Describing Images in Natural Language Part II

CVPR tutorial
Julia Hockenmaier
University of Illinois
juliahmr@illinois.edu

Overview

Part 1: High-Level Introduction to Sentence-Based Image Description

- What do we mean by image description?
- What kind of data sets are available?
- What kind of tasks have been proposed?
- How do we evaluate image description systems?
- A proposal for a shared task

Part 2: A brief intro to NLP for image description

- What is language understanding?
- Why is it difficult?

NLP for image description

What do you need to know about natural language processing/understanding for image description?

- Basic concepts from linguistics and NLP/NLU

To understand image captions:

- The structure of sentences
- The meaning of sentences

To understand how captions might relate to an image:

- Discourse models

What do you need to know about language/NLP?

What does it mean to *“understand simple sentences”*?

Task: Linguistic inference

People are shopping
in a supermarket

They are sitting at desks.

They are walking
on the street.

They are buying clothes.

They are at home.

No

They are standing or walking.

They are pushing
shopping carts.

They are in an indoor space.

There are aisles of shelves.

Yes

Understanding language

=

Being able to draw

(logical or commonsense)

inferences

Task: Image Description

People are shopping
in a supermarket

No

Yes

Understanding language

=

Being able to connect language to the external world

Yes

A woman wearing a purple tank top and a quilt skirt stands at the stand for kettle korn.

No

No

No

A woman wearing a purple tank top and a quilt skirt stands at the stand for kettle corn.

Semantics

for Image Description:

A caption is either
a true (correct) description
of the image or not.

The lady wears a black t-shirt.

Pragmatics

for Image Description:

Even a correct caption may be inappropriate.

A for kettle korn purple quilt skirt stand tank top wearing woman.

Syntax

for Image Description:

Captions are not just word salad

Who: a woman
What: standing
wearing a purple tank top
wearing a quilt skirt
Where: at a stand for kettle korn

Yes

A woman wearing a purple tank top and a quilt skirt stands at the stand for kettle korn.

Who: a woman,
a quilt skirt
What: standing for kettle korn
wearing a purple tank top
Where: at a stand

No

A woman wearing a purple tank top and a quilt skirt stands at the stand for kettle korn.

More Syntax

for Image Description:

The syntactic structure of a caption determines its meaning.

Language understanding requires knowledge of syntax, semantics, and pragmatics

Why is language understanding difficult?

1. Language is **ambiguous**:
   Every sentence has many possible interpretations.

2. Language is **productive**:
   We will always encounter new words or new constructions (‘kettle korn’?)

(Lexical) semantics:

What does each symbol mean?

Syntax:

How do these symbols fit together?

Compositional semantics:

What does this sentence mean?

Discourse:

What does this text mean?

How does it relate to the world?

The NLP pipeline

Tokenizer:
identify words and sentences
Part-of-speech (POS) tagger:
identify the parts of speech of the words
Chunker or Syntactic Parser:
obtain the grammatical structure of sentences
Semantic Parser:
obtain the predicate-argument structure (meaning) of sentences
Named Entity Recognition:
identify names of people, organizations, locations, dates etc.
Coreference resolution:
keep track of the mentioned entities throughout text

Syntax

What is the structure of this caption?

POS Tagger
Chunker
Syntactic Parser

Semantics

What does this caption mean?

Word Sense Disambiguation
Semantic Parser

Pragmatics

When is it appropriate to use this caption?

Referring Expressions
Discourse Model

Publicly available tools

Natural Language Toolkit (NLTK)

Python-based libraries (originally developed for teaching purposes) http://www.nltk.org

OpenNLP:

Java APIs: https://opennlp.apache.org

Illinois NLP software:

http://cogcomp.cs.illinois.edu/page/software

Stanford NLP software:

http://nlp.stanford.edu/software/

Many others, especially for individual components of the NLP pipeline

The structure

of sentences

Part-of-speech tagging

Part-of-speech ambiguity

A woman wearing a purple tank top and a quilt skirt ***stands*** *at the* ***stand*** *for kettle korn.*

POS Tagging

Words often have more than one part of speech (POS):

- *The **back** door* (adjective)
- *On my **back*** (noun)
- *Win the voters **back*** (particle)
- *Promised to **back** the bill* (verb)

The POS tagging task is to determine the POS tag for a *particular instance* of a word.
Since there is **ambiguity**, we cannot simply look up the correct POS in a dictionary.

These examples from Dekang Lin

Word classes

Open classes (many, possibly very rare, words):
- Nouns,
- Verbs,
- Adjectives,
- Adverbs

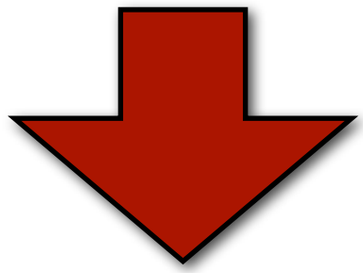
Closed classes (few, mostly very common, words):
- Auxiliaries and modal verbs
- Prepositions, Conjunctions
- Pronouns, Determiners
- Particles, Numerals

Penn Treebank tag set

CC	Coordinating Conjunction	PRP$	Possessive Pronoun
CD	Cardinal Number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign Word	RP	Particle
IN	Preposition/Subordinating Conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal verb	VBG	Verb, present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, present tense, not 3rdPers. Singular
NNP	Proper Noun, singular	VBZ	Verb, present tense, 3rd Pers. Singular
NNPS	Proper Noun, plural	WDT	Wh-determiner
PDT	Predeterminer	WP	Wh-pronoun
POS	Possessive ‘s	WP$	Possessive Wh-pronoun
PRP	Personal pronoun	WRB	Wh-adverb

POS tagging

```
Pierre Vinken , 61 years old , will join IBM ‘s board
as a nonexecutive director Nov. 29 .
```



```
Pierre_NNP Vinken_NNP ,_, 61_CD years_NNS old_JJ ,_,
will_MD join_VB IBM_NNP ‘s_POS board_NN as_IN a_DT
nonexecutive_JJ director_NN Nov._NNP 29_CD ._.
```

Task: assign POS tags to words

POS-tagging methods

POS tagging is a sequence-labeling task:

Standard techniques use
  Hidden Markov Models
  Chain Conditional Random Fields

Chunking (Shallow Parsing)

Chunking/Shallow Parsing

[NP *A woman*] [VP *wearing*] [NP *a purple tank top*] *and* [NP *a quilt skirt*] [VP *stands*] [PP *at*] [NP *the stand*] [PP *for*] [NP *kettle korn*].

Chunking/Shallow Parsing

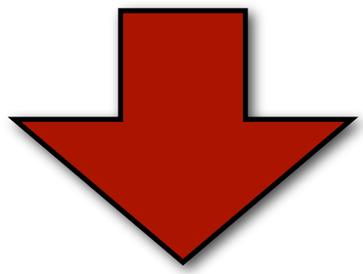
Identifies **non-recursive phrases** (aka chunks):

[NP *A woman*]
[VP *wearing*]
[NP *a purple tank top*]
and
[NP *a quilt skirt*]
[VP *stands*]
[PP *at*]
[NP *the stand*]
[PP *for*]
[NP *kettle korn*]

Can therefore be treated as a **sequence-labeling task** (typically after POS-tagging)

Shallow parsing

```
Pierre Vinken , 61 years old , will join IBM ‘s board
as a nonexecutive director Nov. 29 .
```



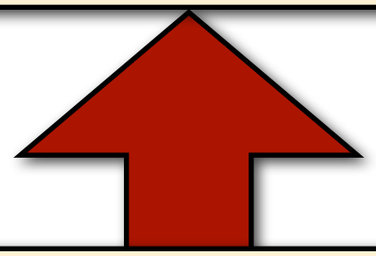
```
[NP Pierre Vinken] , [NP 61 years] old , [VP will join]
[NP IBM] ‘s [NP board] [PP as] [NP a nonexecutive
director] [NP Nov. 2] .
```

Task: identify all non-recursive NP,
verb (“VP”) and preposition (“PP”) chunks

The BIO encoding

B-NP B-VP B-PP: beginning of a chunk
I-NP: inside of a chunk
O: outside of any chunk

```
[NP Pierre Vinken] , [NP 61 years] old , [VP will join]
[NP IBM] 's [NP board] [PP as] [NP a nonexecutive
director] [NP Nov. 2] .
```



```
Pierre_B-NP Vinken_I-NP ,_O 61_B-NP years_I-NP
old_O ,_O will_B-VP join_I-VP IBM_B-NP 's_O board_B-NP
as_B-PP a_B-NP nonexecutive_I-NP director_I-NP Nov._B-
NP 29_I-NP ._O
```

Syntactic Parsing

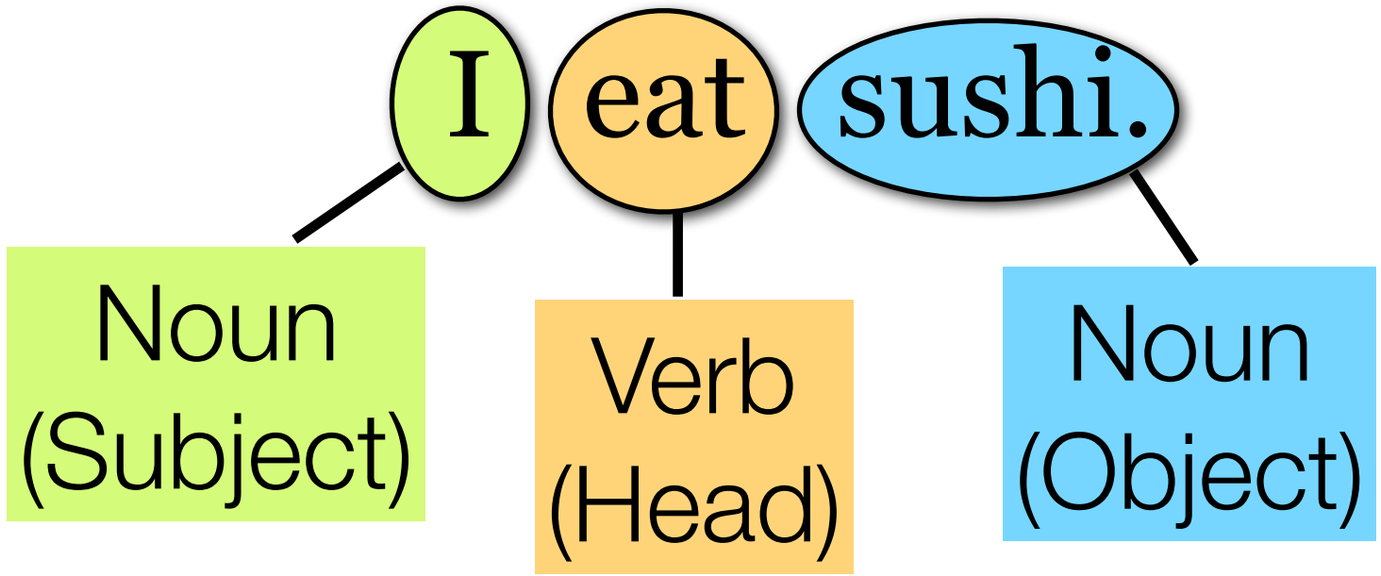
PP attachment ambiguity

*A man walks **down** a field **with** a crowd **in** the stands **behind** him.*

Does the man walk with the crowd?
Is the man in the stands?
Is this a field with a crowd?

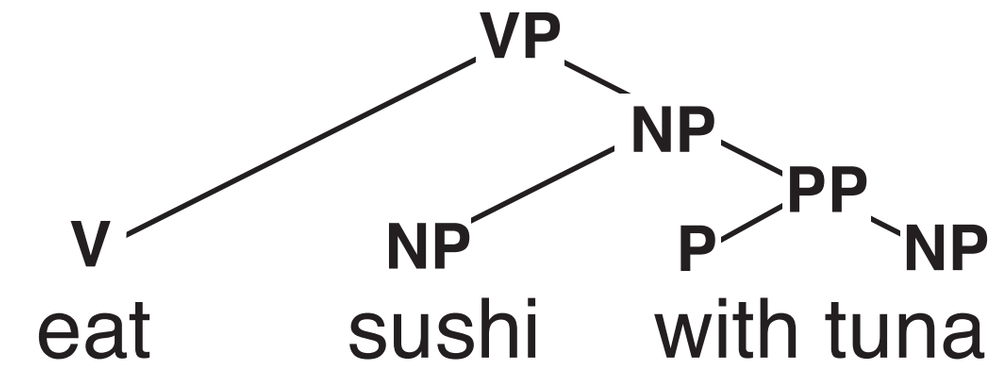
There is a combinatorial explosion of possible analyses (most of which are wildly implausible for people).

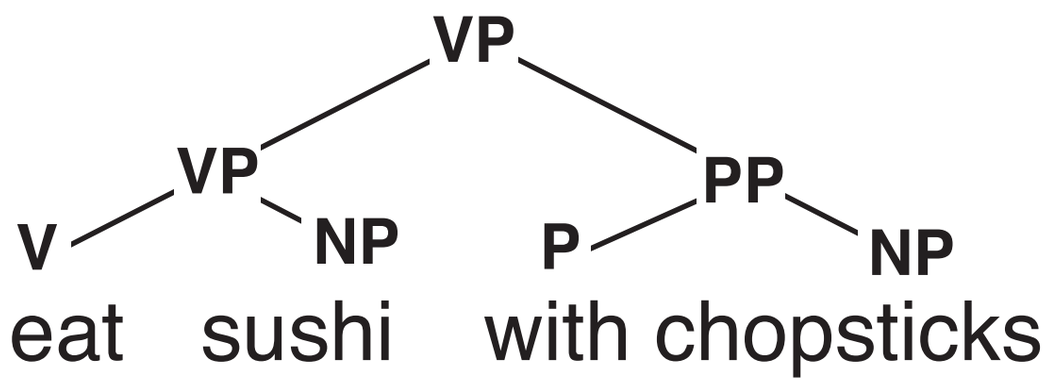
Basic sentence structure



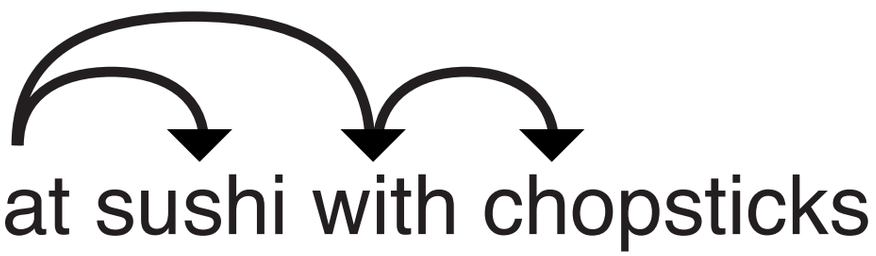
Two ways to represent syntactic structure

Phrase structure trees

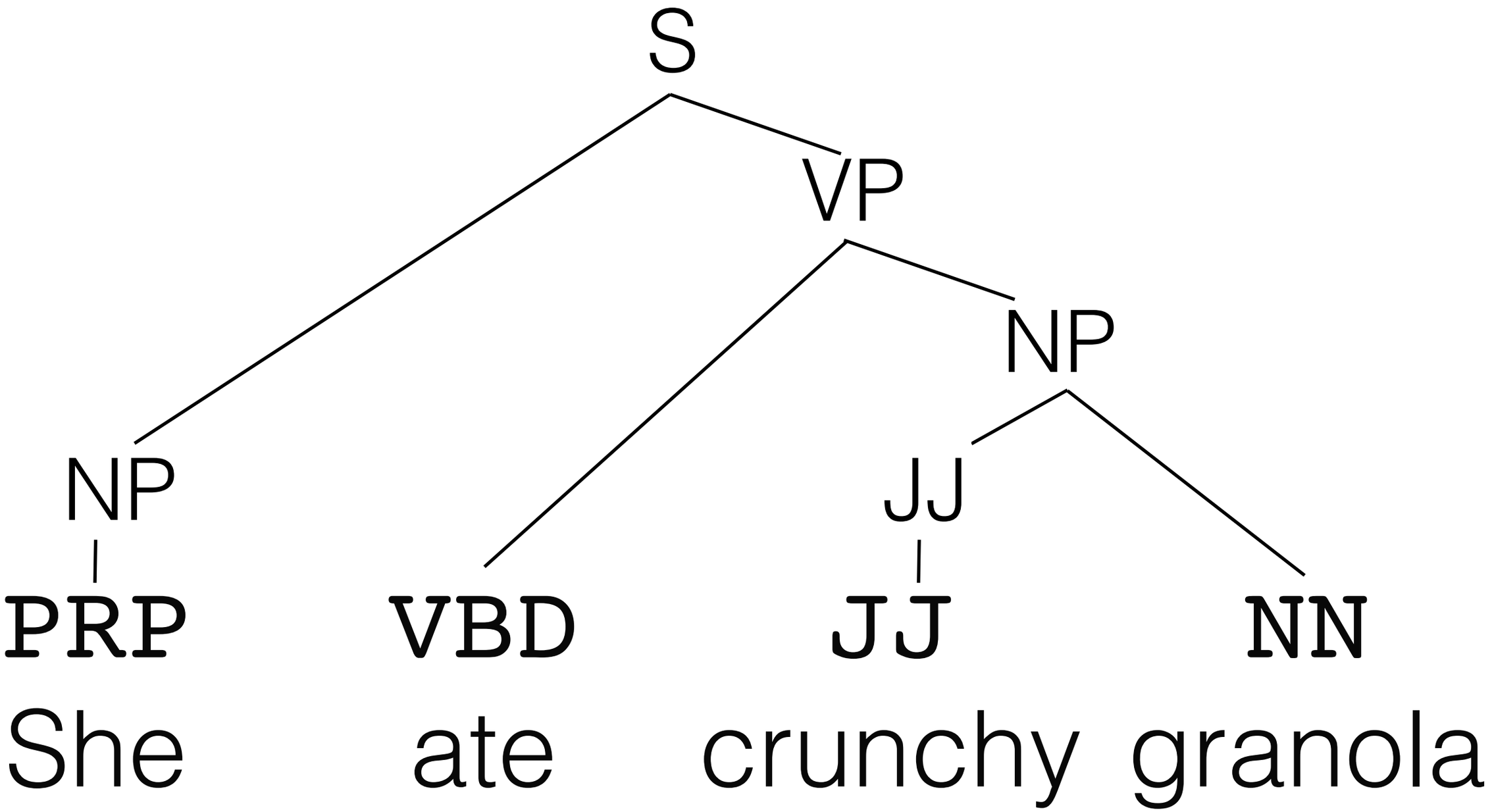




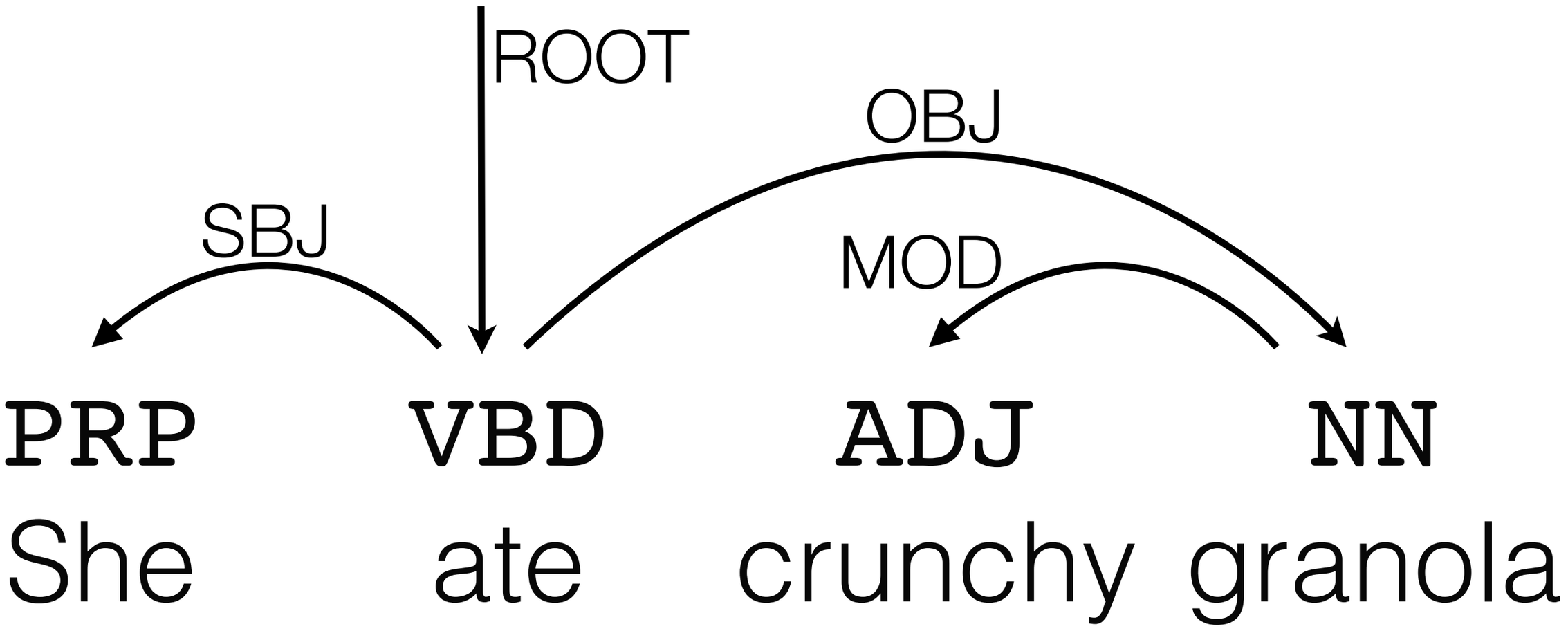
Dependency trees



Penn Treebank CFGs



Dependency Grammar



Syntactic dependencies

Syntactic dependencies capture **grammatical roles** (subject, object, modifier, etc.).

For simple sentences, it is often easy to map between grammatical and **semantic roles** (who did what), but that is not true in general.

There are many different standards and conventions for role sets, and how to handle different kinds of constructions.

Statistical Parsing

Grammar

Defines the sentences of the language and their possible structures (trees τ)

Probability model

Assigns a score $P(\tau)$ to each tree τ

Parsing algorithm

Returns the best tree $\tau^* = \text{argmax } P(\tau)$ for each sentence

Probabilistic Context-Free Grammars

For every nonterminal X, define a distribution $P(X \rightarrow \alpha \mid X)$ over all rules with the same LHS symbol X:

S	→	NP VP	0.8
S	→	S conj S	0.2
NP	→	Noun	0.2
NP	→	Det Noun	0.4
NP	→	NP PP	0.2
NP	→	NP conj NP	0.2
VP	→	Verb	0.4
VP	→	Verb NP	0.3
VP	→	Verb NP NP	0.1
VP	→	VP PP	0.2
PP	→	P NP	1.0

Non-local dependencies

A little boy being amazed by a giant ***bubble*** *he just* ***created****.*

In general, this requires richer grammatical representations.

In English, non-local dependencies arise in relative clauses, questions, and coordination

The interpretation of sentences

Semantics

In order to understand language, we need to know its meaning.

- What is the meaning of a word?
  (**Lexical semantics**)

- What is the meaning of a sentence?
  ([Compositional] semantics)

- What is the meaning of a longer piece of text?
  (Discourse semantics)

Lexical semantics

Distributional similarity

What is ***tezgüino*** **?**

A bottle of ***tezgüino*** *is on the table.*
Everybody likes ***tezgüino****.*
Tezgüino *makes you drunk.*
We make ***tezgüino*** *out of corn.*

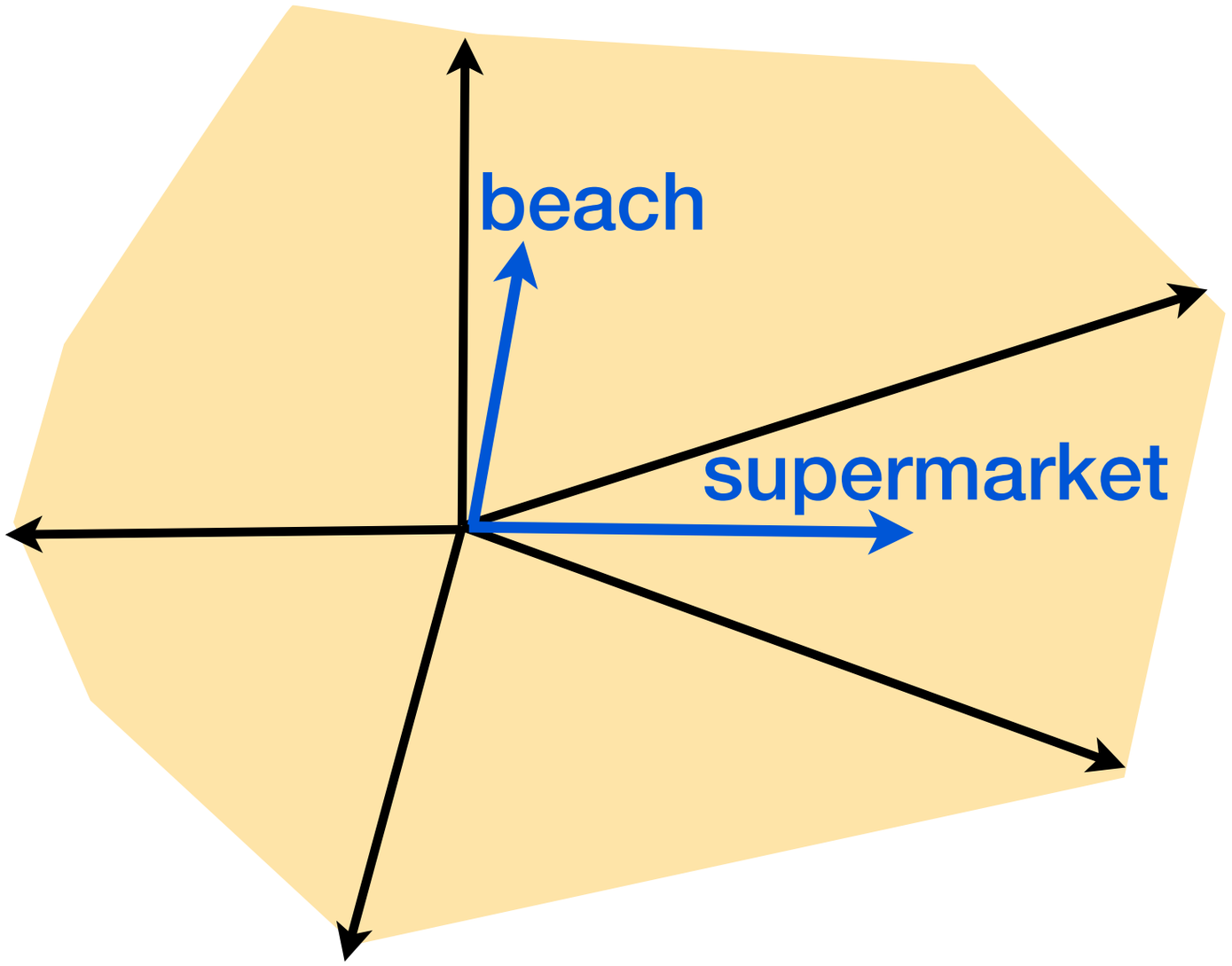
(Lin, 1998; Nida, 1975)

The Distributional Hypothesis:

You shall know a word by the company it keeps.

(*Firth 1957*)

Vector-space semantics



Vector-space semantics

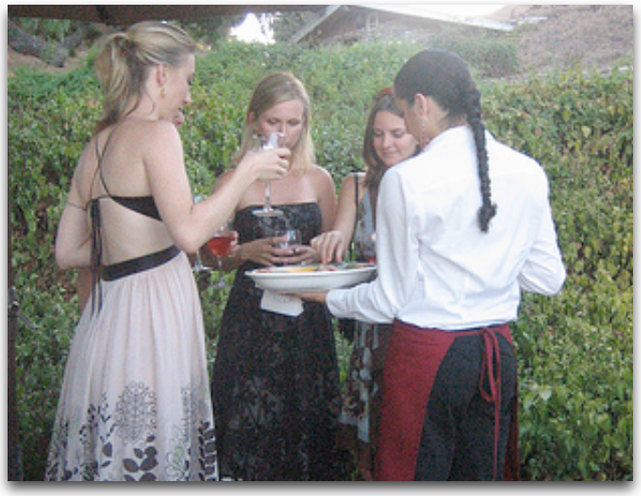
Traditionally: Distributional similarities

Define a set of contexts in which words can occur (e.g. adjacent words, grammatical relations).
Each context = one dimension in the vector space.
Count how often each word appears in each context.
Compute point-wise mutual information between words and contexts to get the vectors.

More recently: Embeddings

e.g based on neural networks.

Word Sense

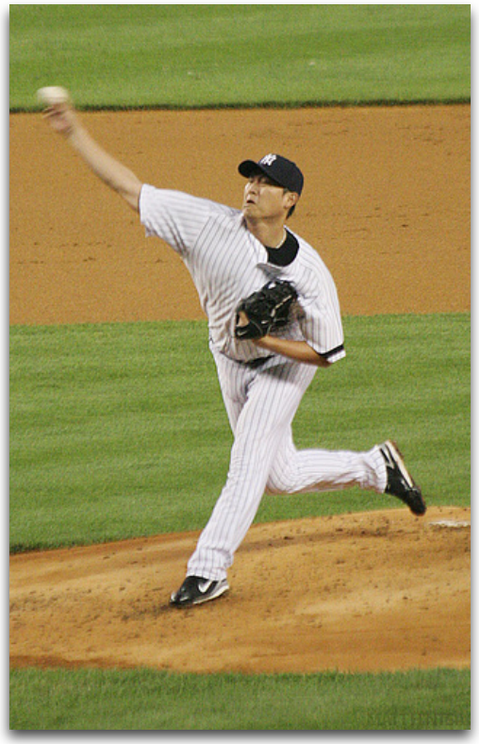


*A waitperson **serving** snacks to three women.*

*A man in red swim trunks is **serving** a beach volleyball.*

Word Sense

A man in a bar drinks from a ***pitcher***



A baseball ***pitcher*** *is in the middle of a throw*

What does this word mean?

This **plant** needs to be watered each day.

⇒ **living plant**

This **plant** manufactures 1000 widgets each day.

⇒ **factory**

Word Sense Disambiguation (WSD):

Identify the sense of content words (noun, verb, adjective) in context (assuming a fixed inventory of word senses)

WordNet: sense = synset

Applications: machine translation, question answering, information retrieval, text classification

WordNet

Very large lexical database of English:
110K nouns, 11K verbs, 22K adjectives, 4.5K adverbs
(WordNets for many other languages exist or are under construction)

Word senses grouped into synonym sets ("synsets") linked into a conceptual-semantic hierarchy

Conceptual-semantic relations: hypernym/hyponym (also holonym/meronym)
Also: lemmatization

A WordNet example

WordNet Search - 3.0 - WordNet home page - Glossary - Help

Word to search for: bass Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) **bass** (the lowest part of the musical range)
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
- S: (n) **bass**, basso (an adult male singer with the lowest voice)
- S: (n) sea bass, **bass** (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, **bass** (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) **bass**, bass voice, basso (the lowest adult male singing voice)
- S: (n) **bass** (the member with the lowest range of a family of musical instruments)
- S: (n) **bass** (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Adjective

- S: (adj) **bass**, deep (having or denoting a low vocal or instrumental range) *"a deep voice"; "a bass voice is lower than a baritone voice"; "a bass clarinet"*

WordNet home page

Hypernyms and hyponyms

- S: (n) **bass** (the lowest part of the musical range)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) pitch (the property of sound that varies with variation in the frequency of vibration)
      - S: (n) sound property (an attribute of sound)
        - S: (n) property (a basic or essential attribute shared by all members of a class) *"a stud*
          - S: (n) attribute (an abstraction belonging to or characteristic of an entity)
            - S: (n) abstraction, abstract entity (a general concept formed by extracting
              - S: (n) entity (that which is perceived or known or inferred to have
- S: (n) **bass**, bass part (the lowest part in polyphonic music)
  - *direct hyponym* / ***full hyponym***
    - S: (n) ground bass (a short melody in the bass that is constantly repeated)
    - S: (n) figured bass, basso continuo, continuo, thorough bass (a bass part written out in full and acc
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) part, voice (the melody carried by a particular voice or instrument in polyphonic music) *"he*
      - S: (n) tune, melody, air, strain, melodic line, line, melodic phrase (a succession of notes forn
        - S: (n) music (an artistic form of auditory communication incorporating instrumental c
          - S: (n) auditory communication (communication that relies on hearing)
            - S: (n) communication (something that is communicated by or to or betwe
              - S: (n) abstraction, abstract entity (a general concept formed by extr
                - S: (n) entity (that which is perceived or known or inferred to

Shallow Semantics

Representing meaning

Lexical semantics: what is the meaning of words?
Sentential semantics: what is the meaning of sentences?

We need to define a meaning representation language.

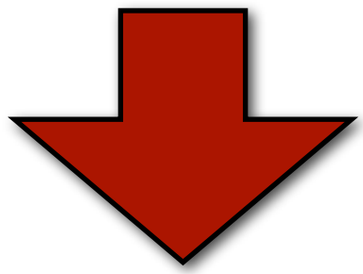
"Shallow" semantic analysis
(information extraction): template-filling
- Named entities: organizations, locations, dates,...
- Event extraction

"Deep" semantic analysis: (variants of) predicate logic

```
∃x∃y(pod_door(x)& Hal(y) & request(open(x,y))
```

Named Entity Recognition

```
Pierre Vinken , 61 years old , will join IBM ‘s board
as a nonexecutive director Nov. 29 .
```



```
[PERS Pierre Vinken] , 61 years old , will join
[ORG IBM] ‘s board as a nonexecutive director
[DATE Nov. 2] .
```

Task: identify all mentions of named entities (people, organizations, locations, dates)

Semantic roles

A large white egg beater *mixes the contents of the silver bowl.*
Someone *is whisking eggs with a handheld mixer.*
Beating an egg with a machine.

Semantic Role Labeling

The arguments of verbs (subjects, objects) have semantic roles (aka. theta roles, thematic roles):

- **Agent** (who is doing the action?)
- **Patient** (who/what is the action being done to?)
- **Theme** (what is the action about?)
- **Instrument** (what is used for the action?)
- **Location, Time, Destination**, etc.

Inventories of roles vary, and standard resources (Propbank) use non-committal names such as Arg0, Arg1, instead.

Compositional Semantics

What do sentences mean?

Declarative sentences (statements) can be **true or false**, depending on the state of the world:
John sleeps.

In the simplest case, the consist of a verb and one or more noun phrase arguments.

Principle of compositionality (Frege):
The meaning of an expression depends on the meaning of its parts and how they are put together.

Formal Semantics

A girl plays on the beach.

$\exists e \exists x \exists y$ girl'(x) & beach'(y) & play'(e) & agent(e, x) & location(e, y)

Denotational Semantics

The **denotation** of a (declarative) sentence is the **set of possible worlds** (situations) in which it is true:

$$[\![\mathbf{s}]\!] = \{\mathbf{w} \in \mathrm{U}: \mathbf{s} \text{ is true in } \mathbf{w}\}$$

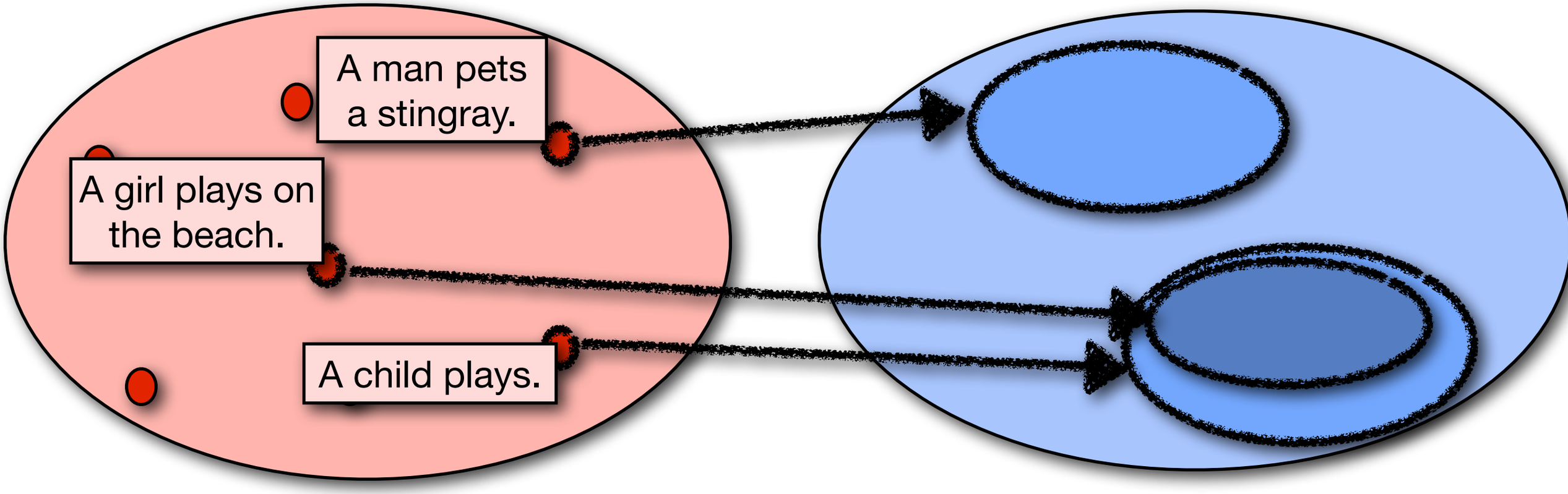
Denotations capture **entailment:**

$\mathbf{s}$ entails $\mathbf{s'}$ if $[\![\mathbf{s}]\!] \subseteq [\![\mathbf{s'}]\!]$

Denotational semantics from image descriptions

Young, Lai, Hodosh, Hockenmaier,
Transactions of the ACL 2014

Denotational Semantics



Language L

Universe U

Our hypothesis:

denotational similarities

would be particularly useful for tasks that require inference.

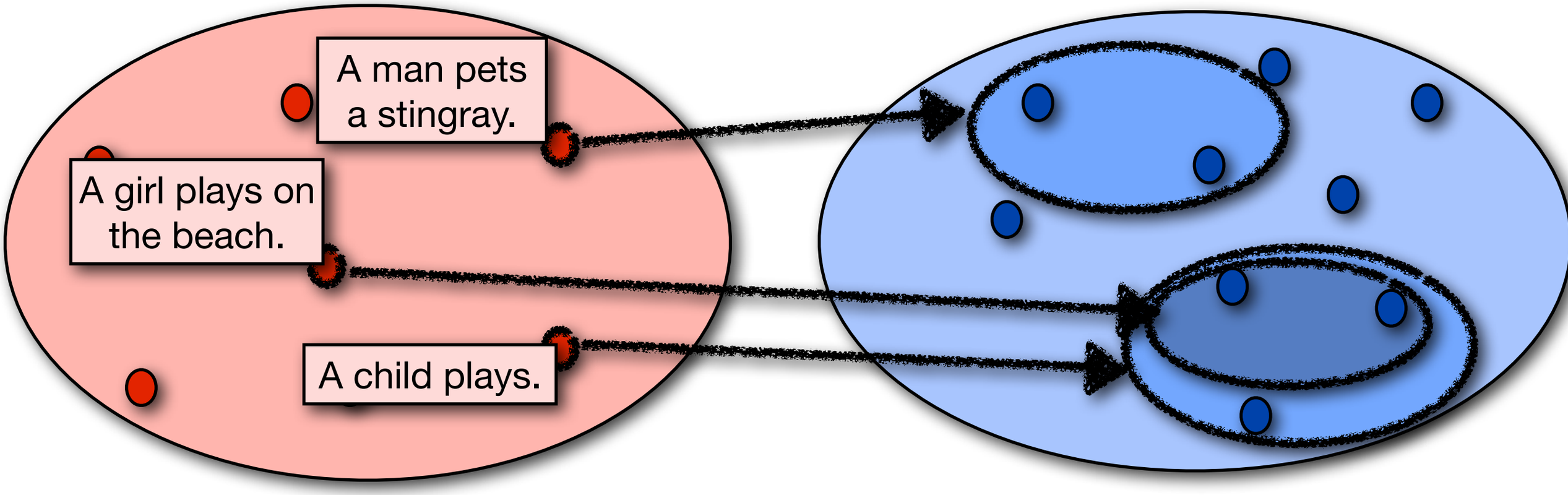
Visual Denotations

The **visual denotation** of a descriptive sentence is the **set of images** for which it is a correct description:

$$[\![\mathbf{s}]\!] = \{\mathbf{i} \in \mathrm{I}: \mathbf{s} \text{ describes (part of) } \mathbf{i}\}$$

Young et al., From image descriptions to visual denotations, TACL 2014

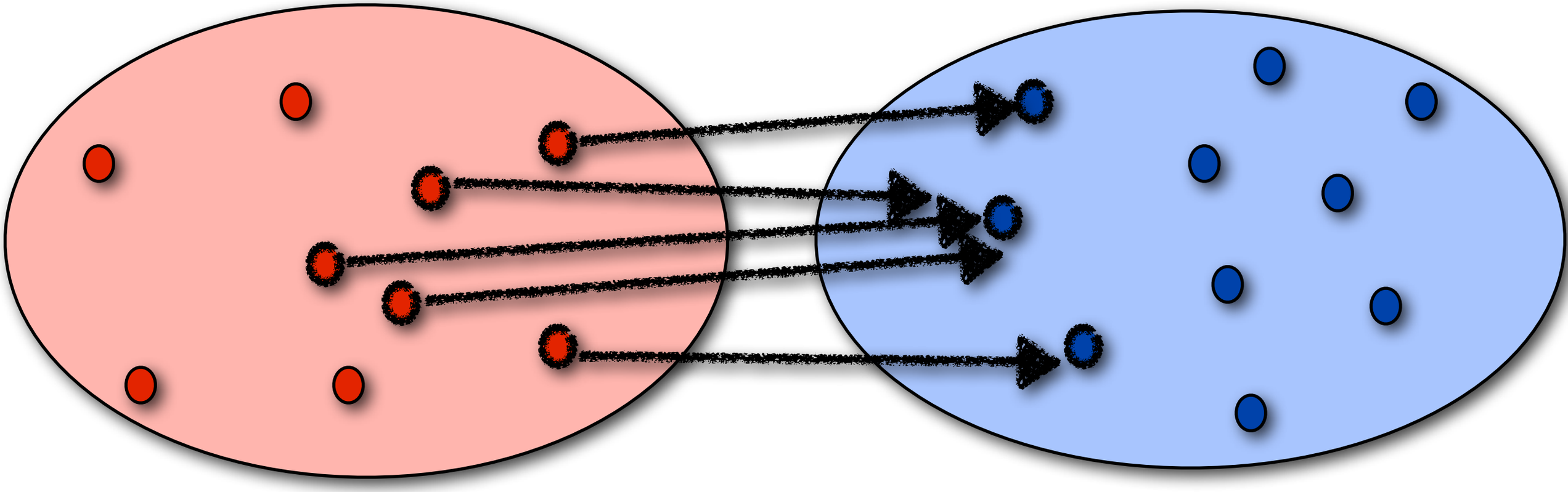
Visual Denotations



Language L

Images I

Our corpus



Language L

Images I

The visual denotation of most captions in our corpus is a singleton:

$$[\![\ \textbf{A player from the white and green highschool team dribbles down court defended by a player from the other team}\ ]\!] = \{\ \}$$

Young et al., From image descriptions to visual denotations, TACL 2014

Each image is in the denotation of multiple captions:

$\in$ ⟦ **A player from the white and green highschool team dribbles down court defended by a player from the other team** ⟧

$\in$ ⟦ **Two boys in green and white uniforms play basketball with two boys in blue and white uniforms** ⟧

This allows us to capture relations between different descriptions of the same event:

$\in$ ⟦ **dribble down court** ⟧

$\in$ ⟦ **play basketball** ⟧

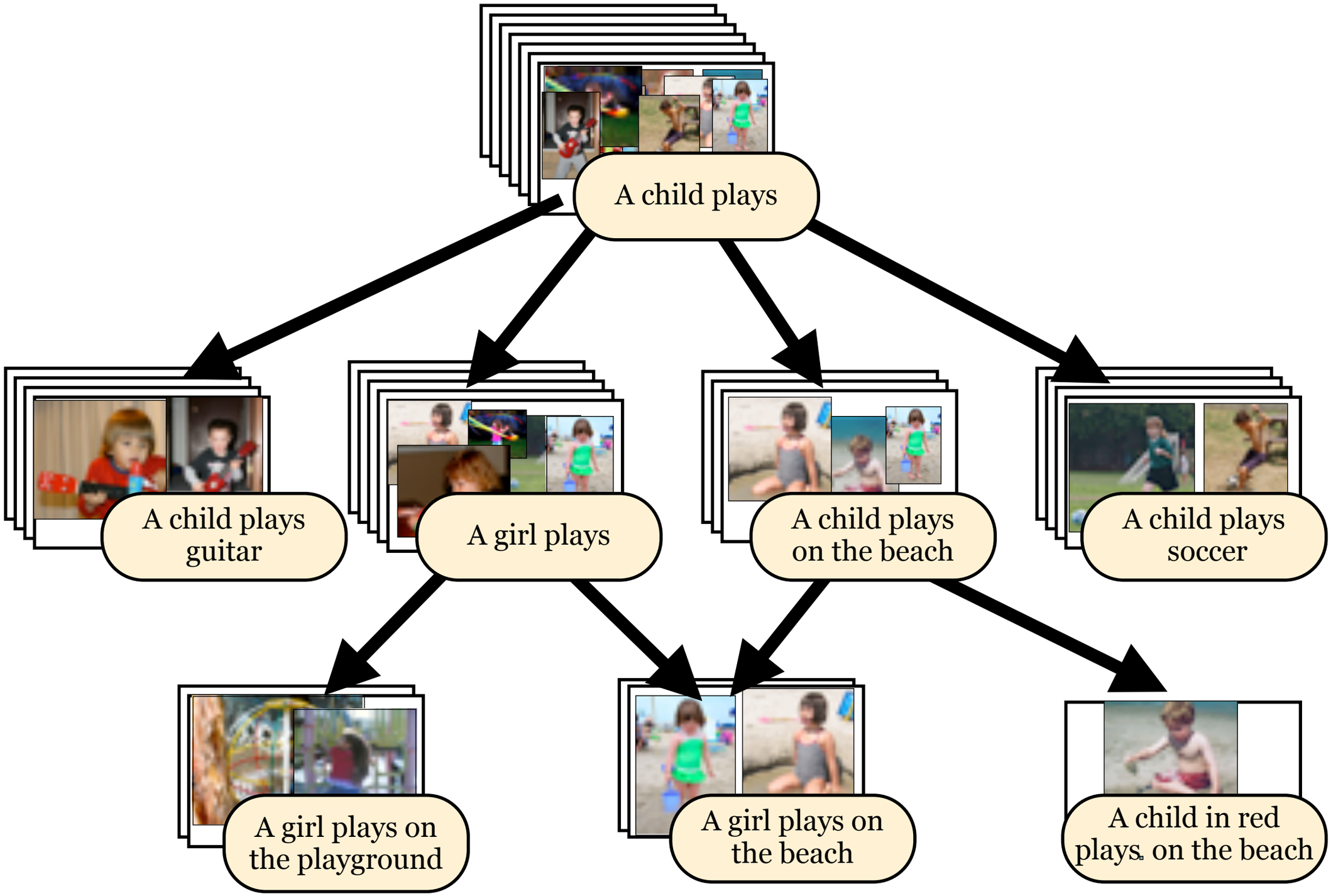
Denotation Graph

Denotations define a **subsumption hierarchy** (lattice) over image descriptions:

$$[\![\text{a girl plays on the beach}]\!] \subset [\![\text{a child plays}]\!]$$

We can build this hierarchy automatically from our corpus of image captions.

Denotation Graph



identify transformations bottom-up

build graph top-down

Parent Node

⟦ girl plays ⟧

drop modifier

add modifier

⟦ blond girl plays ⟧

Child Node

Parent Node

⟦ girl plays ⟧

drop modifier

add modifier

⟦ girl plays **on the beach** ⟧

Child Node

Parent Node

⟦ **child** plays ⟧

replace head noun by hypernym

replace head noun by hyponym

⟦ **girl** plays on the beach ⟧

Child Node

Parent Node

⟦ girl ⟧

extract simple constituent

insert into sentence

⟦ girl plays on the beach ⟧

Child Node

Statistics

Original data (~32,000 images)

~160K distinct captions

Denotation graph:

~1750K distinct captions:

~230K captions with ⟦s⟧ ≥ 2

~53K captions with ⟦s⟧ ≥ 5

~22K captions with ⟦s⟧ ≥ 10

~1.9K captions with ⟦s⟧ ≥ 100

161 captions with ⟦s⟧ ≥ 1000

e.g. person play instrument, woman standing, ...

The denotation graph allows us to estimate the **denotational similarity** of sentences/phrases.

Denotational Similarities

$$P_{[\![]\!]}(x) = | [\![x]\!] | / | U |$$

$$P_{[\![]\!]}(x, y) = | [\![x]\!] \cap [\![y]\!] | / | U |$$

Two kinds of denotational similarities:

Conditional probabilities $P_{[\![]\!]}(x \mid y)$

Normalized Pointwise Mutual Information $nPMI_{[\![]\!]}(x, y)$

Denotational similarities

$p(VP_1 \mid VP_2)$	
p(talk | engage in conversation)	= 0.79
p(play tennis | swing racket)	= 0.82
p(stand | wait for subway)	= 0.58
p(sit | ride subway)	= 0.56
p(stand | lean against building)	= 0.53
p(shave | look in mirror)	= 0.41
p(dig hole | use shovel)	= 0.38
p(make face | stick out tongue)	= 0.38

Young et al., From image descriptions to visual denotations, TACL 2014

Denotational similarities

		$\text{nPMI}_{[\![]\!]}(VP_1, VP_2)$
open present	unwrap	0.84
lasso	try to rope	0.83
get ready to kick	run towards ball	0.79
try to tag	slide into base	0.79
shave face	look into mirror	0.77

Young et al., From image descriptions to visual denotations, TACL 2014

The structure of discourse

What is discourse?

On Monday, John went to Bevande. He wanted to buy lunch. But the cafe was closed. That made him angry, so the next day he went to Green Street instead.

'Discourse':
any linguistic unit that consists of **multiple sentences**

Speakers describe "some situation or state of the real or some hypothetical world" (Webber, 1983)

Speakers attempt to get the **listener**
to construct a similar **model of the situation**.

Discourse models

An explicit representation of:

- the **events and entities** that a discourse talks about
- the **relations** between them
  (and to the real world).

This representation is often written
in some form of logic.

What does this logic need to capture?

Discourse models should capture…

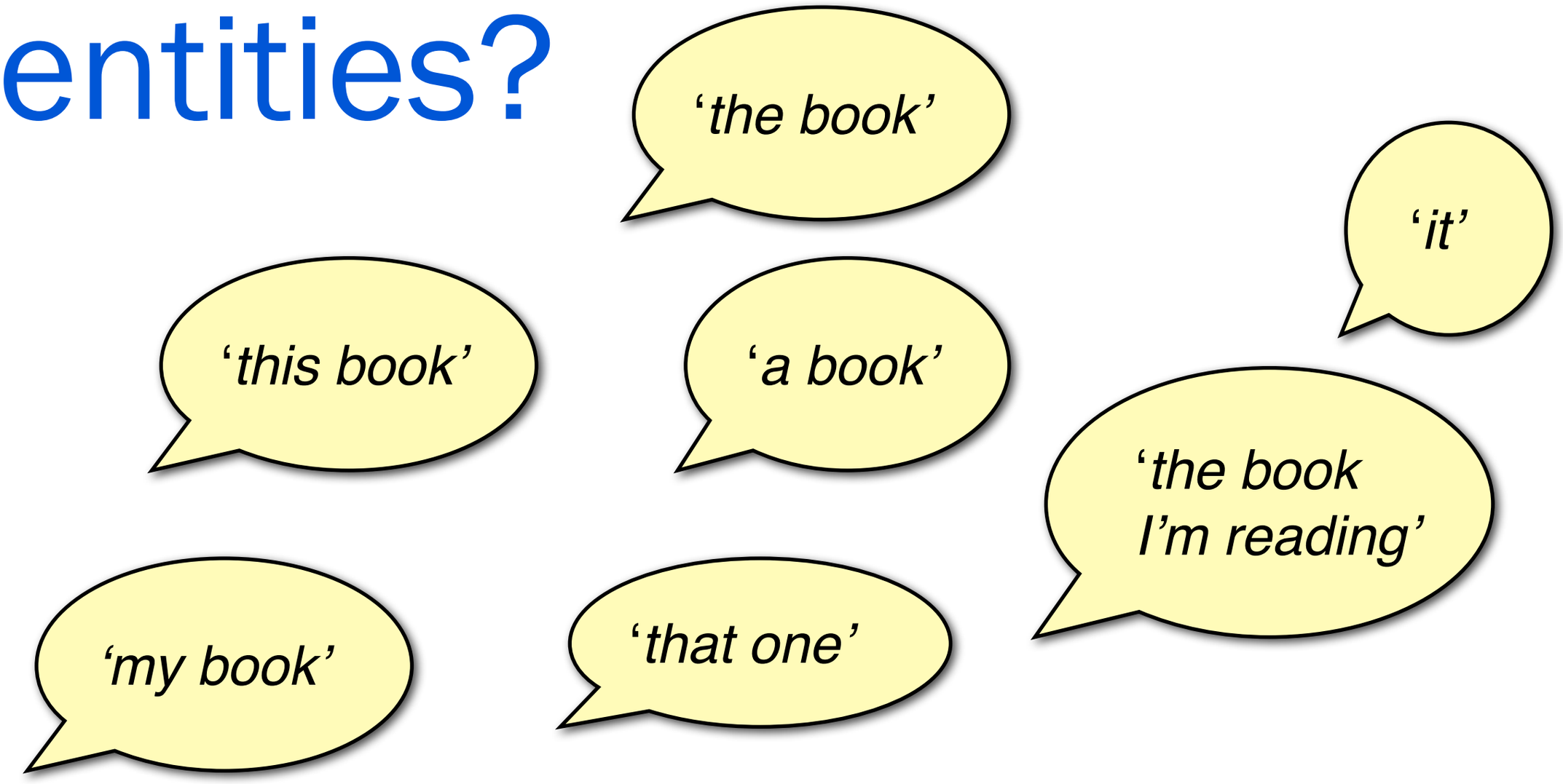
Physical entities: John, Bevande, lunch
Events: On Monday, John went to Bevande
  involve entities, take place at a point in time
States: It was closed.
  involve entities and hold for a period of time

images

Temporal relations: afterwards
  between events and states

Rhetorical ('discourse') relations: … so … instead
between events and states

How do we refer to entities?



This depends on what the speaker assumes about what the hearer knows, and what they have previously talked about.

Some terminology

Referring expressions (‘*this book*’, ‘it’) refer to some entity (e.g. a book), which is called the **referent.**

Co-reference: two referring expressions that refer to the same entity **co-refer** (are co-referent).

I saw a movie last night. I think you should see it too!

In multi-sentence text, **anaphora resolution** is important (and often difficult)
In single captions, it is pretty straightforward:
***A man** is walking **his** dog*.

Cross-caption Coreference resolution

There is a bride and groom with two children, a woman and a man carrying a flag, standing on a stony place.

A bride holding a bouquet of flowers is standing next to a man in a tuxedo.

A bride and groom stand in front of a brick building with others.

A man and woman at their wedding and little children playing.

A bride and her groom prepare to say their vows.

Cross-caption Coreference resolution

*There is a **bride** and **groom** with **two children**, **a woman** and **a man** carrying a flag, standing on a stony place.*

***A bride** holding a bouquet of flowers is standing next to **a man in a tuxedo**.*

*A **bride** and **groom** stand in front of a brick building with **others**.*

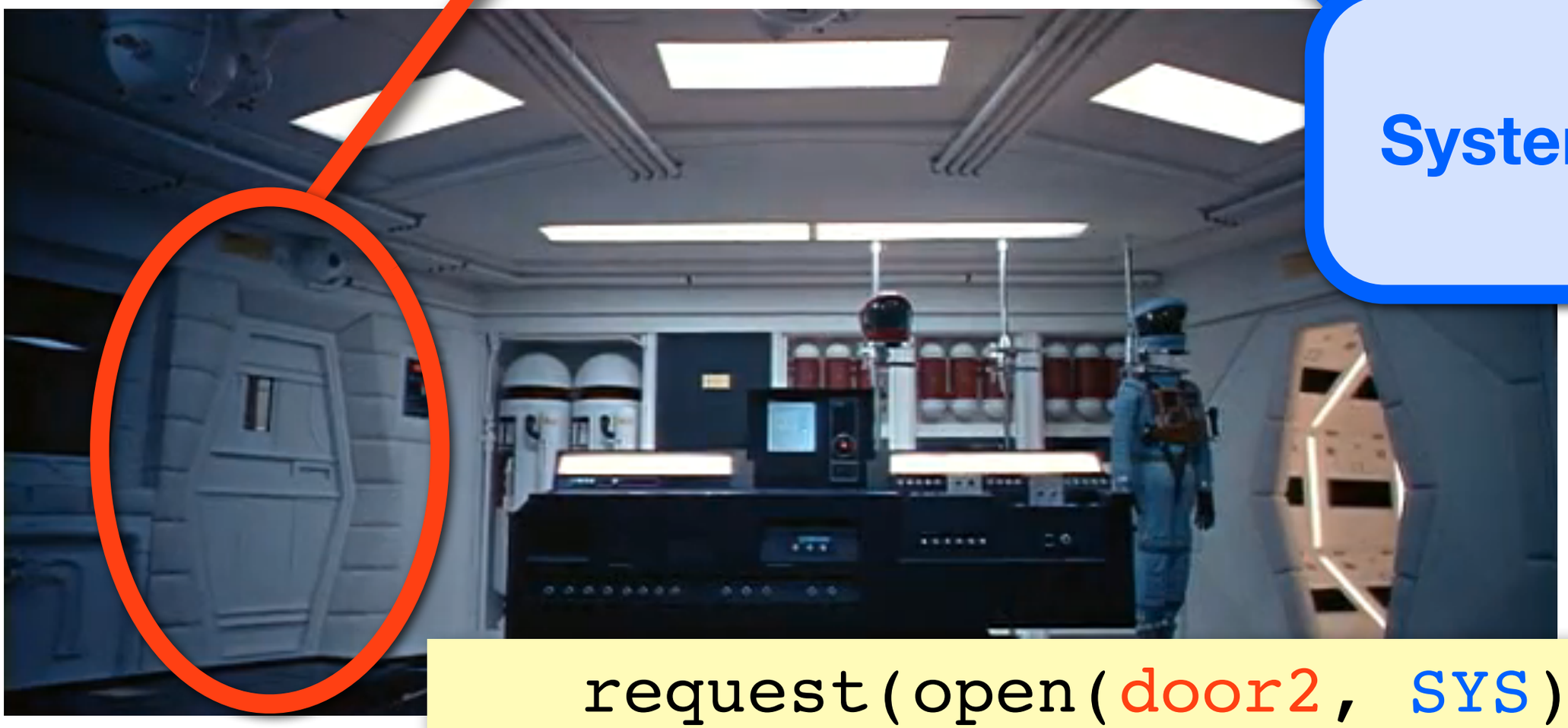
*A **man** and **woman** at their wedding and **little children** playing.*

***A bride** and **her groom** prepare to say their vows.*

Language usage and understanding in context

Multimodal NLP:

$\exists x \exists y$(pod_door(x) & Hal(y) & request(open(x, y)))



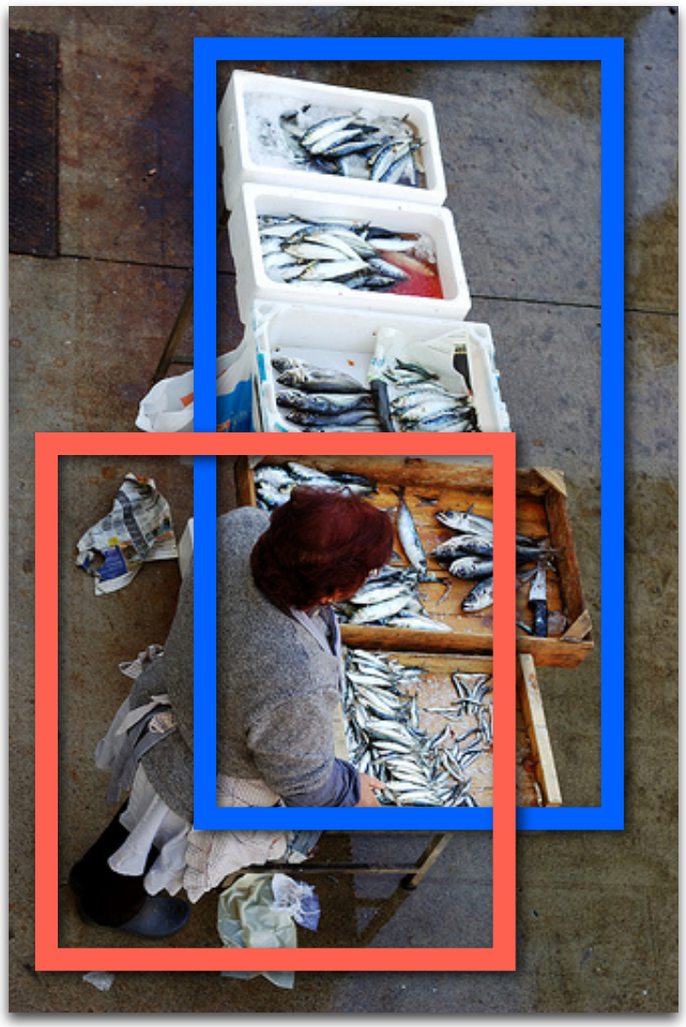
System

request(open(door2, SYS))

Spatial relations

A young child *plays a musical instrument* ***in front of*** ***another boy.***

Spatial relations



Woman in gray sweater *stands* ***beside*** ***display of fish***.