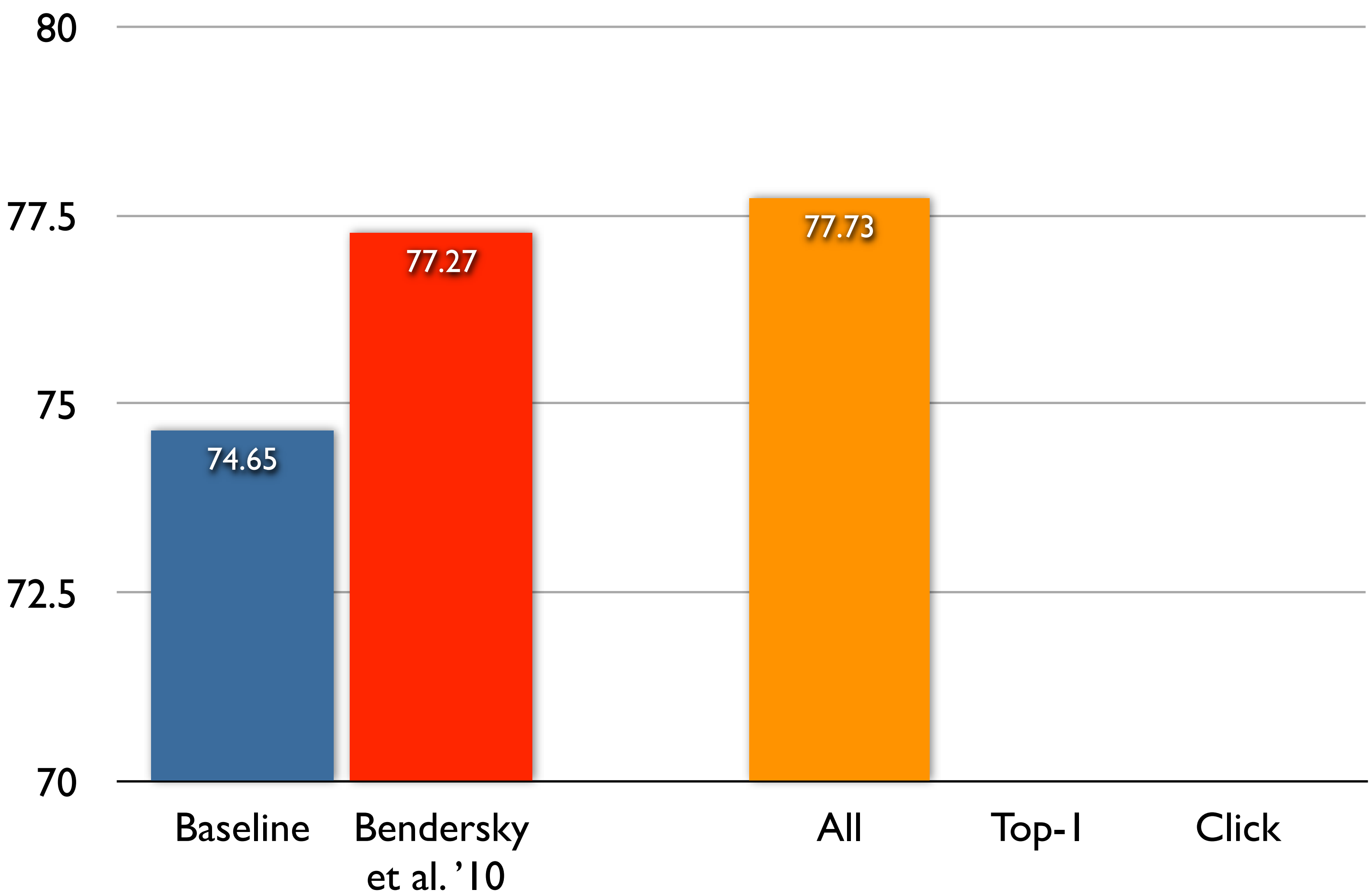
Query Tagging Results

[Ganchev et al. '12]



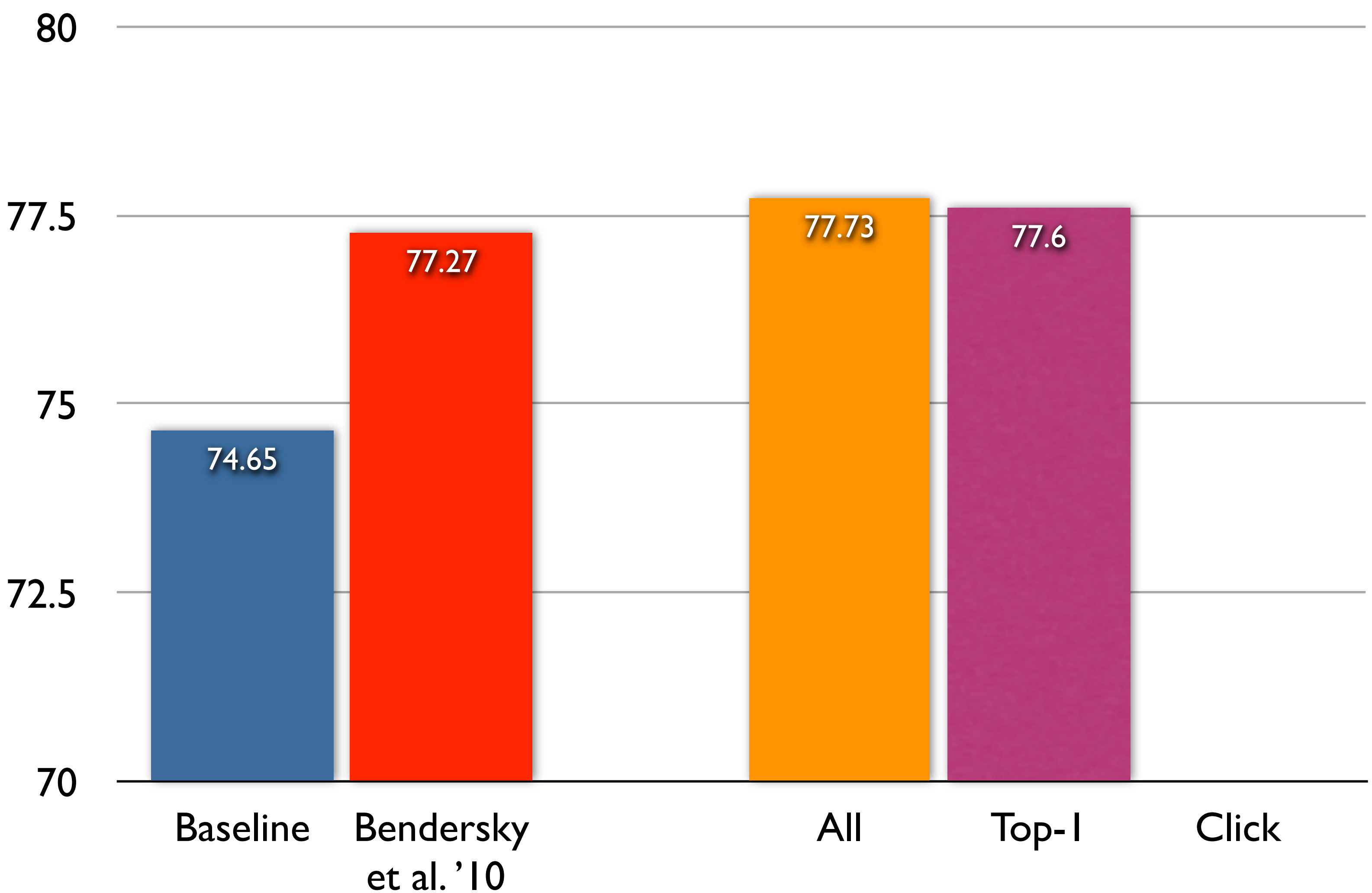
Query Tagging Results

[Ganchev et al. '12]



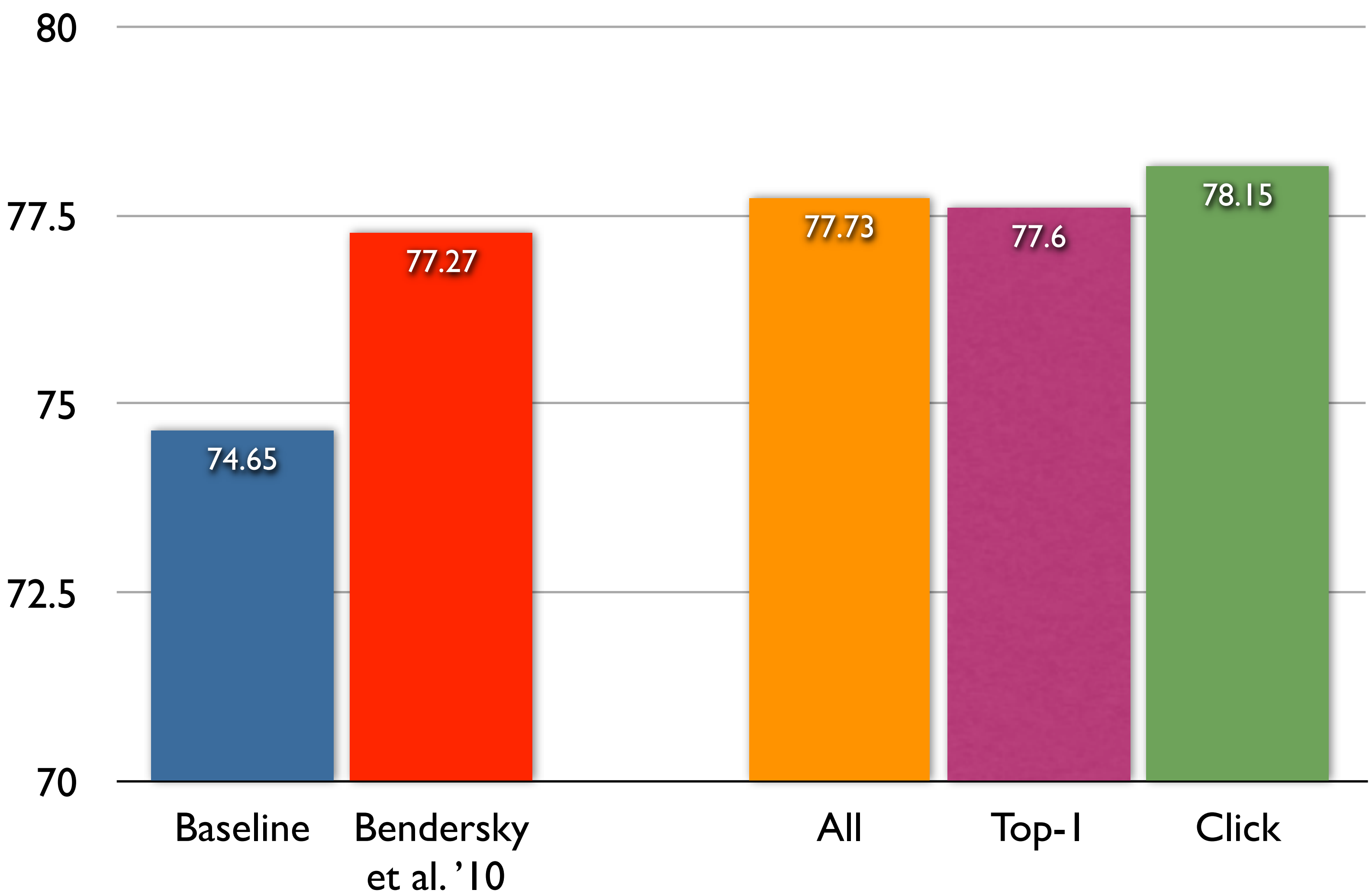
Query Tagging Results

[Ganchev et al. '12]

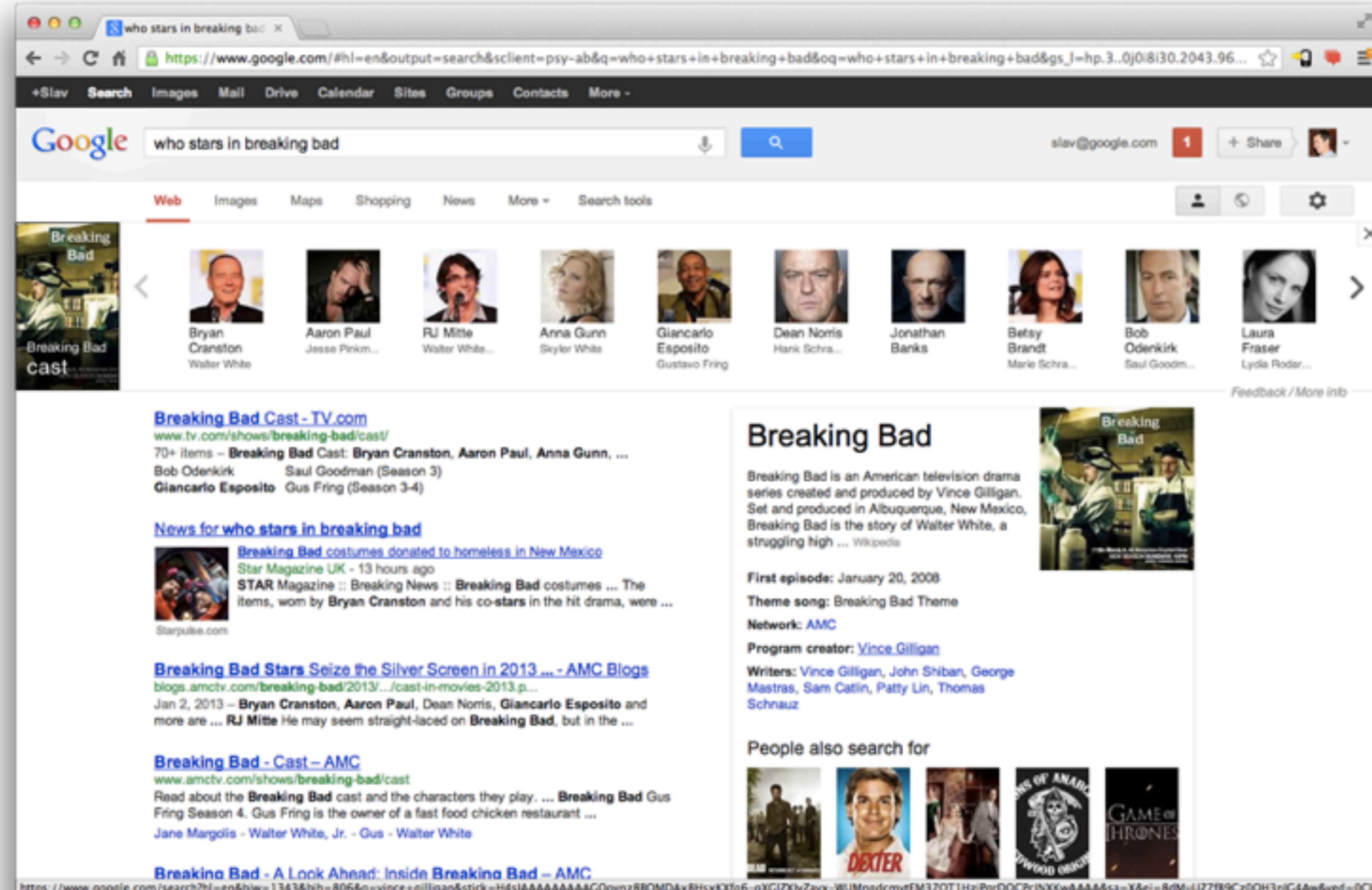


Query Tagging Results

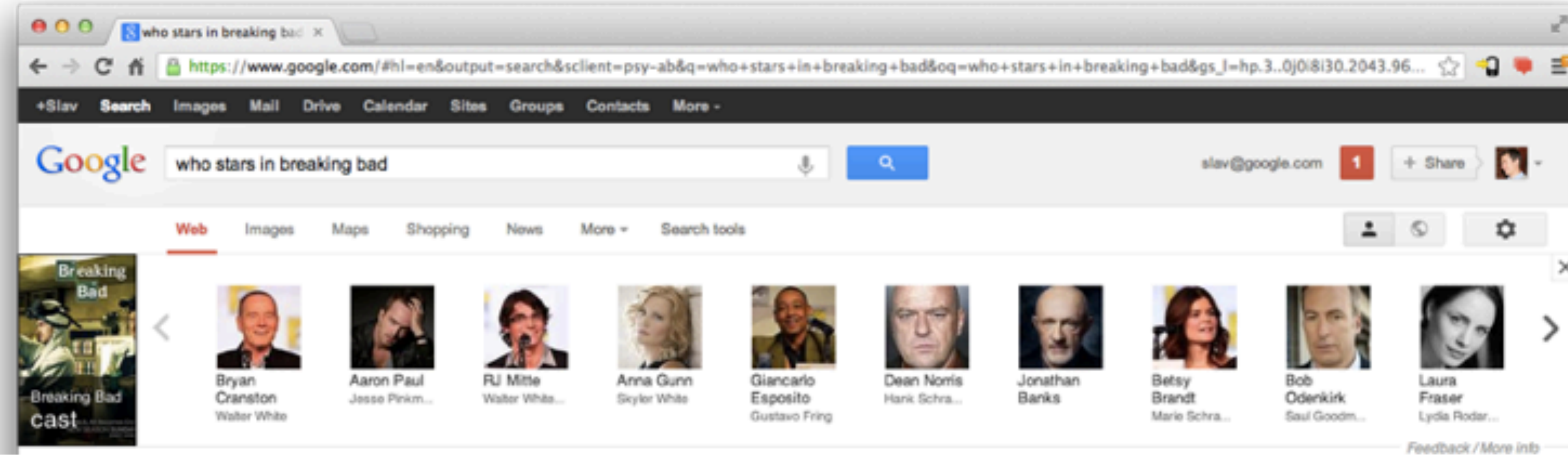
[Ganchev et al. '12]

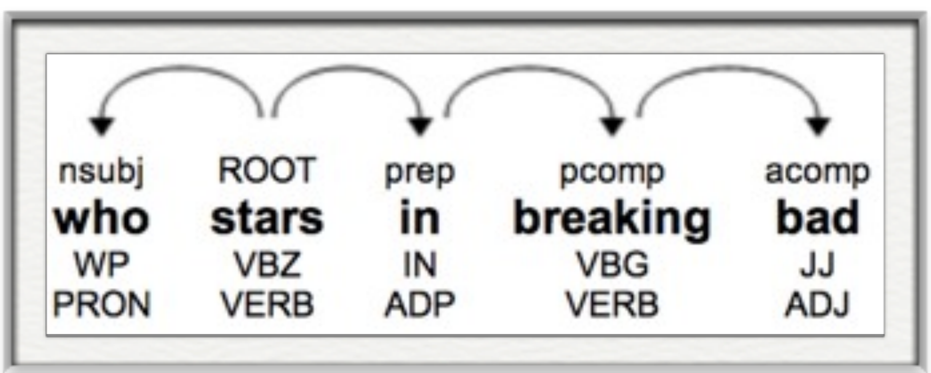


The breaking bad case...

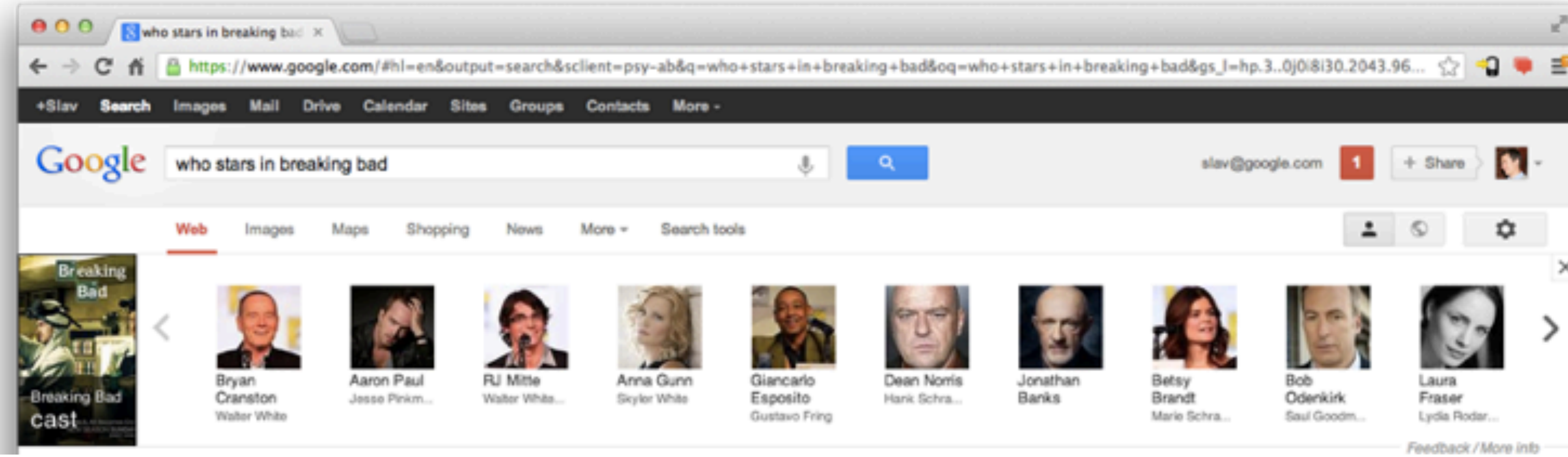


The breaking bad case...

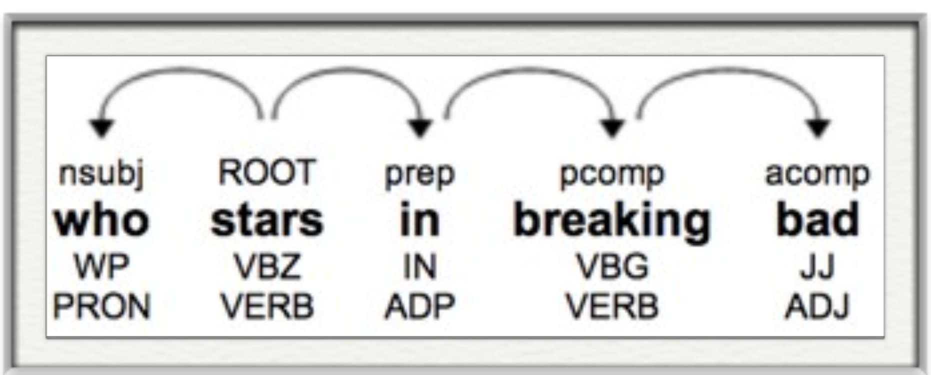




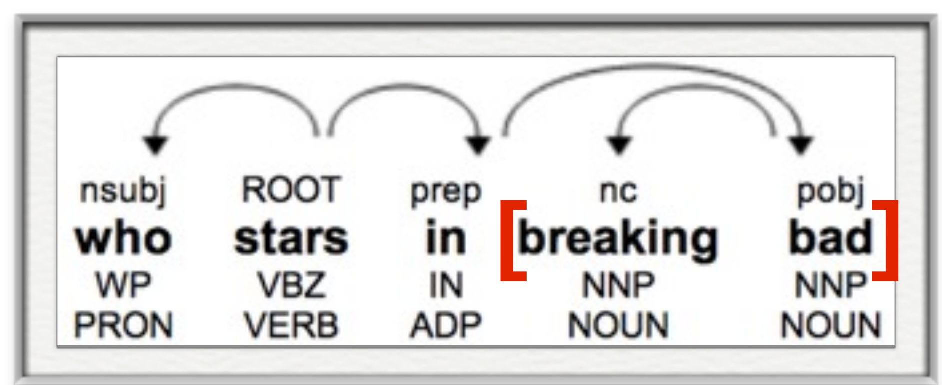
The breaking bad case...



Before

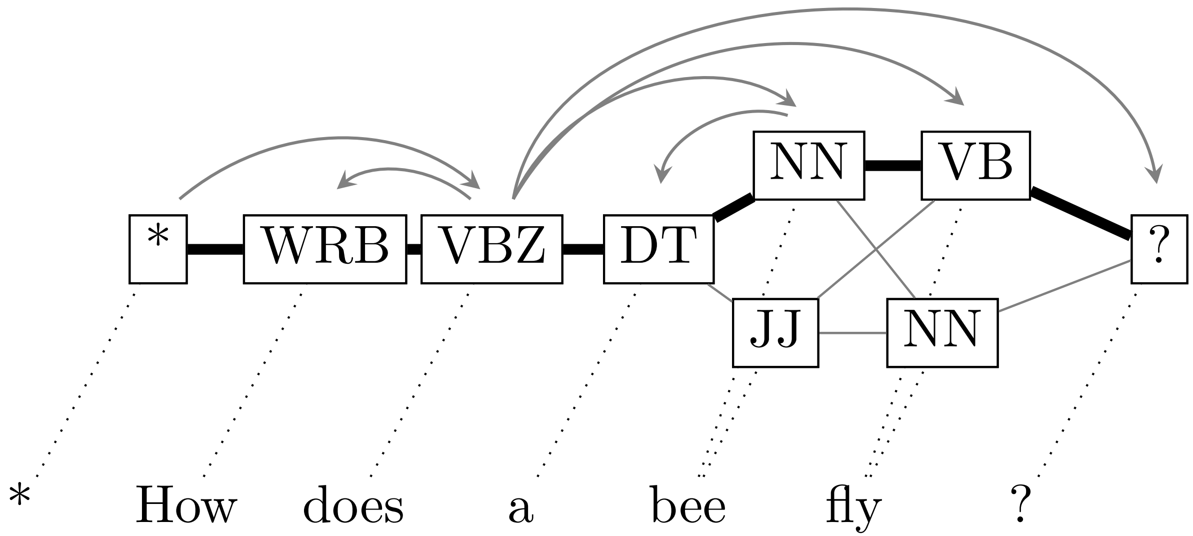


After



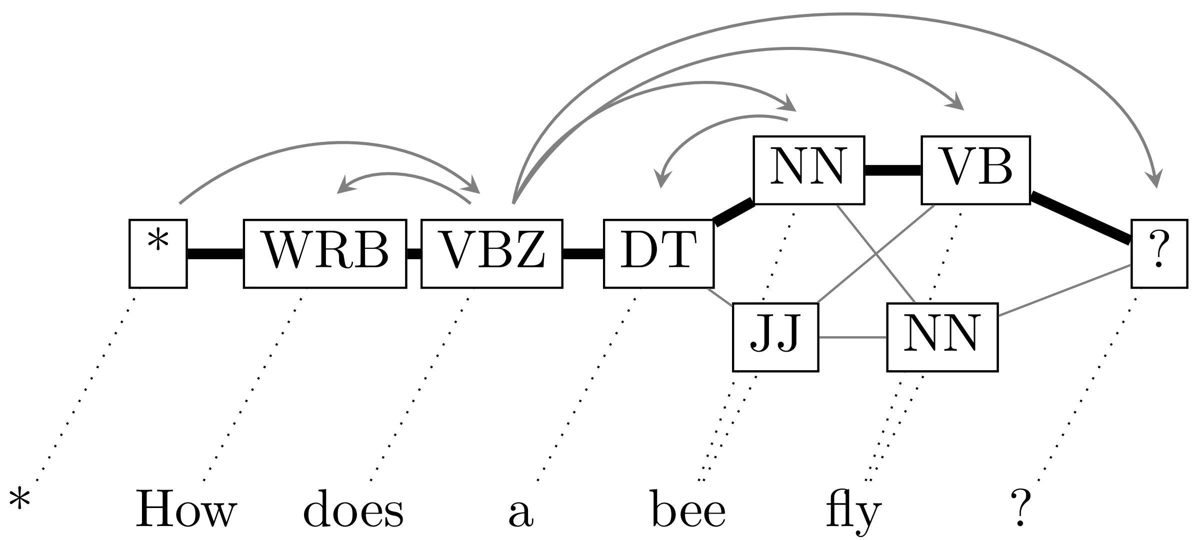
Use freebase entries as features

Integrated Tagging & Parsing



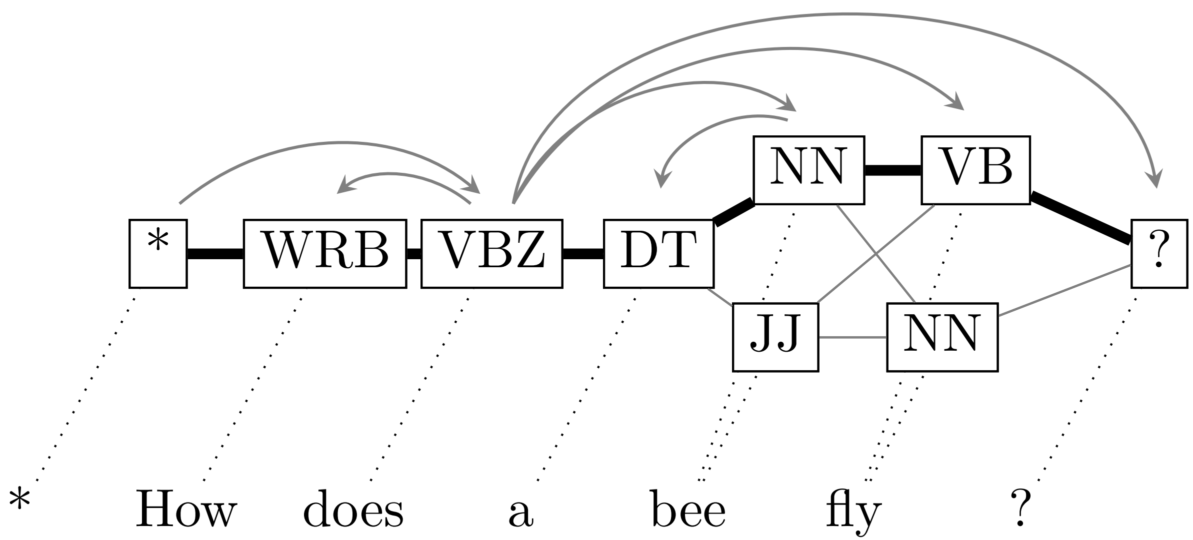
Integrated Tagging & Parsing

- Sequential Markov Models have limitations:



Integrated Tagging & Parsing

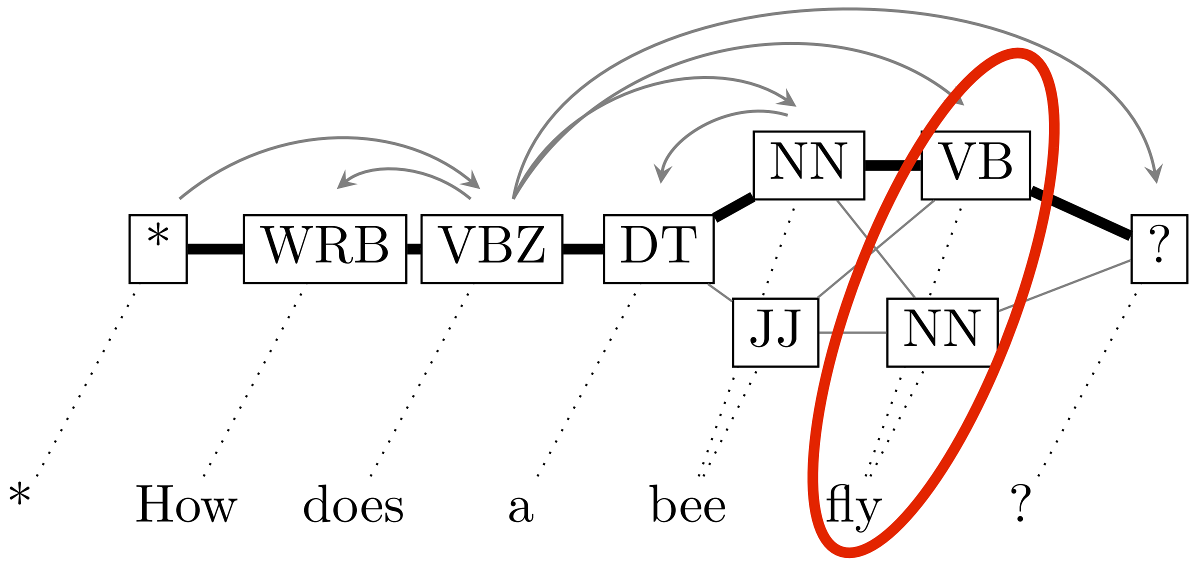
- Sequential Markov Models have limitations:



... 'to coach kim'

Integrated Tagging & Parsing

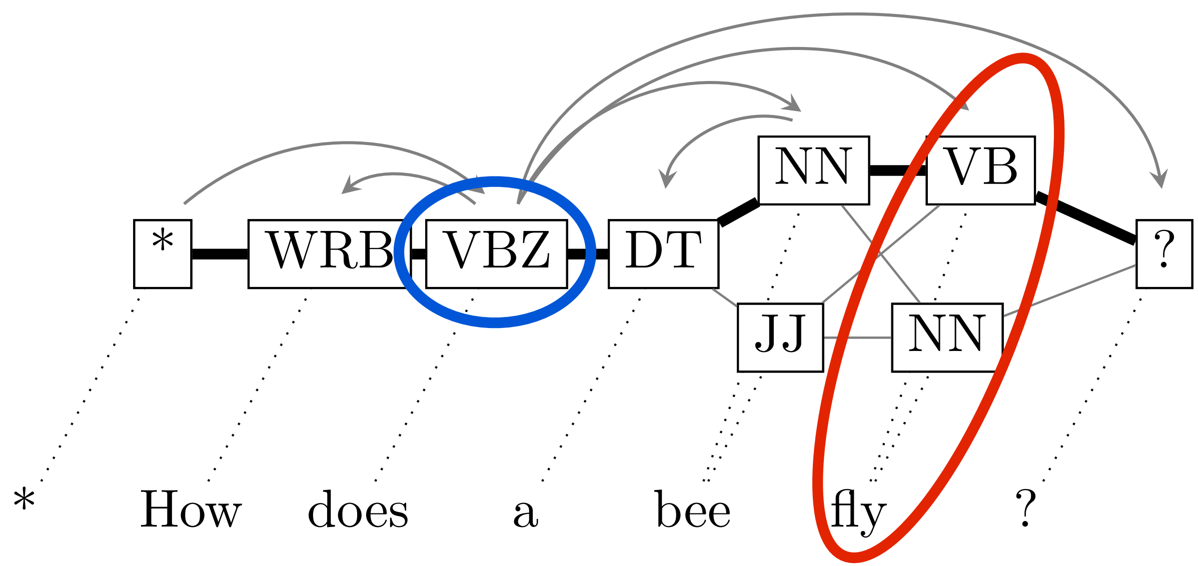
- Sequential Markov Models have limitations:



… 'to coach kim'

Integrated Tagging & Parsing

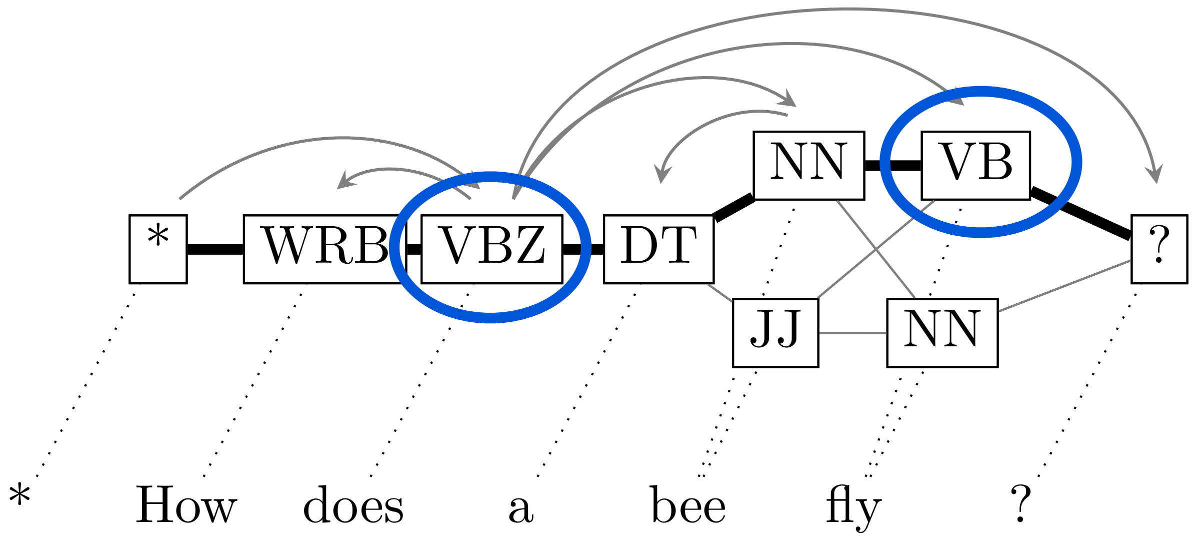
- Sequential Markov Models have limitations:



... 'to coach kim'

Integrated Tagging & Parsing

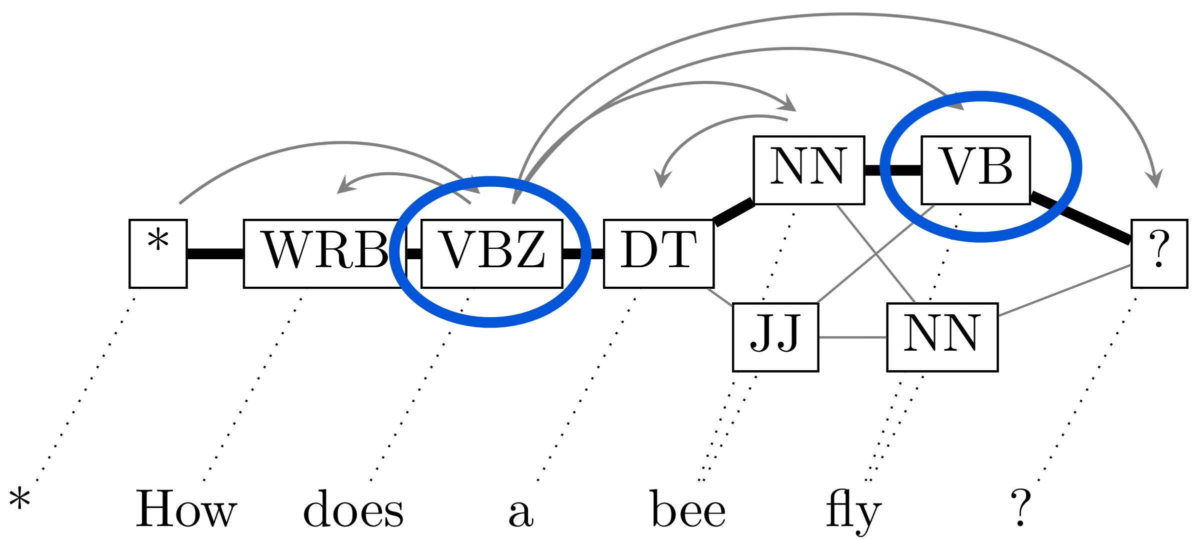
- Sequential Markov Models have limitations:



... 'to coach kim'

Integrated Tagging & Parsing

- Sequential Markov Models have limitations:

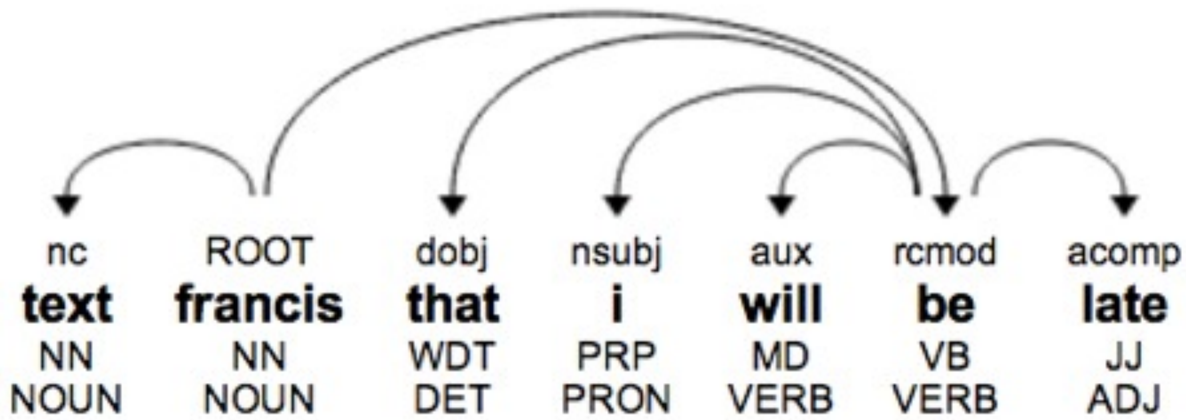


... 'to coach kim'

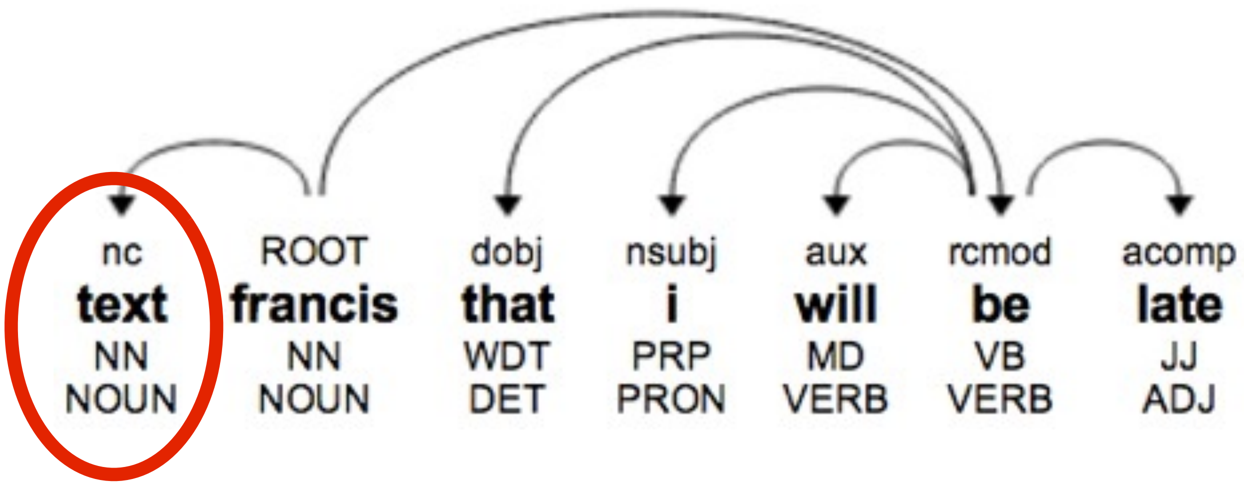
- Ongoing work:
  Search jointly over POS tags and parse trees.

Manual Intervention

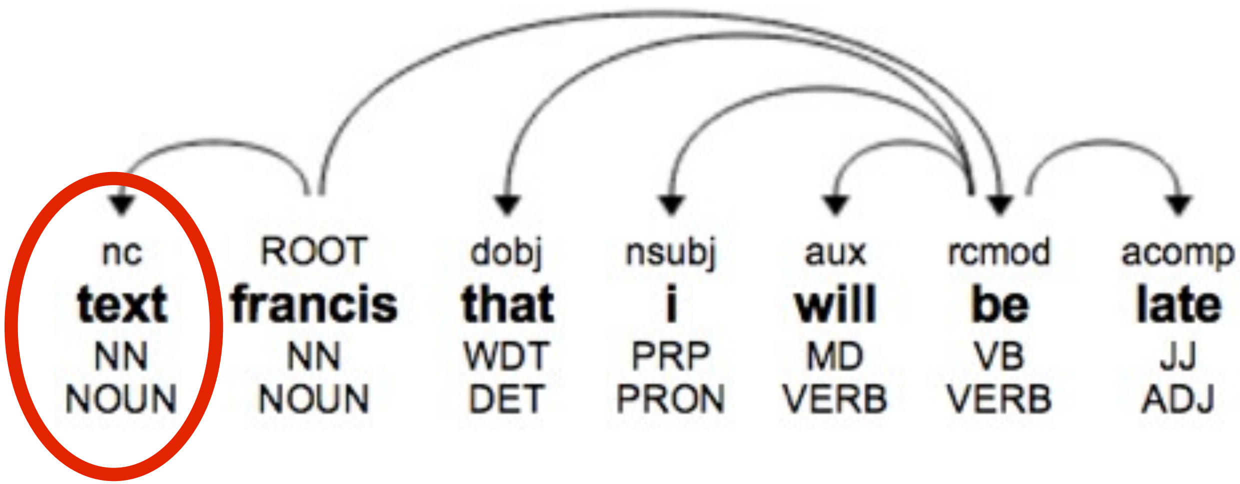
Manual Intervention

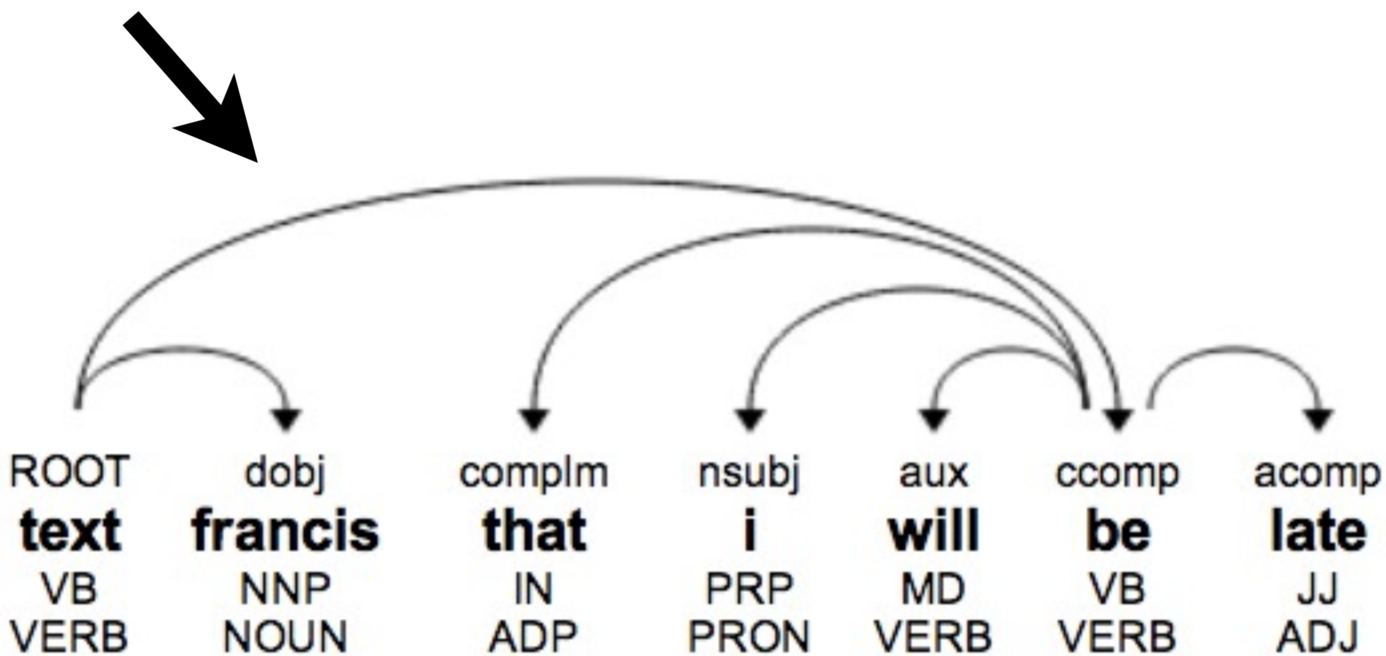


Manual Intervention



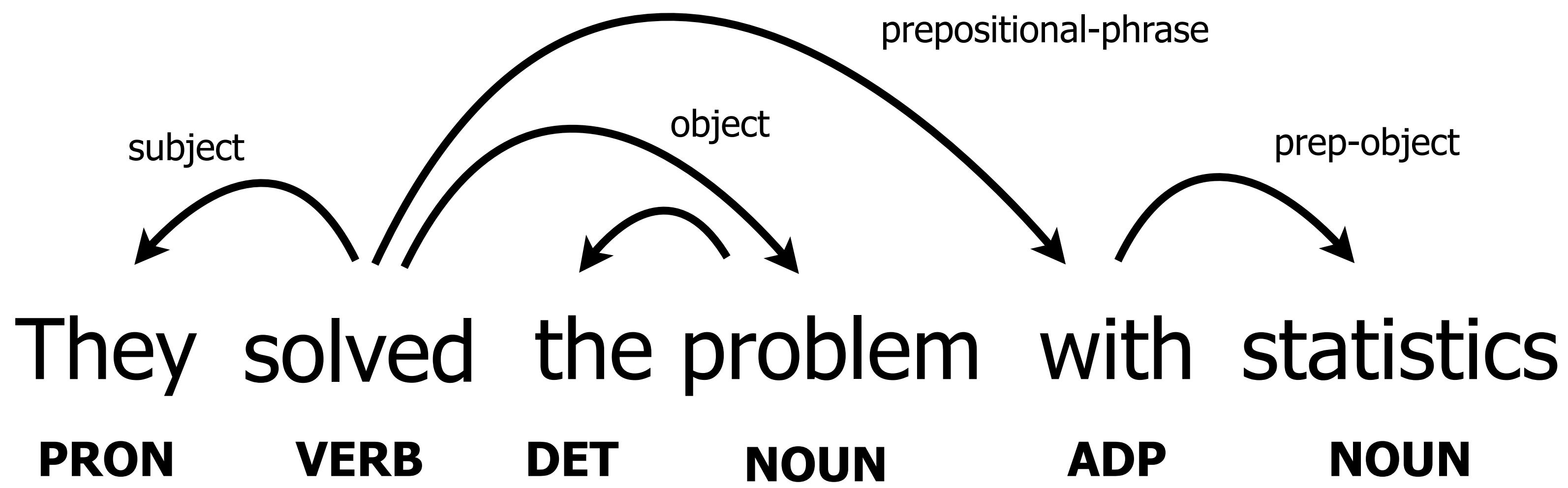
Manual Intervention





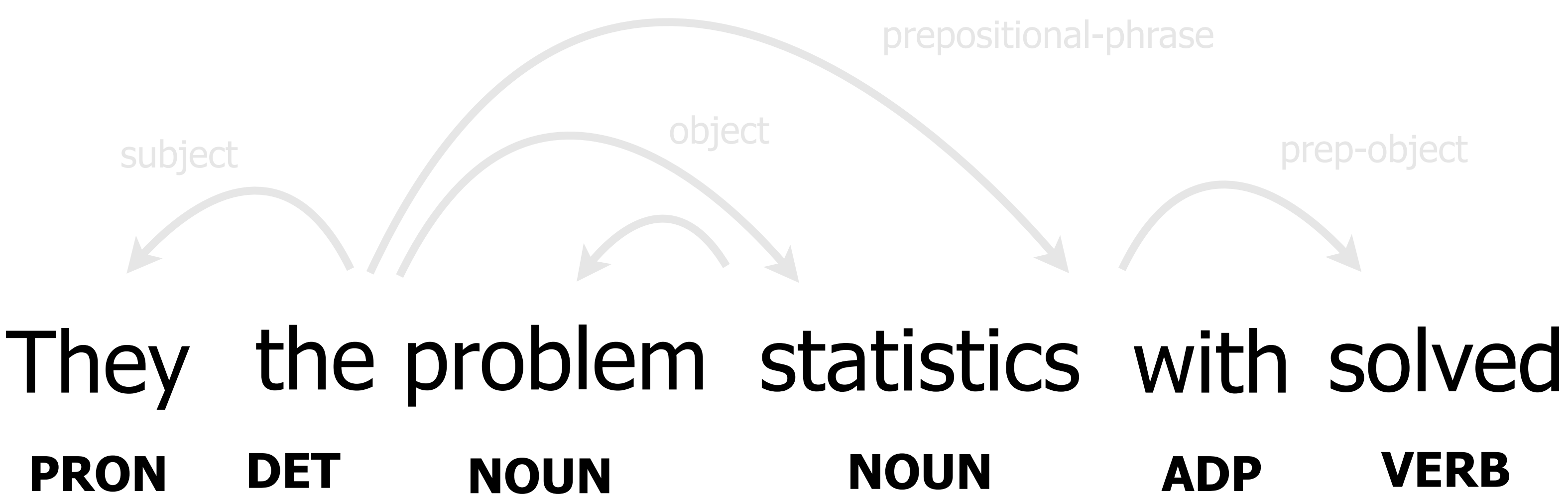
Machine Translation Reordering

- Source-side reordering for Japanese-English MT [Collins et al. '05]
- Dependency-based reordering for English-Japanese [Xu et al. '09]



Machine Translation Reordering

- Source-side reordering for Japanese-English MT [Collins et al. '05]
- Dependency-based reordering for English-Japanese [Xu et al. '09]



Reordering Score

[Talbot et al. '11]

- Source-side reordering for Japanese-English MT

  [Collins et al. '05]

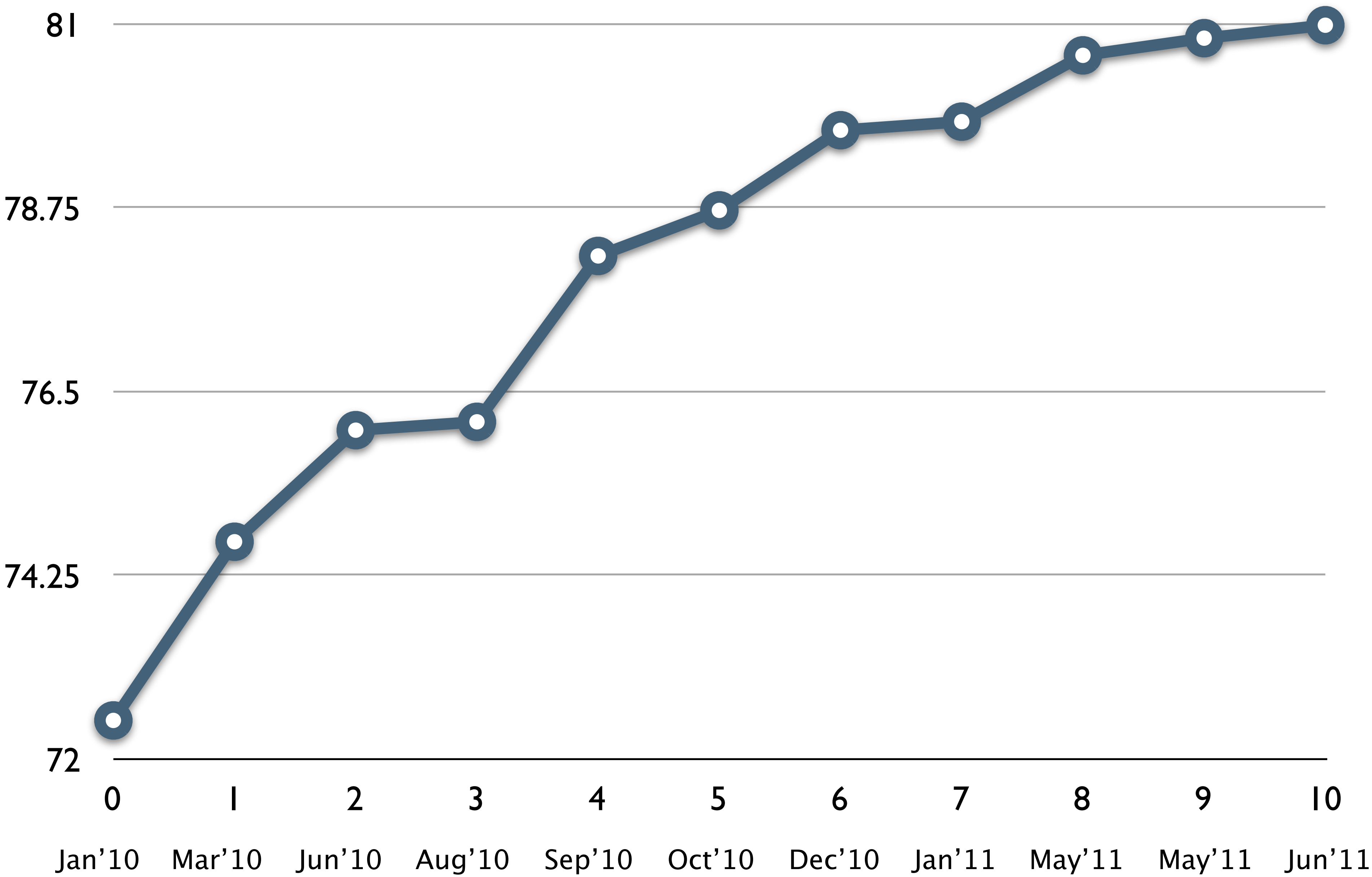
- Hand generated reordering data (English + Jenglish)
  - ~ 10k sentences for Augmented Loss training
  - ~ 6k evaluation sentences
- Score based on reordering penalty of METEOR

$$\text{reorder-score} = \frac{\#\ \text{chunks} - 1}{\#\ \text{unigrams matched} - 1}$$

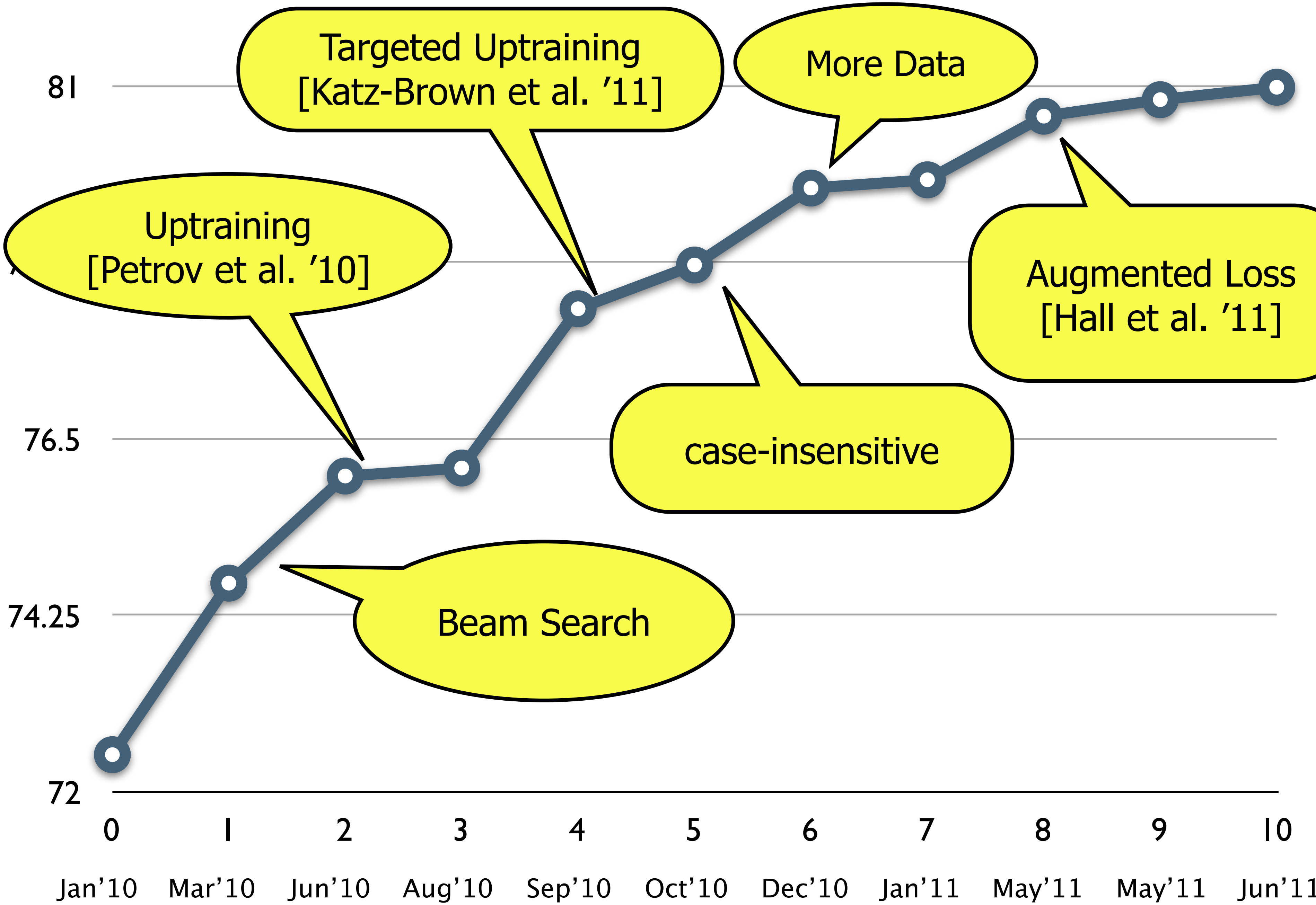
$$\text{reorder-loss} = 1 - \text{reorder-score}$$

- Very well correlated with human eval scores.

EnJa Fuzzy Reordering Evolution



EnJa Fuzzy Reordering Evolution



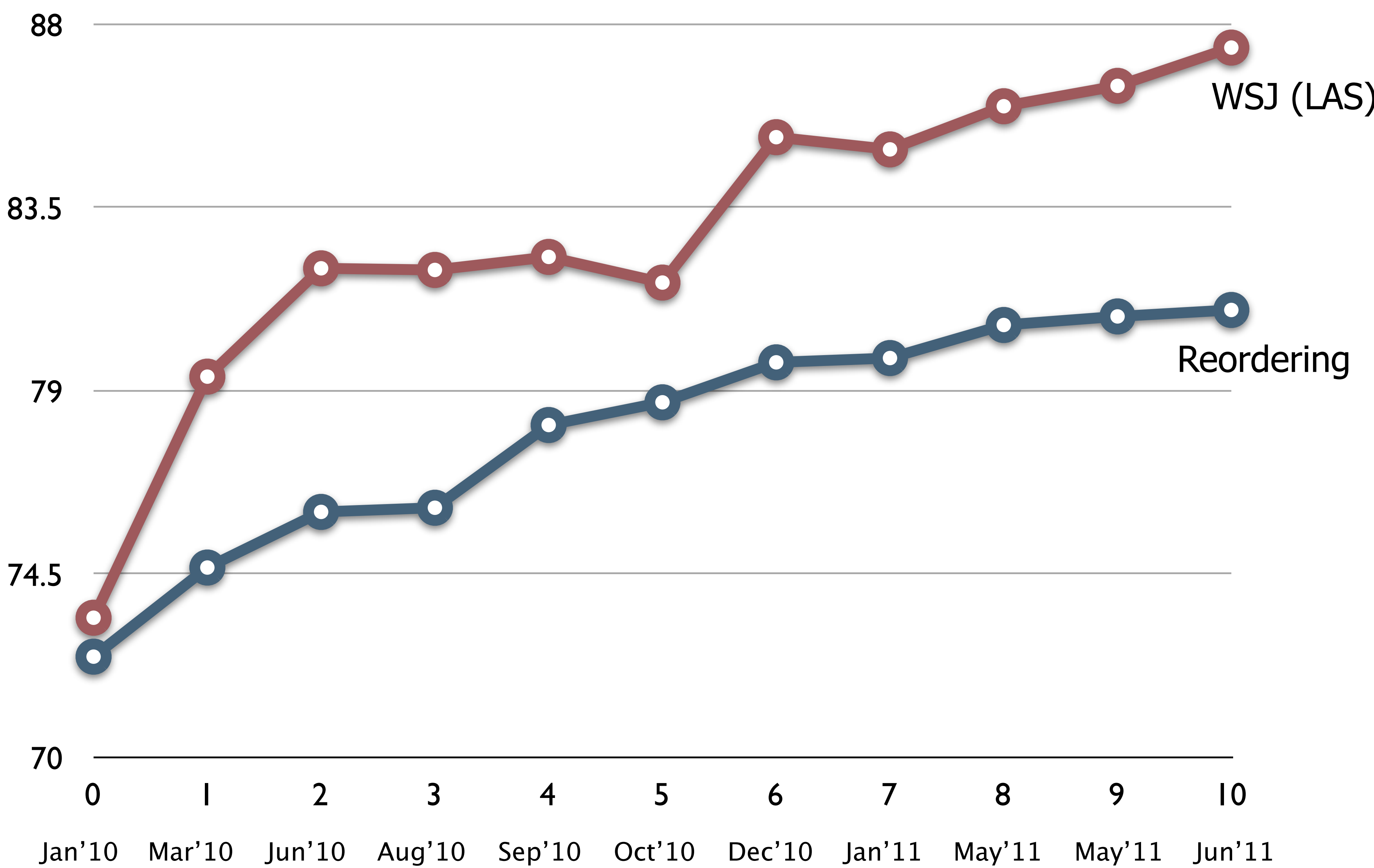
EnJa Fuzzy Reordering Evolution

WSJ (LAS)

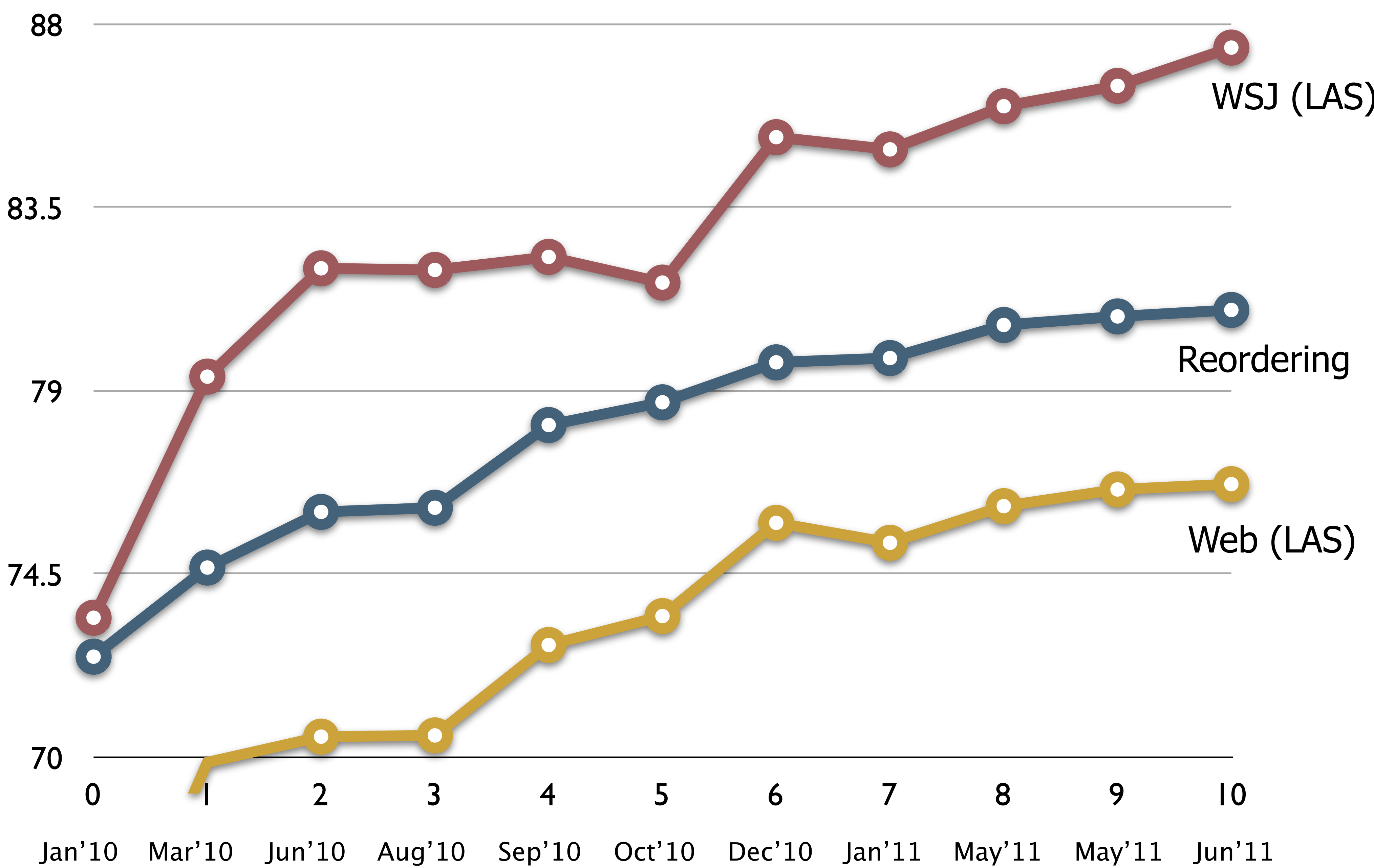
Reordering

Jan’10 Mar’10 Jun’10 Aug’10 Sep’10 Oct’10 Dec’10 Jan’11 May’11 May’11 Jun’11

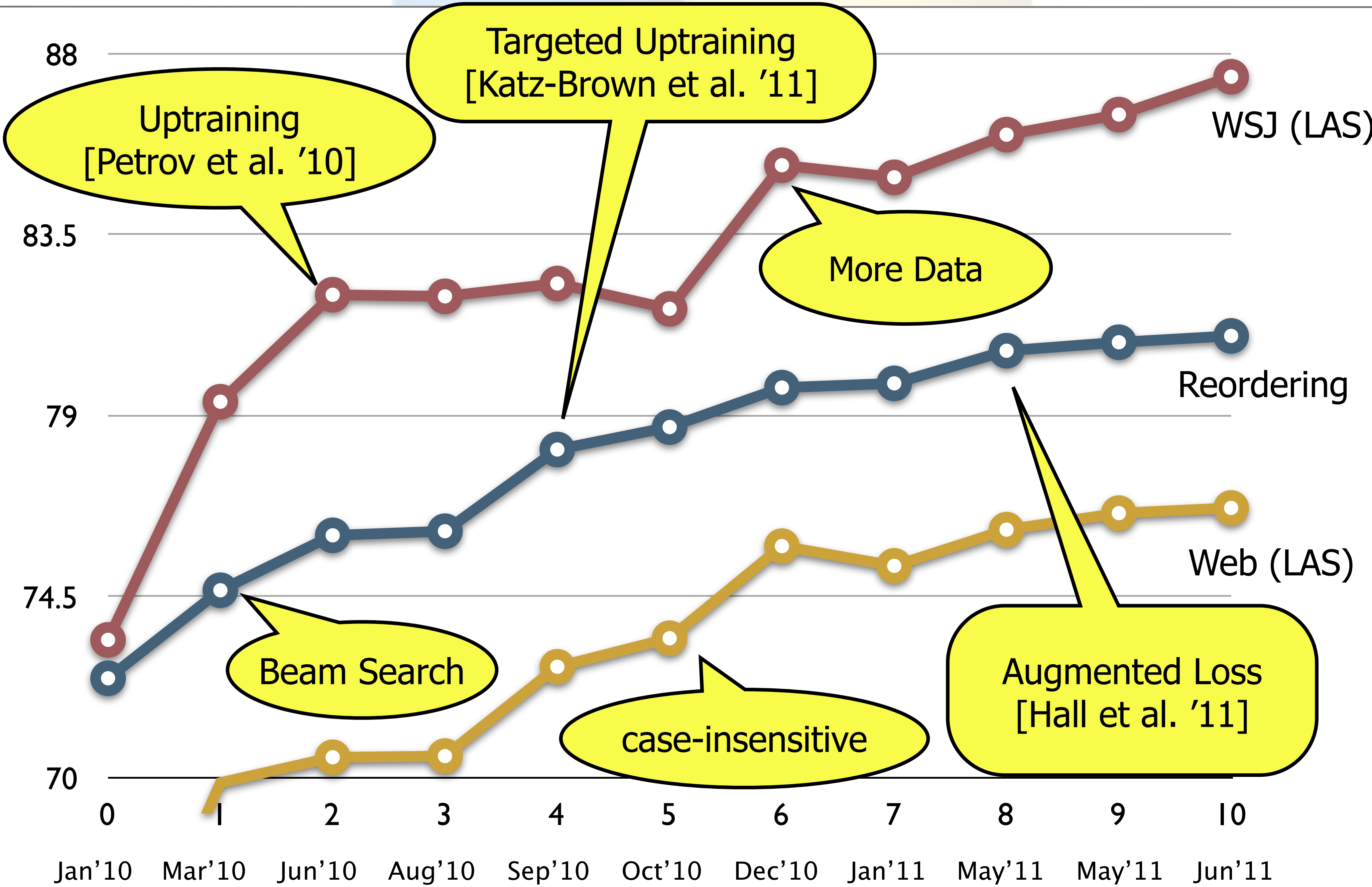
EnJa Fuzzy Reordering Evolution



EnJa Fuzzy Reordering Evolution



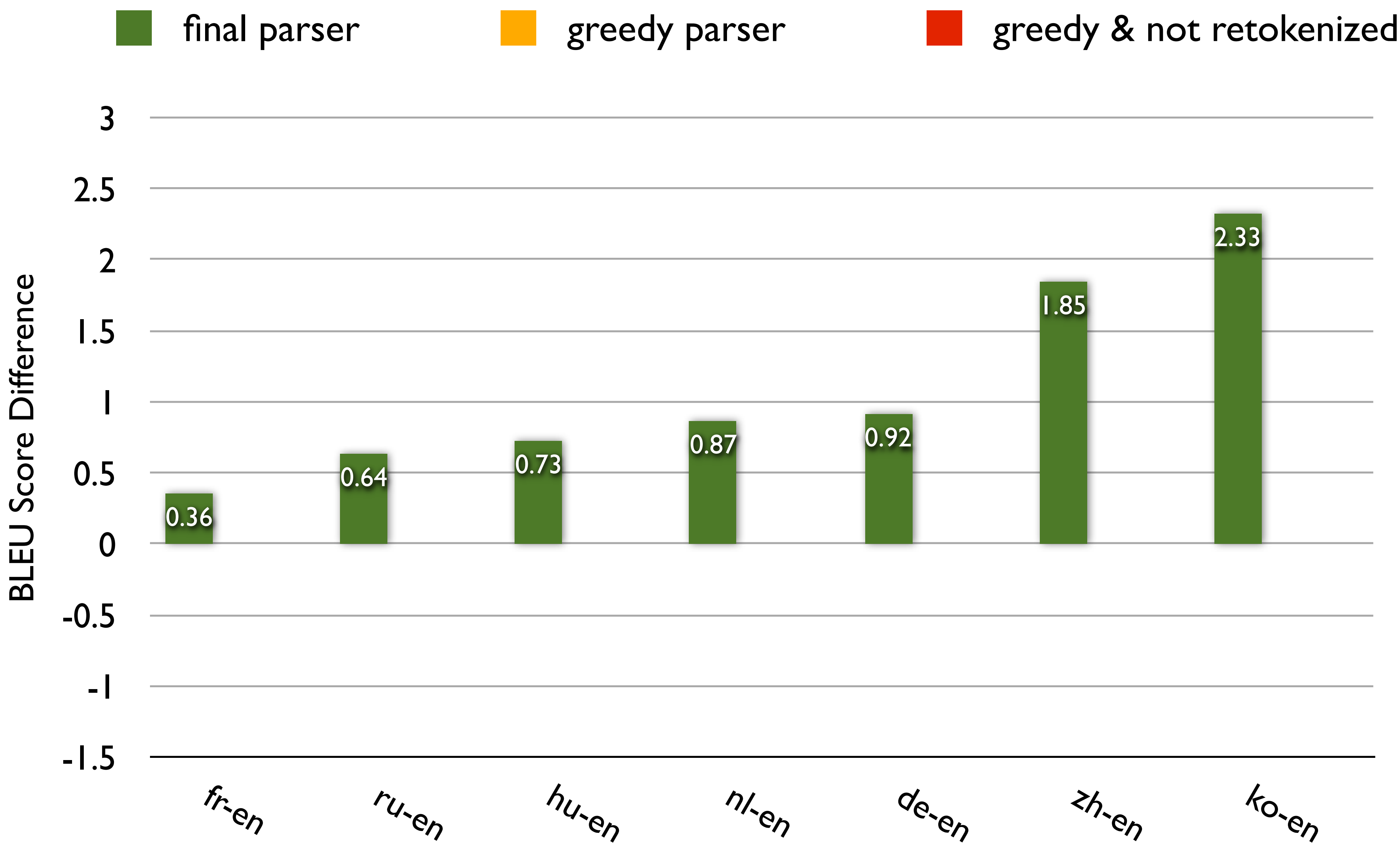
EnJa Fuzzy Reordering Evolution



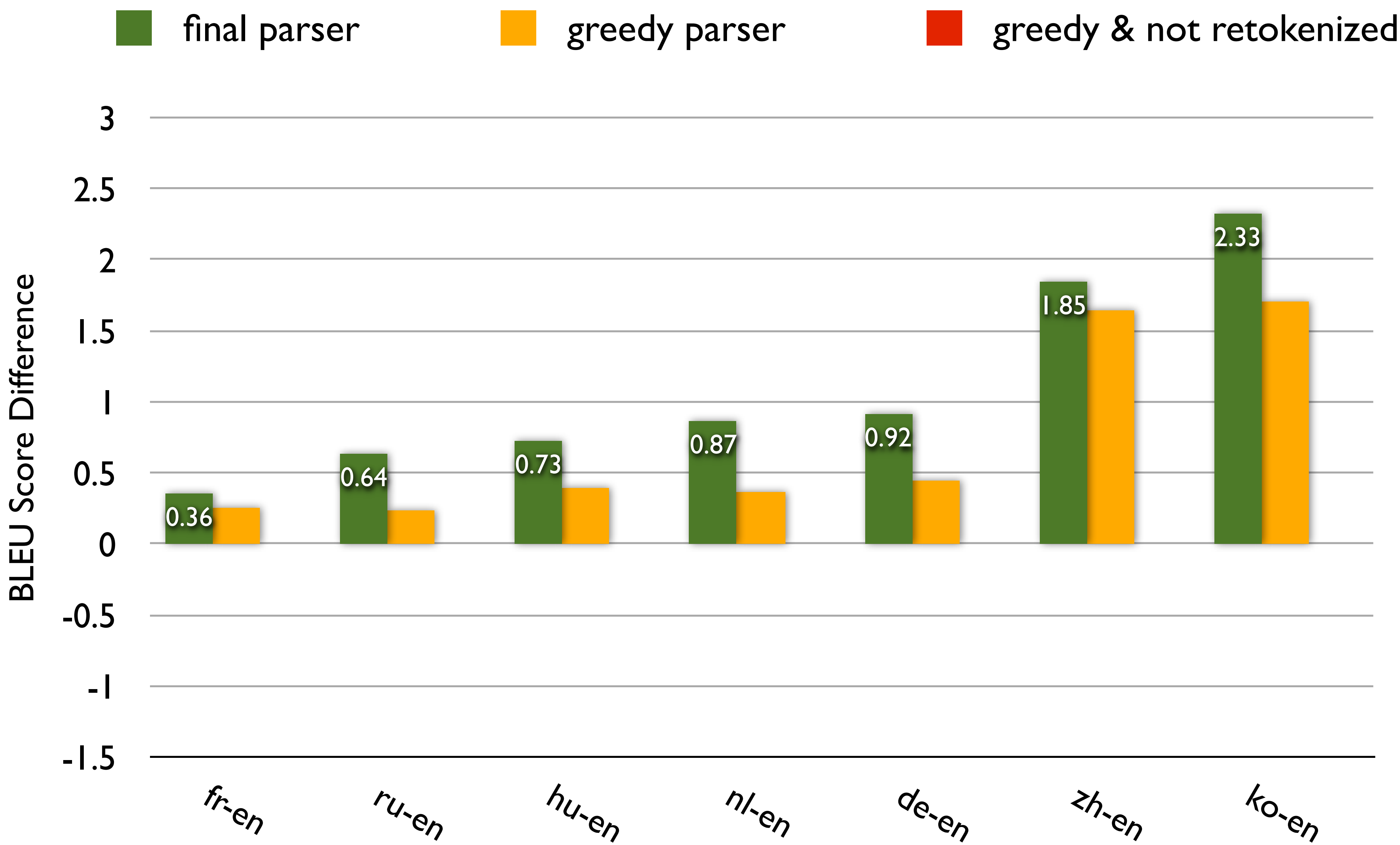
Retrospective Analysis

final parser   greedy parser   greedy & not retokenized

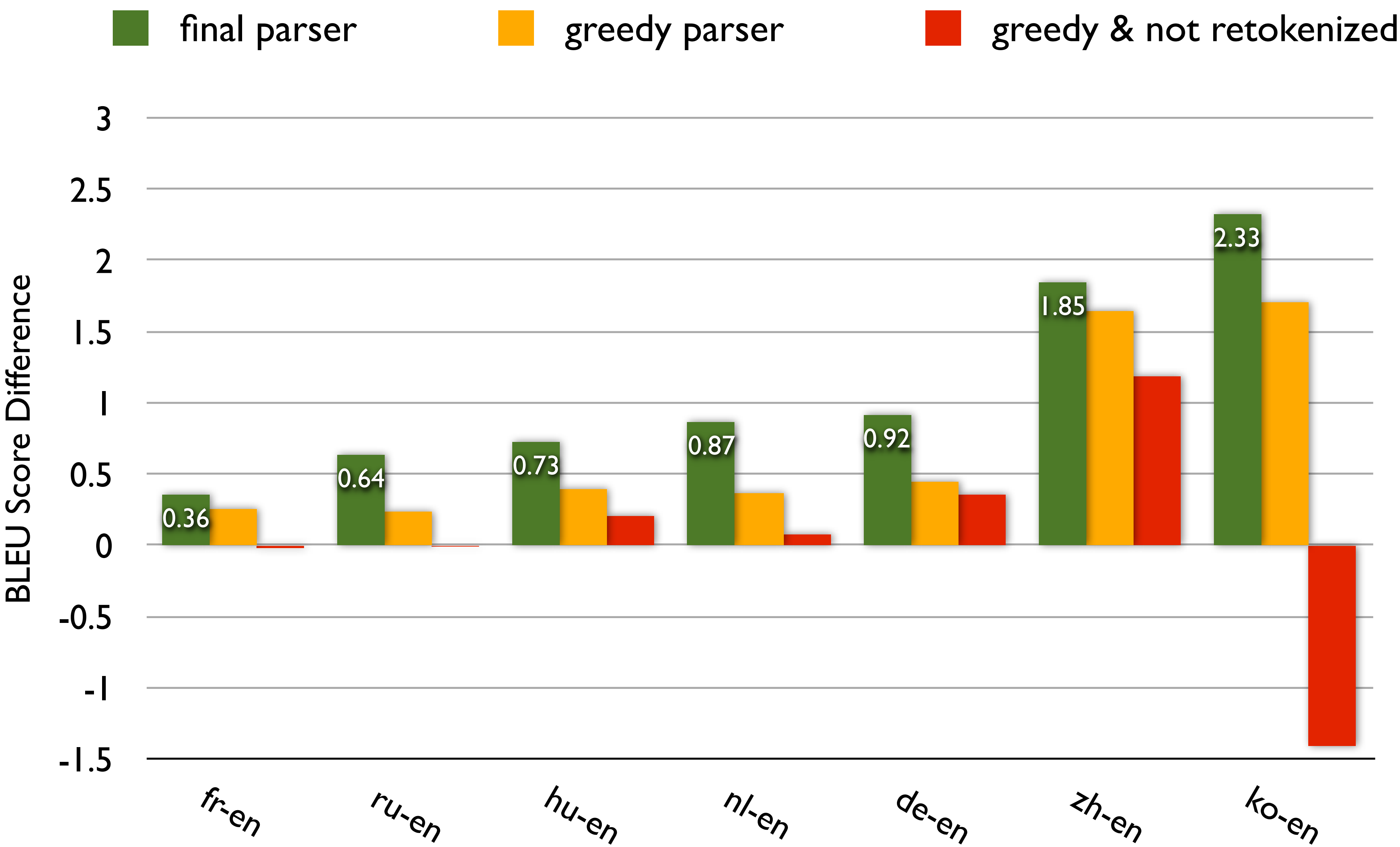
Retrospective Analysis



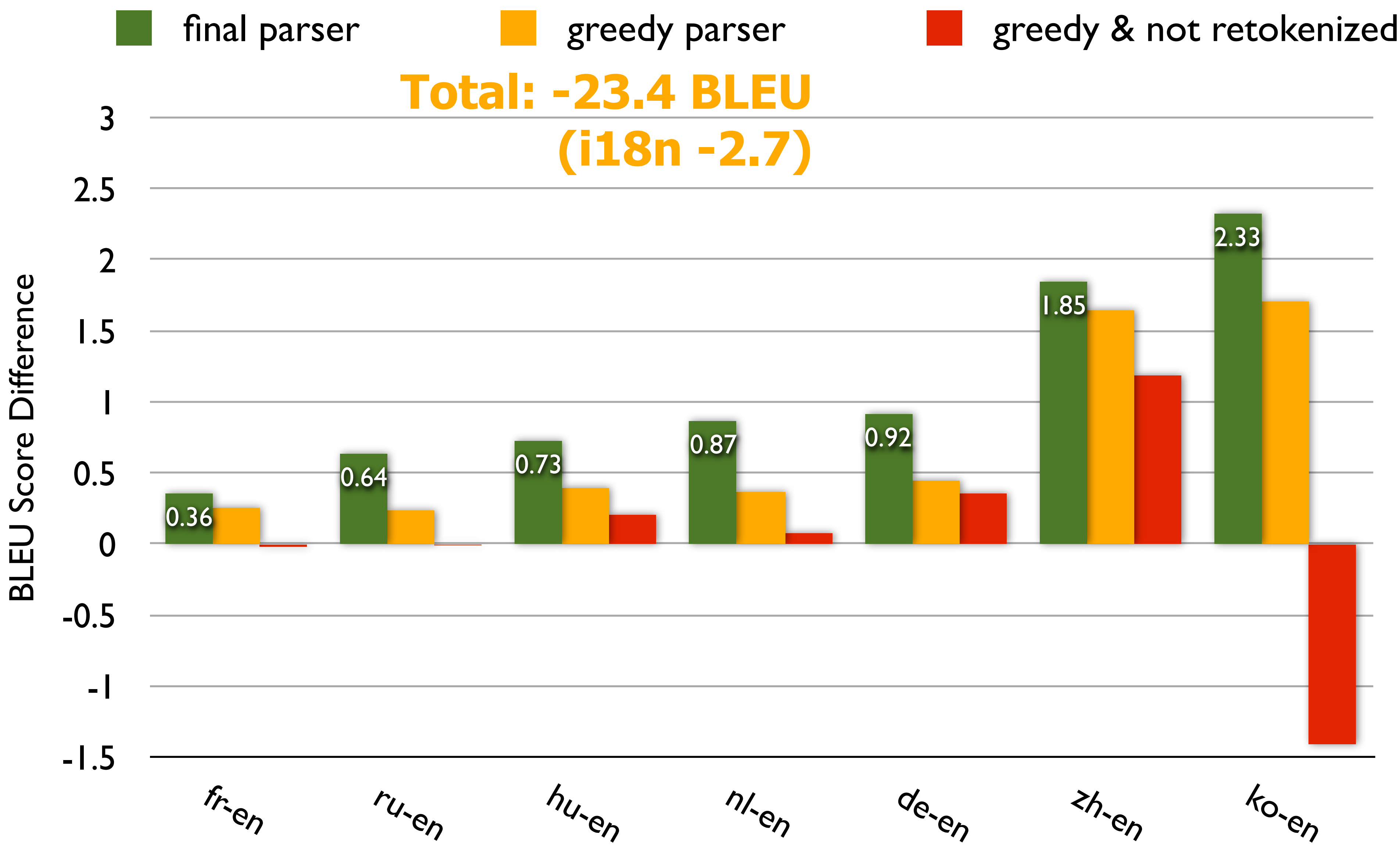
Retrospective Analysis



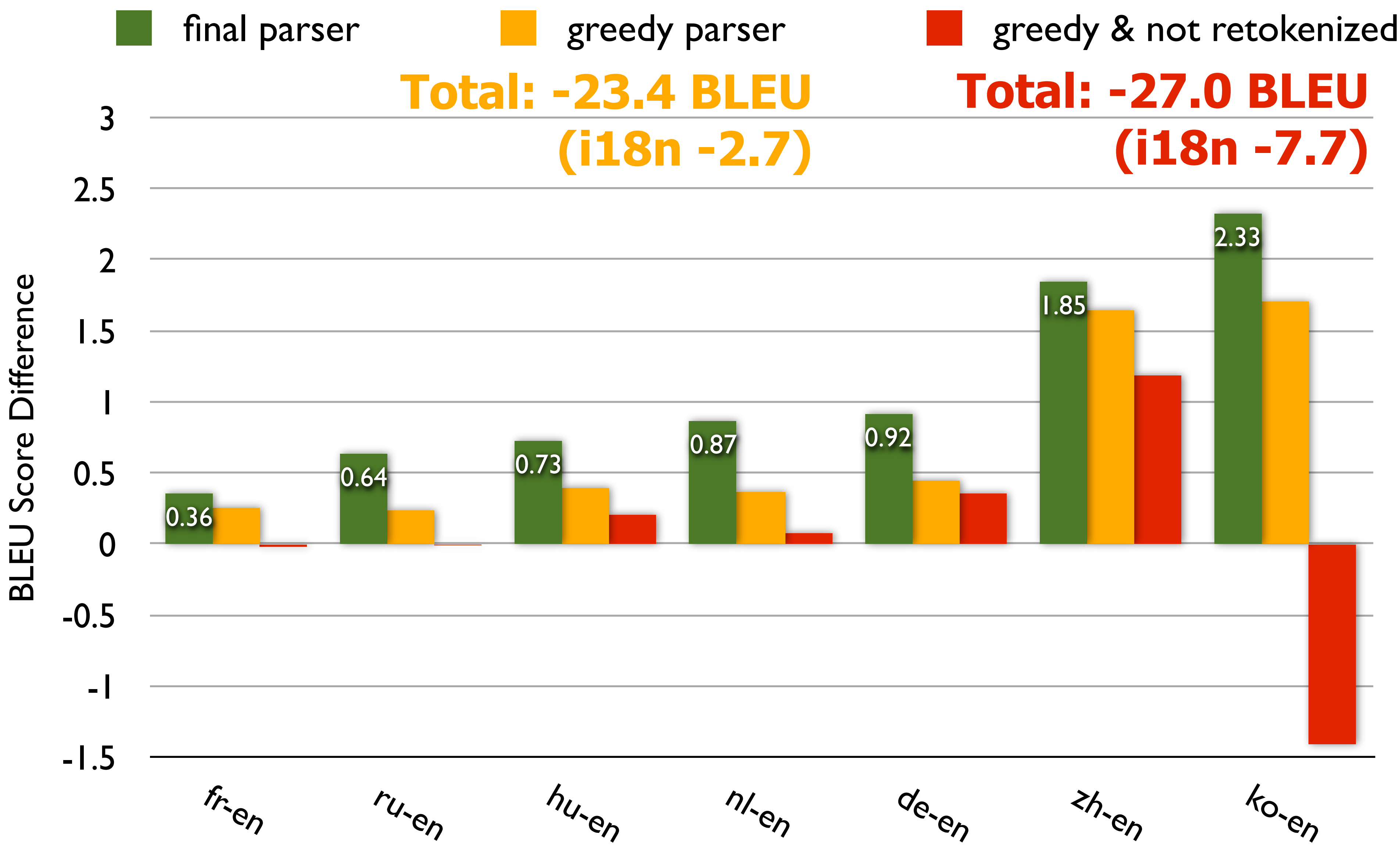
Retrospective Analysis



Retrospective Analysis



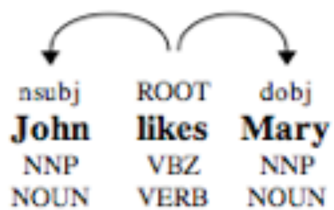
Retrospective Analysis

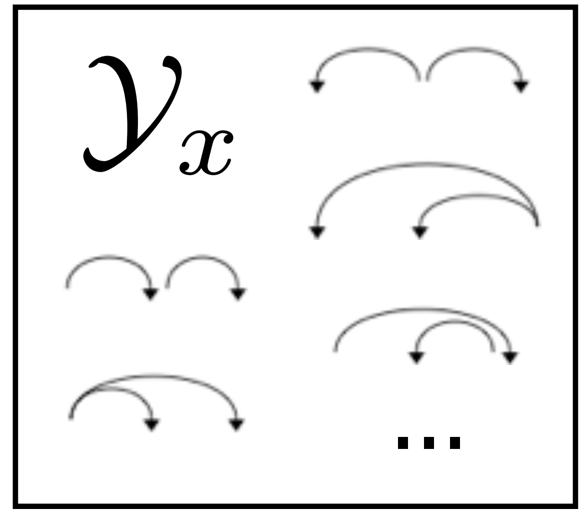


Standard Perceptron Training

Standard Perceptron

x = John likes Mary

y* = 



$$y = \arg\max_{y \in \mathcal{Y}_x} w \cdot \phi(y)$$

$$\theta = \theta + \phi(y^*) - \phi(y)$$

Augmented Loss (Online) Training

[Hall et al. '11]

Intrinsic Data

x_1, y_1

x_2, y_2

⋮

x_i, y_i

⋮

x_n, y_n $\quad y \in \mathcal{Y}$

Trainer

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss (Online) Training

[Hall et al. '11]

Intrinsic Data

⋮

Trainer

⋮

x_n, y_n $y \in \mathcal{Y}$

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss (Online) Training

[Hall et al. '11]

Intrinsic Data

x_1, y_1

x_2, y_2

$\vdots$

x_i, y_i

$\vdots$

x_n, y_n $\quad y \in \mathcal{Y}$

Trainer

Extrinsic Data

x'_1, y'_1

x'_2, y'_2

$\vdots$

x'_i, y'_i

$\vdots$

x'_n, y'_n $\quad y' \in \mathcal{Y}'$

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss (Online) Training

[Hall et al. '11]

Intrinsic Data

x_1, y_1
x_2, y_2
$\vdots$
x_i, y_i
$\vdots$
x_n, y_n $\quad y \in \mathcal{Y}$

Trainer

Extrinsic Data

x'_1, y'_1 $\quad l(\hat{y}_i, y'_i) \quad \hat{y} \in \mathcal{Y}$
x'_2, y'_2
$\vdots$
x'_i, y'_i
$\vdots$
x'_n, y'_n $\quad y' \in \mathcal{Y}'$

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss (Online) Training

[Hall et al. '11]

Intrinsic Data		Extrinsic Data	
x_1, y_1		x'_1, y'_1	$l(\hat{y}_i, y'_i)$ $\hat{y} \in \mathcal{Y}$
x_2, y_2	Trainer	x'_2, y'_2	
⋮		⋮	
x_i, y_i		x'_i, y'_i	
⋮		⋮	
x_n, y_n	$y \in \mathcal{Y}$	x'_n, y'_n	$y' \in \mathcal{Y}'$

$$\mathcal{Y} \neq \mathcal{Y}'$$

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss (Online) Training

[Hall et al. '11]

<u>Intrinsic Data</u>

⋮

x_i, y_i

⋮

x_n, y_n $\quad y \in \mathcal{Y}$

Trainer

<u>Extrinsic Data</u>

x'_1, y'_1 $\quad l(\hat{y}_i, y'_i)$ $\quad \hat{y} \in \mathcal{Y}$

x'_2, y'_2

⋮

x'_i, y'_i

⋮

x'_n, y'_n $\quad y' \in \mathcal{Y}'$

$$\mathcal{Y} \neq \mathcal{Y}'$$

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss (Online) Training

[Hall et al. '11]

Intrinsic Data

Trainer

⋮

⋮

x_n, y_n $\quad y \in \mathcal{Y}$

$\mathcal{Y} \neq \mathcal{Y}'$

Extrinsic Data

$l(\hat{y}_i, y'_i) \quad \hat{y} \in \mathcal{Y}$

x'_2, y'_2

⋮

x'_i, y'_i

⋮

x'_n, y'_n $\quad y' \in \mathcal{Y}'$

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss (Online) Training

[Hall et al. '11]

Intrinsic Data

⋮

⋮

x_n, y_n $\quad y \in \mathcal{Y}$

Extrinsic Data

$l(\hat{y}_i, y'_i) \quad \hat{y} \in \mathcal{Y}$

x'_2, y'_2

⋮

x'_i, y'_i

⋮

x'_n, y'_n $\quad y' \in \mathcal{Y}'$

Trainer

$$\mathcal{Y} \neq \mathcal{Y}'$$

Intrinsic: supervised training data/objective

Extrinsic: downstream task which "uses" outputs of model

Similar to [Chang et al. '08, McAllester et al. '10]

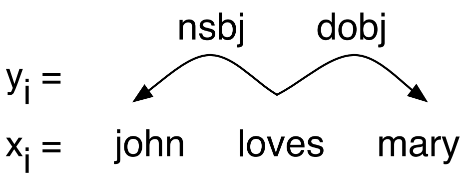
Augmented Loss Perceptron

x_1, y_1

x_2, y_2

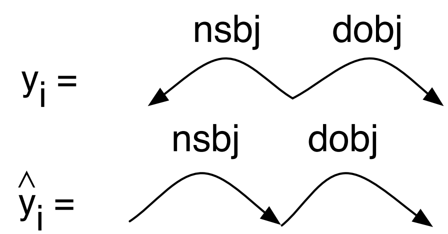
$\vdots$

x_i, y_i



$\vdots$

x_n, y_n



$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$

$$\text{if } y_i \neq \hat{y}_i \text{ then}$$

$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

Similar to [Chang et al. ’08, McAllester et al. ’10]

Augmented Loss Perceptron

x_1, y_1

x_2, y_2

$\vdots$

x_i, y_i

$\vdots$

x_n, y_n

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$

if $y_i \neq \hat{y}_i$ then

$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss Perceptron

x_1,y_1

x_2,y_2

⋮

x_i,y_i

⋮

x_n,y_n

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$

if $y_i \neq \hat{y}_i$ then

$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

Extrinsic Data

x'_1,y'_1

x'_2,y'_2

⋮

x'_i,y'_i

⋮

x'_n,y'_n

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss Perceptron

x_1,y_1

x_2,y_2

⋮

x_i,y_i

⋮

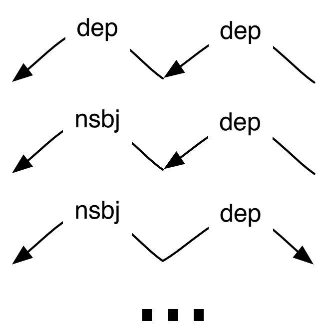
x_n,y_n

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$

if $y_i \neq \hat{y}_i$ then

$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

Parse(x'_i){



<u>Extrinsic Data</u>

x'_1,y'_1

x'_2,y'_2

⋮

⋮

x'_n,y'_n

Similar to [Chang et al. ’08, McAllester et al. ’10]

Augmented Loss Perceptron

x_1, y_1

x_2, y_2

⋮

x_i, y_i

⋮

x_n, y_n

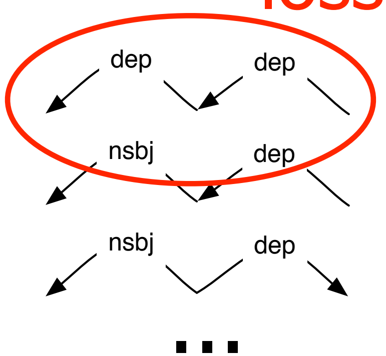
$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$

if $y_i \neq \hat{y}_i$ then

$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

$\text{loss}(y'_1, \hat{y}_1)$

$\text{Parse}(x'_i)\{$



Extrinsic Data

x'_1, y'_1

x'_2, y'_2

⋮

⋮

x'_n, y'_n

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss Perceptron

x_1, y_1

x_2, y_2

⋮

x_i, y_i

⋮

x_n, y_n

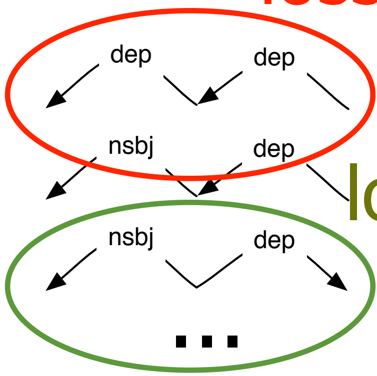
$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$

if $y_i \neq \hat{y}_i$ then

$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

loss(y‘$_1$,ŷ$_1$)

Parse(x’$_i$){



loss(y’$_1$,y$_1$*)

Extrinsic Data

x'_1, y'_1

x'_2, y'_2

⋮

⋮

x'_n, y'_n

Similar to [Chang et al. ’08, McAllester et al. ’10]

Augmented Loss Perceptron

x_1, y_1
x_2, y_2
$\vdots$
x_i, y_i
$\vdots$
x_n, y_n

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$
if $y_i \neq \hat{y}_i$ then
$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_{x'_i}} \theta \cdot \phi(x'_i, y)$$
$$y^*_i = \arg\min_{y \in \text{k-best-parse}(x'_i)} \text{loss}(y, y'_i)$$
if $\text{loss}(y^*_i, y'_i) < \text{loss}(\hat{y}_i, y'_i)$ then
$$\theta^t = \theta^{t-1} + \phi(y^*_i) - \phi(\hat{y}_i)$$

Extrinsic Data

x'_1, y'_1
x'_2, y'_2
$\vdots$
$\vdots$
x'_n, y'_n

Similar to [Chang et al. '08, McAllester et al. '10]

Augmented Loss Perceptron

x_1, y_1
x_2, y_2
$\vdots$
x_i, y_i
$\vdots$
x_n, y_n

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_x} \theta \cdot \phi(x_i, y)$$
$$\text{if } y_i \neq \hat{y}_i \text{ then}$$
$$\theta^t = \theta^{t-1} + \phi(y_i) - \phi(\hat{y}_i)$$

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}_{x'_i}} \theta \cdot \phi(x'_i, y)$$
$$y^*_i = \arg\min_{y \in \text{k-best-parse}(x'_i)} \text{loss}(y, y'_i)$$
$$\text{if loss}(y^*_i, y'_i) < \text{loss}(\hat{y}_i, y'_i) \text{then}$$
$$\theta^t = \theta^{t-1} + \phi(y^*_i) - \phi(\hat{y}_i)$$

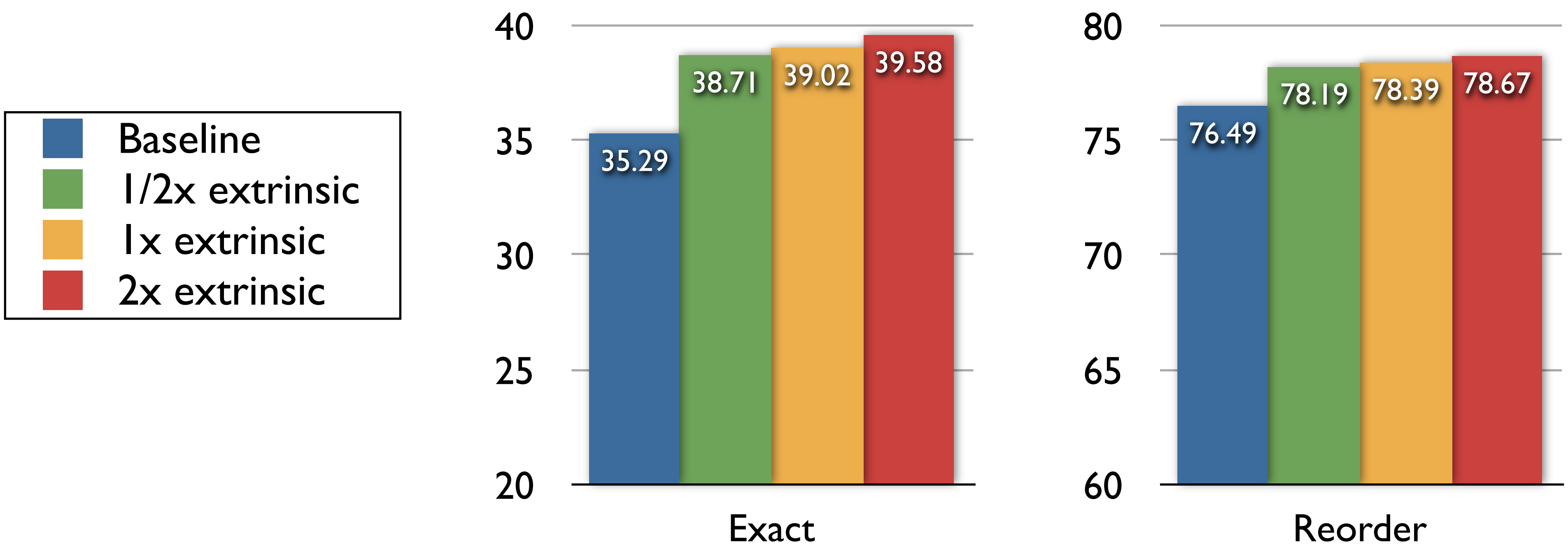
Extrinsic Data

x'_1, y'_1
x'_2, y'_2
$\vdots$
$\vdots$
x'_n, y'_n

Similar to [Chang et al. '08, McAllester et al. '10]

Example MT-Reordering

<u>"Clean" experiment</u>



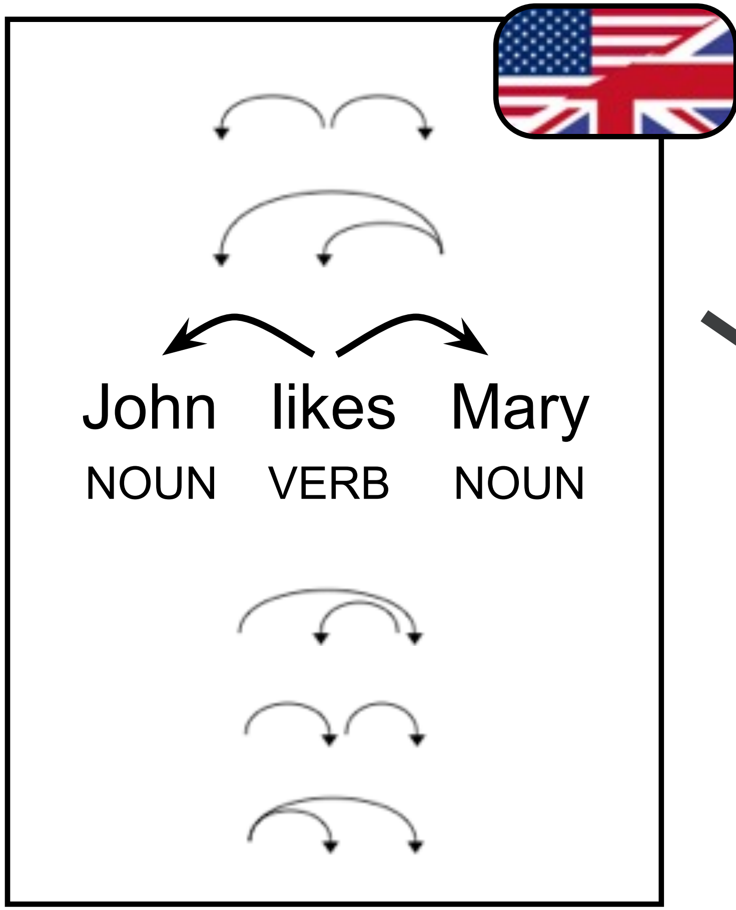
Production: 0.786 -> 0.792 enja fuzzy score

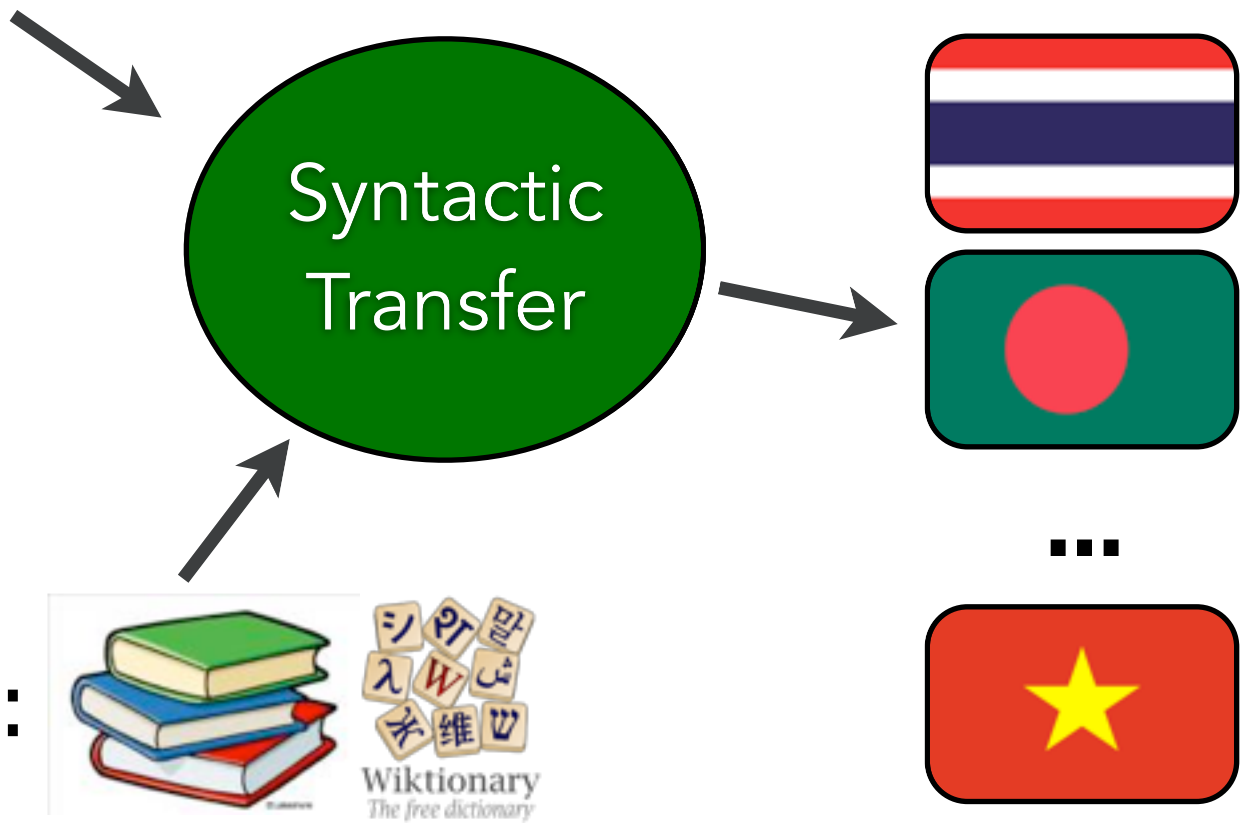
On top of targeted up/self-training [Katz-Brown et al. '11]

Syntactic Transfer

- Learn parsers for resource-poor languages from resource-rich languages
- Hwa et al. 2005 and earlier

English Treebank

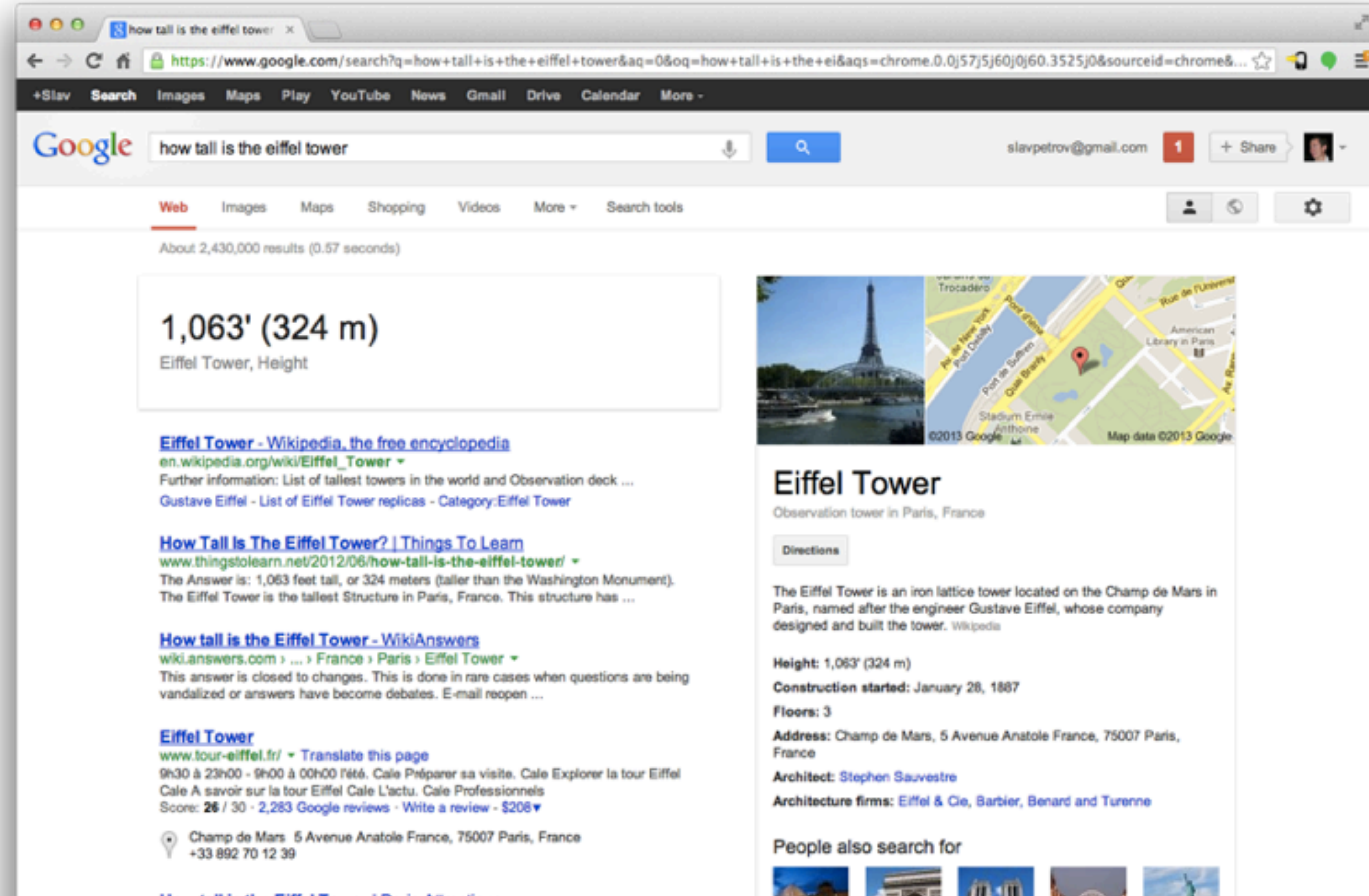




Dictionaries:

Confidence Estimation

Confidence Estimation



Confidence Estimation

- Sometimes it is possible to say: “I don’t know”
- Ten blue links are better than triggering an incorrect answer
- The (English) web is redundant:
  - Skip examples with low confidence predictions
  - Discount low-confidence contradicting predictions

In Summary

- Efficiency:
  - Exploit problem structure and domain knowledge
  - Train models specifically for pruning
- Adaptation:
  - Use unsupervised data
  - Use indirect signals
  - “Manual intervention”
- Confidence:
  - ?

Learning from Indirect Signals

Learning from Indirect Signals

Labeled Data

Learning from Indirect Signals

Labeled Data

Learning from Indirect Signals

Structured Data

Labeled Data

Unlabeled Data

Learning from Indirect Signals

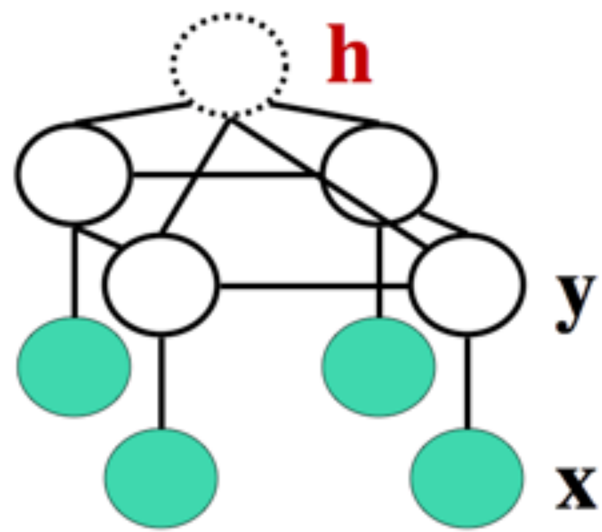
Structured Data

Wiktionary
The free dictionary

urban DICTIONARY

Labeled Data

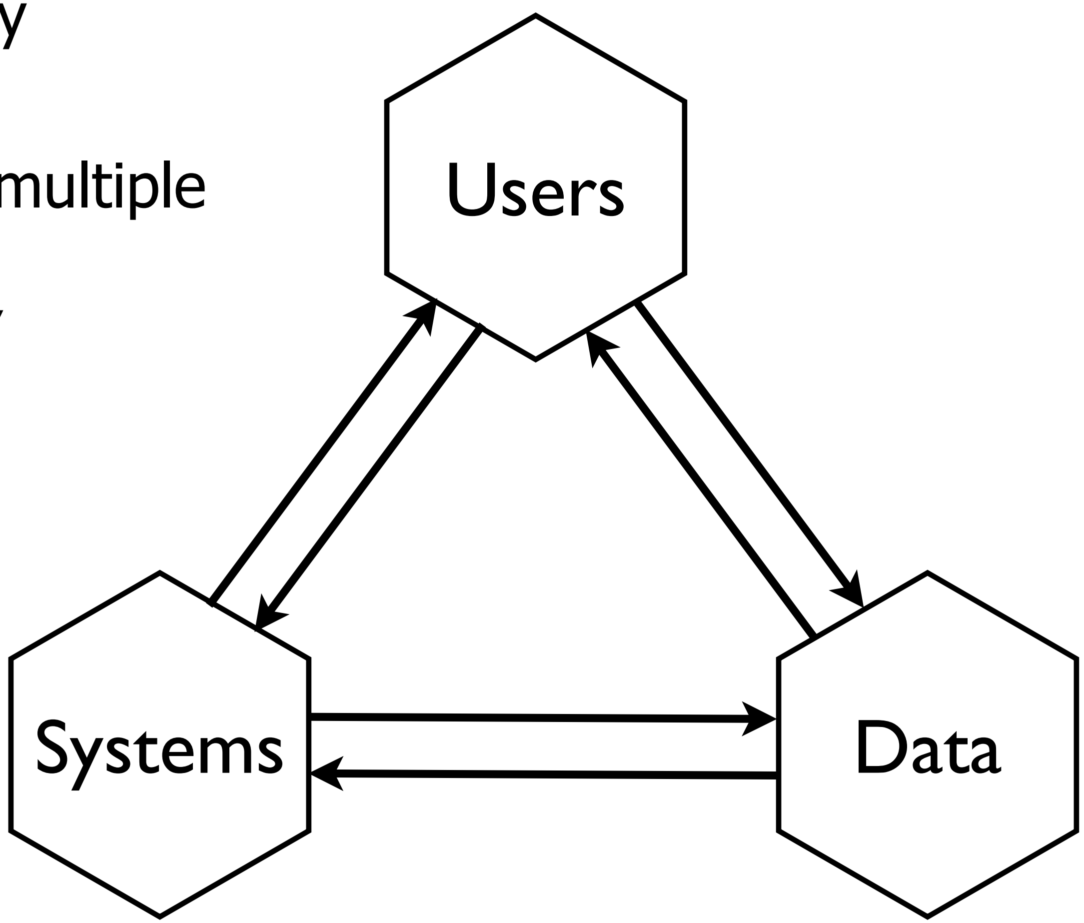
Unlabeled Data



Learning & Modeling

Hypothesis

- Understanding arises from machine learning of relationships implicit in web content and use
  - Some expert annotation may be needed to start
  - Most evidence is not explicitly annotated: text "in the wild"
  - Aggregate information from multiple unstructured sources into a structured "knowledge base"
  - Exploit user interactions and implicit user feedback



[Based on slide from Fernando Pereira]

Thank you!

Research at Google

slav@google.com