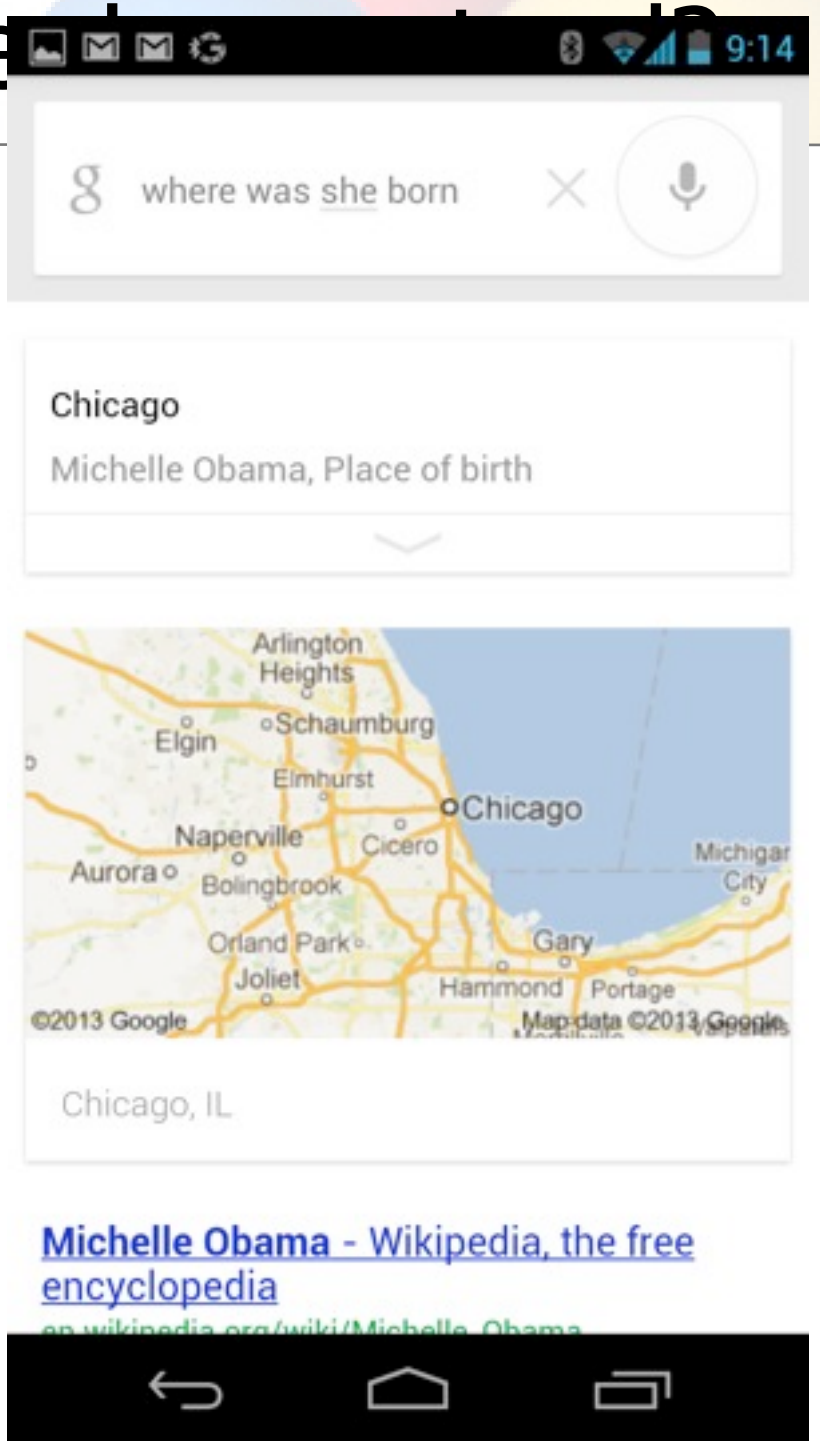
Understanding all of the World's Languages

Research at Google

Slav Petrov

Thanks to
Dipanjan Das, Ryan McDonald, Keith Hall, Fernando Pereira,
Kuzman Ganchev, Hao Zhang, Michael Ringgaard,
and many more

NLP: Where

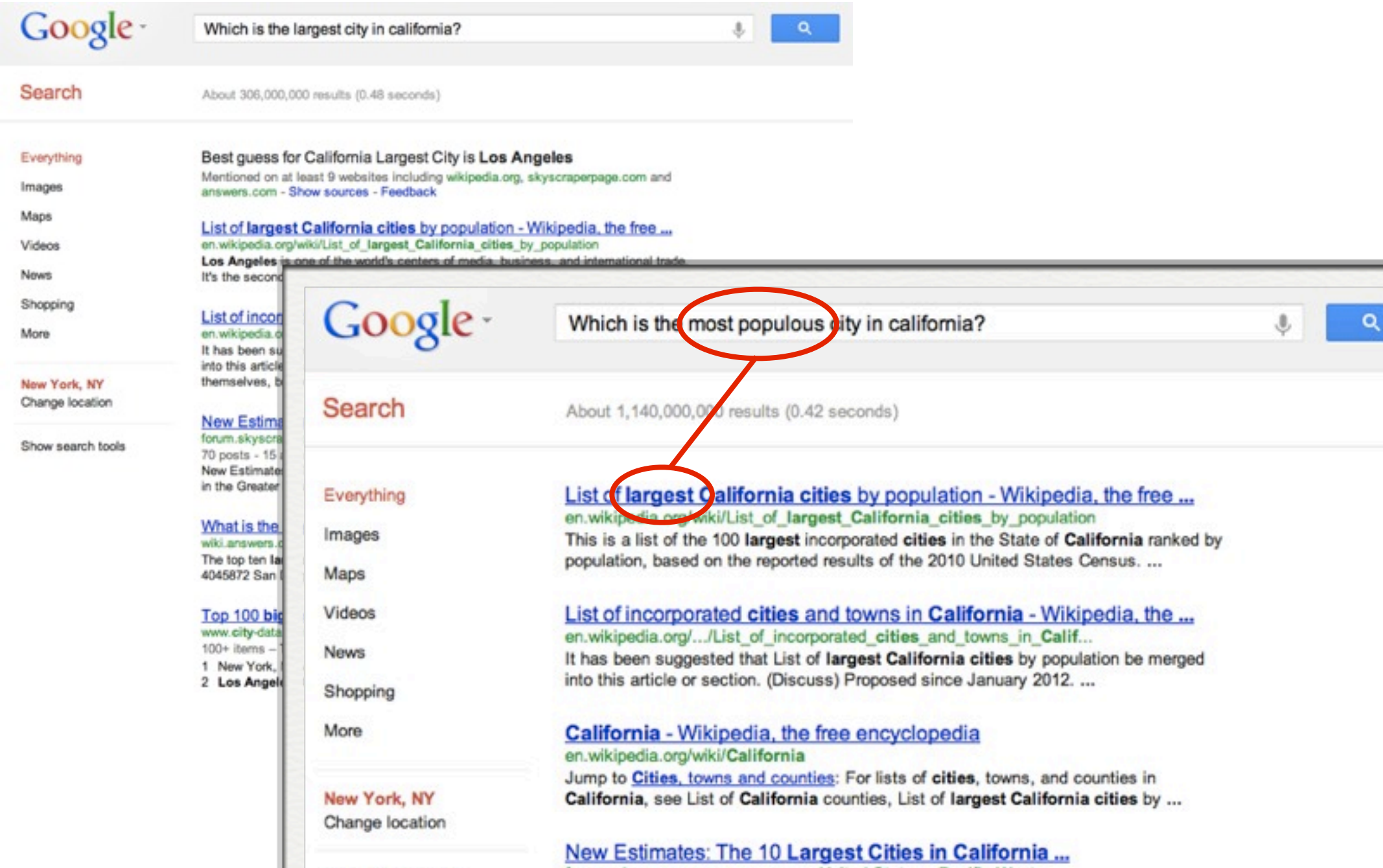

9:14
where was she born
Chicago
Michelle Obama, Place of birth
Arlington Heights
Schaumburg
Elgin
Elmhurst
Chicago
Naperville
Cicero
Aurora
Bolingbrook
Orland Park
Joliet
Gary
Hammond
Portage
©2013 Google
Map data ©2013 Google
Chicago, IL
Michelle Obama - Wikipedia, the free encyclopedia

Question Answering

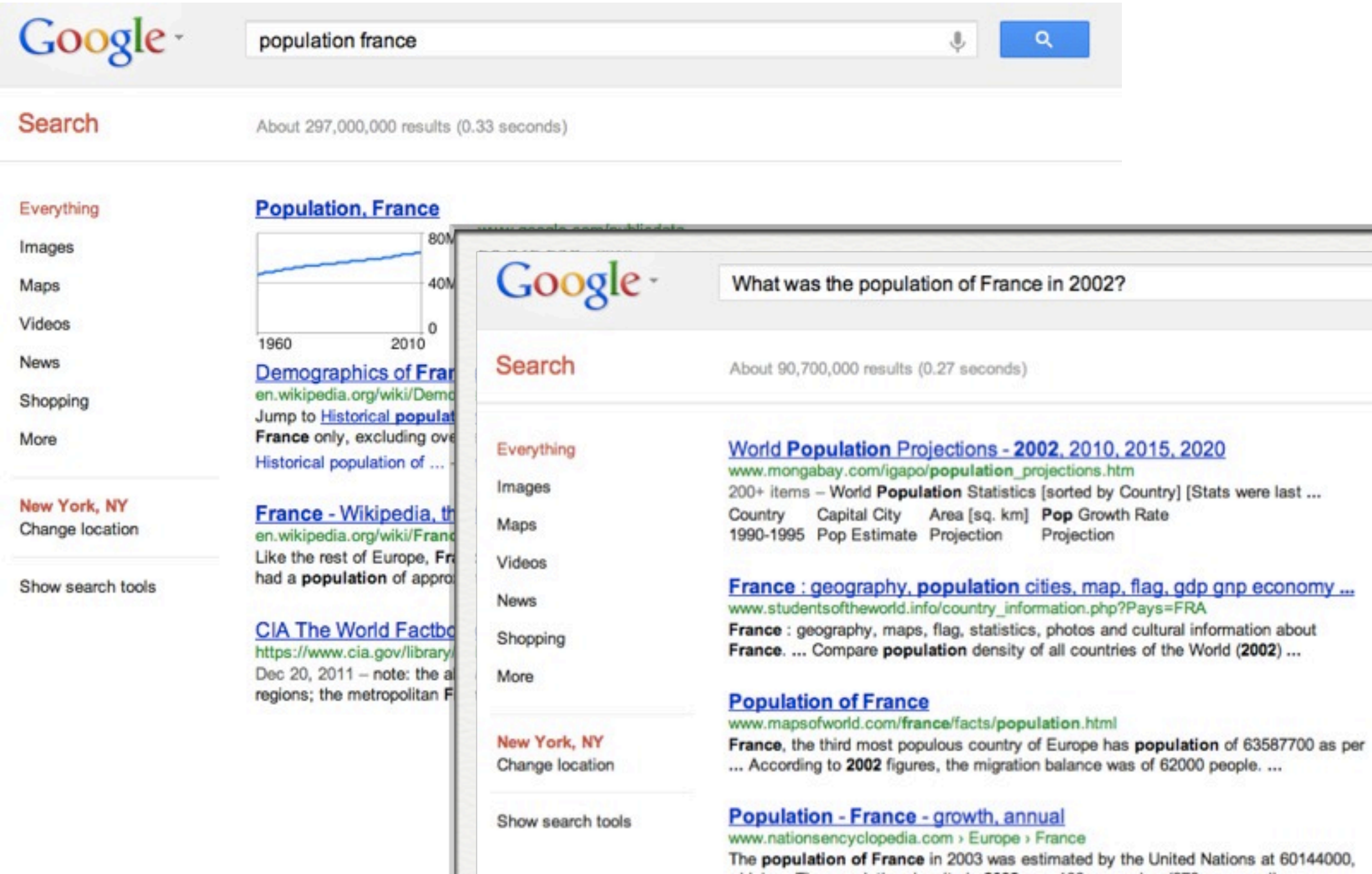
- More than search
- Ask general comprehension questions of a document collection
- Can be really easy: "What's the capital of Wyoming?"
- Can be harder: "Which painkillers don't upset the stomach?"
- Can be open ended: "What are the main issues in the global warming debate?"

In 2012:

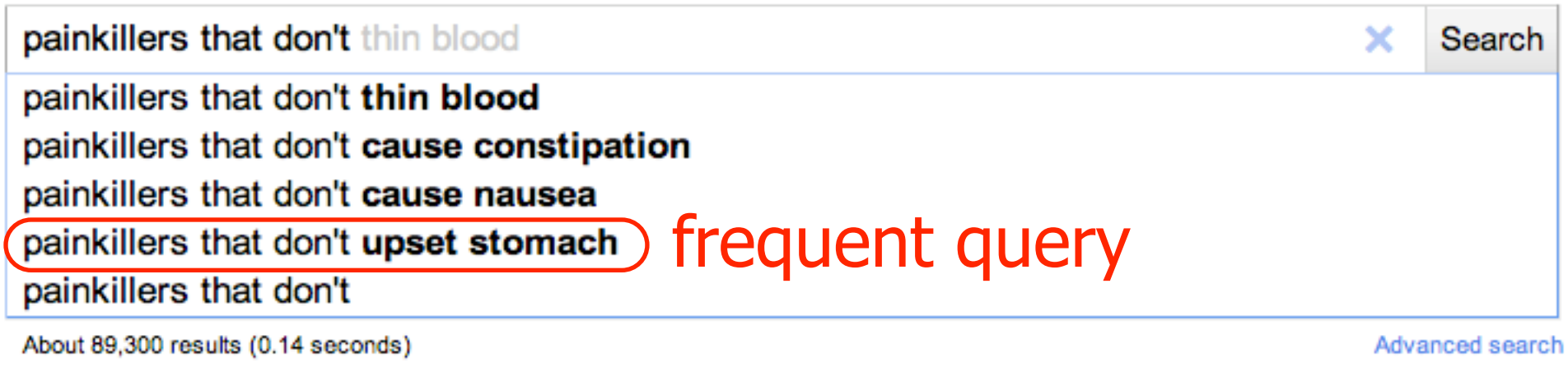
Question Answering



No Understanding



A “more involved” query…



… that fails

class negation synonyms

painkillers that don't upset stomach

Search

About 29,700 results (0.15 seconds) Advanced search

Sponsored link

Symptoms of Crohn's
www.crohnsonline.com Do You Know the Signs of Crohn's? Read About Symptoms and More Now.

Best Pain Reliever, Over the Counter Pain Relievers, Prescription ...
For **stomach upset**, eating or drinking milk before you take a painkiller can help. ... **Don't** expect **painkillers** to relieve all your pain, but the best pain ...
www.healthblurbs.com/best-pain-reliever-over-the-counter-pain-relievers-prescription-analgesic-**painkillers**/ - Cached - Similar

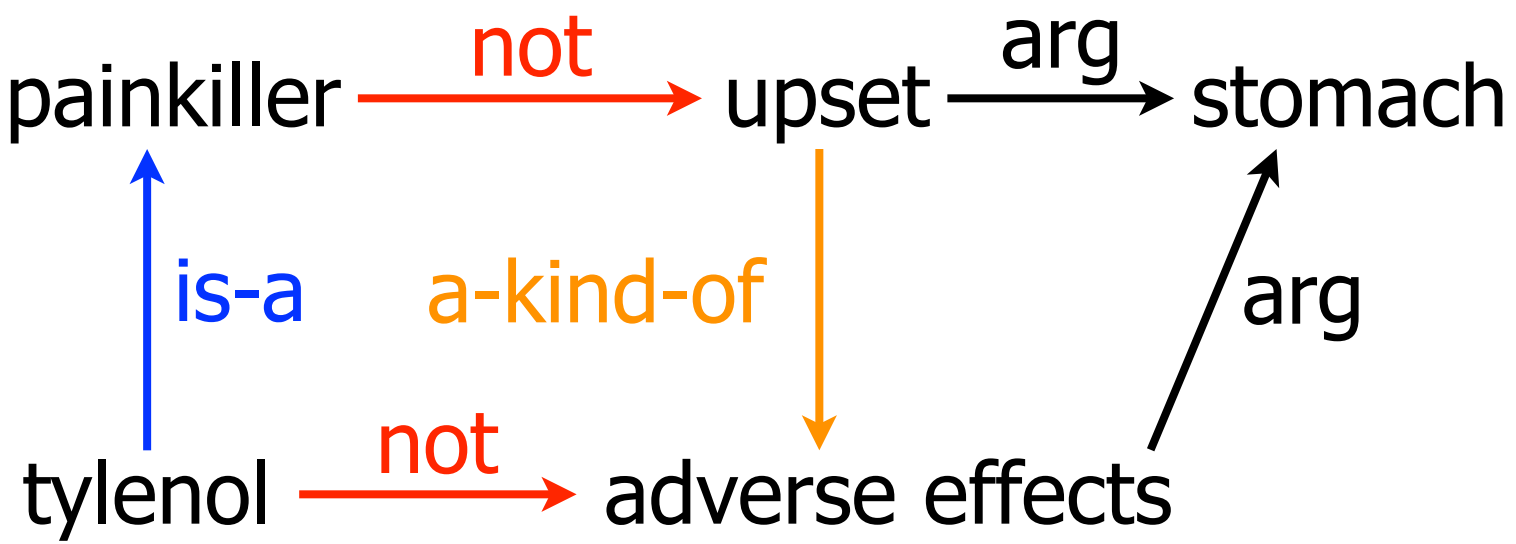
Why do painkillers make your stomach hurt? - Yahoo! Answers
Aug 30, 2008 ... **stomach upset**, reflux/indigestion, or if severe ulcers ... I often take **painkillers** FOR **stomach** aches lol. **Don't** do this again. 2 years ago ...
answers.yahoo.com › Health › Other - Health - Cached - Similar

- How are **painkillers** bad for you? - Apr 13, 2010
- What effect do **painkillers** have on you? - Sep 30, 2009
- What's the difference between OTC **painkillers**? (Advil, Aleeve ... - Nov 19, 2008
- Is it bad to take **painkillers** on an empty **stomach** ? - Aug 20, 2008

More results from answers.yahoo.com »

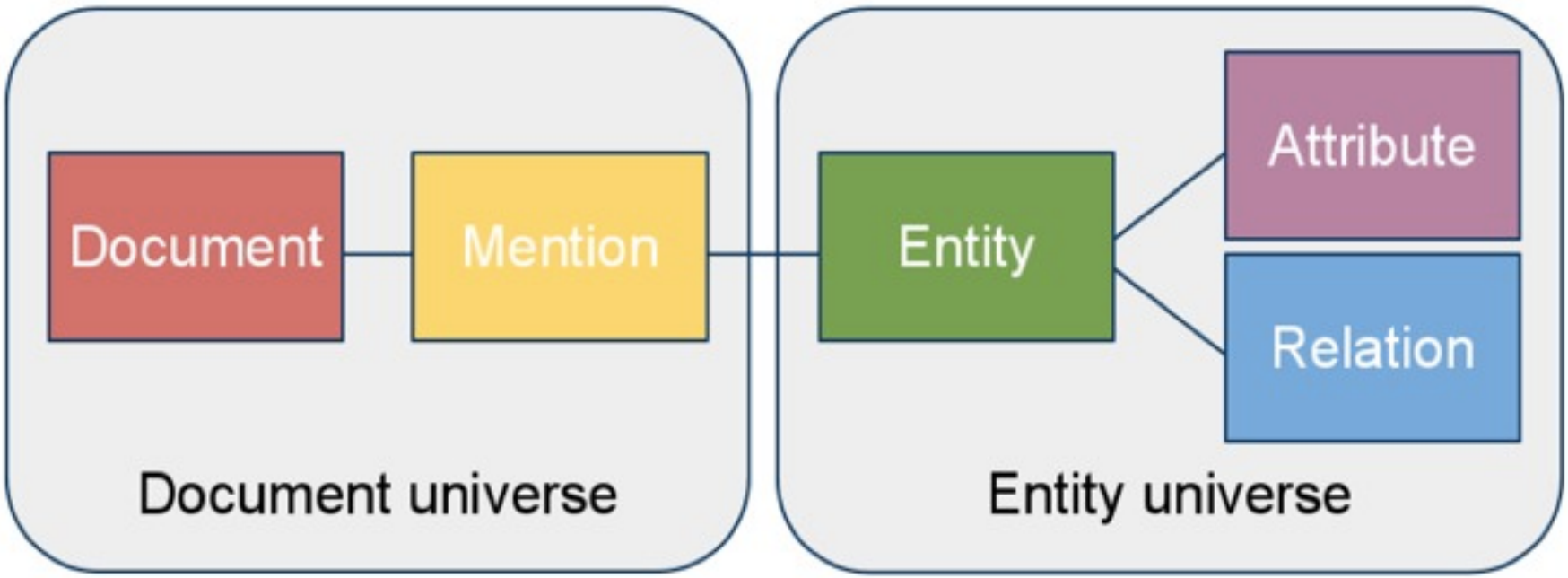
Phys Ed: Does Ibuprofen Help or Hurt During Exercise? - NYTimes.com
Sep 1, 2009 ... When are ibuprofen and other anti-inflammatory **painkillers** justified? ... Can"t take NSAIDS due to **upset stomach**. — bonemri ... Perhaps part of the problem is that you **don't** always have the option to rest your injury if ...
well.blogs.nytimes.com/.../phys-ed-does-ibuprofen-help-or-hurt-during-exercise/ - Cached - Similar

From Search to Inference



- The answers are there for search to find
- … if we can find and combine reliably
- … all the relevant evidence

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

$_s$ ⟨President⟩ $_1$ [Barack Obama]$_1$ has been re-elected to a second ⟨term⟩ $_2$, defeating [Republican]$_4$ ⟨challenger⟩ $_3$ [Mitt Romney]$_3$. $_s$ [America]$_6$'s first black ⟨president⟩ $_5$ secured more than the 270 ⟨votes⟩ $_7$ in the ⟨electoral college⟩ $_8$ needed to win. $_s$ 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". $_s$ ⟨Mr⟩ $_1$ [Obama]$_1$ prevailed despite lingering ⟨dissatisfaction⟩ $_{13}$ with the ⟨economy⟩ $_{14}$ and a hard-fought ⟨challenge⟩ $_{15}$ by ⟨Mr⟩ $_3$ [Romney]$_3$.

$_s$ {His}$_1$ [Democrats]$_{16}$ also retained {their}$_{16}$ ⟨majority⟩ $_{17}$ in the [Senate]$_{18}$, {which}$_{18}$ {they}$_{16}$ have held since 2007. $_s$ 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⟩ $_{27}$.

$_s$ In {his}$_{28}$ ⟨address⟩ $_{29}$, the ⟨president⟩ $_{28}$ challenged {his}$_{28}$ ⟨opponents⟩ $_{30}$, asking {them}$_{30}$ to work with {him}$_{28}$. $_s$ 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. $_s$ The popular ⟨vote⟩ $_{34}$, {which}$_{34}$ is symbolically and politically important but not decisive in the ⟨race⟩ $_{35}$, remains very close.

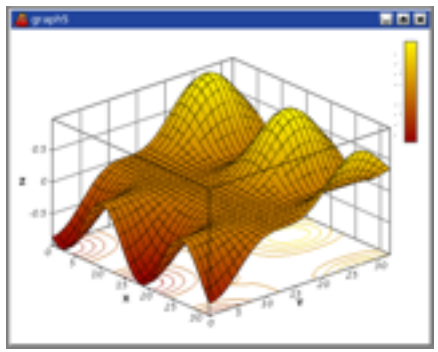
$_s$ ⟨Mr⟩ $_1$ [Obama]$_1$ congratulated ⟨Mr⟩ $_3$ [Romney]$_3$ and [Republican]$_4$ ⟨vice-presidential candidate⟩ $_{36}$ [Paul Ryan]$_{36}$ on {their}$_{33}$ hard-fought ⟨campaign⟩ $_{37}$. $_s$ "{We}$_{33}$ have picked {ourselves}$_{33}$ up, {we}$_{33}$ have fought {our}$_{33}$ ⟨way⟩ $_{38}$ back 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. $_s$ ⟨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". $_s$ {He}$_1$ pledged to work with [Republican]$_4$ ⟨leaders⟩ $_{44}$ in [Congress]$_{45}$ to reduce the

Traditional Supervised NLP

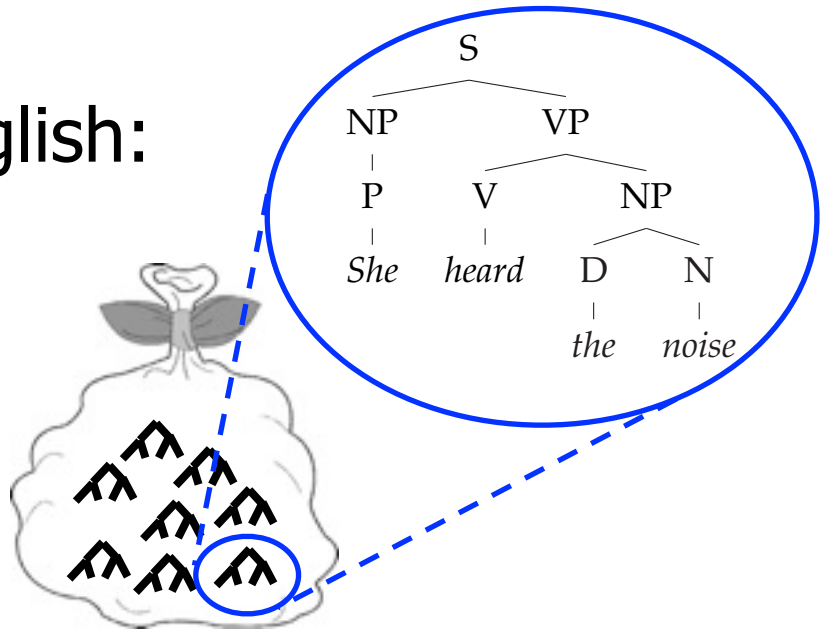
New Task

Linguists

Machine Learning



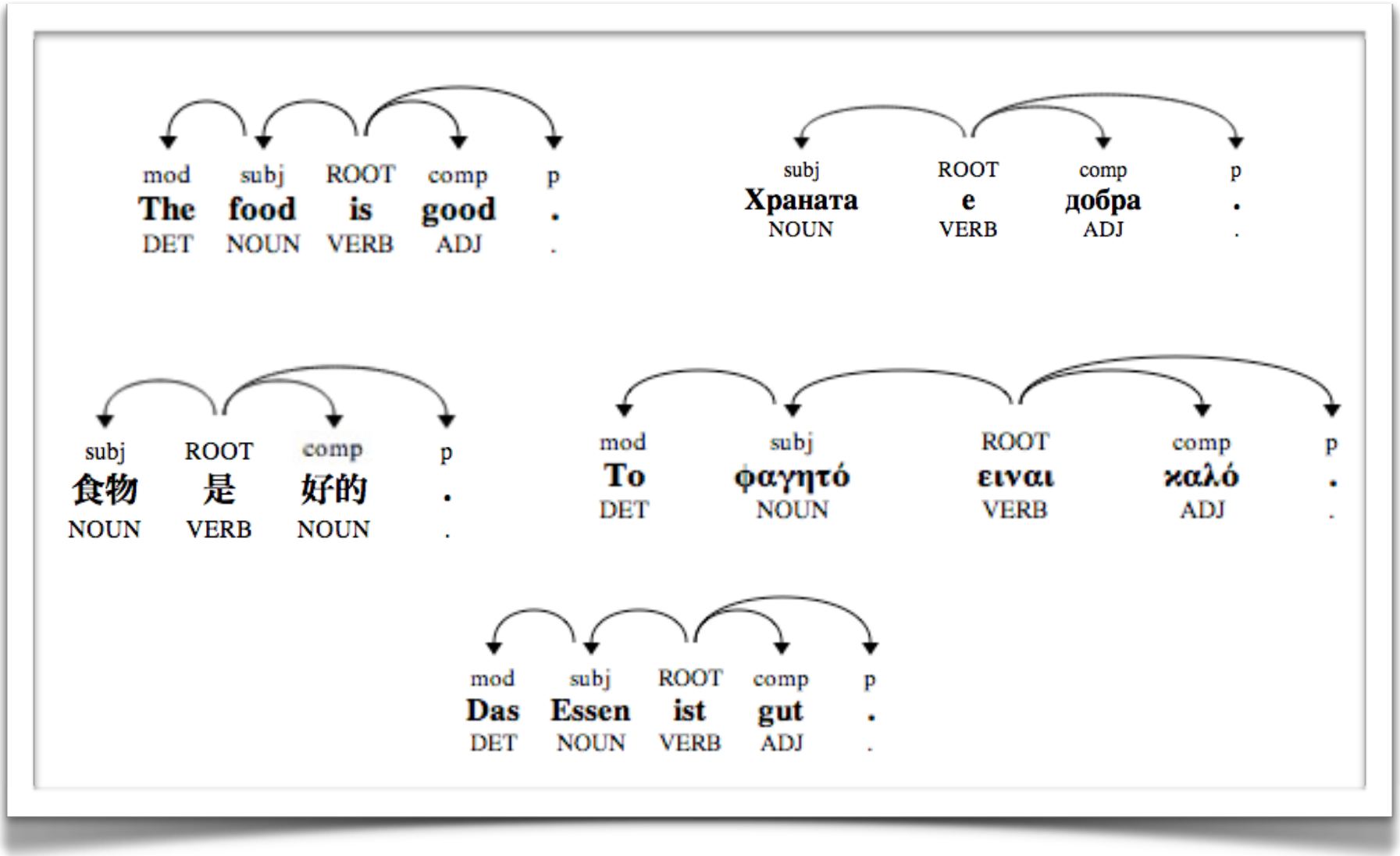
- Example: Syntactic analysis for English:
  - Not solved, but accuracies are high
  - 97% **(or 90%?)** for parts-of-speech
  - 93% **(or 83%?)** for parse trees



Treebank:
2 million words

Universal Tagging/Parsing

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



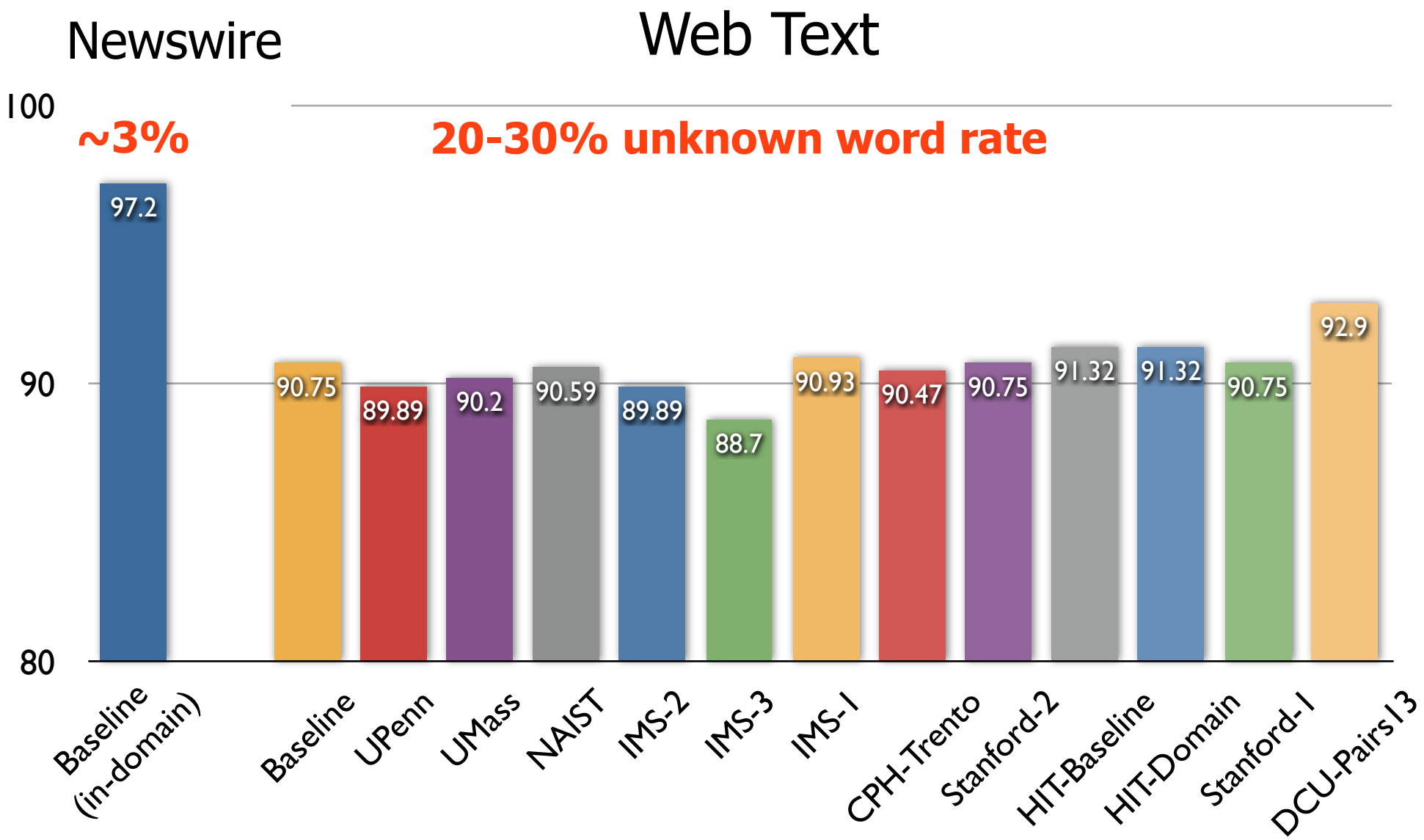
Syntactic Processing for the Web

- 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]

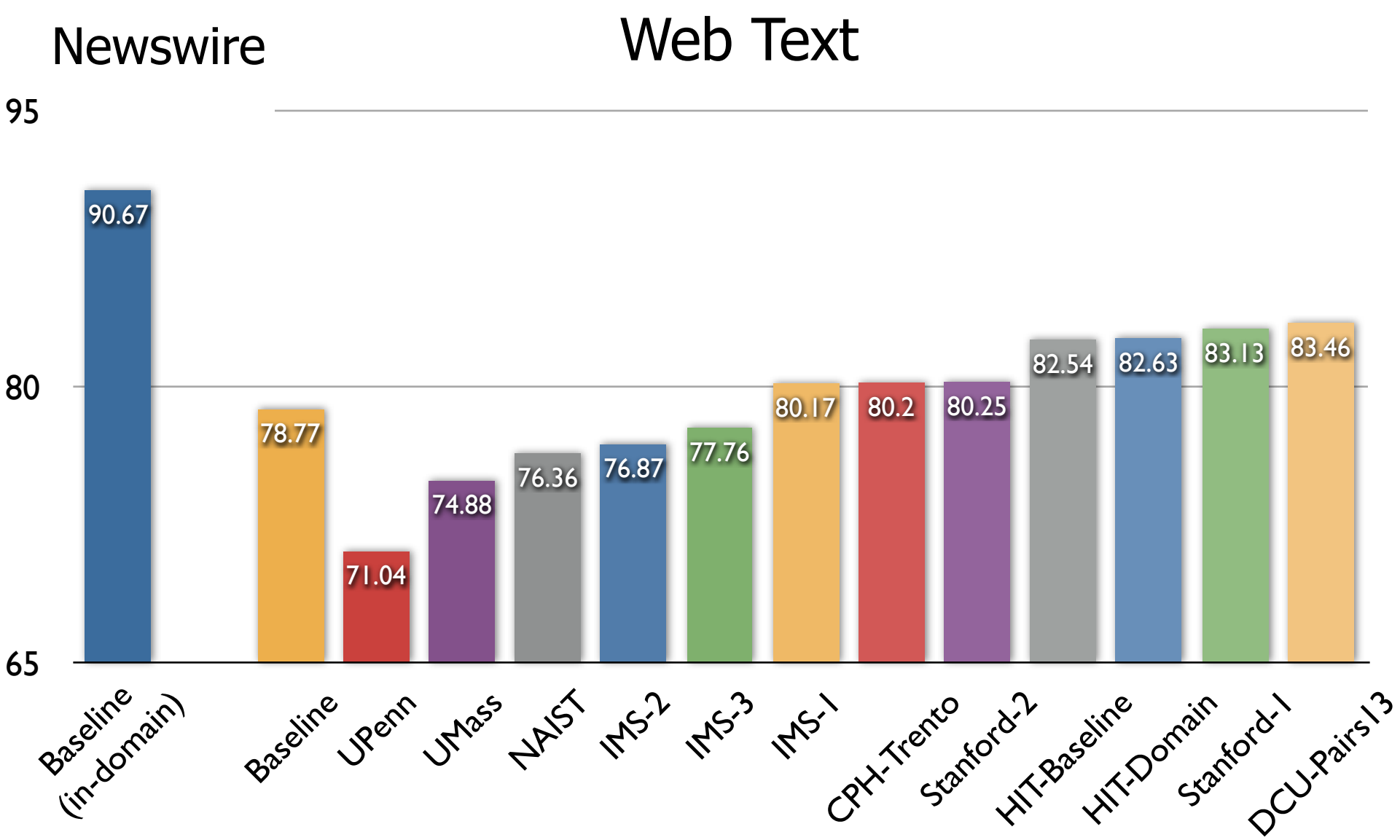
- Multilingual tagger and parser projection
  - Will never have labeled resources for all languages
  - Use parallel data to project information [Das & Petrov '11, McDonald et al. '11]

POS Accuracy (SANCL Shared Task)



Baseline = StanfordTagger v2.0 [Manning '11]

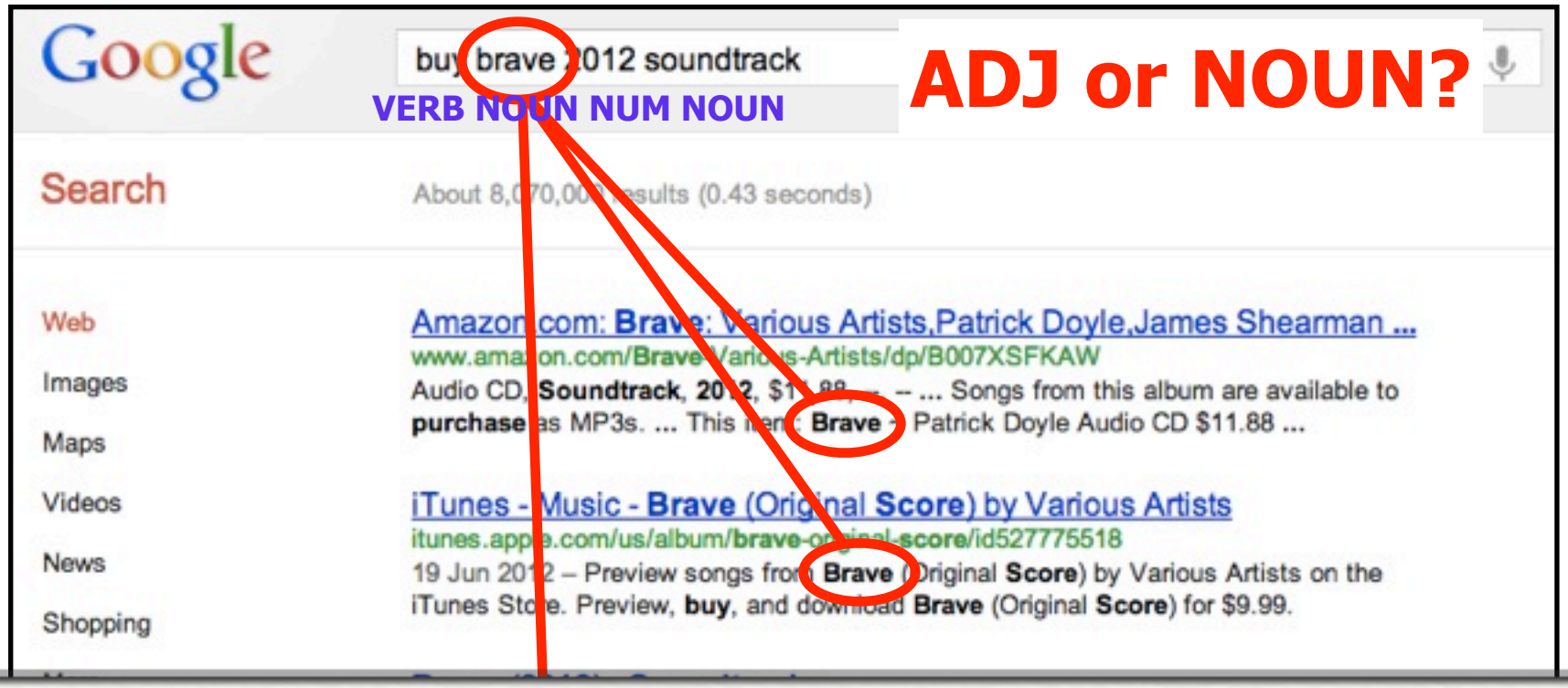
Parsing Accuracy (SANCL Shared Task)



Baseline = MaltParser [Zhang & Nivre '11]

Tagging Search Queries

[Ganchev et al. '12]



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	.

Baseline

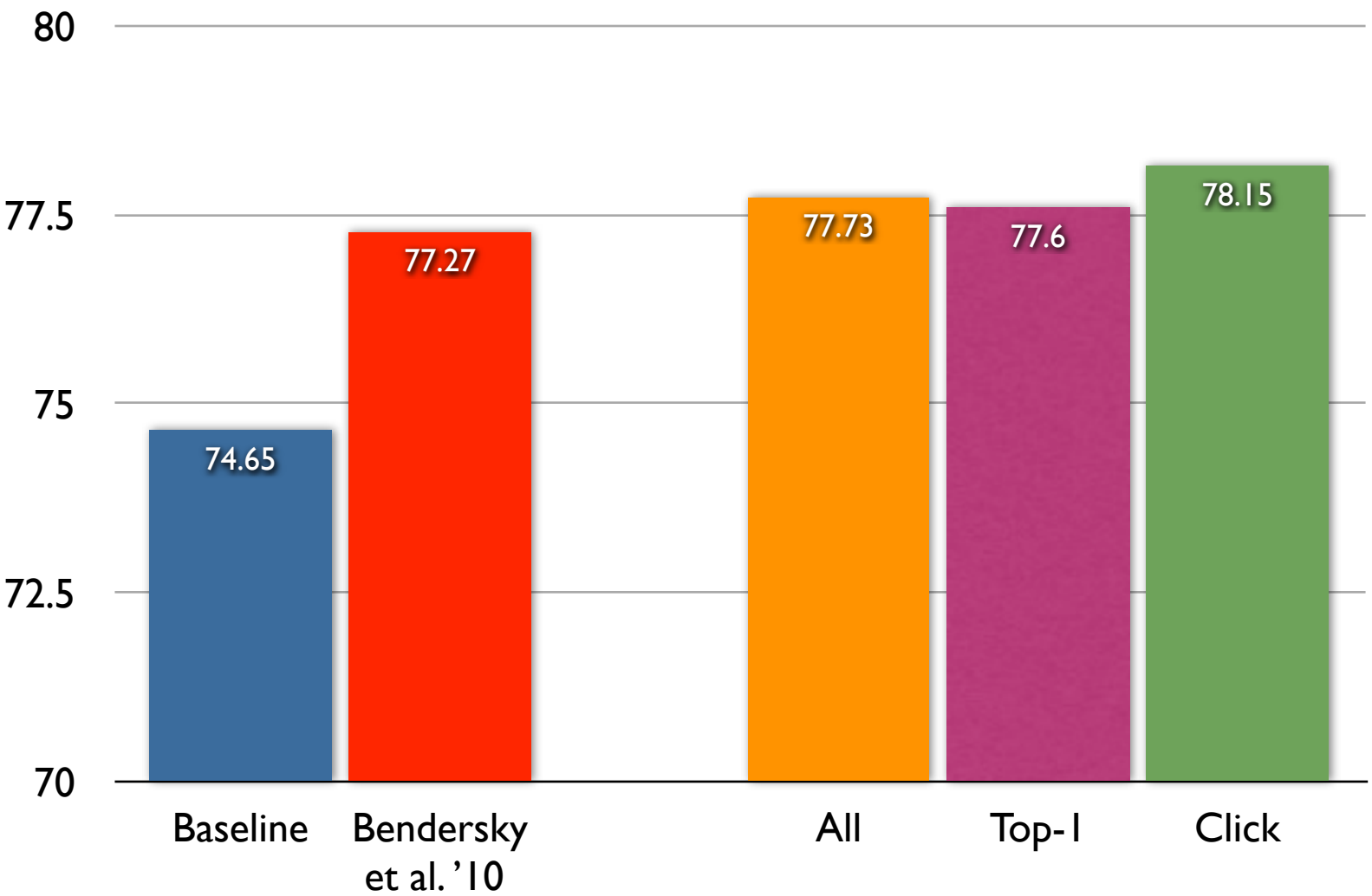
buy	brave	2012	soundtrack
VB	JJ	CD	NN
VERB	ADJ	NUM	NOUN

Retrained

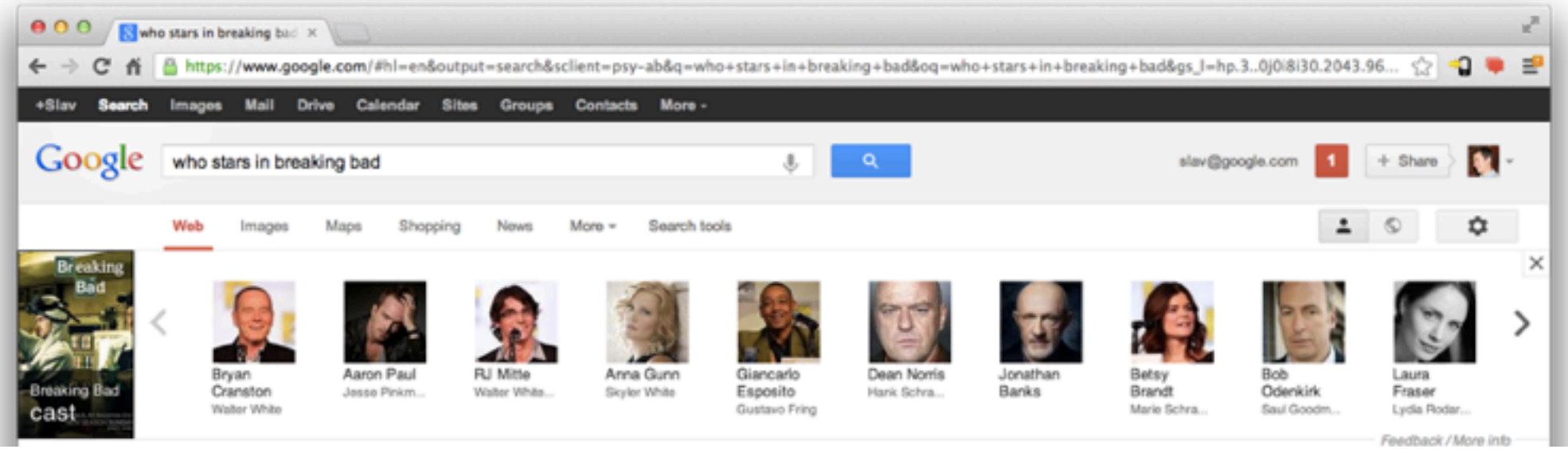
buy	brave	2012	soundtrack
VB	NNP	CD	NN
VERB	NOUN	NUM	NOUN

Query Tagging Results

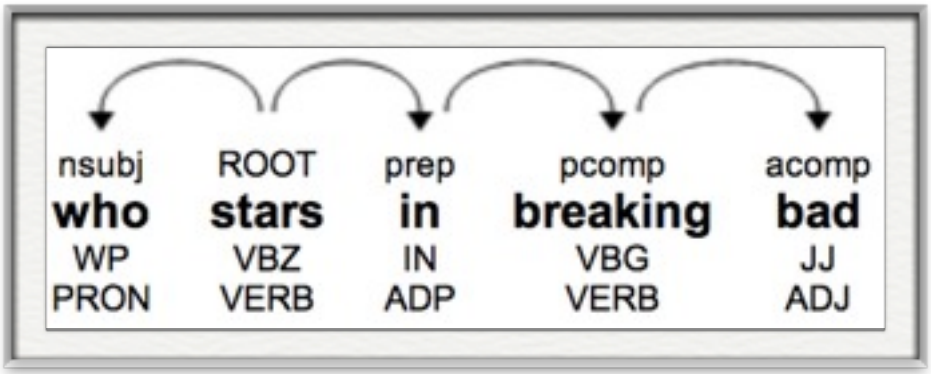
[Ganchev et al. '12]



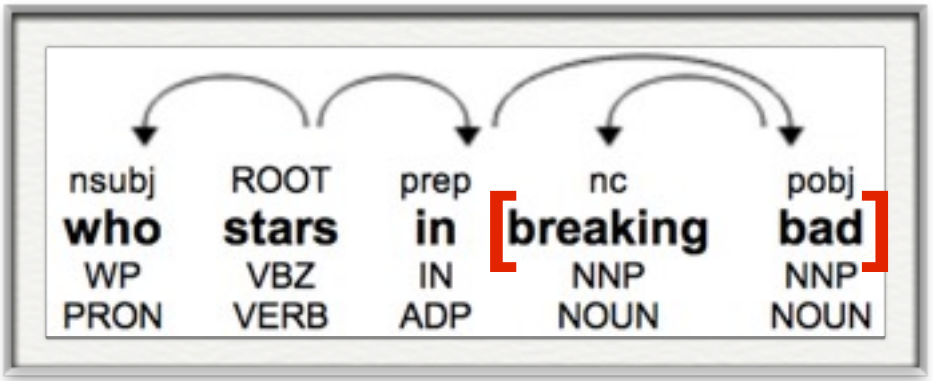
The breaking bad case...



Before



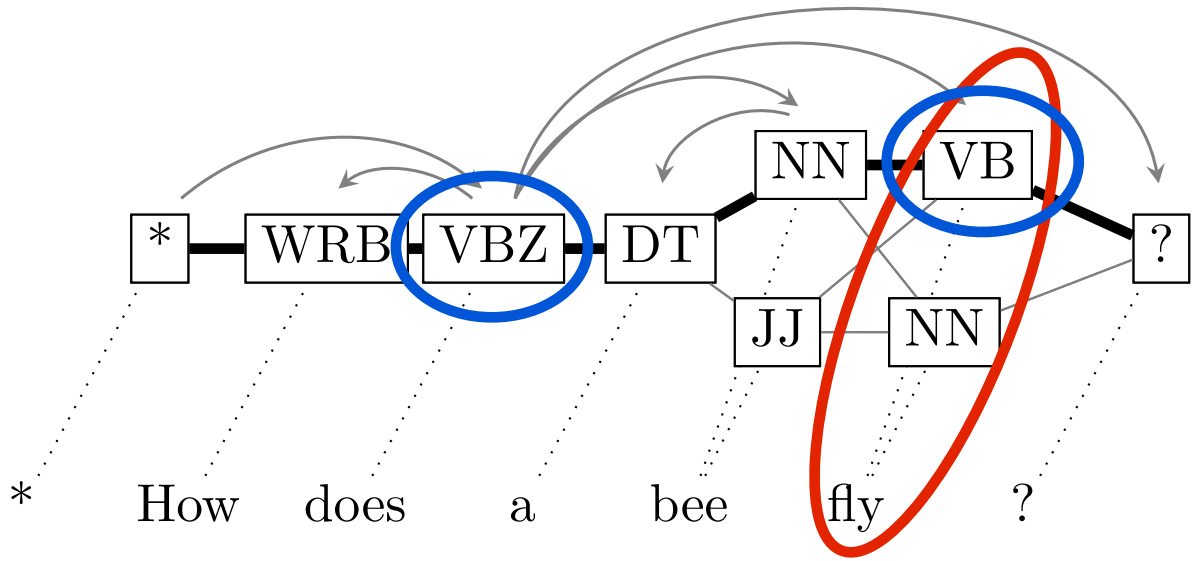
After



Use freebase entries as features

Integrated Tagging & Parsing

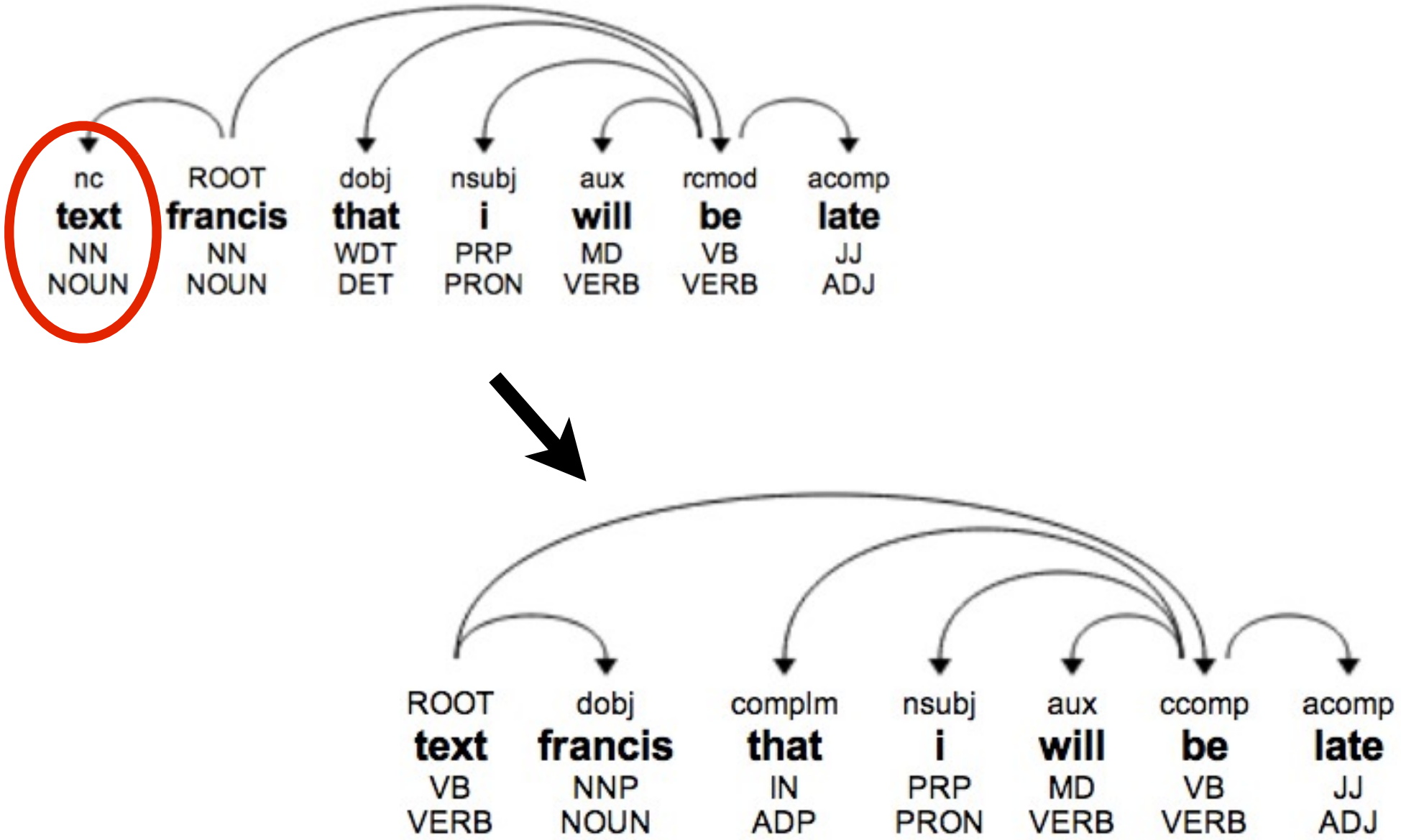
- Sequential Markov Models have limitations:



… 'to coach kim'

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

Manual Intervention



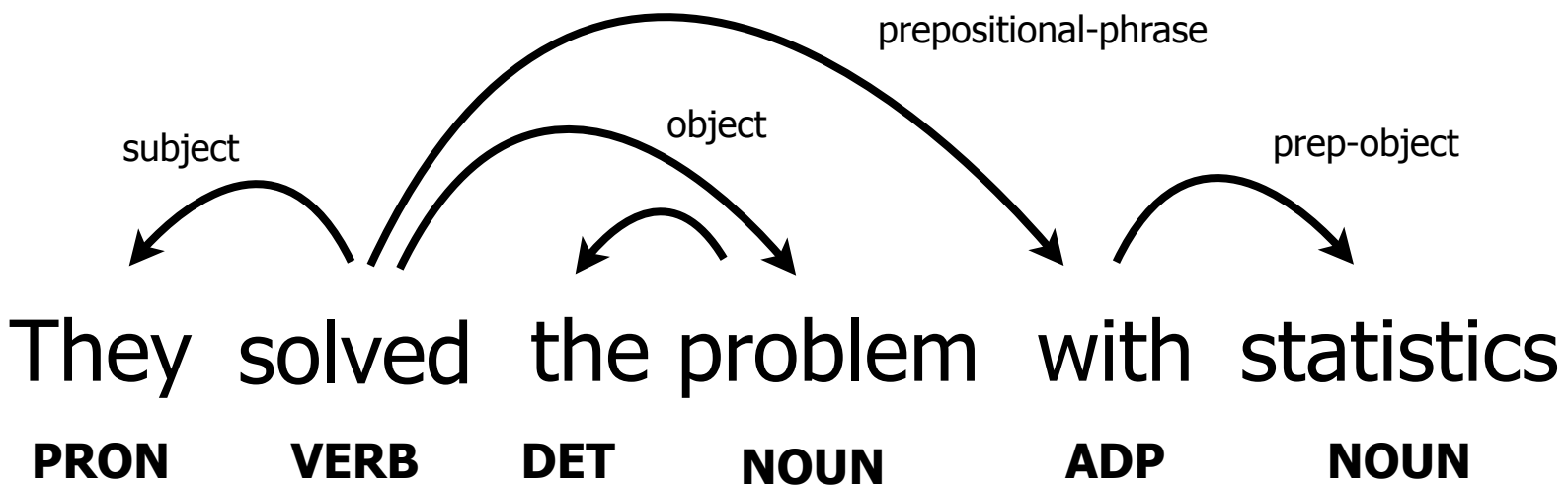
Machine Translation Reordering

- Source-side reordering for Japanese-English MT

[Collins et al. ’05]

- Dependency-based reordering for English-Japanese

[Xu et al. ’09]



Reordering Score

[Talbot et al. '11]

- Source-side reordering for Japanese-English MT

  [Collins et al. '05]

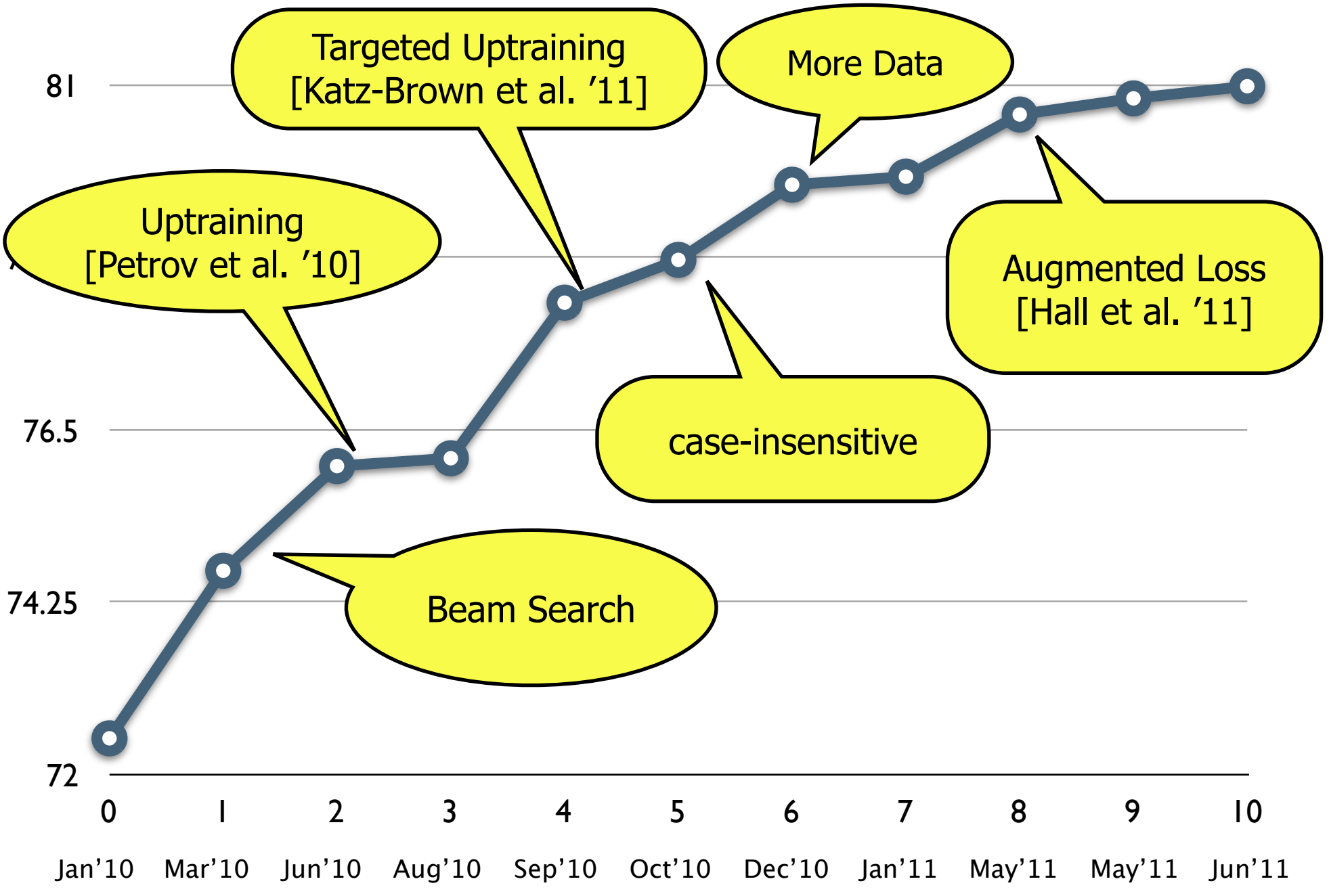
- Hand generated reordering data (English + Jenglish)
  - ~ 10k sentences for Augmented Loss training
  - ~ 6k evaluation sentences
- Score based on reordering penalty of METEOR

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

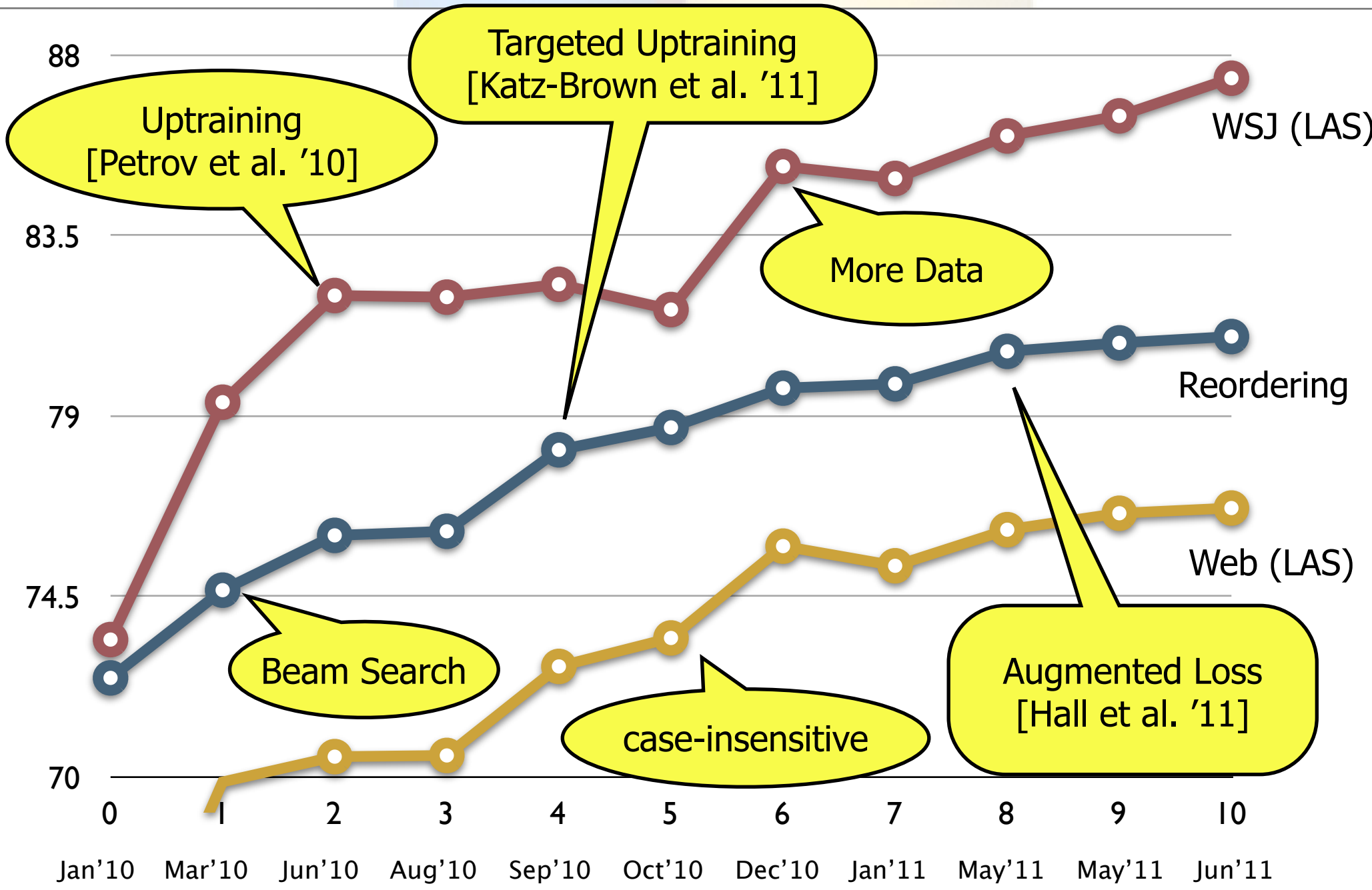
$$\text{reorder-loss} = 1 - \text{reorder-score}$$

- Very well correlated with human eval scores.

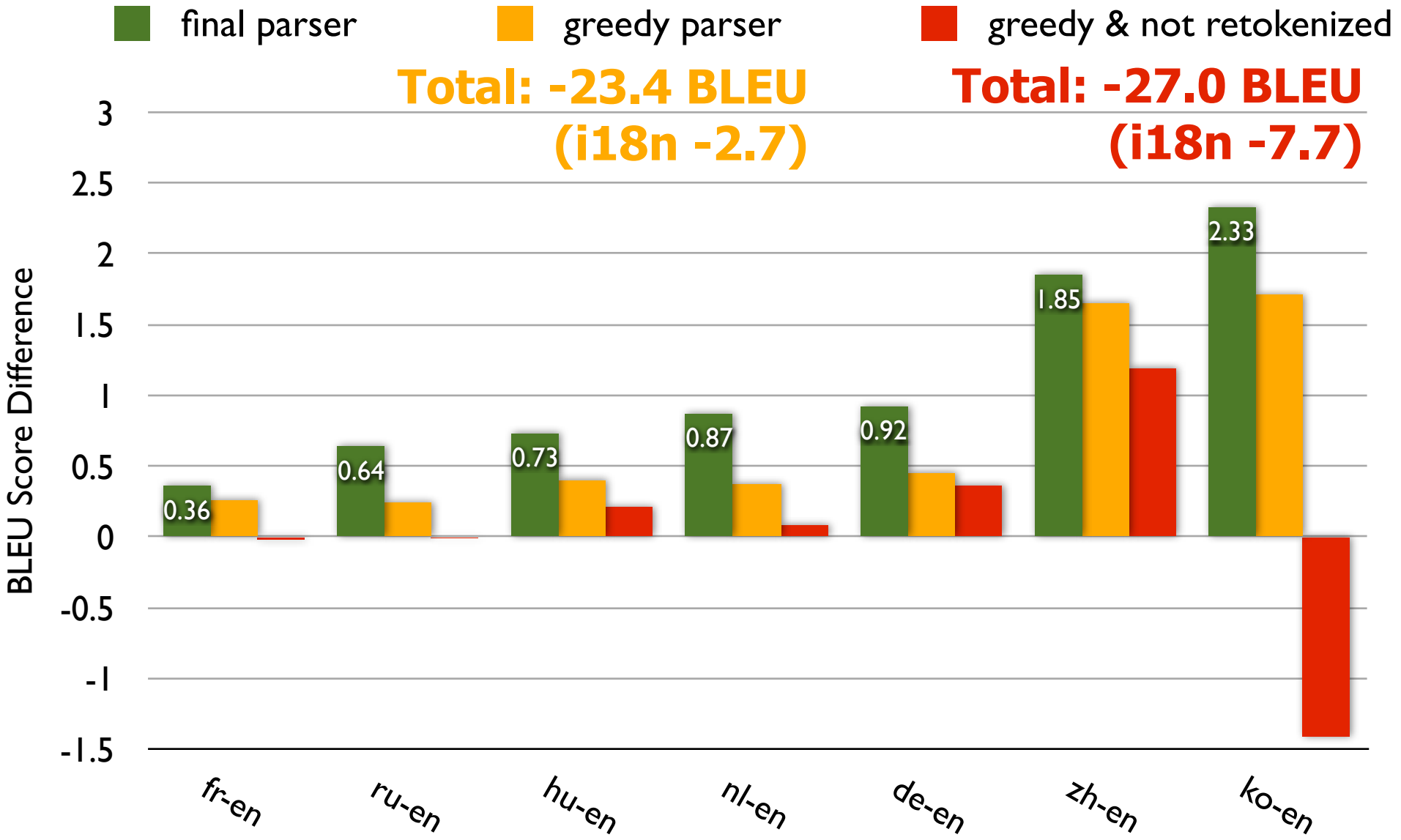
EnJa Fuzzy Reordering Evolution



EnJa Fuzzy Reordering Evolution



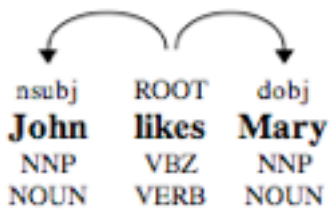
Retrospective Analysis

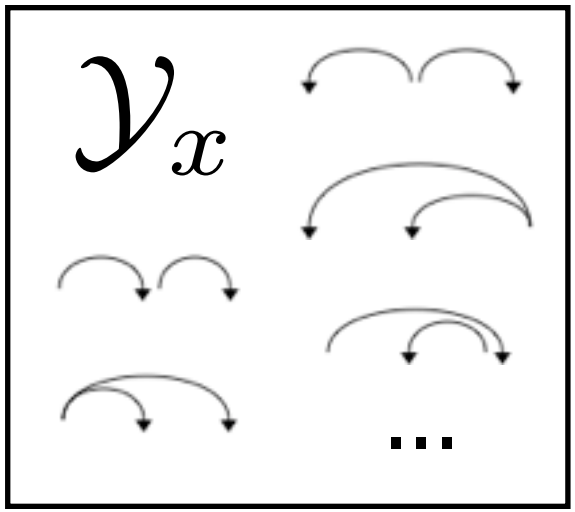


Standard Perceptron Training

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
$\vdots$
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 $\quad y' \in \mathcal{Y}'$

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

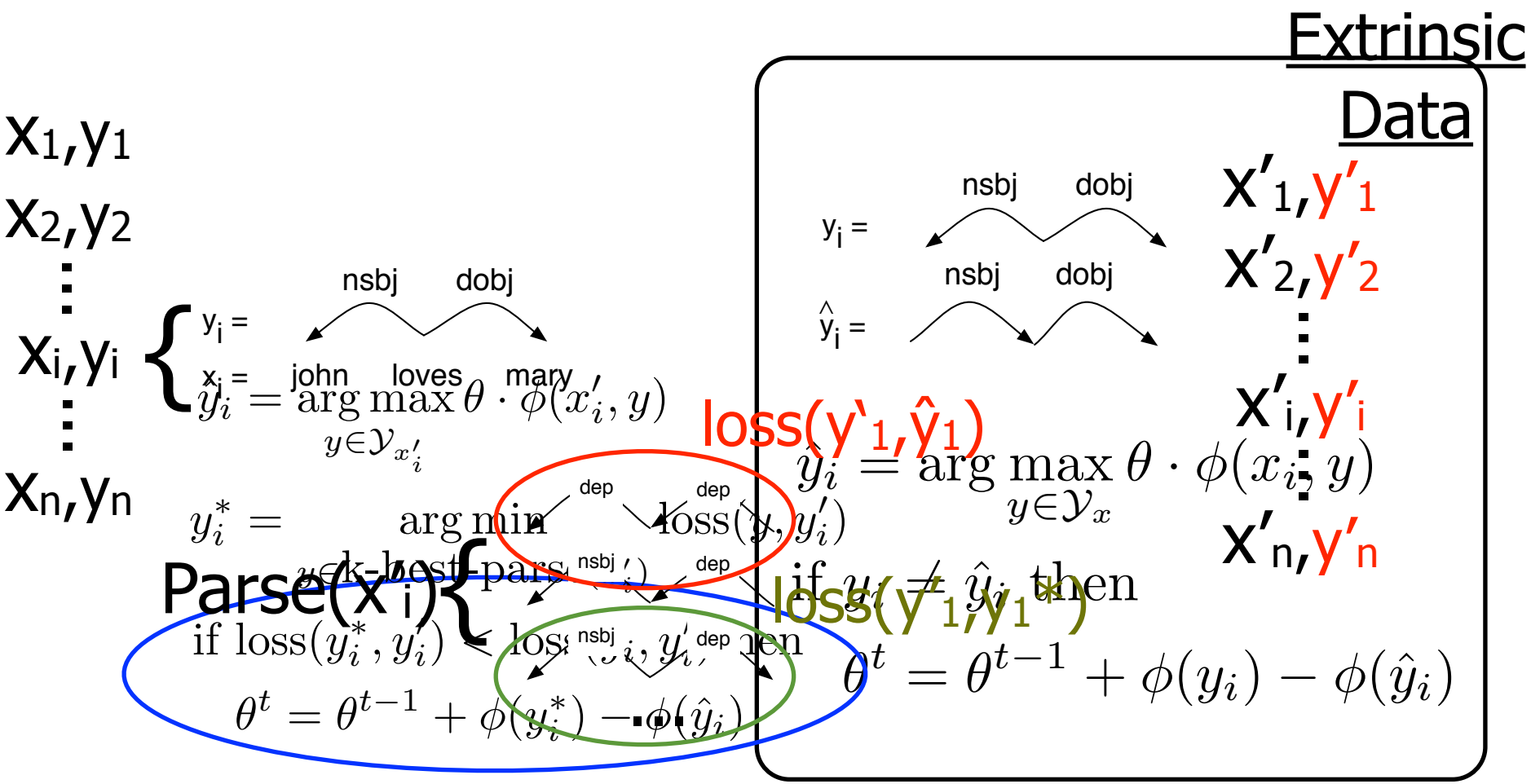
$\mathcal{Y} \neq \mathcal{Y}'$

Intrinsic: supervised training data/objective

Extrinsic: downstream task which "uses" outputs of model

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

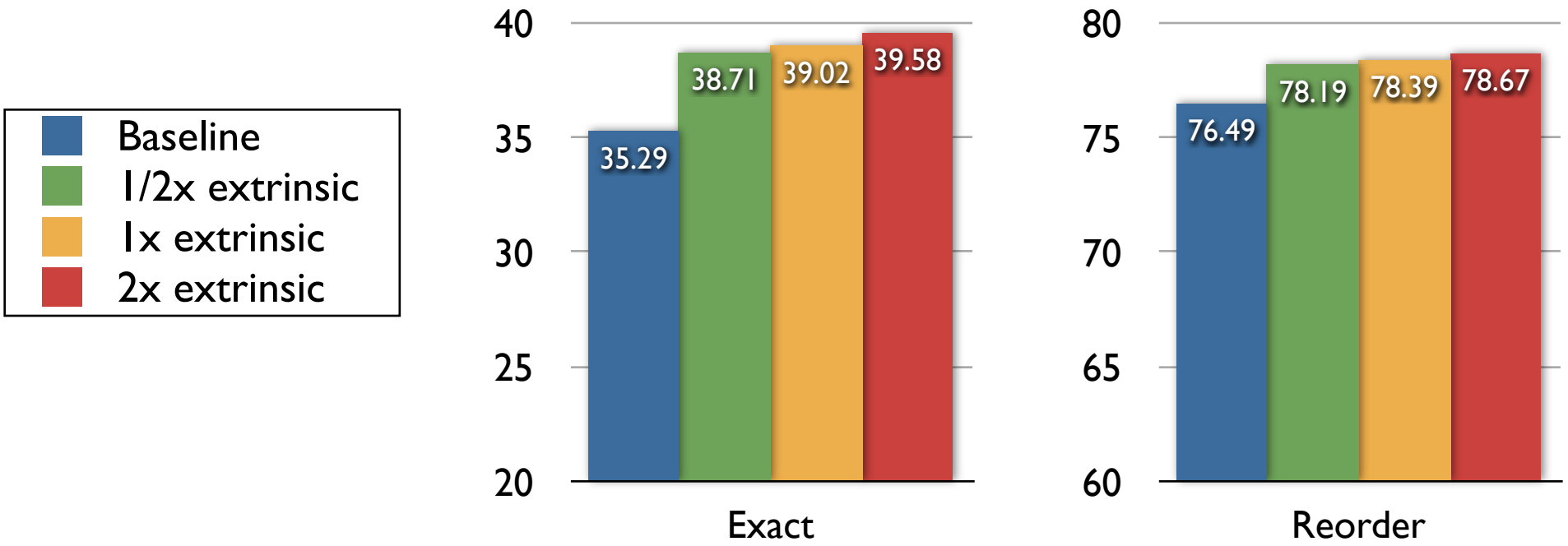
Augmented Loss Perceptron



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

Example MT-Reordering

"Clean" experiment



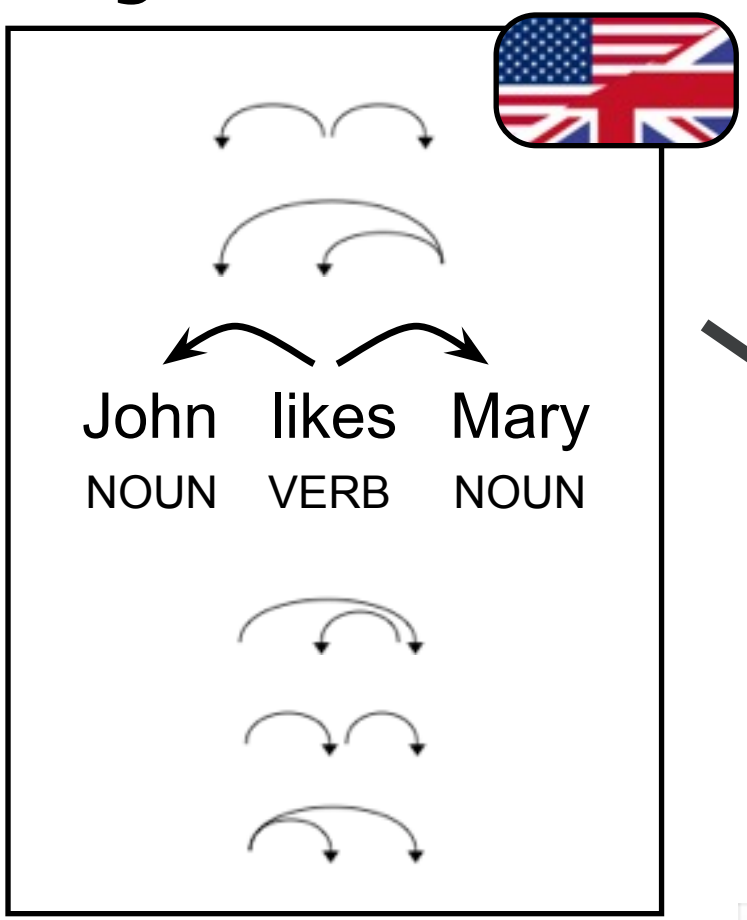
Production: 0.786 -> 0.792 enja fuzzy score

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

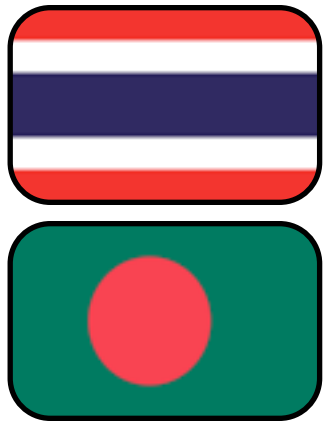
Syntactic Transfer

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

English Treebank



Syntactic Transfer

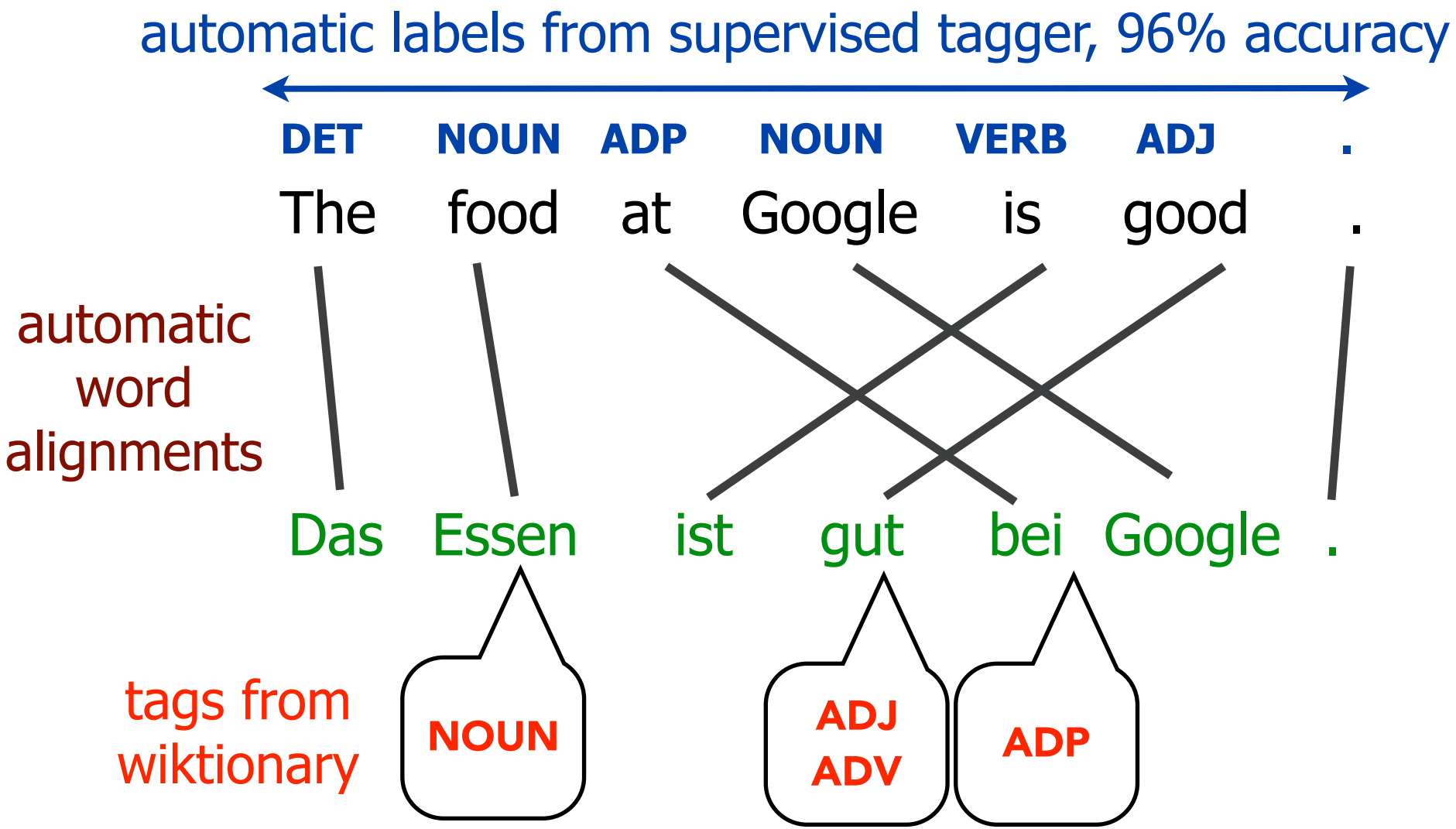


...

Dictionaries:

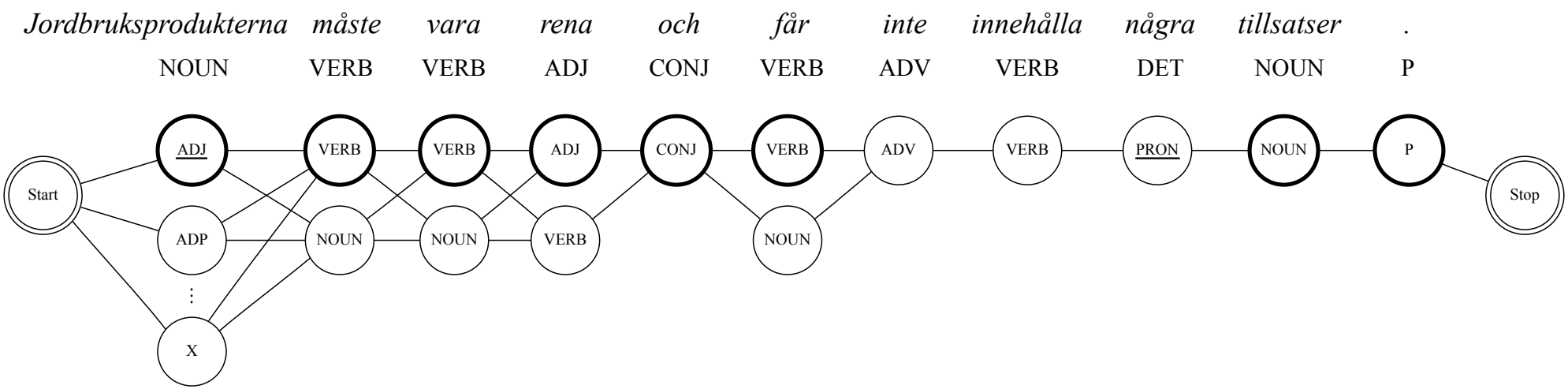
Cross-Lingual Projection

[Das & Petrov '11, Täckström et al. '13]



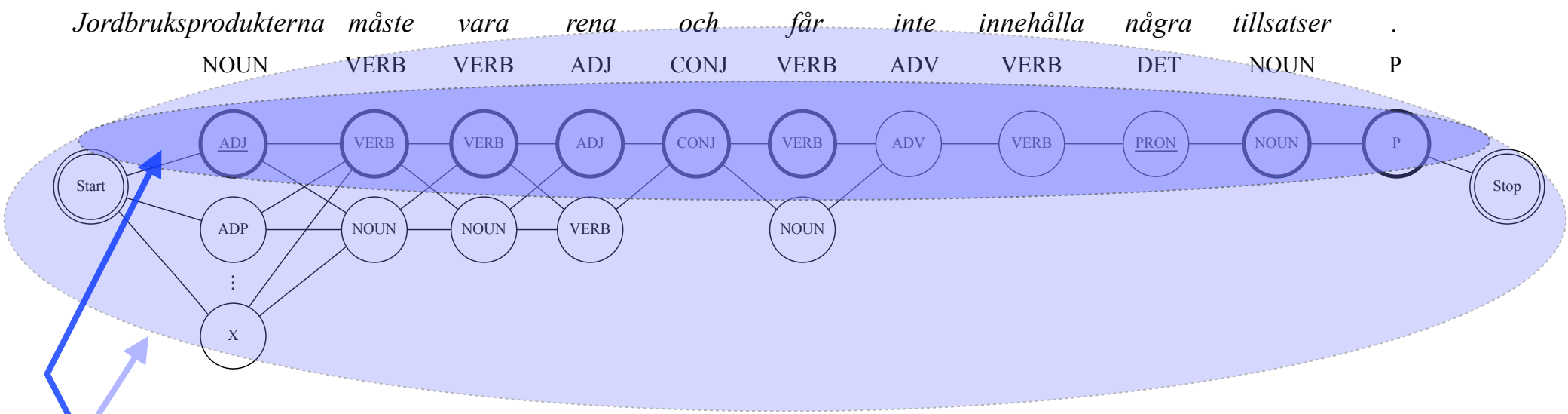
Type and Token Constraints

[Täckström et al. ’13]



- the lattice is constructed using a tag dictionary
- bold nodes show projected tokens
- underlines indicate incorrect tags

Hidden CRF Model



- train using:
  - type- and token-constrained lattice as observation
  - type-constrained lattice as the normalizer

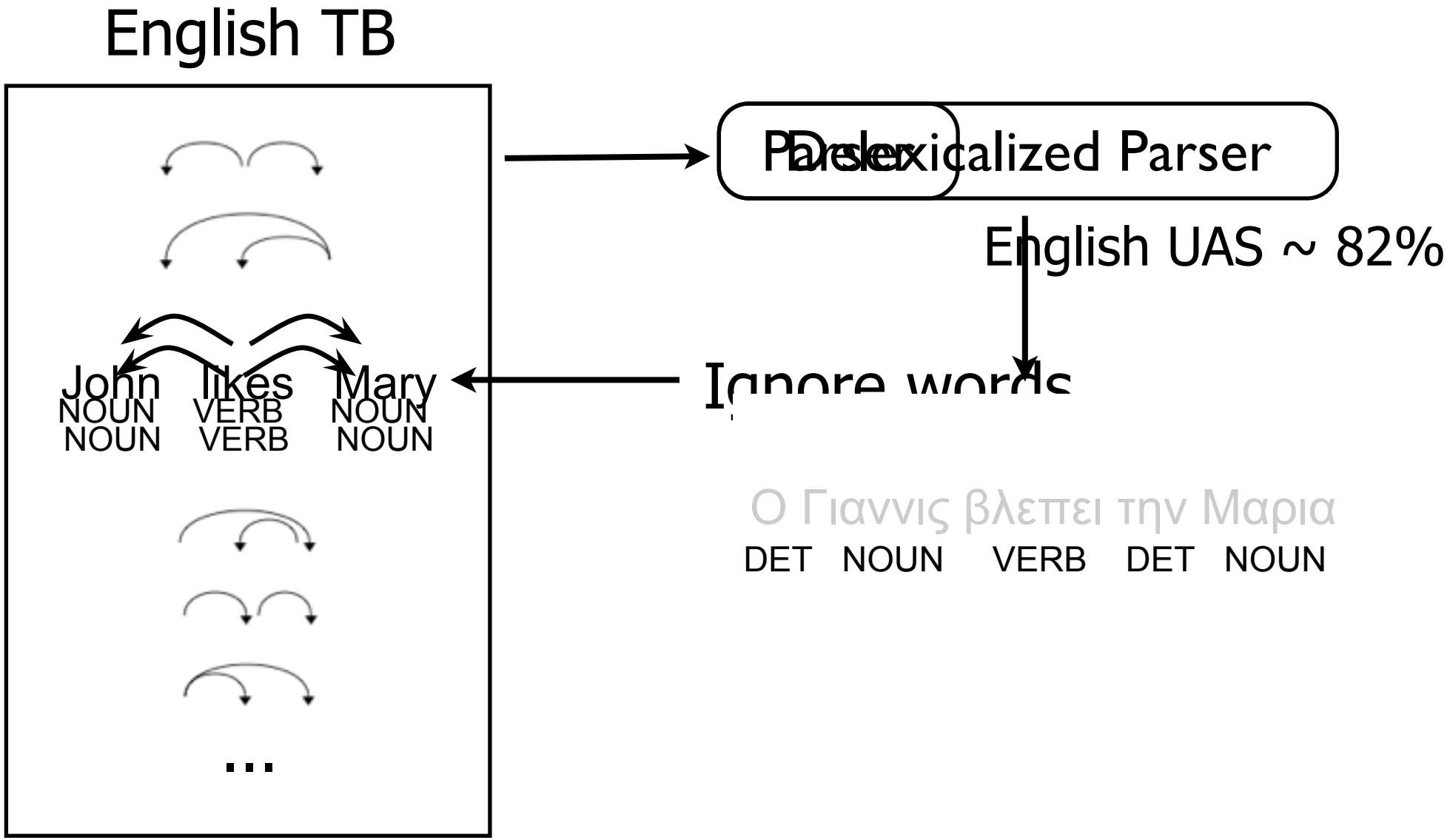
POS Projection Results

	Danish	Dutch	German	Greek	Italian	Portugues	Spanish	Swedish	**Average**
Feature-HMM	69.1	65.1	81.3	71.8	68.1	78.4	80.2	70.1	**73.0**
Direct Projection	73.6	77.0	83.2	79.3	79.7	82.6	80.1	74.7	**78.8**
(a) Graph-Based Projections	83.2	79.5	82.8	82.5	86.8	87.9	84.2	80.5	**83.4**
(b) **Type&Token Constraints**	88.2	85.9	90.5	89.5	89.3	91.0	87.1	88.9	**88.8**
Oracle Dictionary	*93.1*	*94.7*	*93.5*	*96.6*	*96.4*	*94.0*	*95.8*	*85.5*	*93.7*
Supervised	*96.9*	*94.9*	*98.2*	*97.8*	*95.8*	*97.2*	*96.8*	*94.8*	*96.6*

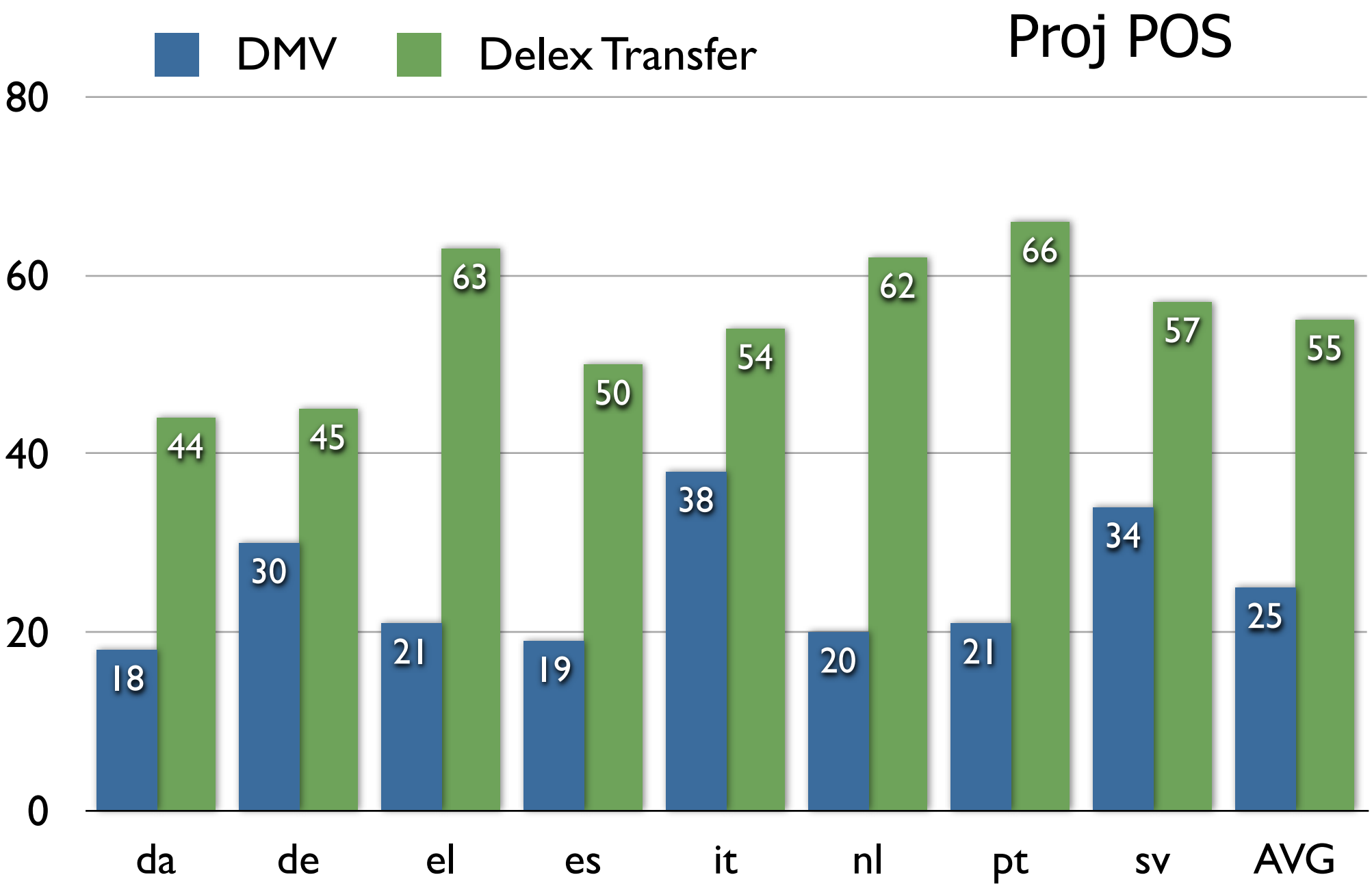
[(a) Das & Petrov '11, (b) Täckström et al. '13]

Delexicalized Parser Transfer

[Zeman and Resnik '08]

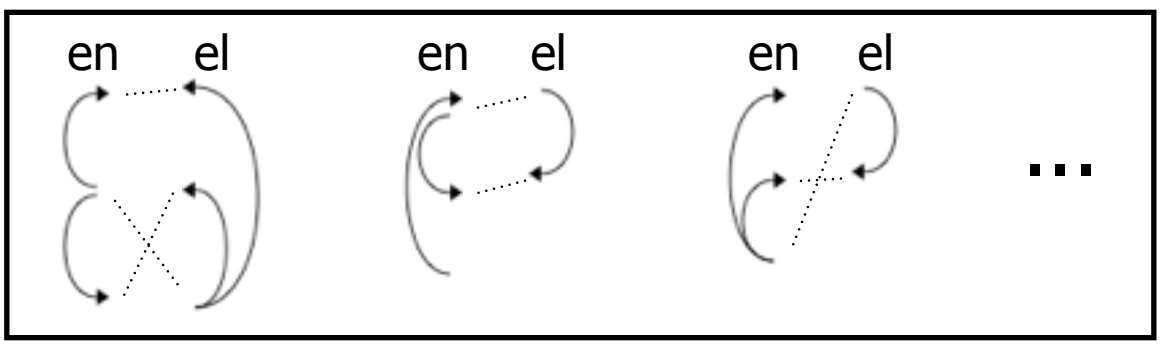


Delexicalized Transfer

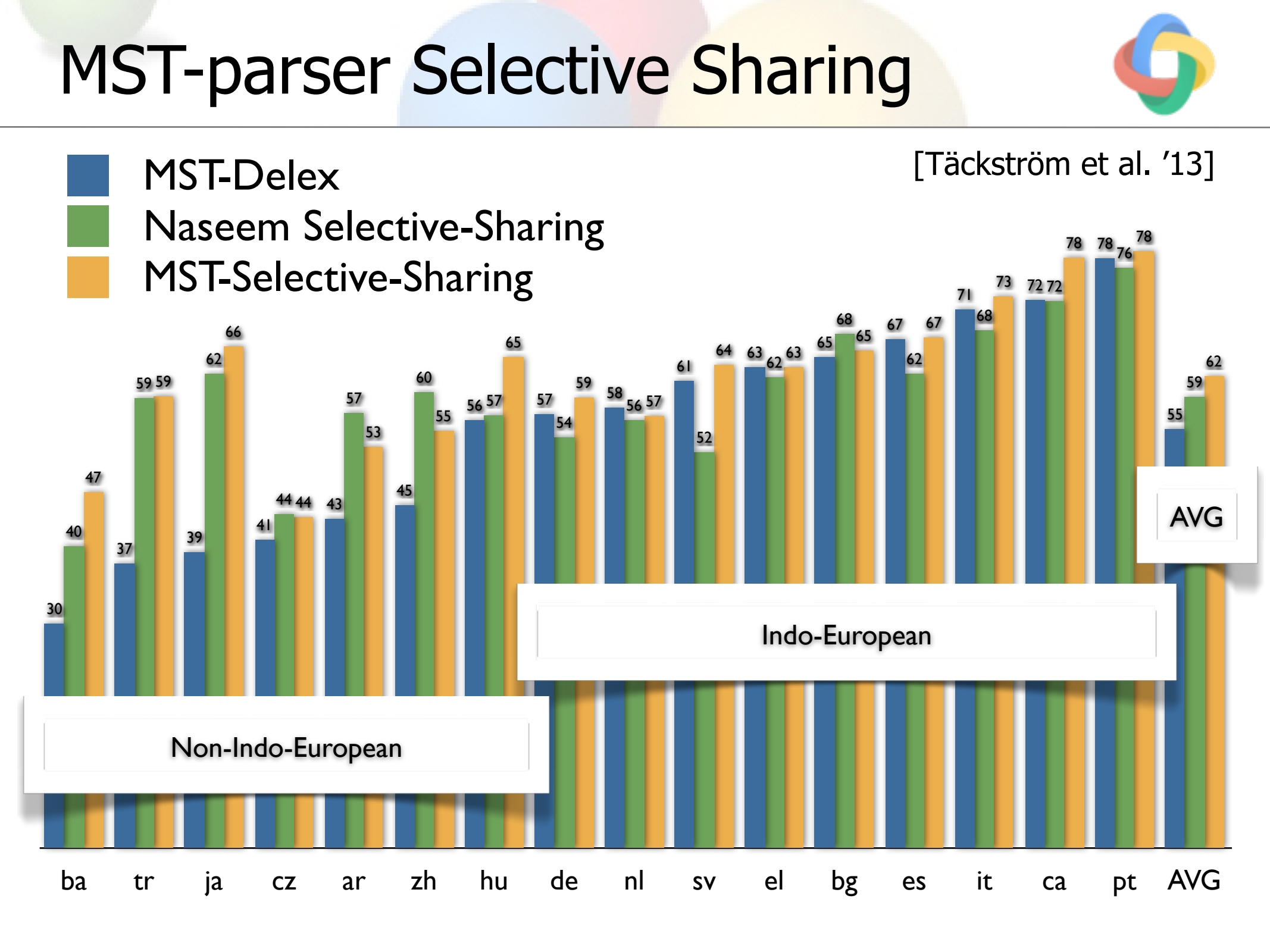


Projected Transfer

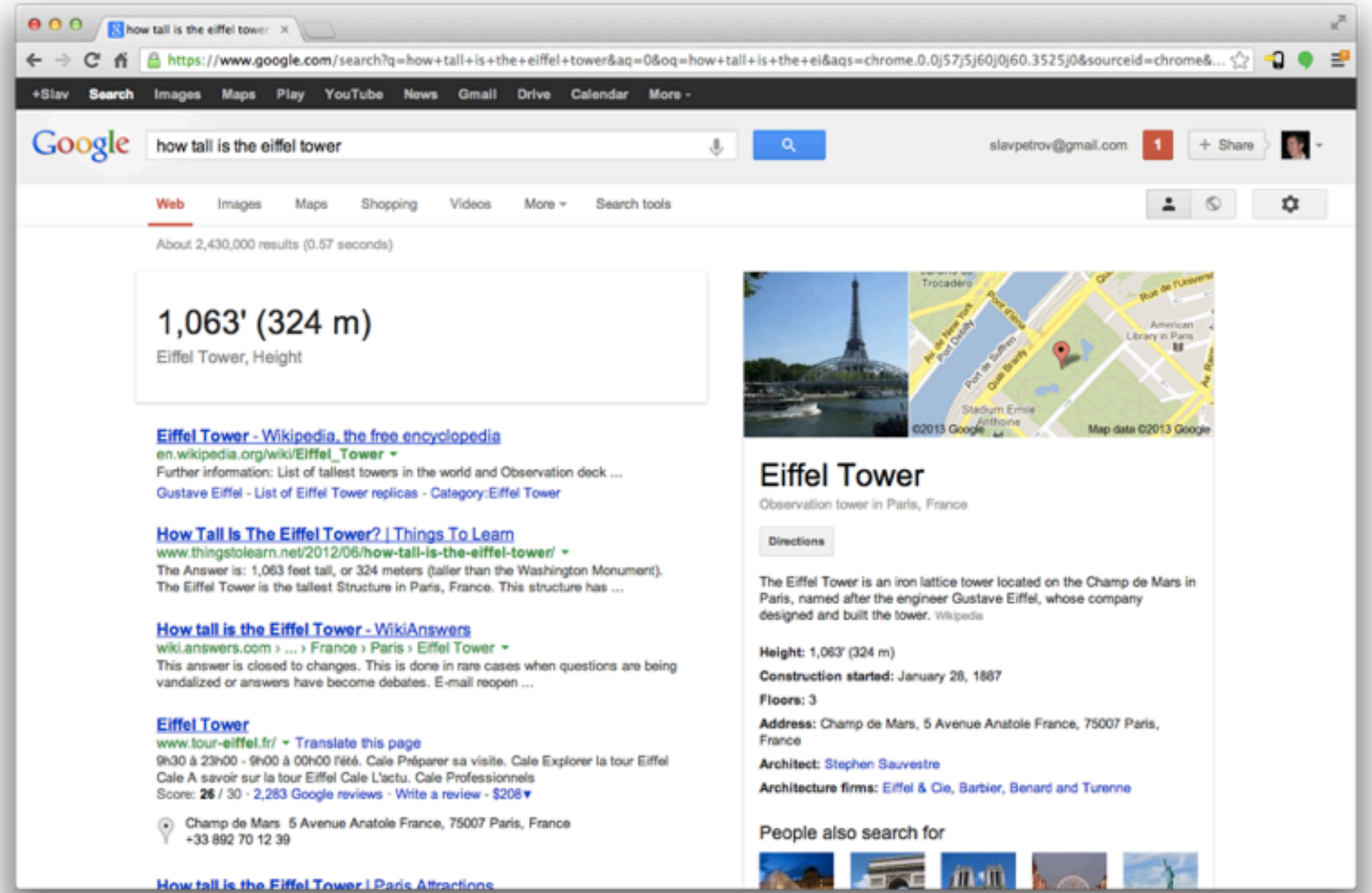
- Pioneered by Hwa et al. ’05
  - Many studies since: Ganchev et al. ’09, Smith & Eisner ’09, Spreyer et al. ’09,’10, etc.
- Project English parses via word aligned parallel data.
  - Use automatic word alignments.
  - Derive new gold trees or partial tree constraints.
- Use information about language typology (SOV vs VOS)



MST-parser Selective Sharing



Confidence Estimation



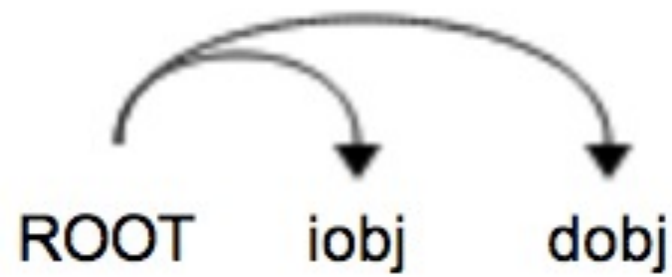
Confidence Estimation

- 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

Segmentation / Morphology

Segmentation / Morphology

POS?



dá - me - lo

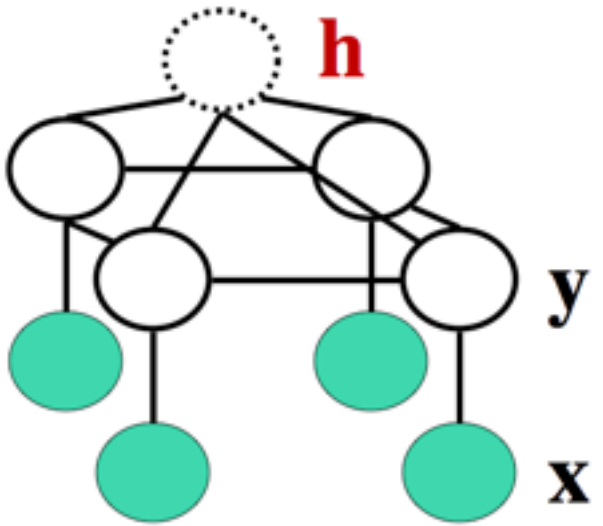
VERB PRON PRON

Russian,
Turkish,
Finnish,
...

Learning from Indirect Signals

Structured Data

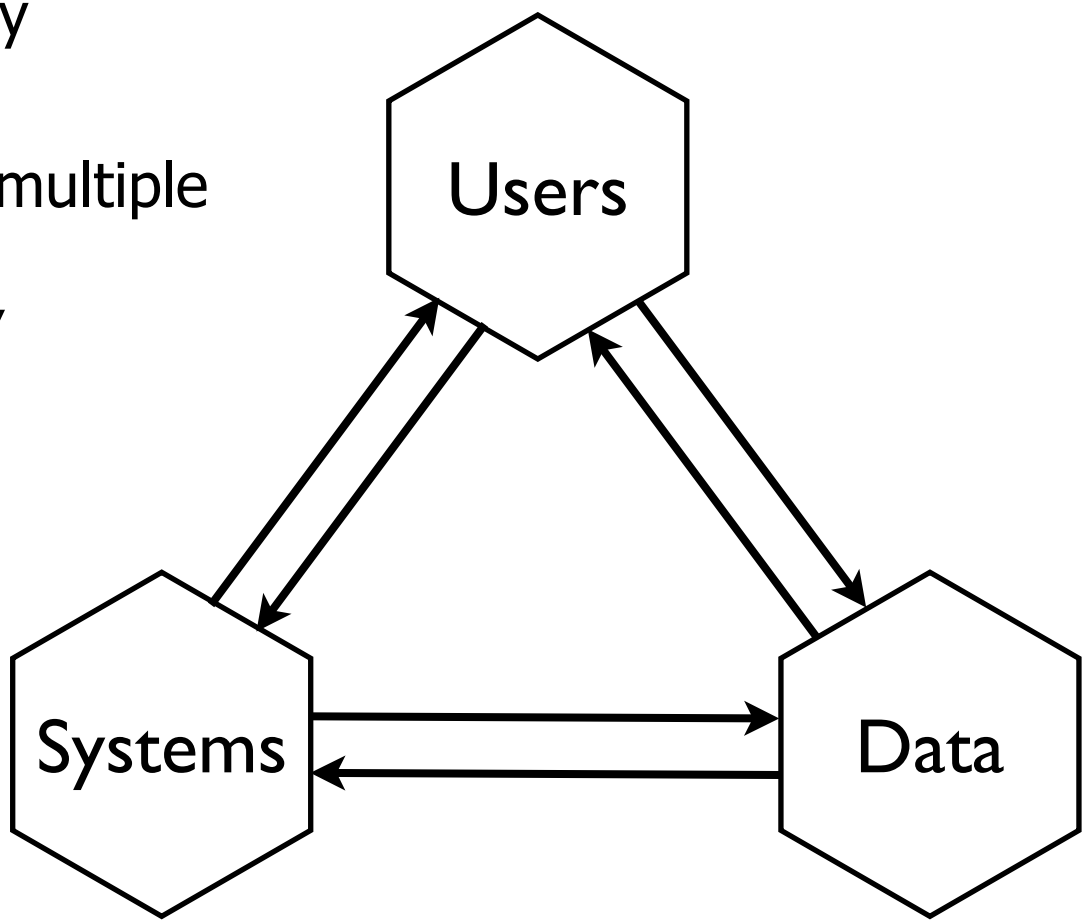
Labeled Data



Learning & Modeling

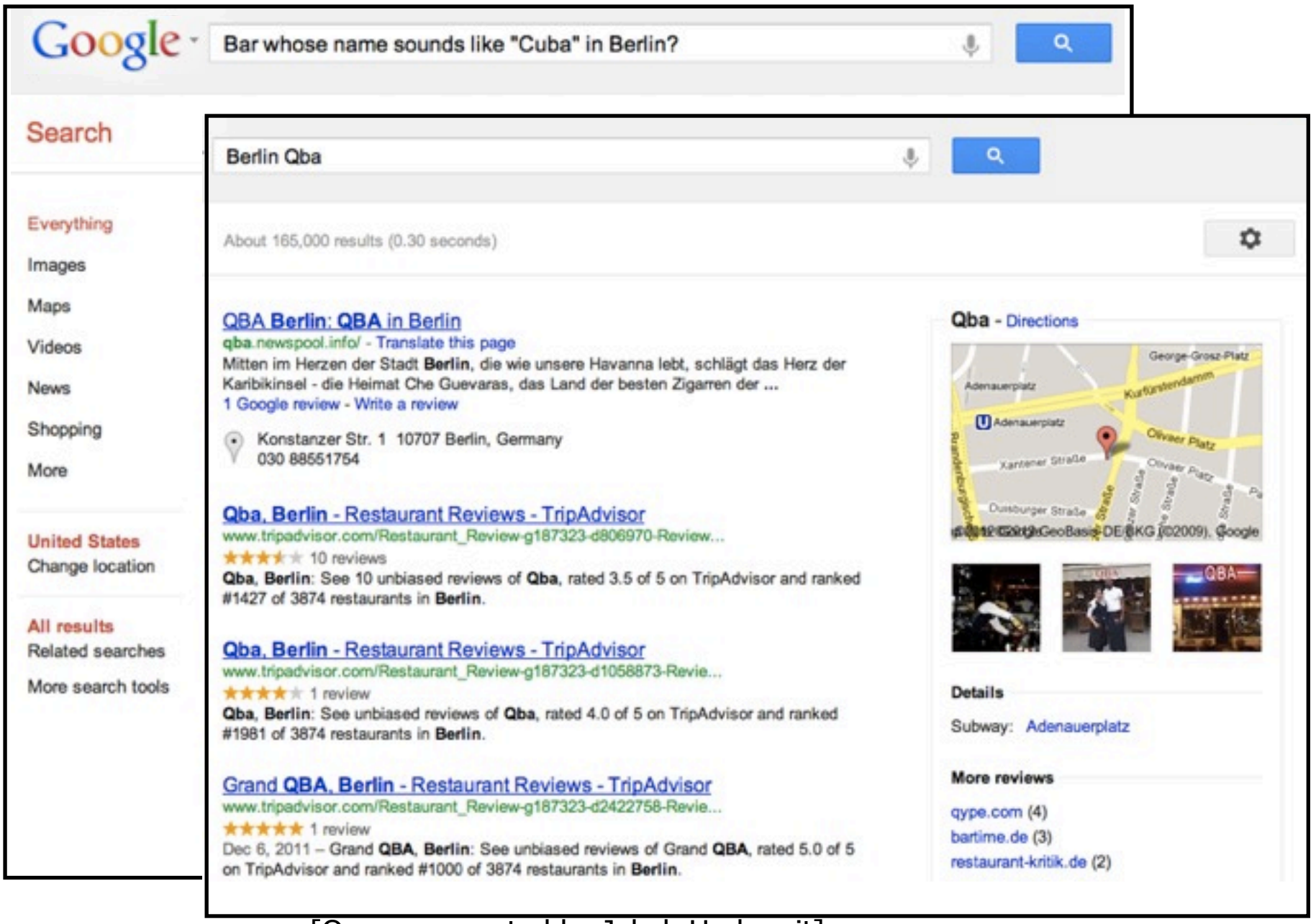
Hypothesis

- 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



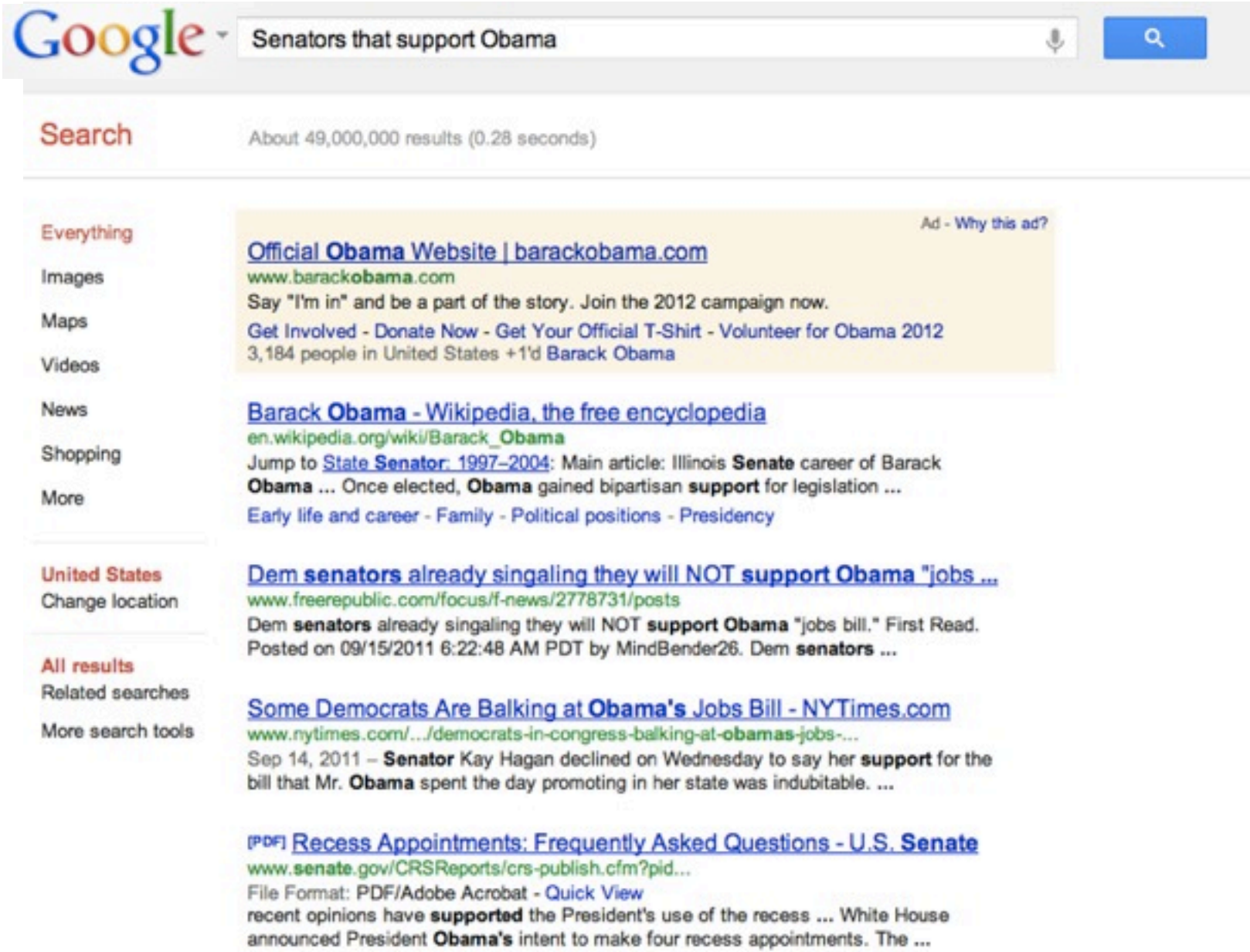
[Based on slide from Fernando Pereira]

Difficult "fuzzy" Queries



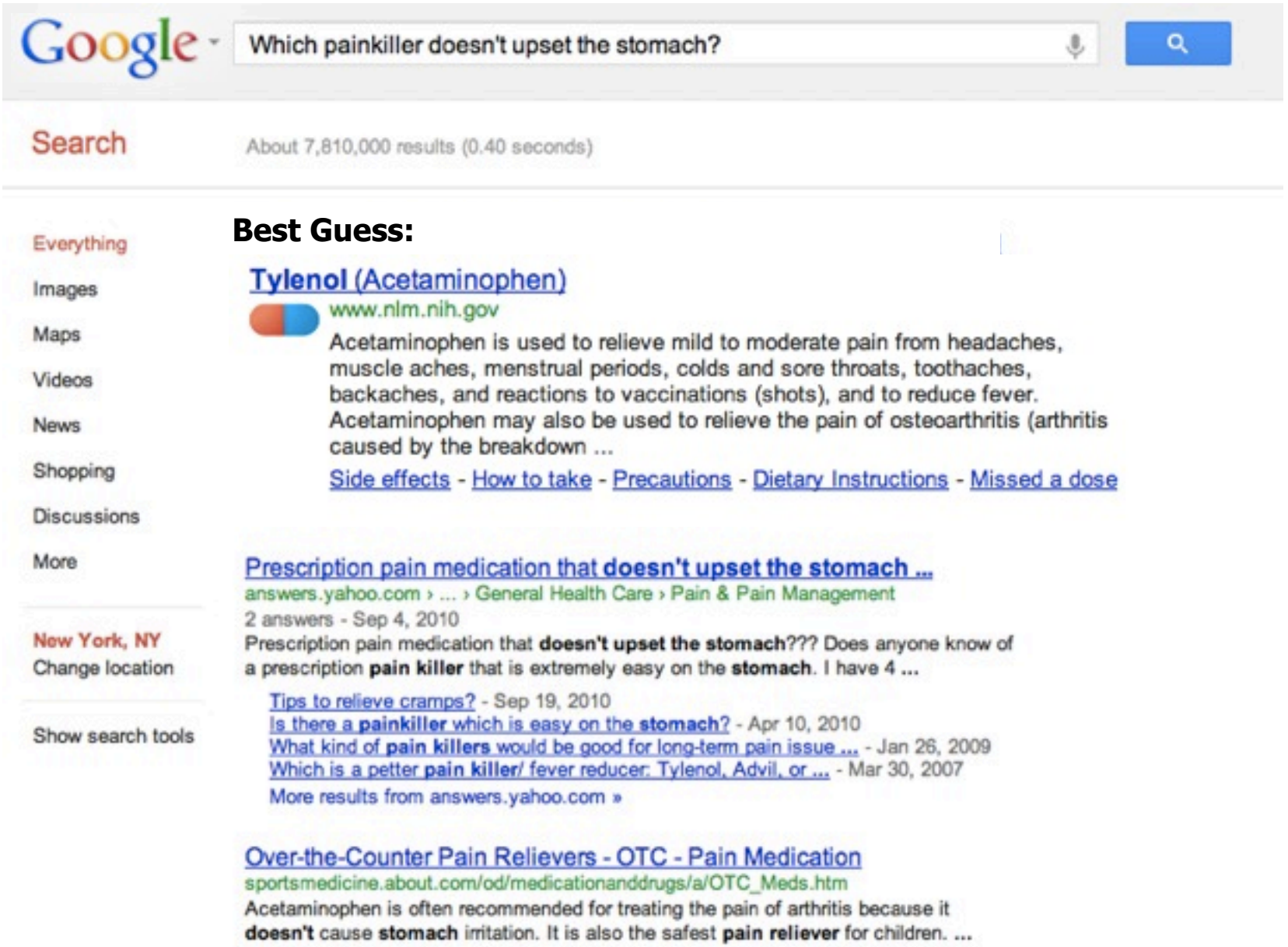
[Query suggested by Jakob Uszkoreit]

Difficult "fuzzy" Queries



[Query suggested by Percy Liang]

One day... (hopefully soon)



Questions?