Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 8: Introduction to NLP
Warning!

I am not an NLP researcher ...
Goal of NLP

**Fundamental goal:** *deep understanding of broad language* (going beyond string processing or keyword matching!)

*slide adopted from Dan Klein*
Goal of **NLP**

**Fundamental goal:** *deep understanding of broad language*

(going beyond string processing or keyword matching!)

**End systems** we want to build:

**Ambitious / Complex:**
- speech recognition
- machine translation
- information extraction
- dialog interfaces / understanding
- question answering

**Modest / Less complex:**
- spelling correction
- parts of speech tagging
- text categorization

*slide adopted from Dan Klein*
Why **NLP** is hard?

1. Human language is **ambiguous**

*slide adopted from Dan Klein*
Why **NLP** is hard?

1. Human language is **ambiguous**

**Task:** pronoun resolution

Jack drank the wine on the table. **It** was red and round.

Example adapted from Wilks (1975)
Why **NLP** is hard?

1. Human language is **ambiguous**

**Task:** pronoun resolution

Jack drank the wine on the table. **It** was red and round.

Example adapted from Wilks (1975)
Why **NLP** is hard?

1. Human language is **ambiguous**

**Task:** pronoun resolution

Jack drank the wine on the table. *It* was red and round.

Jack saw Sam at the party. *He* went back to the bar to get another drink.

Example adapted from Wilks (1975)
Why **NLP** is hard?

1. Human language is **ambiguous**

**Task:** pronoun resolution

Jack drank the wine on the table. *It* was red and round.

Jack saw Sam at the party. *He* went back to the bar to get another drink.

Jack saw Sam at the party. *He* clearly had drunk too much.

Example adapted from Wilks (1975)
Why NLP is hard?

1. Human language is ambiguous

**Task:** preposition attachment

I ate the bread with pecans.

I ate the bread with fingers.
Why **NLP** is hard?

1. Human language is **ambiguous**

**Task:** preposition attachment

I ate the **bread** with pecans.

I **ate** the bread with fingers.

Despite the structure of the two sentences being identical, the two prepositional phrases relate to different POS (noun vs. verb)

*slide from Yejin Choi*
Why NLP is hard?

1. Human language is **ambiguous**

2. Requires **reasoning** beyond what is explicitly mentioned (a, b) and some of reasoning requires **world knowledge** (c)

*slide from Yejin Choi*
Why **NLP** is hard?

1. Human language is **ambiguous**

2. Requires **reasoning** beyond what is explicitly mentioned (a, b) and some of reasoning requires **world knowledge** (c)

**Example:** I couldn’t submit the homework because my horse ate it.

*slide from Yejin Choi*
Why NLP is hard?

1. Human language is **ambiguous**

2. Requires **reasoning** beyond what is explicitly mentioned (a, b) and some of reasoning requires **world knowledge** (c)

**Example:** I couldn’t submit the homework because my horse ate it.

(a) I have a horse.

(b) I did my homework.

(c) My homework was done on soft material (like paper) as opposed to on hard/heavy object (like a computer).

**Reasoning:** It is more likely horse ate paper than a computer.

*slide from Yejin Choi*
Why **NLP** is hard?

1. Human language is **ambiguous**

2. Requires **reasoning** beyond what is explicitly mentioned (a, b) and some of reasoning requires **world knowledge** (c)

3. Language is difficult even for humans

---

Learning **native language** you may think is easy (but compare 5 / 10 / 20 year old)

Learning **foreign language(s)** — even harder

*slide from Yejin Choi*
Is **NLP** really this hard?

In the back of your mind, if you’re thinking …

“My native language is so easy. How hard could it be to type all the grammar rules, and idioms, etc. into software program? Sure it might take a while, but with enough people and money, it should be doable!”

… you are not alone!

*slide from Yejin Choi*
Short History of NLP

Birth of NLP and Linguistics

- Initially people thought NLP was easy
- Predicted “machine translation” can be solved in 3 years
- Hand-coded rules / linguistic oriented approaches
- The 3 year project continued for 10 years with no good results (despite significant expenditures)

*slide adopted from Yejin Choi*
**Short History of NLP**

**Dark Era**

- After initial hype, people believed NLP was impossible
- NLP research is mostly abandoned

*slide adopted from Yejin Choi*
Short History of NLP

Slow Revival of NLP

- Some research activities resumed
- Still emphasis on linguistically oriented approaches
- Working on small toy problems with weak empirical evaluation

*slide adopted from Yejin Choi*
Short History of NLP

Statistical Era / Revolution

– Computational power has increased substantially
– Data-driven, statistical approaches with simple representations win over complex hand-coded linguistic rules

*slide adopted from Yejin Choi
**Short History of NLP**

**Statistical Era / Revolution**

- Computational power has increased substantially
- Data-driven, statistical approaches with simple representations win over complex hand-coded linguistic rules
- “Whenever I fire a linguist our machine translation performance improves”

*Jelinek 1988*
Short History of NLP

Statistics Powered by Linguistic Insights

- More sophisticated statistical models
- Focus on new richer linguistic representations

*slide adopted from Yejin Choi*
Ambiguity is **Explosive**

**Ambiguities compound** to generate enormous number of interpretations

In English, sentence ending in N propositional phrases has over $2^N$ syntactic interpretations

*slide from Ray Mooney*
Ambiguity is **Explosive**

**Ambiguities compound** to generate enormous number of interpretations

In English, sentence ending in N propositional phrases has over $2^N$ syntactic interpretations

**Example:**

— I saw a man with the telescope.

*slide from Ray Mooney*
Ambiguity is **Explosive**

**Ambiguities compound** to generate enormous number of interpretations

In English, sentence ending in N propositional phrases has over \(2^N\) syntactic interpretations

**Example:**

- I saw a man with the telescope. -> 2 parses

*slide from Ray Mooney*
Ambiguity is **Explosive**

**Ambiguities compound** to generate enormous number of interpretations

In English, sentence ending in N propositional phrases has over \(2^N\) syntactic interpretations

**Example:**

— I saw a man with the telescope.
Ambiguity is **Explosive**

**Ambiguities compound** to generate enormous number of interpretations

In English, sentence ending in N propositional phrases has over $2^N$ syntactic interpretations

**Example:**

- I saw a man with the telescope. -> 2 parses
- I saw a man on the hill with the telescope. -> 5 parses
- I saw a man on the hill in Texas with the telescope. -> 14 parses
- I saw a man on the hill in Texas with the telescope at noon. -> 42 parses
- I saw a man on the hill in Texas with the telescope at noon on Monday. -> 132 parses

*slide from Ray Mooney*
Many **jokes rely on ambiguity** of language:

— Groucho Marx: “One morning I shot an elephant in my pajamas. How he got into my pajamas, I’ll never know”.

— Noah took all of the animals on the ark in pairs. Except the worms, they came in apples.

— Policeman to little boy: “We are looking for a thief with a bicycle.” Little boy: “Wouldn’t you be better using your eyes.”

— Why is the teacher wearing sun-glasses. Because the class is so bright.
Why is Language Ambiguous?

— Having a **unique linguistic expression** for every possible conceptualization that could be conveyed would make language **overly complex** and linguistic expressions unnecessarily long.

— Allowing **resolvable ambiguity** permits shorter linguistic expression, i.e., data compression.

— Language relies on people’s ability to use their **knowledge and inference abilities to properly resolve ambiguities**.

— Infrequently, disambiguation fails, i.e., the **compression is lossy**.
Natural vs. Computer Languages

— Ambiguity is the primary difference between natural and computer languages

— Formal programming languages are designed to be unambiguous, i.e., they can be defined by a grammar and produce a unique parse for each sentence (line of code) in the language.

— Programming languages are also designed for efficient (deterministic) parsing

*slide from Ray Mooney*
1. Word **segmentation**

   - Breaking a string of characters into a sequence of words.
   - In some written languages (e.g., Chinese) words are not separated by spaces.

*slide adopted from Ray Mooney
Syntactic NLP Tasks

1. Word **segmentation**

2. **Morphological** analysis
   - **Morphology** - field of linguistics that studies the internal structure or words
   - A **morpheme** is the smallest linguistic unit that has semantic meaning
   - Morphological analysis is the task of segmenting a word into morphemes

<table>
<thead>
<tr>
<th>carried</th>
<th>carry + ed (past tense)</th>
</tr>
</thead>
<tbody>
<tr>
<td>independently</td>
<td>in + (depend + ent) + ly</td>
</tr>
</tbody>
</table>
Syntactic NLP Tasks

1. Word **segmentation**

2. **Morphological** analysis

3. Parts of Speech (**POS**) tagging
   - Annotate each word in a sentence with a pat-of-speech

   - Useful for other language (e.g., syntactic parsing) and vision + language tasks
Syntactic NLP Tasks

1. Word segmentation

2. Morphological analysis

3. Parts of Speech (POS) tagging
   - Annotate each word in a sentence with a pat-of-speech

   I ate the spaghetti with meatballs.
   Pro V Det N Prep N

   John saw the saw and decided to take it to the table.
   PN V Det N Con V Part V Pro Prep Det N

   - Useful for other language (e.g., syntactic parsing) and vision + language tasks

*slide adopted from Ray Mooney*
Syntactic NLP Tasks

1. Word **segmentation**

2. **Morphological** analysis

3. Parts of Speech (**POS**) tagging

4. **Phrase** Chunking
   - Find all noun phrases (NPs) and verb phrases (VPs) in a sentence

   - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].

   - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP 1.8 billion].

*slide adopted from Ray Mooney*
Syntactic NLP Tasks

1. Word **segmentation**

2. **Morphological** analysis

3. Parts of Speech (**POS**) tagging

4. **Phrase** Chunking

5. **Syntactic** parsing

*slide adopted from Ray Mooney*
Semantic NLP Tasks

1. Word **Sense Disambiguation** (WSD)

- Words in language can have multiple meanings

  - Ellen has strong **interest** in computational linguistics.
  
  - Ellen pays a large amount of **interest** on her credit card.

- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined

*slide adopted from Ray Mooney*
Semantic NLP Tasks

1. Word **Sense Disambiguation** (WSD)

2. **Semantic Role** Labeling (SRL)

   — For each clause, determine the semantic role played by each noun phrase that is an argument to the verb

   — John drove Mary from Austin to Dallas in his Toyota Prius.

   — The hammer broke the window.

*slide adopted from Ray Mooney*
Semantic NLP Tasks

1. Word **Sense Disambiguation** (WSD)

2. **Semantic Role** Labeling (SRL)

   — For each clause, determine the semantic role played by each noun phrase that is an argument to the verb

   — John drove Mary from Austin to Dallas in his Toyota Prius.

   — The hammer broke the window.
Semantic NLP Tasks

1. Word **Sense Disambiguation** (WSD)

2. **Semantic Role** Labeling (SRL)

3. Textural **Entailment**

   — Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.
Semantic NLP Tasks

1. Word **Sense Disambiguation** (WSD)

2. **Semantic Role** Labeling (SRL)

3. Textural **Entailment**
   - Determine whether one natural language sentence entails (implies) another under an ordinary interpretation.
   - Note, you can think of images entailing captions … [Vendrov et al, 2015]

Sign with a spray paint over it.

*slide adopted from Ray Mooney*
<table>
<thead>
<tr>
<th>Text</th>
<th>Hypothesis</th>
<th>Entailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</td>
<td>Yahoo bought Overture.</td>
<td>TRUE</td>
</tr>
<tr>
<td>Microsoft’s rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</td>
<td>Microsoft bought Star Office.</td>
<td>FALSE</td>
</tr>
<tr>
<td>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</td>
<td>Israel was established in May 1971.</td>
<td>FALSE</td>
</tr>
<tr>
<td>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</td>
<td>Israel was established in 1948.</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

*slide adopted from Ray Mooney*
## Textual Entailment

<table>
<thead>
<tr>
<th>TEXT</th>
<th>HYPOTHESIS</th>
<th>ENTAILMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</td>
<td>Yahoo bought Overture.</td>
<td>TRUE</td>
</tr>
<tr>
<td>Microsoft’s rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</td>
<td>Microsoft bought Star Office.</td>
<td></td>
</tr>
<tr>
<td>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</td>
<td>Israel was established in May 1971.</td>
<td></td>
</tr>
<tr>
<td>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</td>
<td>Israel was established in 1948.</td>
<td></td>
</tr>
</tbody>
</table>

*slide adopted from Ray Mooney*
<table>
<thead>
<tr>
<th>TEXT</th>
<th>HYPOTHESIS</th>
<th>ENTAILMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</td>
<td>Yahoo bought Overture.</td>
<td>TRUE</td>
</tr>
<tr>
<td>Microsoft’s rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</td>
<td>Microsoft bought Star Office.</td>
<td>FALSE</td>
</tr>
<tr>
<td>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</td>
<td>Israel was established in May 1971.</td>
<td>TRUE</td>
</tr>
<tr>
<td>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</td>
<td>Israel was established in 1948.</td>
<td></td>
</tr>
</tbody>
</table>

*slide adopted from Ray Mooney*
## Textual Entailment

<table>
<thead>
<tr>
<th>TEXT</th>
<th>HYPOTHESIS</th>
<th>ENTAILMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</td>
<td>Yahoo bought Overture.</td>
<td>TRUE</td>
</tr>
<tr>
<td>Microsoft’s rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</td>
<td>Microsoft bought Star Office.</td>
<td>FALSE</td>
</tr>
<tr>
<td>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</td>
<td>Israel was established in May 1971.</td>
<td>FALSE</td>
</tr>
<tr>
<td>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</td>
<td>Israel was established in 1948.</td>
<td></td>
</tr>
<tr>
<td>TEXT</td>
<td>HYPOTHESIS</td>
<td>ENTAILMENT</td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>------------</td>
</tr>
<tr>
<td>Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.</td>
<td>Yahoo bought Overture.</td>
<td>TRUE</td>
</tr>
<tr>
<td>Microsoft’s rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.</td>
<td>Microsoft bought Star Office.</td>
<td>FALSE</td>
</tr>
<tr>
<td>The National Institute for Psychobiology in Israel was established in May 1971 as the Israel Center for Psychobiology by Prof. Joel.</td>
<td>Israel was established in May 1971.</td>
<td>FALSE</td>
</tr>
<tr>
<td>Since its formation in 1948, Israel fought many wars with neighboring Arab countries.</td>
<td>Israel was established in 1948.</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

*slide adopted from Ray Mooney*
Pragmatics and Discourse Tasks

Determine which phrases in a document refer to the same underlying entity

– John put the carrot on the plate and ate it.

– Bush started the war in Iraq. But the president needed the consent of Congress.

*slide adopted from Ray Mooney
Pragmatics and **Discourse** Tasks

Determine which phrases in a document refer to the same underlying entity

– John put the **carrot** on the **plate** and ate **it**.

– Bush started the war in Iraq. But the **president** needed the consent of Congress.

Some cases require difficult reasoning

– Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."