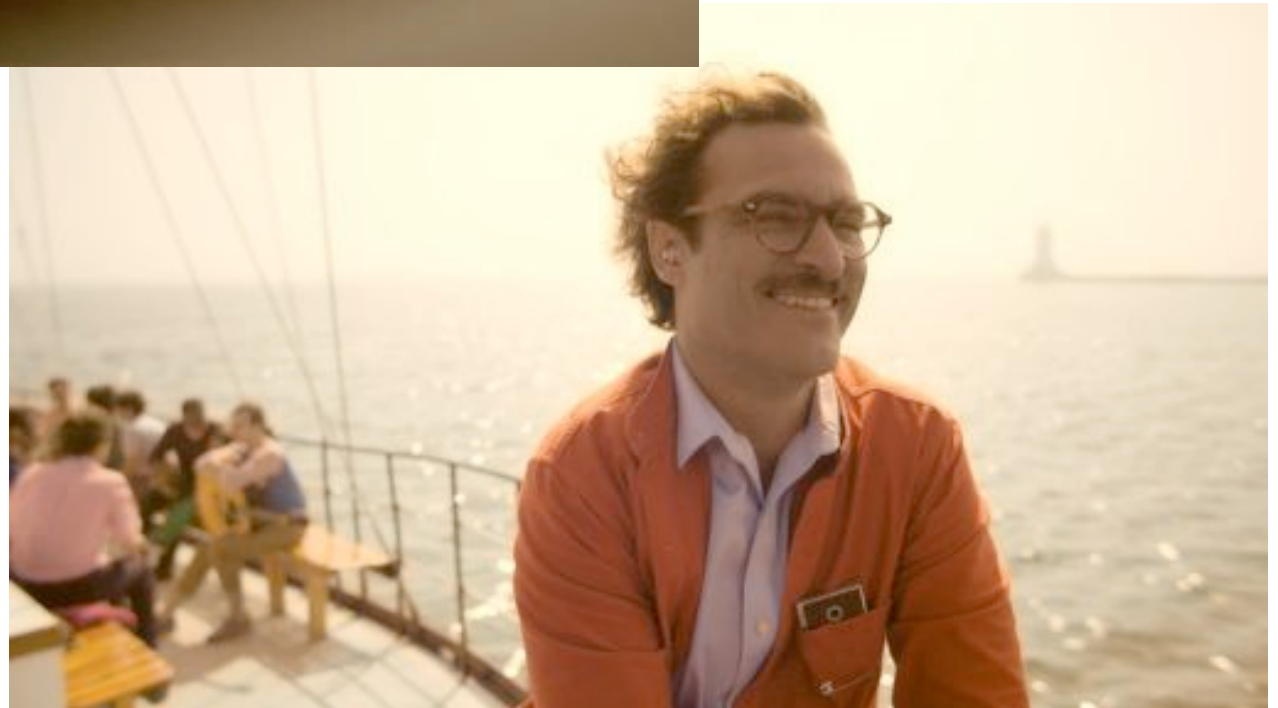


# LARGE-SCALE LANGUAGE GROUNDING WITH VISION

Yejin Choi

Computer Science & Engineering

**W** UNIVERSITY *of* WASHINGTON





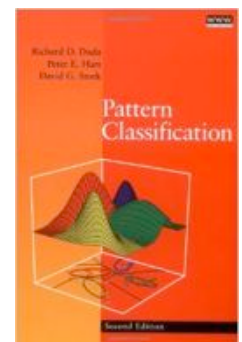
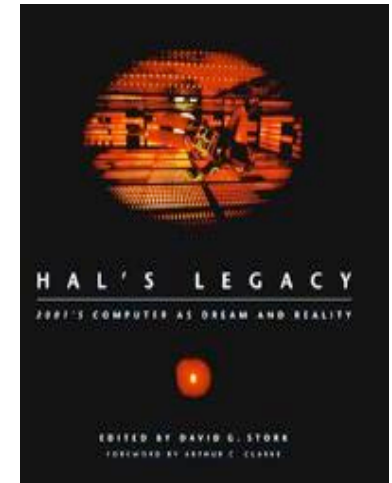
# HAL (a space odyssey, 1968)

- David Stork (HAL's Legacy, 1998)

*"Imagine, for example, a computer that could look at an arbitrary scene anything from a sunset over a fishing village to Grand Central Station at rush hour and produce a verbal description.*

*This is a problem of overwhelming difficulty, relying as it does on **finding solutions to both vision and language** and then **integrating them**.*

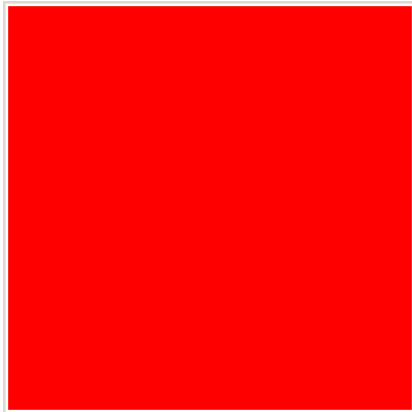
*I suspect that scene analysis will be one of the last cognitive tasks to be performed well by computers"*



# Language grounding with vision

- Understanding the meaning of language with perceptual signals
- What does red mean?
  - red --- having a color resembling that of blood
- What does blood mean?
  - blood – the red fluid that circulates through the heart...

• red :=



- Not just words, but phrases and sentences.

# Not just words, but descriptions

- “playing a soccer” vs. “playing a piano”



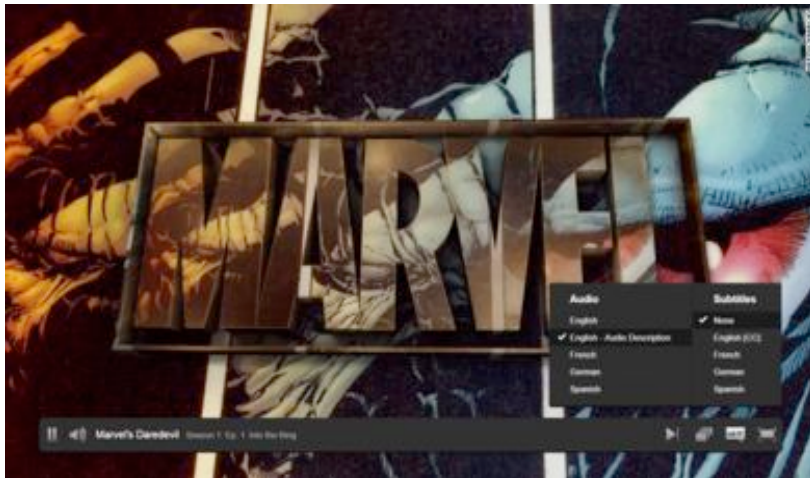
“enjoying the ride”



# Automatic Image Captioning

Can be useful for:

- AI agent that can see and *talk*
- automatic summary of your photo album
- image search with complex natural language queries
  - e.g., find all images with a man with a backpack entering a red car
- equal web access for visually impaired



"In the middle of flickering pages of action comics, appears the logo 'Marvel' in bold letters."

- from the opening credit of "Daredevil"

# Automatic Image Captioning

Can be useful for:

- AI agent that can see and *talk*
- automatic summary of your photo album
- image search with complex natural language queries
  - e.g., find all images with a man with a backpack entering a red car
- equal web access for visually impaired



In this painting, dozens of irises rise up in waves of color, like green and blue flames fanned by a wind that blows them, now flattens them, ... On the left, a solitary white iris commands the swirl of purple and green from its outpost ...

- example from [artbeyondsight.org](http://artbeyondsight.org)



# How to obtain rich annotations?

- Label them all (by asking human workers)
  - Flickr 30K
  - MSR CoCo --- 100K images with 5 captions each
- Learn from data in the wild
  - Facebook alone has over 250 billion images as of Jun 2013, with 350 million images added daily by over 1 billion users
  - Flickr has over 2 billion images
  - Data available at a significantly larger scale
  - And significantly noisier

## Example annotations in the CoCo dataset



- the man, the young girl, and dog are on the surfboard.
- a couple of people and a dog in the water.
- people and a dog take a ride on a surfboard
- a man holding a woman and a dog riding the same surfboard.
- a man holding a woman by her inner thighs on top of a surfboard over a small dog in a pink life jacket in the ocean.

# Flickr captions are noisier (some better examples)



- Dad, daughter and doggie tandem surf ride
- I believe this was a world record with two humans and 7 dogs...

- Oh no... here we go
- Surrounded by splash
- Pulling through
- Tada!
- Nani having a good time
- Digging deep



# Related Work

Compose using only detected words

- Yao et al. (2010)
- Kulkarni et al. (2011)
- Yatkarn et al (2014)
- Thomason et al (2014)
- Guadarrama et al (2013)

Compose using detected words + hallucinated words

- Yang et al. (2011)
- Li et al. (2011)
- Kuznetsova et al. (2012)
- Elliot and Keller (2013)
- Mitchell et al (2012)

Generation as whole sentence retrieval

- Farhadi et al. (2010)
- Ordonez et al. (2011)
- Socher et al. (2013)
- Hodosh et al. (2013)

Compose using retrieved text

- Kuznetsova et al. (2012)
- Mason (2013)
- Feng and Lapata (2013)

Deep learning variants

- Kiros et al 2014
- Fang et al 2015
- Chen et al 2015
- Xu et al 2015
- Donahue et al. 2015
- Karpathy et al 2015
- Mao et al 2014
- Vinyals et al 2015



More precise  
Fixed/small vocabulary  
Fixed / formulaic language

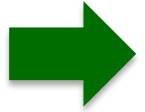
CVPR 2011

CoNLL 2011

ACL 2012,  
ACL 2013,  
TACL 2014

More expressive  
Open vocabulary  
Everyday people's language

# Plan for the talk



- BabyTalk
  - [CVPR 2011]
- TreeTalk
  - [TACL 2014, ACL 2013, ACL 2012]





“This picture shows one person,





“This picture shows one person, one grass,



“This picture shows one person, one grass, one chair,



“This picture shows one person, one grass, one chair, and one potted plant.



“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass,



“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair.



“This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant.”

# Methodology Overview



Input Image



a) dog

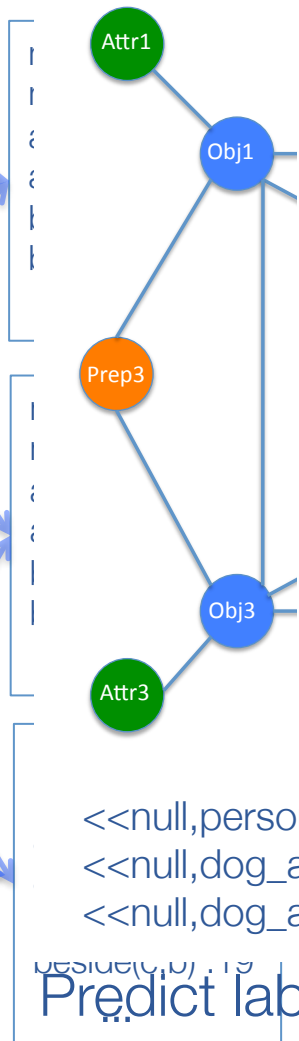


b) person



c) sofa

Extract Objects/stuff, smoothed with text potentials  
Predict attributes



This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

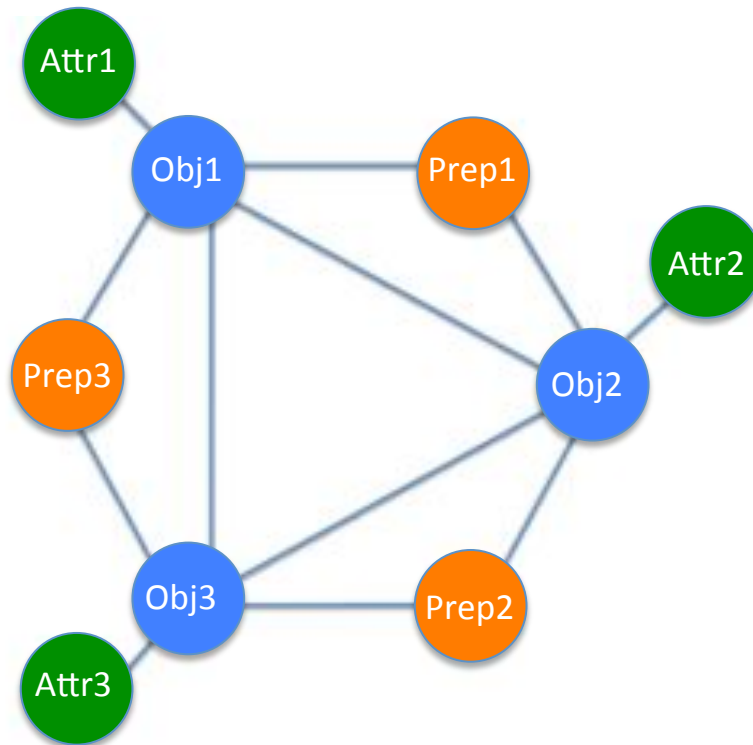
<<null, person\_b>, against, <brown, sofa\_c>>  
 <<null, dog\_a>, near, <null, person\_b>>  
 <<null, dog\_a>, beside, <brown, sofa\_c>>

Generate natural language description

Predict labeling – vision potentials

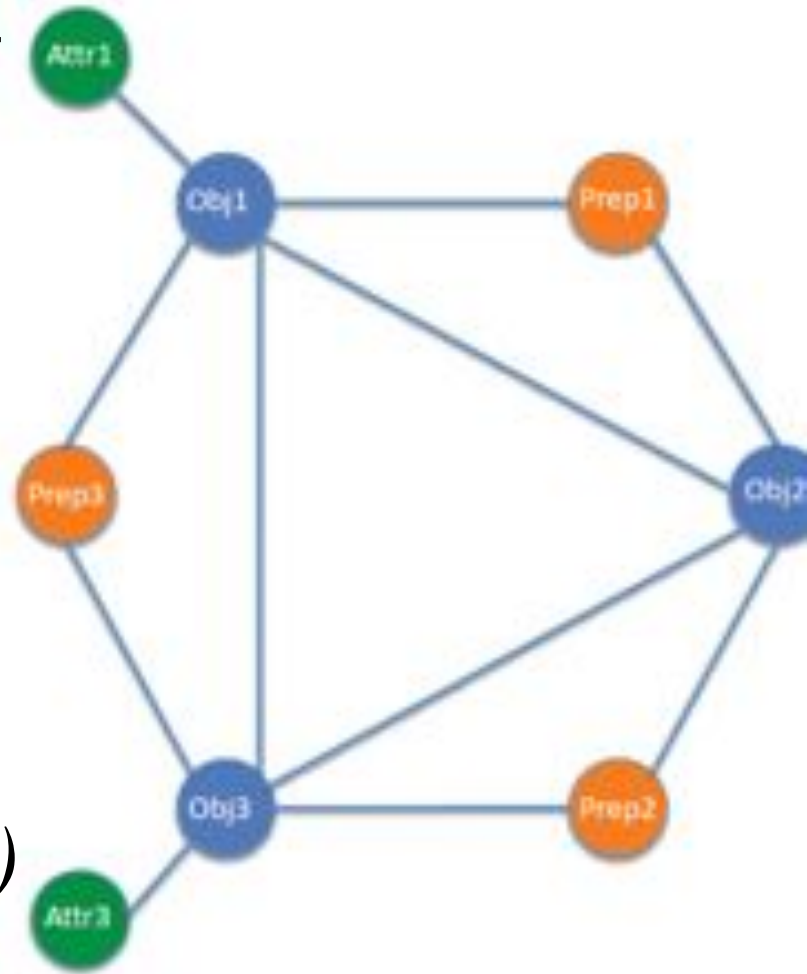
smoothed with text potentials

# Conditional Random Fields (CRF)





# Potential Functions for CRF



unary  
potentials

$$\psi(\text{object}_i)$$

$$\psi(\text{attribute}_i)$$

$$\psi(\text{preposition}_{ij})$$

relational  
(binary &  
ternary)  
potentials

$$\psi(\text{attribute}_i, \text{object}_i)$$

$$\psi(\text{object}_i, \text{preposition}_{ij}, \text{object}_j)$$

# Potential Functions for CRF

Practical challenge of relational potentials:



observing all possible combinations of variables unlikely  
(limited corpus with detailed visual annotations)

unary  
potentials

$$\psi(\text{object}_i)$$

$$\psi(\text{attribute}_i)$$

$$\psi(\text{preposition}_{ij})$$

relational  
(binary &  
ternary)  
potentials

$$\psi(\text{attribute}_i, \text{object}_i)$$

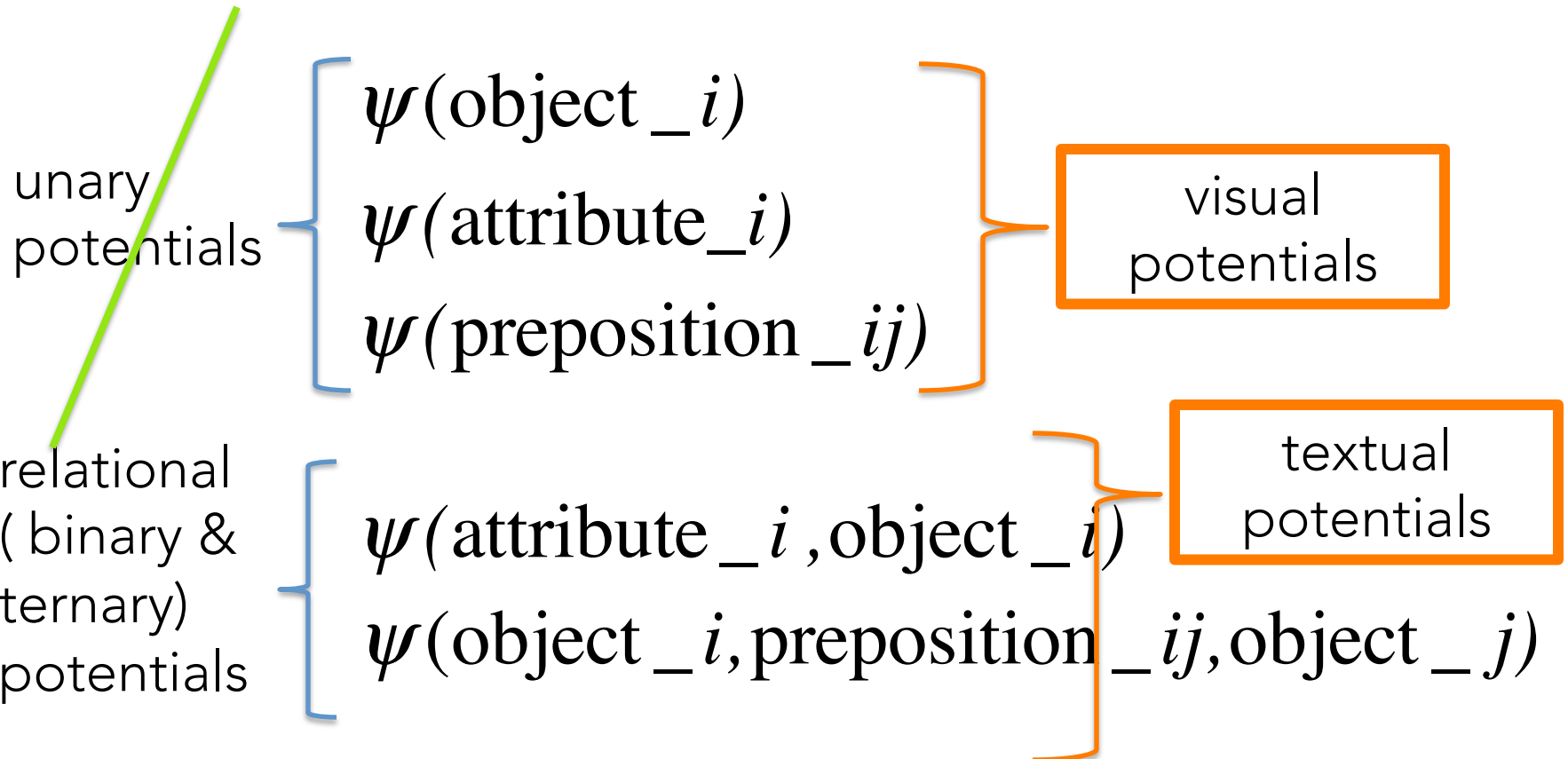
$$\psi(\text{object}_i, \text{preposition}_{ij}, \text{object}_j)$$

# Potential Functions for CRF

Practical challenge of relational potentials:



observing all possible combinations of variables unlikely  
(limited corpus with detailed visual annotations)



**Learning:** mixture coefficients of different types of potentials (grid search)

**Inference:** Tree Re-Weighted message passing (TRW-S) (Kolmogorov 2006)

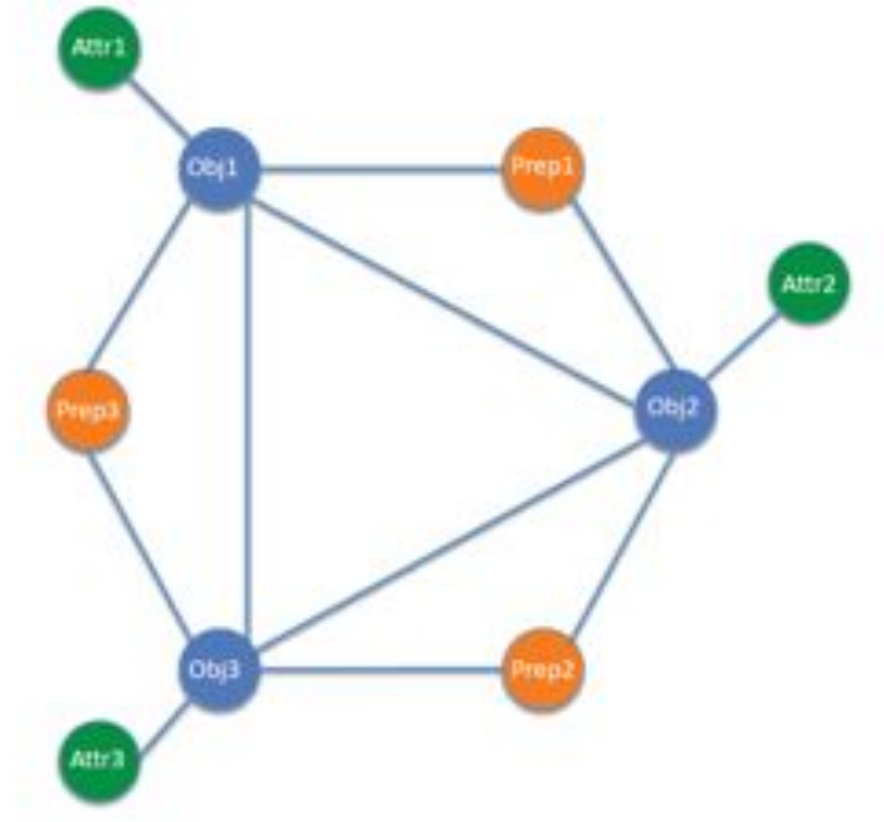


unary potentials {  $\psi(\text{object}_i)$   
 $\psi(\text{attribute}_i)$   
 $\psi(\text{preposition}_{ij})$

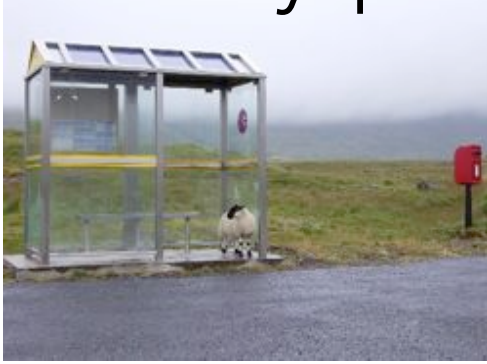
relational (binary & ternary) potentials {  $\psi(\text{attribute}_i, \text{object}_i)$   
 $\psi(\text{object}_i, \text{preposition}_{ij}, \text{object}_j)$

# Generation (aka "surface realization")

Template filling (traversing the graph and reading off the detected objects, attributes, and their spatial relations in sequence)



# Cherry-picked examples



This is a picture of one sky, one road and one sheep. The gray sky is over the gray road. The gray sheep is by the gray road.



Here we see one road, one sky and one bicycle. The road is near the blue sky, and near the colorful bicycle. The colorful bicycle is within the blue sky.

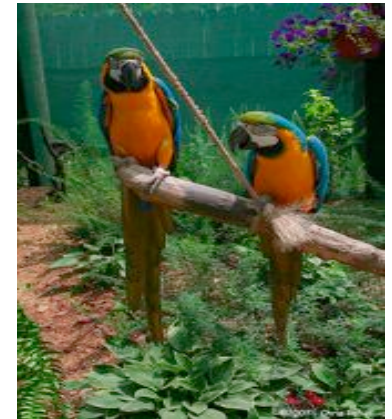


This is a picture of two dogs. The first dog is near the second furry dog.

# Lemons



There are one road and one cat. The furry road is in the furry cat.



This is a picture of one tree, one road and one person. The rusty tree is under the red road. The colorful person is near the rusty tree, and under the red road.

# Computer vs Human Generated Caption



**Computer:** "This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant."



**Human (UIUC Pascal dataset):**

- A. A Lemonade stand **is manned by** a blonde child with a cookie.
- B. A small child at a lemonade and cookie stand **on a city corner.**
- C. Young child behind lemonade stand **eating a cookie.**

- (1) formulaic, robotic and unnatural  
(2) limited semantic expressiveness, especially, no verb except "be" verb



How can we reduce the gap between these two?

Computer: "This picture shows one person, one grass, one chair, and one potted plant. The person is near the green grass, and in the chair. The green grass is by the chair, and near the potted plant."

Human (UIUC Pascal dataset):

- A. A Lemonade stand **is manned by** a blonde child with a cookie.
- B. A small child at a lemonade and cookie stand **on a city corner.**
- C. Young child behind lemonade stand **eating a cookie.**



# How can we scale up the range of descriptive words and phrases?



"Butterfly and flower"

"A butterfly **attracted** to flowers"

"Butterfly **sipping nectar** from the flower"

"Butterfly **feeding on** a flower"

"A butterfly **having lunch**"



# How can we scale up the range of descriptive words and phrases?

Two Challenges:

- **recognition**: we don't have descriptive-verb recognizers at scale.  
e.g., "attracted\_recognizer", "feeding\_on\_recognizer"
- **formalism**: not easy for humans to formalize all these variations of meanings in symbolic meaning representation and annotate them

"Butterfly **feeding on** a flower"

"A butterfly **having lunch**"



# How can we scale up the range of descriptive words and phrases?

Two Challenges:

- **recognition**: we don't have descriptive-verb recognizers at scale.  
e.g., "attracted\_recognizer", "feeding\_on\_recognizer"
- **formalism**: not easy for humans to formalize all these variations of meanings in symbolic meaning representation and annotate them

Reflection on BabyTalk:

Humans decide

{what can be described} = {what can be recognized}

"Butterfly **feeding on** a flower"

"A butterfly **having lunch**"



Web  
in 1995

**MONEY & INVESTING UPDATE**  
*from THE WALL STREET JOURNAL*

Front Page | **STOCKS** | U.S. | Small U.S. | Americas | Asia | Europe | Heard on the Street | Credit Markets | Foreign Exchange | Commodities | Mutual Funds

Wednesday, September 6, 1995

### What's News —

Business and Finance

#### MARKETS DIARY 5 p.m. EDT

DJIA	9082.81	+ 31.78
S&P 500	401.38	+ 0.38
Russell Composite	401.38	+ 0.40
Nikkei (Tokyo)	17911	+ 0.90
London (FTSE 100)	3728	+ 0.20
30-Day Treasury T-bill	4.50%	
Japanese yen (per \$100)	98.82	
German mark (per \$100)	1.4768	

### Computer Shares Lift Stocks Again; Bonds Are Weak

By DAVE PETTIT  
*Money & Investing Update*



## Welcome to Amazon.com Books!

*One million titles,  
consistently low prices.*

(If you explore just one thing, make it our personal notification service. We think it's very cool!)

### SPOTLIGHT! -- AUGUST 16TH

These are the books we love, offered at Amazon.com low prices. The spotlight moves EVERY day so please come often.

### ONE MILLION TITLES

Search Amazon.com's [million title catalog](#) by author, subject, title, keyword, and more... Or take a look at the [books we recommend](#) in over 20 categories... Check out our [customer reviews](#) and the [award winners](#) from the Hugo and Nebula to the Pulitzer and Nobel... and [bestsellers](#) are 30% off the publishers list...

### EYES & EDITORS, A PERSONAL NOTIFICATION SERVICE

Like to know when that book you want comes out in paperback or when your favorite author

# Web Today: Increasingly Visual

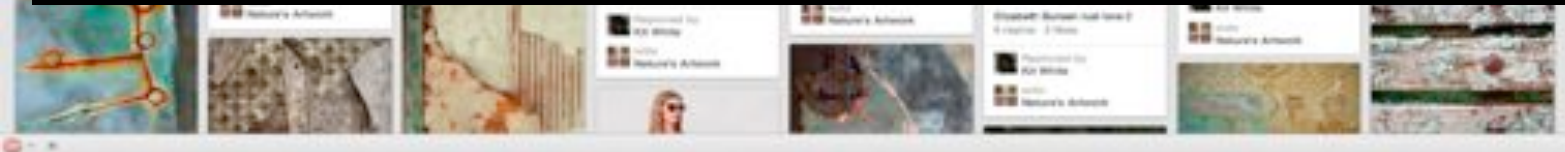
-- social media, news media, online shopping

flickr

Pinterest



- Facebook.com has over 250 billion images uploaded as of Jun 2013
- 1.15 billion users uploading 350 million images a day on average



# How can we scale up the range of descriptive words and phrases?

Two Challenges:

- **recognition**: we don't have descriptive-verb recognizers at scale.  
e.g., "attracted\_recognizer", "feeding\_on\_recognizer"
- **formalism**: not easy for humans to formalize all these variations of meanings in symbolic meaning representation and annotate them

Reflection on BabyTalk:

{what can be described} = {what can be recognized}

Humans decide

"Butterfly **feeding on** a flower"

"A butterfly **having lunch**"



# How can we scale up the range of descriptive words and phrases?

Two Challenges:

- **recognition**: we don't have descriptive-verb recognizers at scale.  
e.g., "attracted\_recognizer", "feeding\_on\_recognizer"
- **formalism**: not easy for humans to formalize all these variations of meanings in symbolic meaning representation and annotate them

Reflection on BabyTalk:

{what can be described} = {what can be recognized}

Key Idea:

{what can be described}  $\supset$  {what can be recognized}

~ Farhadi et al. 2010

Data decides

Distributional Hypothesis (Harris 1954)

"Butterfly **feeding on** a flower"

"A butterfly **having lunch**"

Humans decide



# Plan for the talk

- BabyTalk
  - [CVPR 2011]
- TreeTalk
  - [TACL 2014, ACL 2013, ACL 2012]





# Operational Overview

Given a query image (& an object)



① **Harvest** tree branches

② **Compose** a new tree by combining tree branches



1,000,000 (image, caption)

SBU Captioned Photo Dataset  
(Ordonez et al. 2011)

# Description Generation

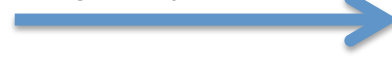


Object appearance



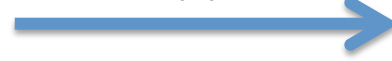
NP: the dirty sheep

Object pose



VP: meandered along a desolate road

Scene appearance



PP: in the highlands of Scotland

Region  
appearance &  
relationship



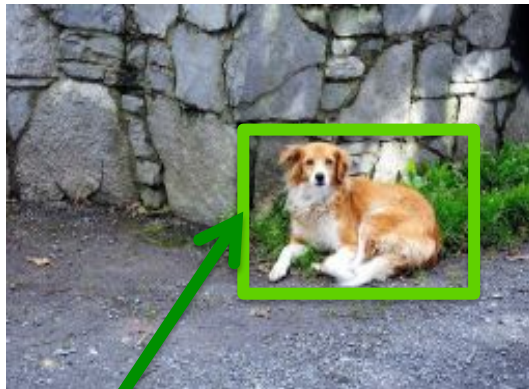
PP: through frozen grass



Example Composition:

the dirty sheep meandered along a desolate road in the highlands of Scotland through frozen grass

# Retrieving VPs



Contented dog just laying on the edge of the road in front of a house..



Peruvian dog sleeping on city street in the city of Cusco, (Peru)

~ Distributional Hypothesis (Harris 1954)

Detect: dog

Find matching detections by pose similarity

--- using **color**, **texton**, **HoG** and **SIFT**



this dog was laying in the middle of the road on a back street in jaco



Closeup of my dog sleeping under my desk.

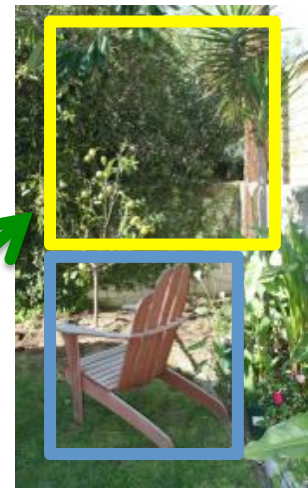
# Retrieving PPstuff

Find matching regions by appearance + arrangement similarity

--- using **color**, **texon**, **HoG** and **SIFT**



Cordoba - lonely elephant **under an orange tree**...



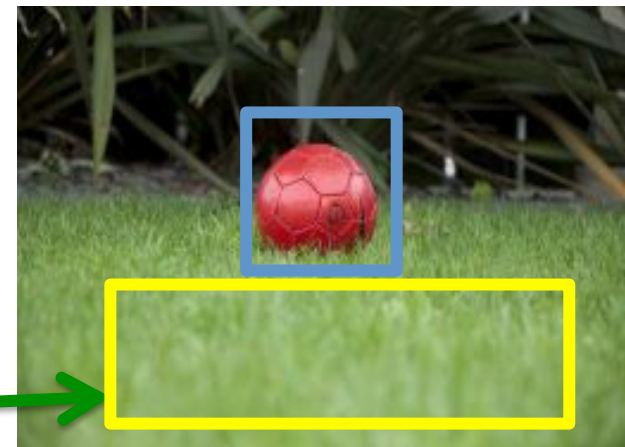
I positioned the chairs **around the lemon tree** -- it's like a shrine



Detect: stuff



Comfy chair **under a tree**.



Mini Nike soccer ball all alone **in the grass**

# Operational Overview

Given a query image

① **Harvest** tree branches

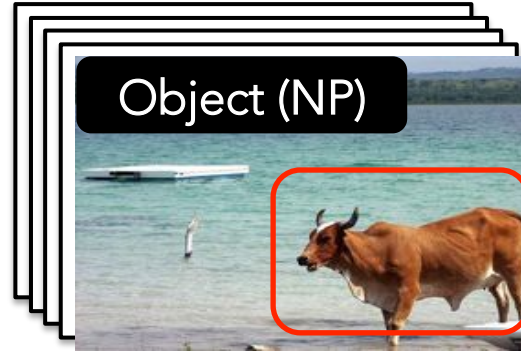
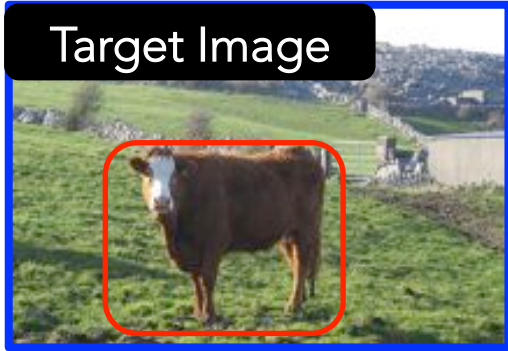
➔ ② **Compose** a new tree by combining tree branches



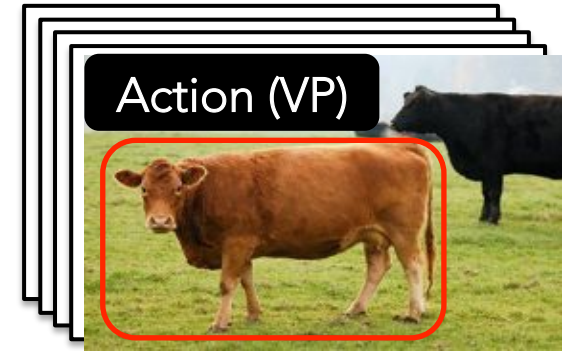
1,000,000 (image, caption)

SBU Captioned Photo Dataset

# Input to Sentence Composition :=



A cow



was staring at me



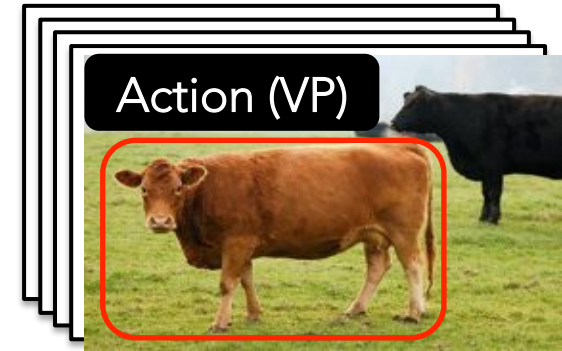
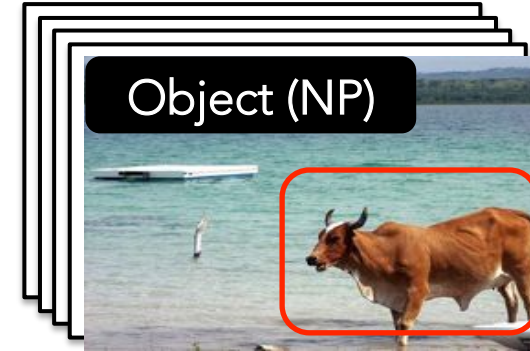
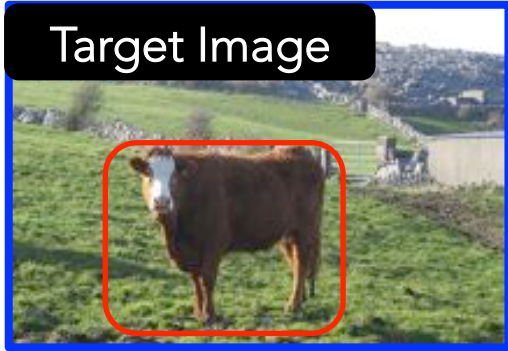
in the grass



in the countryside

# Sentence Composition :=

1. Select a subset of harvested phrases
2. Decide the ordering of the selected phrases



A cow  
in the grass  
was staring at me  
in the countryside

A cow

was staring at me



A cow  
was staring at me  
~~in the grass~~  
in the countryside

in the grass

in the countryside

# Sentence Composition :=

1. Select a subset of harvested phrases
2. Decide the ordering of the selected phrases

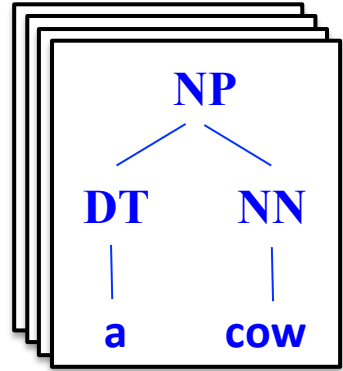
Tree Structure --- Probabilistic Context Free Grammars (PCFG)



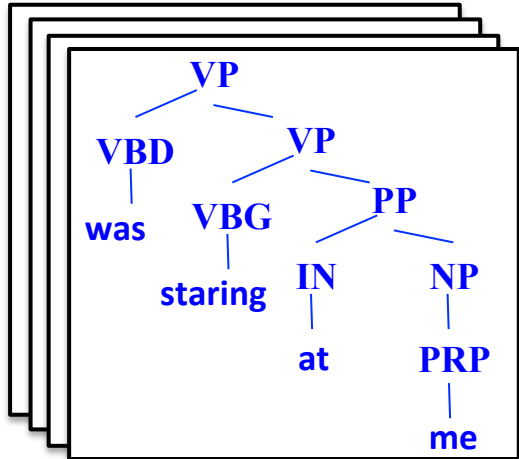
A cow  
in the grass  
was staring at me  
in the countryside

A cow  
was staring at me  
~~in the grass~~  
in the countryside

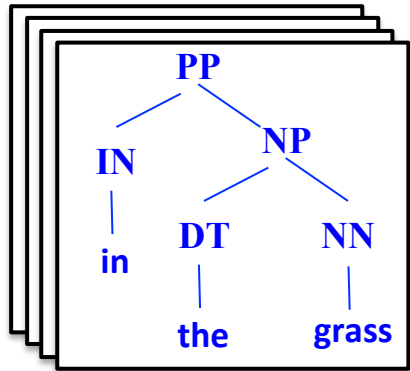
## Object (NP)



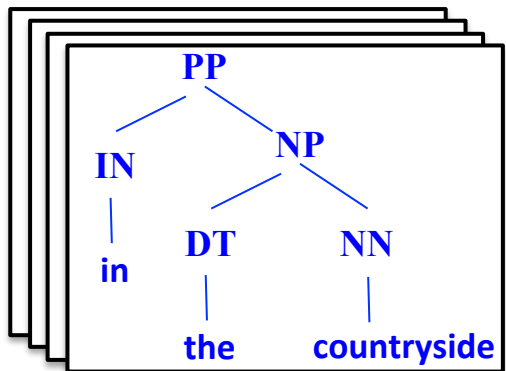
## Action (VP)



## Stuff (PP)



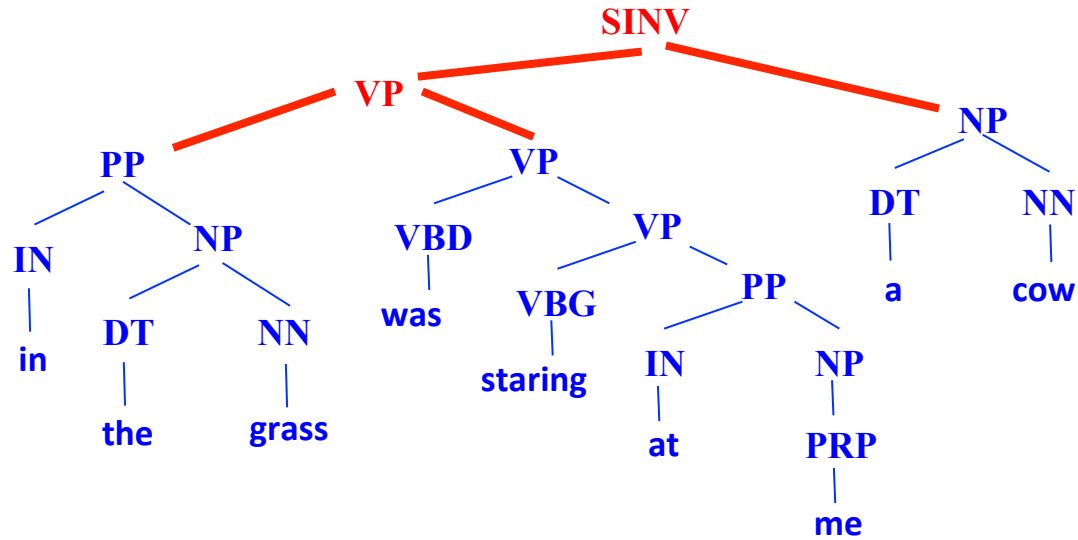
## Scene (PP)





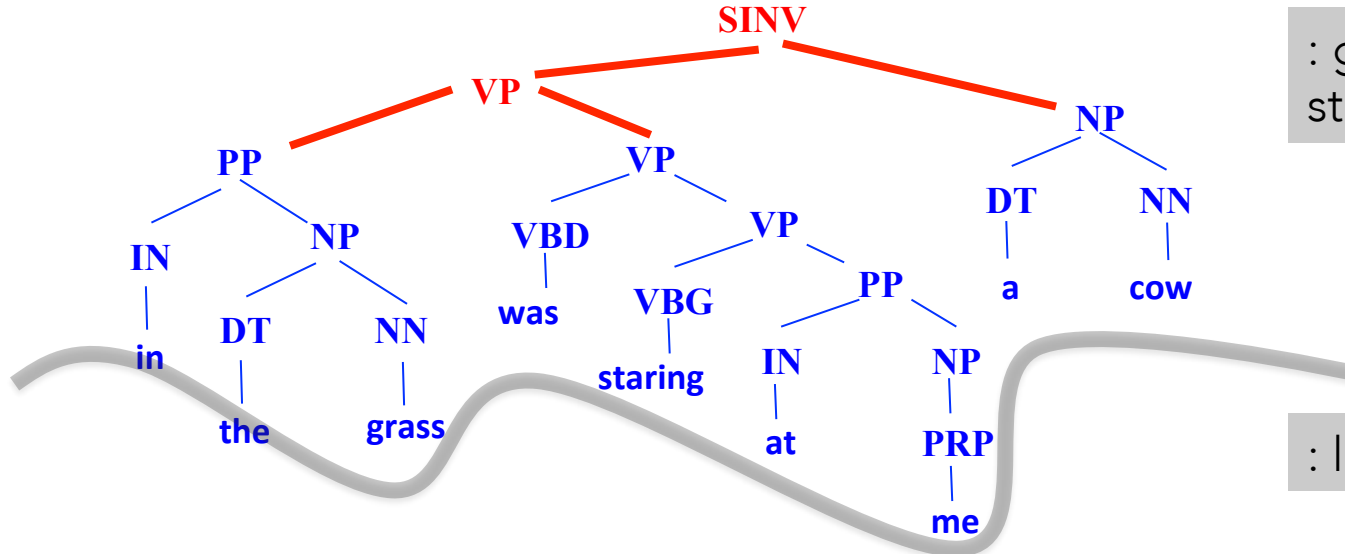
# Sentence Composition :=

In the grass --- was staring at me --- a cow



# Sentence Composition :=

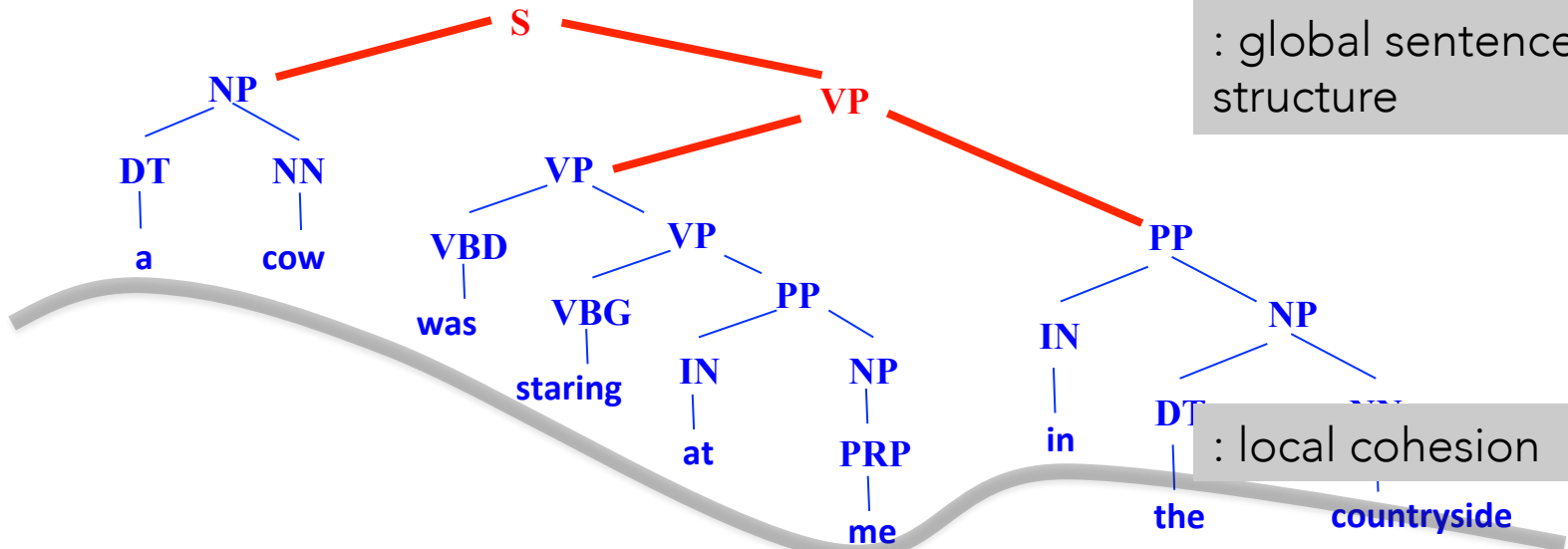
In the grass --- was staring at me --- a cow



: global sentence structure

: local cohesion

A cow --- was staring at me --- in the countryside

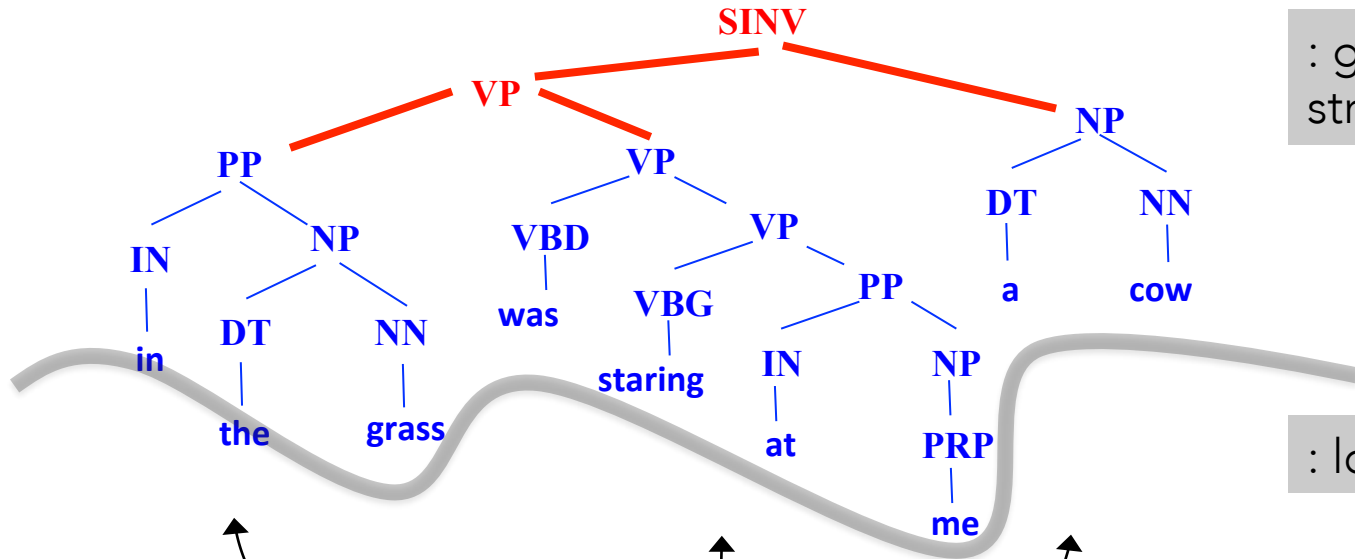


: global sentence structure

: local cohesion

# Sentence Composition :=

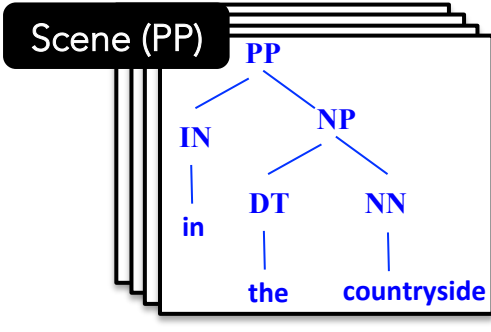
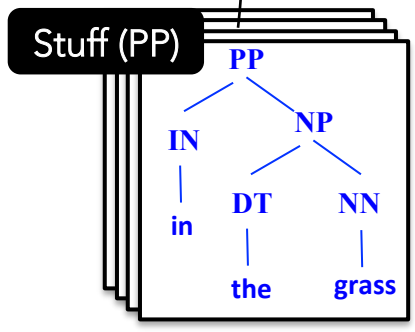
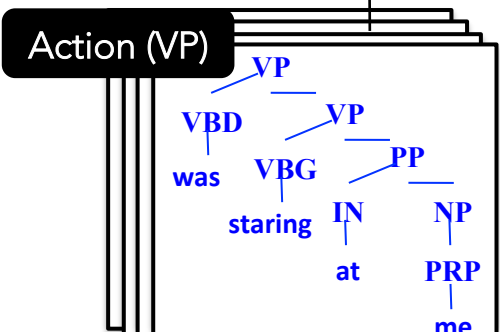
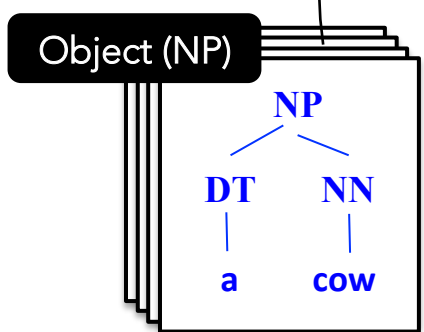
In the grass --- was staring at me --- a cow



: global sentence structure

: local cohesion

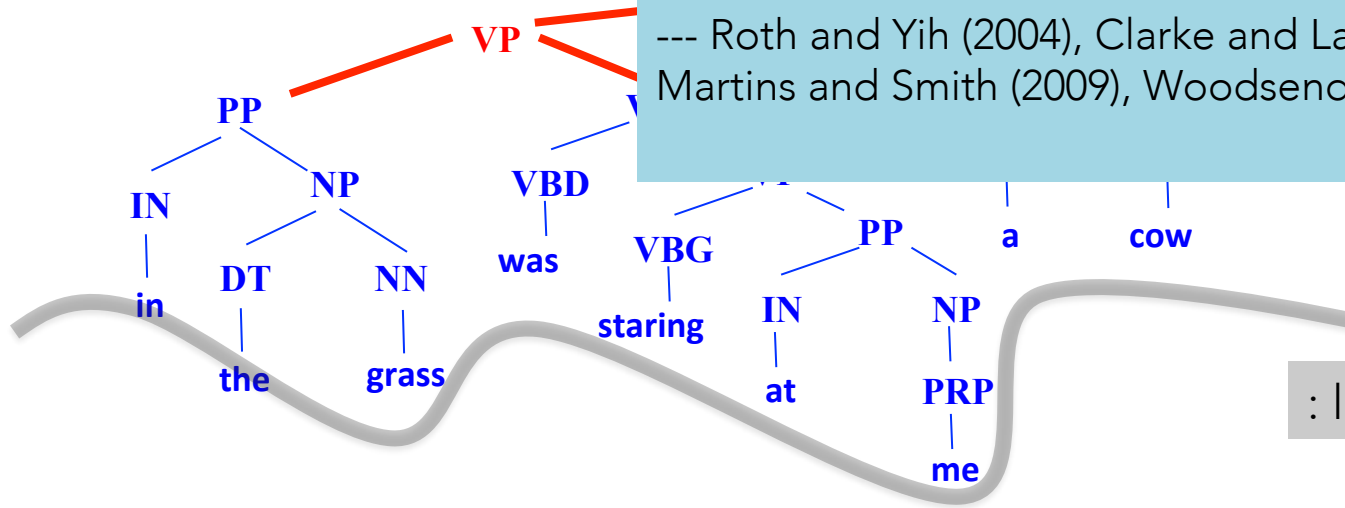
→ different from parsing because we must consider different choices of subtree selection and re-ordering simultaneously



# Sentence Composition as Constraint Optimization using Integer Linear Programming

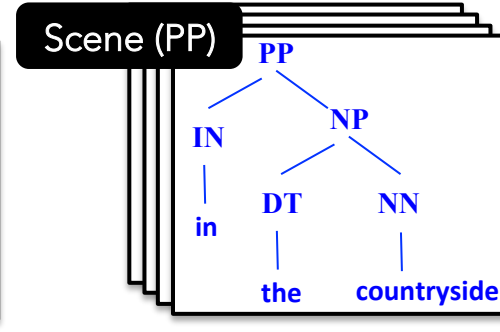
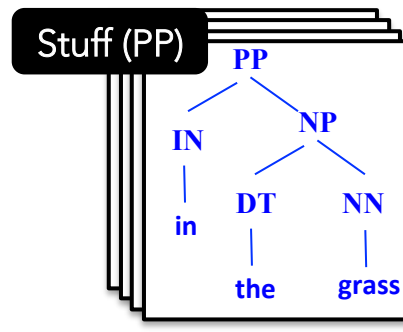
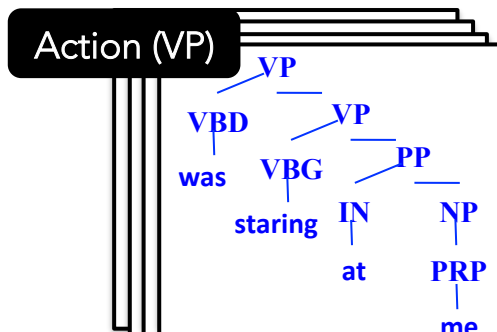
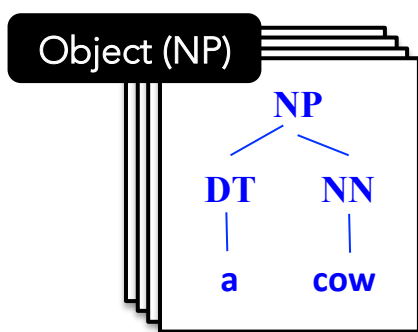
In the grass --- was staring

--- Roth and Yih (2004), Clarke and Lapata (2006), Martins and Smith (2009), Woodsend and Lapata(2010)

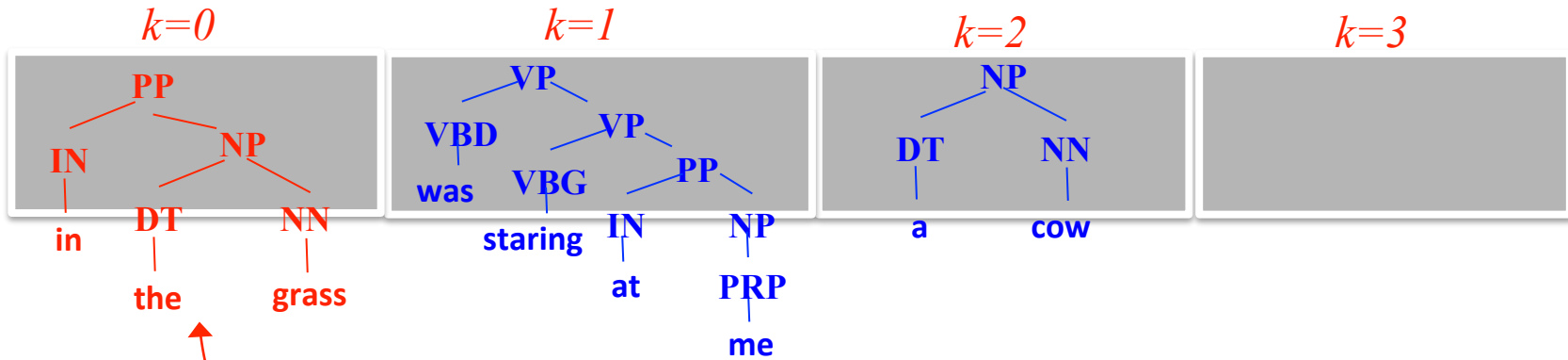


: local cohesion

- different from parsing because we must consider different choices of subtree selection and re-ordering simultaneously
- finding the optimum selection+ordering = NP-hard (~= TSP)



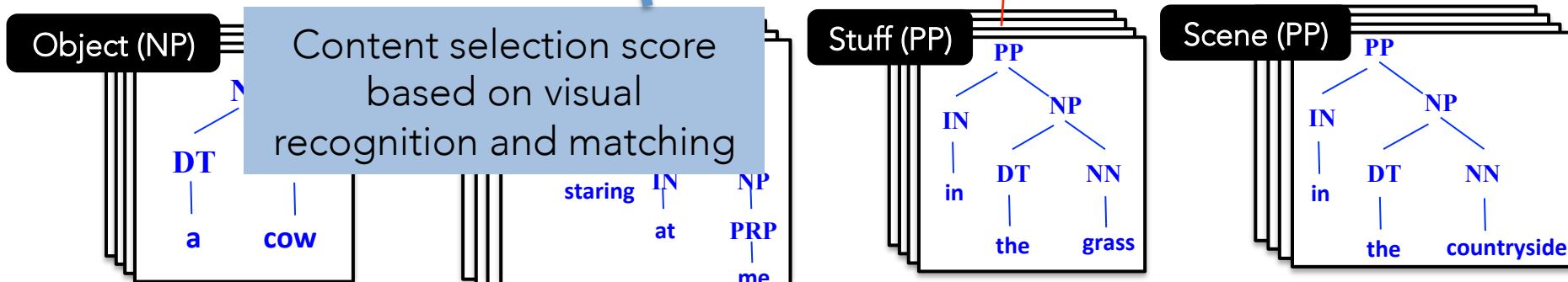
# Sentence Composition as Constraint Optimization using Integer Linear Programming



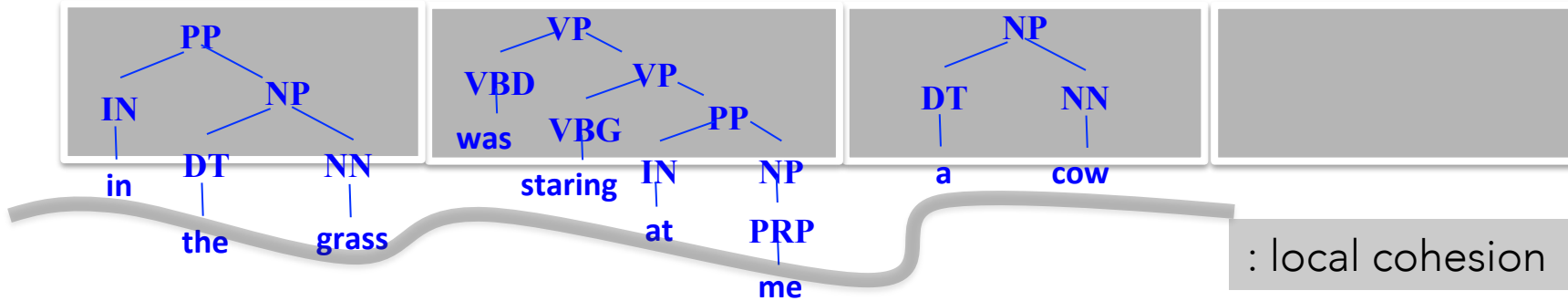
decision variable:  $\alpha_{ijk} = 1$  iff phrase  $i$  of type  $j$  selected for position  $k \in [0, N)$

objective function: 
$$F = \sum_{ij} F_{ij} \times \sum_{k=0}^{N-1} \alpha_{ijk}$$

*i*'th phrase from *Stuff(PP)*-type



# Sentence Composition as Constraint Optimization using Integer Linear Programming

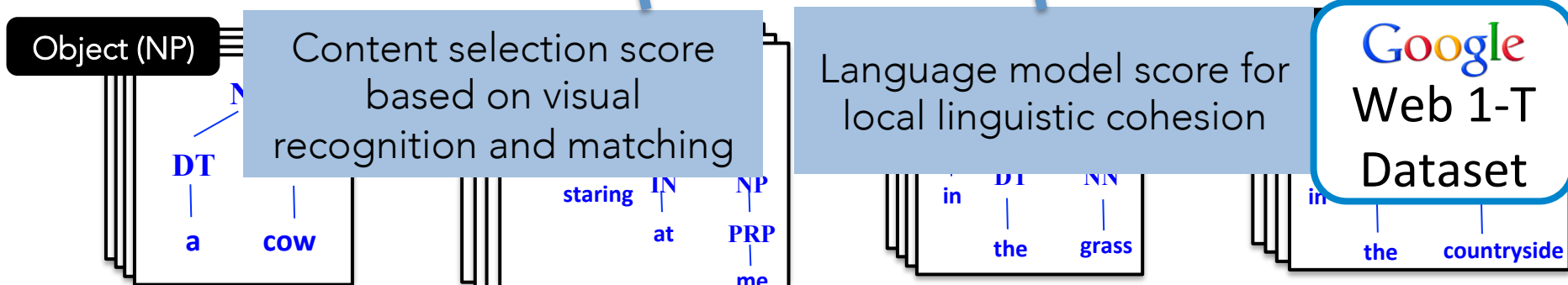


decision variable:  $\alpha_{ijk} = 1$  iff phrase  $i$  of type  $j$  selected for position  $k \in [0, N)$

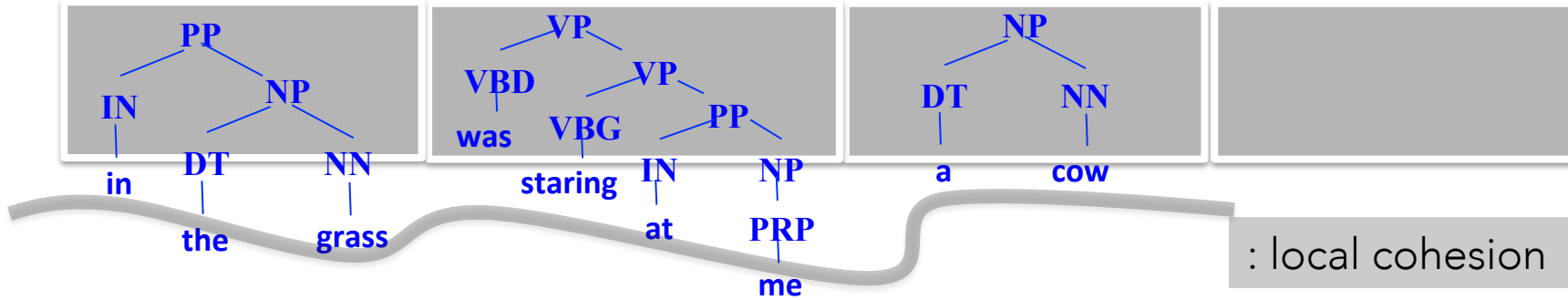
~ ACL 2012 system

$$\alpha_{ijkpq(k+1)} = 1 \text{ iff } \alpha_{ijk} = \alpha_{pq(k+1)} = 1$$

objective function:  $F = \sum_{ij} F_{ij} \times \sum_{k=0}^{N-1} \alpha_{ijk} + \sum_{ijpq} F_{ijpq} \times \sum_{k=0}^{N-2} \alpha_{ijkpq(k+1)}$



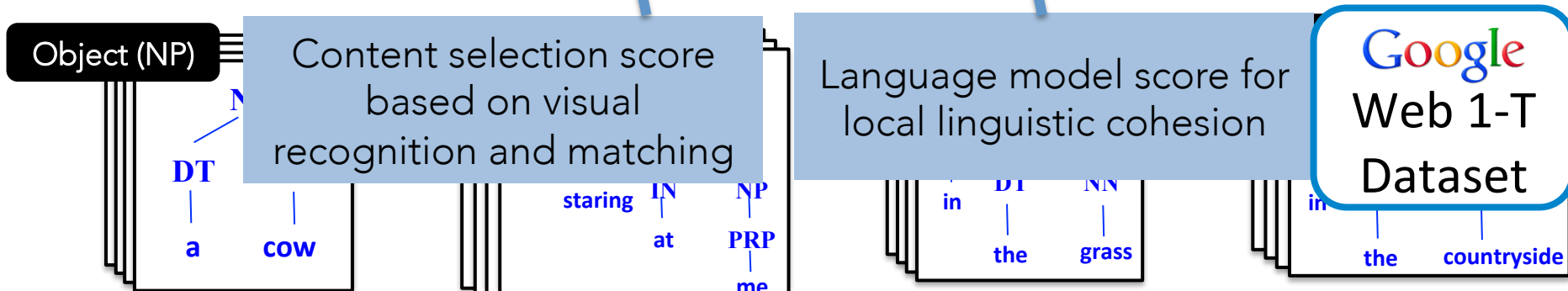
# Sentence Composition as Constraint Optimization using Integer Linear Programming



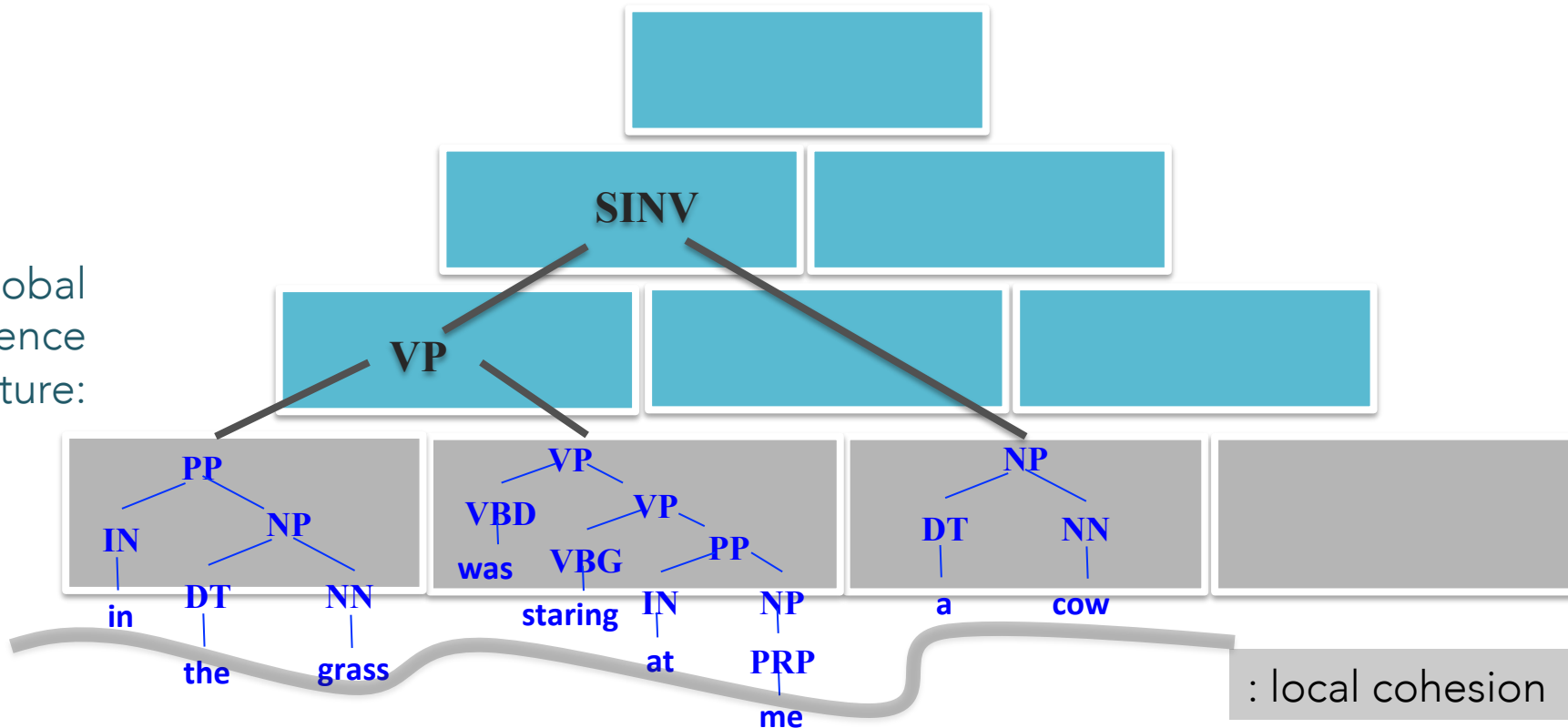
decision variable:  $\alpha_{ijk} = 1$  iff phrase  $i$  of type  $j$  selected for position  $k \in [0, N)$

$$\alpha_{ijkpq(k+1)} = 1 \text{ iff } \alpha_{ijk} = \alpha_{pq(k+1)} = 1$$

objective function: 
$$F = \sum_{ij} F_{ij} \times \sum_{k=0}^{N-1} \alpha_{ijk} + \sum_{ijpq} F_{ijpq} \times \sum_{k=0}^{N-2} \alpha_{ijkpq(k+1)}$$



global sentence structure:



decision variable:  $\alpha_{ijk} = 1$  iff phrase  $i$  of type  $j$  selected for position  $k \in [0, N)$

$\alpha_{ijkpq(k+1)} = 1$  iff  $\alpha_{ijk} = \alpha_{pq(k+1)} = 1$

objective function: 
$$F = \sum_{ij} F_{ij} \times \sum_{k=0}^{N-1} \alpha_{ijk} + \sum_{ijpq} F_{ijpq} \times \sum_{k=0}^{N-2} \alpha_{ijkpq(k+1)}$$

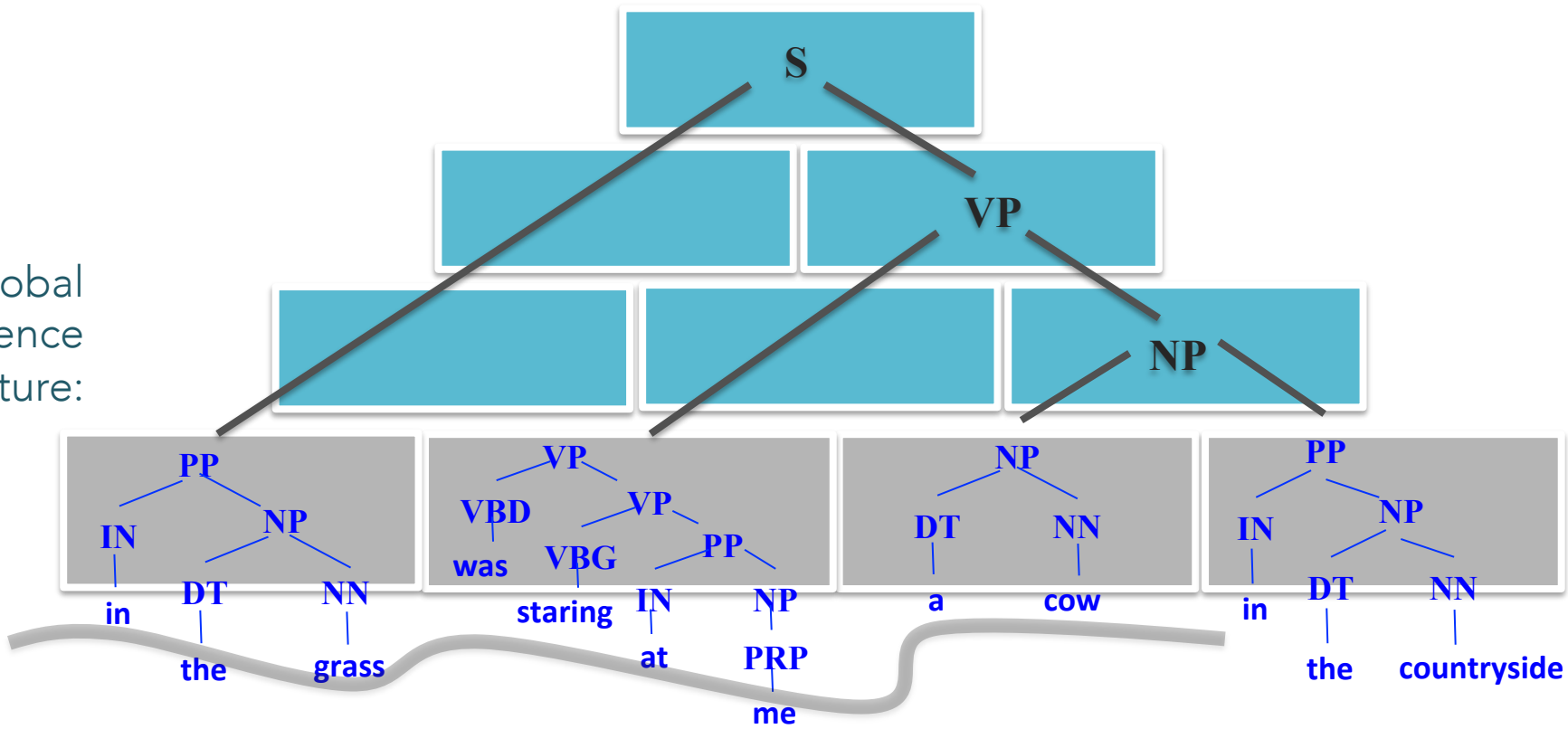
Object (NP)

Content selection score based on visual

Language model score for local linguistic cohesion



global sentence structure:



decision variable:  $\alpha_{ijk} = 1$  iff phrase  $i$  of type  $j$  selected for position  $k \in [0, N)$

$\alpha_{ijkpq(k+1)} = 1$  iff  $\alpha_{ijk} = \alpha_{pq(k+1)} = 1$

objective function: 
$$F = \sum_{ij} F_{ij} \times \sum_{k=0}^{N-1} \alpha_{ijk} + \sum_{ijpq} F_{ijpq} \times \sum_{k=0}^{N-2} \alpha_{ijkpq(k+1)}$$

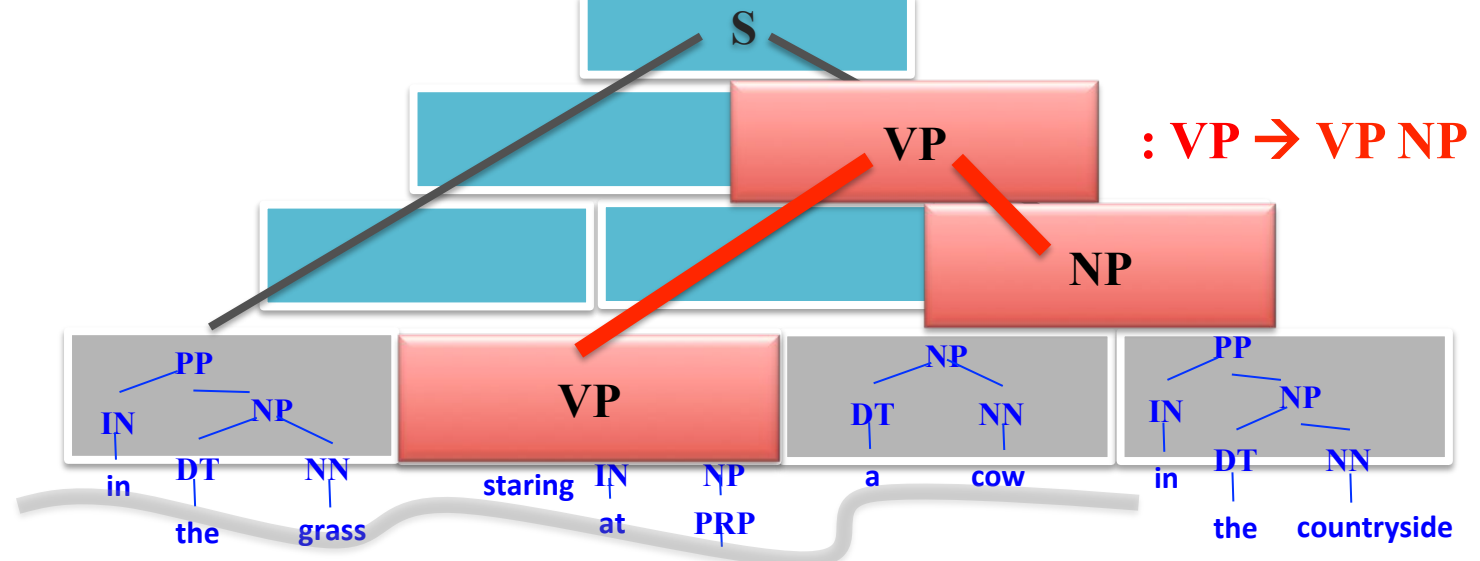
Object (NP)

Content selection score based on visual

Language model score for local linguistic cohesion



global sentence structure:



: VP → VP NP

decision variable:

$$\alpha_{ijk} = 1 \quad \text{iff} \quad \text{phrase } i \text{ of type } i \text{ selected for position } k$$

$$\alpha_{ijkpq(k+1)} = 1 \quad \text{iff} \quad \alpha_{ijk} =$$

Language model score for global parse tree structure

$$\beta_{ijs} = 1 \quad \text{iff} \quad \text{cell } ij \text{ of the matrix is assigned with PCFG tag } s$$

$$\beta_{ijk} = 1 \quad \text{iff} \quad \beta_{ijh} = \beta_{ikp} = \beta_{(k+1)jq} = 1$$

$$+ \sum_{ij} \sum_{k=i}^{j-1} \sum_{r \in R} F_r \times \beta_{ijk}$$

$$\text{objective function: } F = \sum_{ij} F_{ij} \times \sum_{k=0}^{N-1} \alpha_{ijk} + \sum_{ijpq} F_{ijpq} \times \sum_{k=0}^{N-2} \alpha_{ijkpq(k+1)}$$

Object (NP)

Content selection score based on visual

Language model score for local linguistic cohesion



# Sentence Composition

# as Constraint Optimization using Integer Linear Programming

## Constraints:

Consistency between  
sequence variables -----  $\alpha_{ijk}$   
& tree leaf variables -----  $\beta_{ijs}$

$$\forall_{ijk}, \alpha_{ijk} \leq \sum_{s \in S^j} \beta_{kks}$$

$$\forall_k, \sum_{ij} \alpha_{ijk} = \sum_{s \in S} \beta_{kks}$$

Valid PCFG parse tree

$$\forall_{ij}, \sum_{s \in S} \beta_{ijs} \leq 1$$

$$\forall_{i,j>i,h}, \beta_{ijh} = \sum_{k=i}^{j-1} \sum_{r \in R_h} \beta_{ijkr}$$

$R_h = \{r \in R : r = h \rightarrow pq\}$

$$\forall_{k \in [1, N)}, \sum_{s \in S} \beta_{kks} \leq \sum_{t=k}^{N-1} \sum_{s \in S} \beta_{0ts}$$

$$\forall_{ij} \sum_k \gamma_{ijk} \leq 1$$

## Objective function:

$$F = \sum_{ij} F_{ij} \times \sum_{k=0}^{N-1} \alpha_{ijk}$$

(Content selection ~ Visual Rec)

$$+ \sum_{ijpq} F_{ijpq} \times \sum_{k=0}^{N-2} \alpha_{ijkpq(k+1)}$$

(Sequential cohesion ~ Lang Model)

$$+ \sum_{ij} \sum_{k=i}^{j-1} \sum_{r \in R} F_r \times \beta_{ijkr}$$

(Tree structure ~ PCFG Model)

## Decision variable:

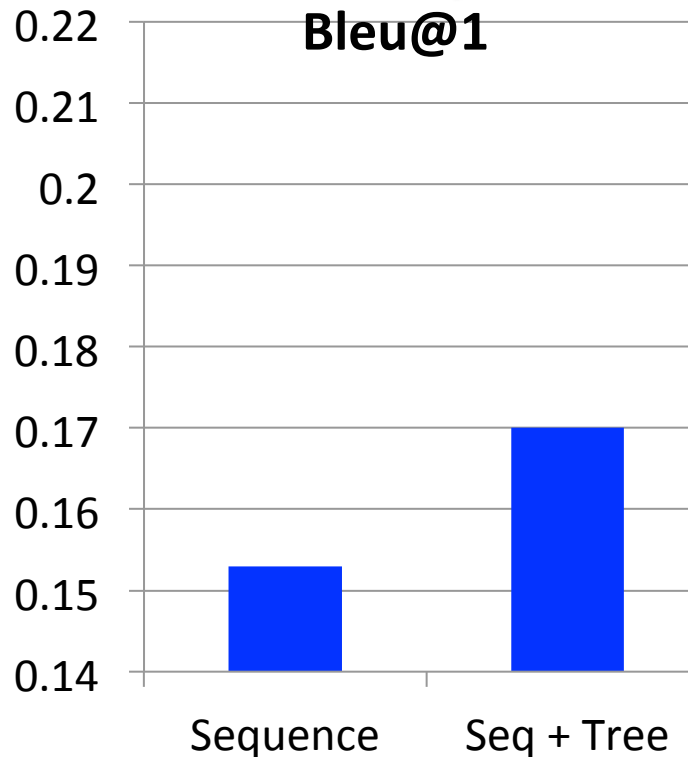
$\alpha_{ijk}$        $\alpha_{ijkpq(k+1)}$       (Sequential)

$\beta_{ijs}$        $\beta_{ijkr}$       (Tree structure)

# Automatic Evaluation

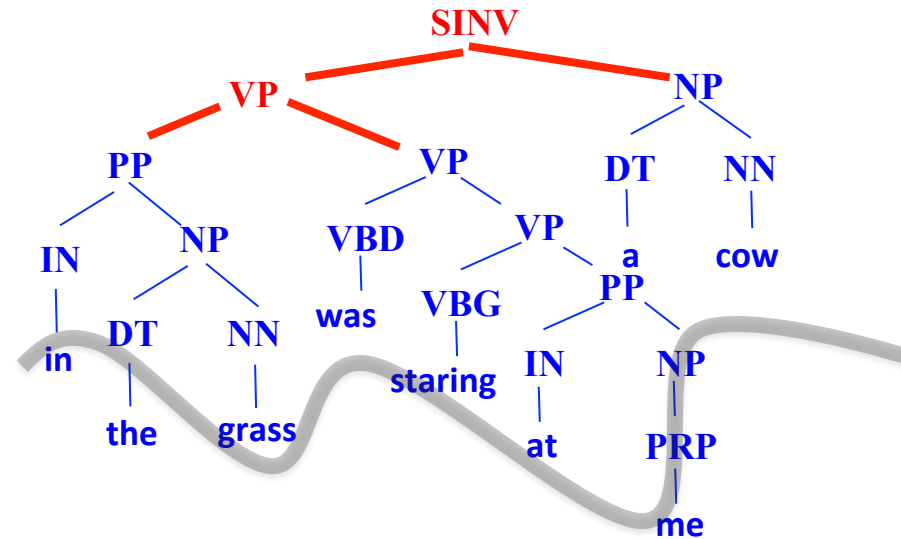
Machine Translation:  
From Images to Text

BLEU – N-gram precision  
(with modifications to handle degenerate cases)



~ ACL 2012 system

TREE: global sentence structure



SEQ: local cohesion

# Half-Successful Examples (to Motivate Tree Pruning)

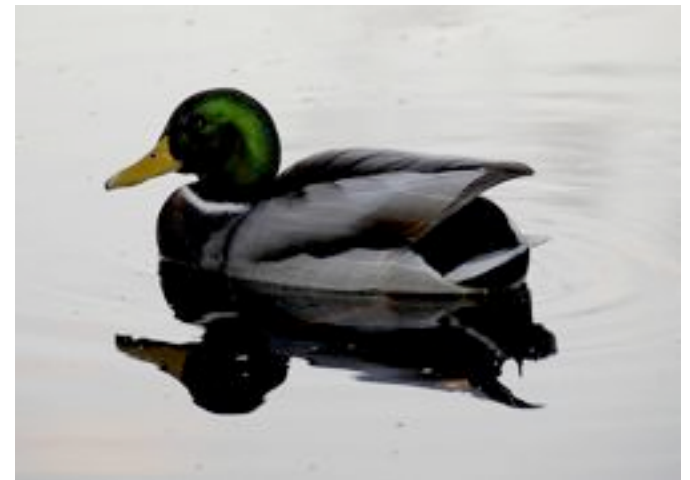


An old clock overlooks the old river bridge in Potes in Cantabria , Spain.

Harvested phrases contain overly extraneous information

→ generalize captions before extracting tree branches

Just a duck swimming in the river Des Peres in Heman Park , UNiversity City , Missouri - May 13 , 2008.



# Operational Overview

Given a query image

① **Harvest** tree branches

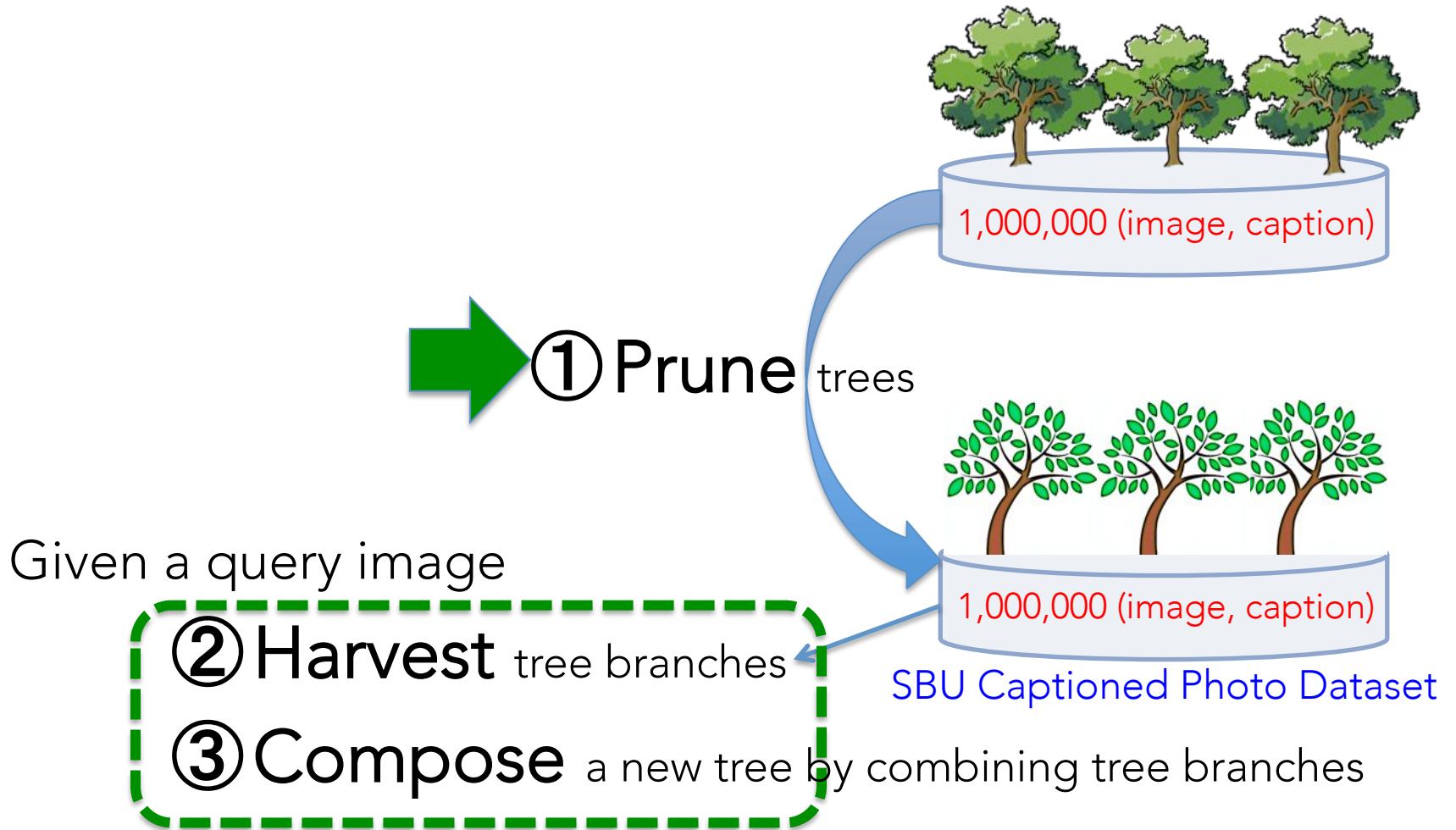
② **Compose** a new tree by combining tree branches



1,000,000 (image, caption)

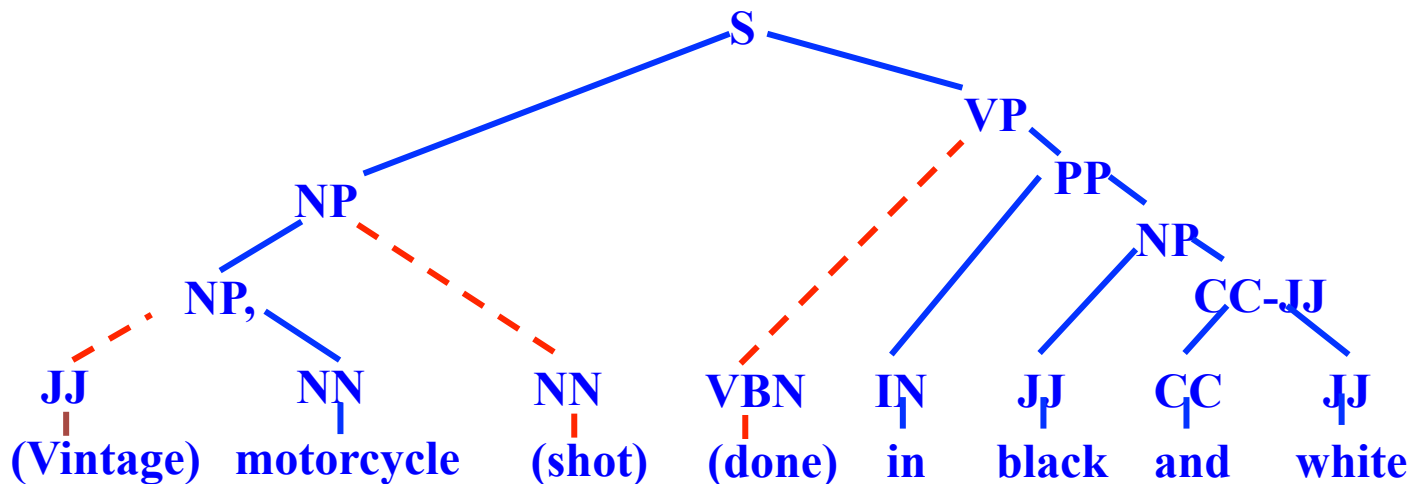
SBU Captioned Photo Dataset

# Operational Overview



# Image Caption Generalization via Tree Compression

Optimization:  $F = \Phi(\text{Visual Salience}) + \Phi(\text{Sequence Cohesion})$   
 $+ \Phi(\text{Tree Structure})$

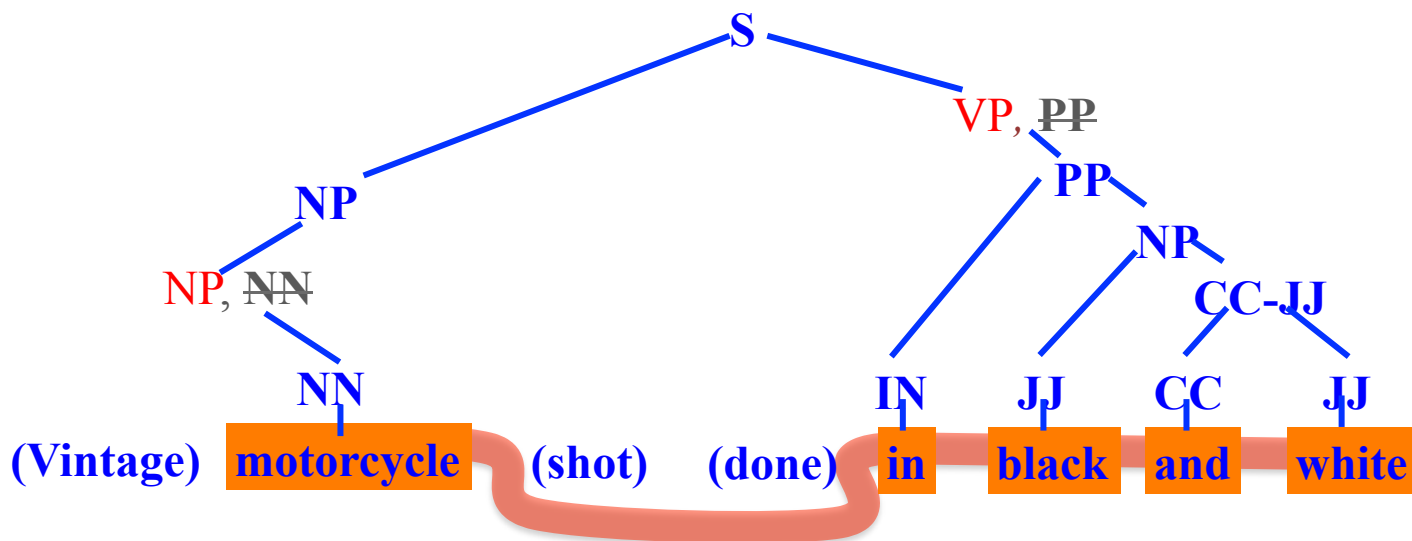




# Image Caption Generalization via Tree Compression

Optimization:  $F = \Phi(\text{Visual Salience}) + \Phi(\text{Sequence Cohesion})$   
 $+ \Phi(\text{Tree Structure})$

- sentence compression with light-weight parsing
- DP algorithm possible (modification to CKY parsing)

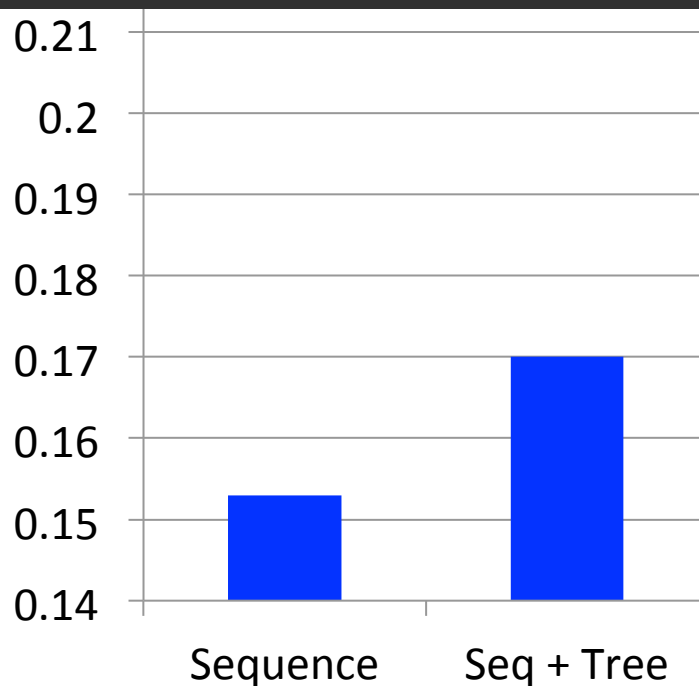




# Machine Caption VS Human Caption (forced choice w/ Amazon Mechanical Turk)



- ACL 2012 system (seq only): 16% win
- Final system (seq + tree + pruning): 24% win

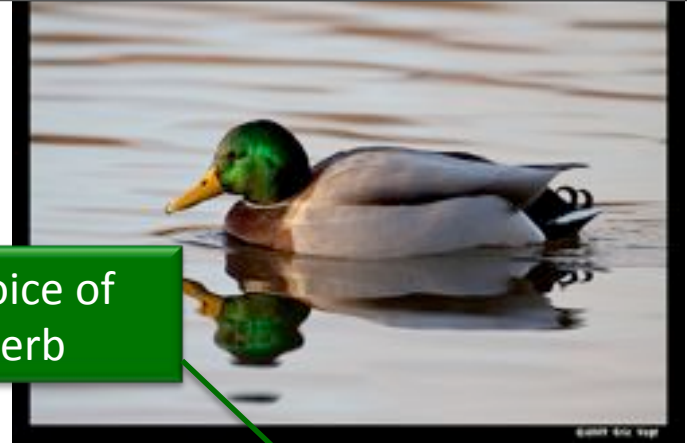


~ ACL 2012 system

# Good Examples



The flower was so **vivid** and attractive.



correct choice of an action verb

The duck **sitting** in the water.



Interesting choice of an abstract verb!

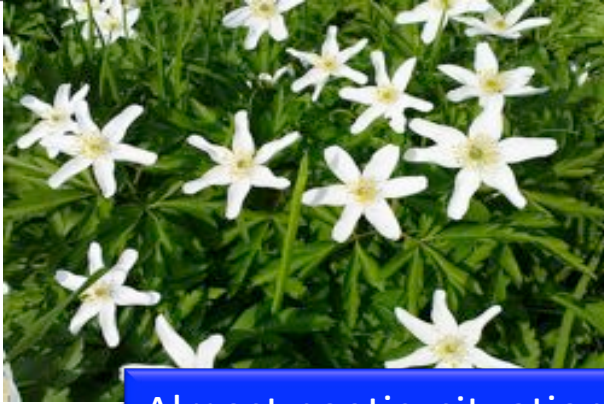
This window **depicts** the church.



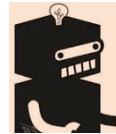
Highly expressive!

Blue flowers are **running rampant** in my garden.

# Mini Turing Test: our system wins in ~ 24 % cases!



Spring in a white dress.



Blue flowers have no scent. Small white flowers have no idea what they are.

Almost poetic, situationally relevant



Scenes around the lake on my bike ride.



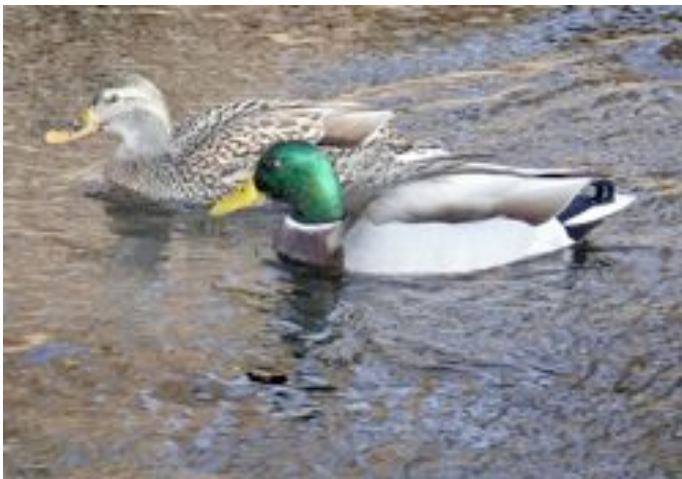
This horse walking along the road as we drove by.



Maybe the most common bird in the neighborhood, not just the most common water fowl in the neighborhood!



The duck was having a feast.



# Examples with Mistakes

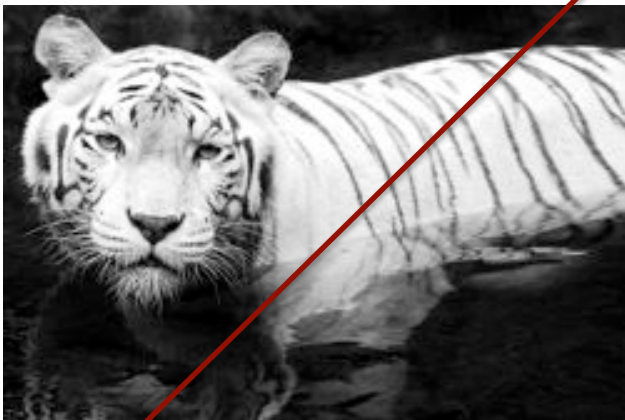


The couch is definitely bigger than it looks in this photo.



Yellow ball suspended in water.

Incorrect Object Recognition



My cat laying in my duffel bag.

Incorrect Scene Matching



Incorrect Composition

A high chair in the trees.

# Examples with Mistakes

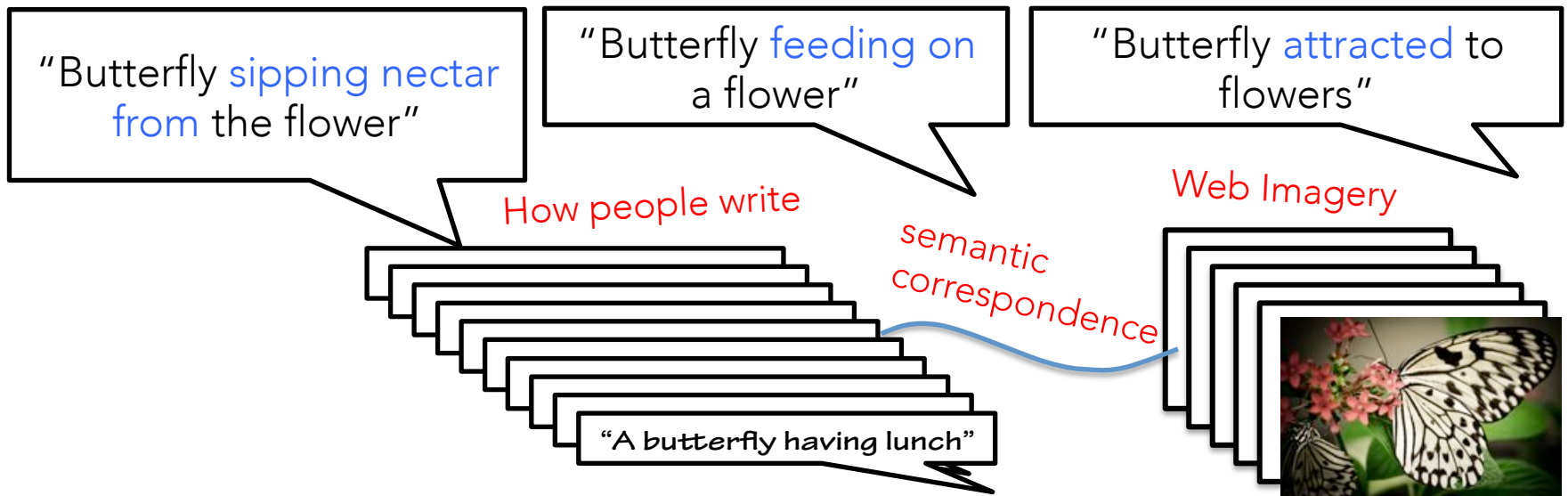
A cat looking for a home.

*The other cats are making  
the computer room. ???*



The castle *known for being  
the home of Hamlet in the  
Shakespeare play.*

Conclusion



### Distributional Hypothesis (Harris 1954)

Data decides

Humans decide

TreeTalk

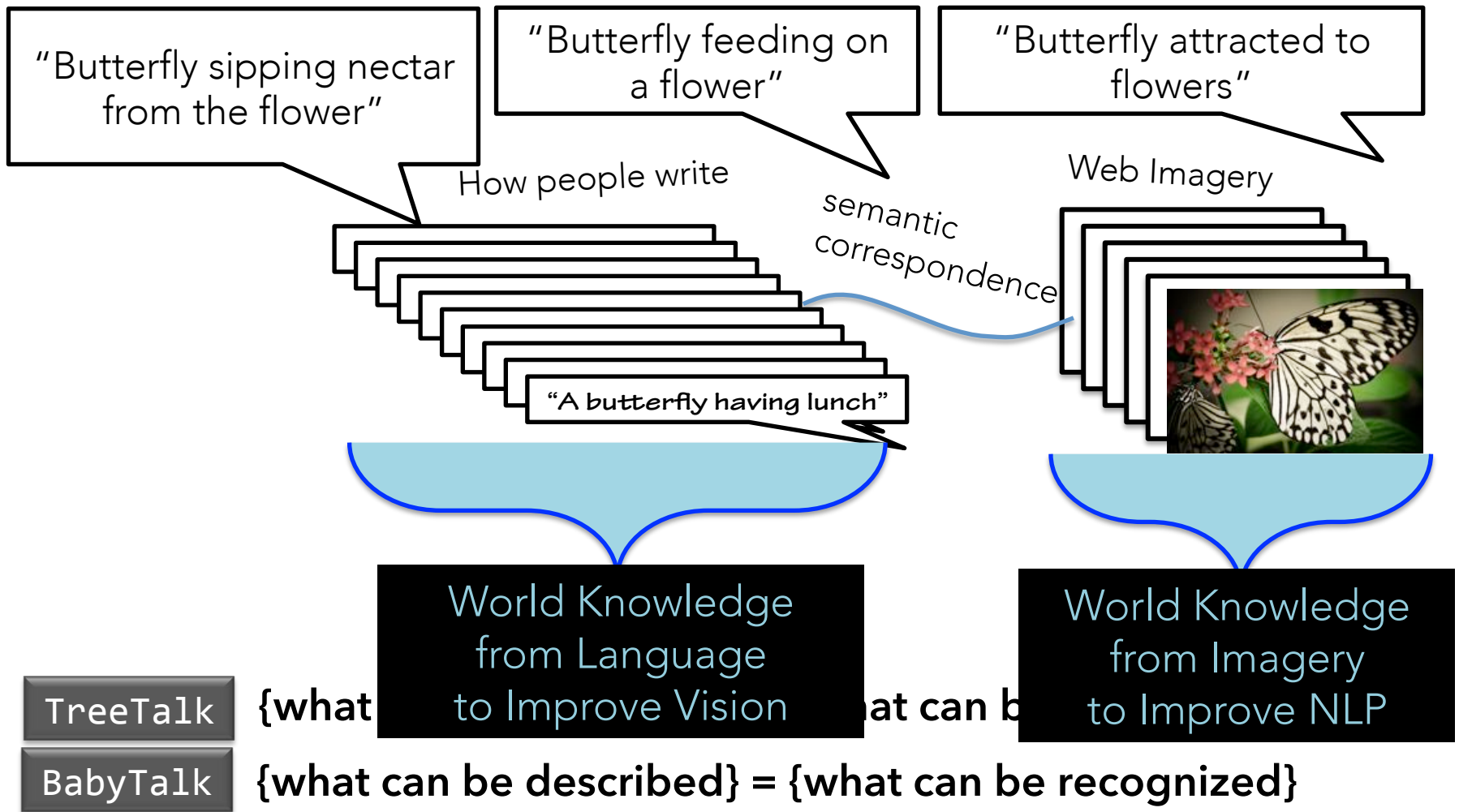
{what can be described}  $\supset$  {what can be recognized}

BabyTalk

{what can be described} = {what can be recognized}

- ◆ Start with a precise (but small) set of {what to recognize}, and increase the set
- ◆ Start with a large (but noisy) set of {what to describe}, and decrease the noise





- ◆ Start with a precise (but small) set of {what to recognize}, and increase the set
- ◆ Start with a large (but noisy) set of {what to describe}, and decrease the noise

# Future: Seeing beyond What's in the Image



- **What's** happening?
- **How / why** did this happen?
- What are the **intent / goal** of the participants?
- **Sentiment**: are they happy?
- **Reaction**: do we need to act on them (e.g., dispatching help)?

# Acknowledgements

-  My PhD Song Feng, Polina Kuznetsova
-  Other PhD Jianfu Chen, Vicente Ordonez, Karl Stratos, Siming Li, Jesse Dodge
-  MS Girish Kulkarni, Sagnik Dhar, Visruth Premraj
-  Undergrad Alyssa Mensch
-  Professor Hal Daumé III, Jia Deng, Alex Berg, Tamara Berg, David Warren
-  Industry Margaret Mitchell, Sujith Ravi, Ravi Kumar, Amit Goyal