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.

- They are standing or walking.
- They are pushing shopping carts.
- They are in an indoor space.
- There are aisles of shelves.

Yes

No
Understanding language = Being able to draw (logical or commonsense) inferences
People are shopping in a supermarket

No

Yes
Understanding language
= 
Being able to connect language to the external world
A woman wearing a purple tank top and a quilt skirt stands at the stand for kettle korn.
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
A woman wearing a purple tank top and a quilt skirt stands at the stand for kettle korn.
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’?)
当天，阿博特在上海举行的“澳大利亚周”欢迎午宴上说，“我们知道马航失联客机黑匣子在几公里范围内”，然而，知道这点并不等同于能找到在海下四公里半深度左右的残骸，从而最终确定在失联航班上到底发生了什么。他称这次搜寻“也许是人类历史上最困难的一次”，并表示澳方将竭尽所能，继续搜寻。
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
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
CD    Cardinal Number
DT    Determiner
EX    Existential *there*
FW    Foreign Word
IN    Preposition/Subordinating Conjunction
JJ    Adjective
JJR   Adjective, comparative
JJS   Adjective, superlative
LS    List item marker
MD    Modal verb
NN    Noun, singular or mass
NNS   Noun, plural
NNP   Proper Noun, singular
NNPS  Proper Noun, plural
PDT   Predeterminer
POS   Possessive ‘s
PRP   Personal pronoun
PRP$  Possessive Pronoun
RB    Adverb
RBR   Adverb, comparative
RBS   Adverb, superlative
RP    Particle
SYM   Symbol
TO    *to*
UH    Interjection
VB    Verb, base form
VBD   Verb, past tense
VBG   Verb, present participle
VBN   Verb, past participle
VBP   Verb, present tense, not 3rd Pers. Singular
VBZ   Verb, present tense, 3rd Pers. Singular
WDT   Wh-determiner
WP    Wh-pronoun
WP$   Possessive Wh-pronoun
WRB   Wh-adverb
Pierre Vinken, 61 years old, will join IBM’s board as a nonexecutive director Nov. 29.

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

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
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

I eat sushi.

- **Noun (Subject)**
- **Verb (Head)**
- **Noun (Object)**
Two ways to represent syntactic structure

Phrase structure trees

```
 VP
  NP
    P
      NP
  V
  eat
  sushi
  with tuna
```

```
 VP
  NP
    P
      NP
  V
  eat
  sushi
  with chopsticks
```

Dependency trees

```
 eat sushi with tuna
```

```
 eat sushi with chopsticks
```
She ate crunchy granola
Dependency Grammar

She ate crunchy granola
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 $\tau$)

**Probability model**
Assigns a score $P(\tau)$ to each tree $\tau$

**Parsing algorithm**
Returns the best tree $\tau^* = \arg\max P(\tau)$ for each sentence
Probabilistic Context-Free Grammars

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

<table>
<thead>
<tr>
<th>Rule</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow NP \ VP$</td>
<td>0.8</td>
</tr>
<tr>
<td>$S \rightarrow S \ conj \ S$</td>
<td>0.2</td>
</tr>
<tr>
<td>$NP \rightarrow \text{Noun}$</td>
<td>0.2</td>
</tr>
<tr>
<td>$NP \rightarrow \text{Det Noun}$</td>
<td>0.4</td>
</tr>
<tr>
<td>$NP \rightarrow NP \ PP$</td>
<td>0.2</td>
</tr>
<tr>
<td>$NP \rightarrow NP \ conj \ NP$</td>
<td>0.2</td>
</tr>
<tr>
<td>$VP \rightarrow \text{Verb}$</td>
<td>0.4</td>
</tr>
<tr>
<td>$VP \rightarrow \text{Verb} \ NP$</td>
<td>0.3</td>
</tr>
<tr>
<td>$VP \rightarrow \text{Verb} \ NP \ NP$</td>
<td>0.1</td>
</tr>
<tr>
<td>$VP \rightarrow VP \ PP$</td>
<td>0.2</td>
</tr>
<tr>
<td>$PP \rightarrow P \ NP$</td>
<td>1.0</td>
</tr>
</tbody>
</table>
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
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

beach

supermarket
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


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 study"
    - **S:** (n) **attribute** (an abstraction belonging to or characteristic of an entity)
      - **S:** (n) **abstraction, abstract entity** (a general concept formed by extracting common characteristic)
      - **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 accompanied)
    - *direct hypernym / inherited hypernym / sister term*
      - **S:** (n) **part, voice** (the melody carried by a particular voice or instrument in polyphonic music) "he
tune, melody, air, strain, melodic line, line, melodic phrase" (a succession of notes forming a unit)
      - **S:** (n) **music** (an artistic form of auditory communication incorporating instrumental or vocal ensemble)
        - **S:** (n) **auditory communication** (communication that relies on hearing)
      - **S:** (n) **communication** (something that is communicated by or to or between)
        - **S:** (n) **abstraction, abstract entity** (a general concept formed by extracting common characteristic)
        - **S:** (n) **entity** (that which is perceived or known or inferred to have)
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  
\[ \exists x \exists y (\text{pod\_door}(x) \& \text{Hal}(y) \& \text{request}(\text{open}(x, y))) \]
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)
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.
A girl plays on the beach.

$$\exists e \exists x \exists y \text{girl}'(x) \& \text{beach}'(y) \& \text{play}'(e) \& \text{agent}(e, x) \& \text{location}(e, y)$$
Denotational Semantics

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

\[
[s] = \{ w \in U : \text{s is true in } w \}
\]

Denotations capture **entailment**:

\[
\text{s entails s'} \text{ if } [s] \subseteq [s']
\]
Denotational semantics from image descriptions

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

Language L

A man pets a stingray.
A girl plays on the beach.
A child plays.

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:

\[
\llbracket s \rrbracket = \{ i \in I : s \text{ describes (part of) } i \}
\]
Visual Denotations

Language L

A man pets a stingray.
A girl plays on the beach.
A child plays.

Images I
Our corpus

Language $L$  Images $I$
The visual denotation of most captions in our corpus is a **singleton**:

\[
\left[ \begin{array}{c}
\text{A player from the white and green highschool team dribbles down court defended by a player from the other team}
\end{array} \right]
\]

\[
\Rightarrow \{\}
\]
Each image is in the denotation of multiple captions:

\[ e \in \begin{cases} \text{A player from the white and green highschool team dribbles down court defended by a player from the other team} \\ \text{Two boys in green and white uniforms play basketball with two boys in blue and white uniforms} \end{cases} \]
This allows us to capture relations between different descriptions of the same event:

\[ \in \left[ \text{dribble down court} \right] \]

\[ \in \left[ \text{play basketball} \right] \]
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

A child plays

A child plays guitar

A girl plays

A child plays on the beach

A child in red plays, on the beach

A girl plays on the playground

A girl plays on the beach
Parent Node

[ girl plays ]

Child Node

[ blond girl plays ]

identify transformations bottom-up

build graph top-down

drop modifier

add modifier
Parent Node

⟦ girl plays ⟧

Child Node

⟦ girl plays on the beach ⟧

drop modifier

add modifier
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 \([s] \geq 2\)
~53K captions with \([s] \geq 5\)
~22K captions with \([s] \geq 10\)
~1.9K captions with \([s] \geq 100\)
161 captions with \([s] \geq 1000\)

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

Young et al., From image descriptions to visual denotations, TACL 2014
The denotation graph allows us to estimate the denotational similarity of sentences/phrases.
Denotational Similarities

\[ P[\square](x) = \frac{|\llbracket x \rrbracket \cup \bigcup U|}{|\llbracket x \rrbracket \bigcup \bigcup U|} \]
\[ P[\square](x, y) = \frac{|\llbracket x \rrbracket \cap \llbracket y \rrbracket |}{|\llbracket x \rrbracket \bigcup \llbracket y \rrbracket |} \]

Two kinds of denotational similarities:
- Conditional probabilities \( P[\square](x \mid y) \)
- Normalized Pointwise Mutual Information \( \text{nPMI}_{\square}(x, y) \)
### Denotational similarities

\[
p( \text{VP}_1 \mid \text{VP}_2 )
\]

<table>
<thead>
<tr>
<th>Action 1</th>
<th>Action 2</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>talk</td>
<td>engage in conversation</td>
<td>0.79</td>
</tr>
<tr>
<td>play tennis</td>
<td>swing racket</td>
<td>0.82</td>
</tr>
<tr>
<td>stand</td>
<td>wait for subway</td>
<td>0.58</td>
</tr>
<tr>
<td>sit</td>
<td>ride subway</td>
<td>0.56</td>
</tr>
<tr>
<td>stand</td>
<td>lean against building</td>
<td>0.53</td>
</tr>
<tr>
<td>shave</td>
<td>look in mirror</td>
<td>0.41</td>
</tr>
<tr>
<td>dig hole</td>
<td>use shovel</td>
<td>0.38</td>
</tr>
<tr>
<td>make face</td>
<td>stick out tongue</td>
<td>0.38</td>
</tr>
</tbody>
</table>
Denotational similarities

<table>
<thead>
<tr>
<th>nPMI<a href="VP_1,VP_2">^</a></th>
<th>open present</th>
<th>unwrap</th>
<th>0.84</th>
</tr>
</thead>
<tbody>
<tr>
<td>lasso</td>
<td>try to rope</td>
<td></td>
<td>0.83</td>
</tr>
<tr>
<td>get ready to kick</td>
<td>run towards ball</td>
<td></td>
<td>0.79</td>
</tr>
<tr>
<td>try to tag</td>
<td>slide into base</td>
<td></td>
<td>0.79</td>
</tr>
<tr>
<td>shave face</td>
<td>look into mirror</td>
<td></td>
<td>0.77</td>
</tr>
</tbody>
</table>
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

**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.*
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.
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
∃x∃y(pod_door(x) & Hal(y) & request(open(x, y)))

Multimodal NLP:

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.*