Introduction to NLP

Why is NLP hard?
Example

Time flies like an arrow.

• How many different interpretations does the above sentence have?
• How many of them are reasonable/grammatical?
Quiz Answer

- The most obvious meaning is
  - time flies very fast; as fast as an arrow.

- This is a metaphorical interpretation.
  - Computers are not really good at metaphors.

- Other interpretations:
  - Flies like honey → flies like an arrow → fruit flies like an arrow
  - Take a stopwatch and time the race → time the flies
More Classic Examples

• Beverly Hills
• Beverly Sills
• The box is in the pen
• The pen is in the box
• Mary and Sue are mothers
• Mary and Sue are sisters
• Every American has a mother
• Every American has a president
• We gave the monkeys the bananas because they were hungry
• We gave the monkeys the bananas because they were over-ripe
Syntax vs. Semantics

* Little a has Mary lamb.
? Colorless green ideas sleep furiously.

[Chomsky 1957]
Ambiguous Words

• ball, board, plant
  – meaning
• fly, rent, tape
  – part of speech
• address, resent, entrance, number, unionized
  – pronunciation – give it a try
Answer to the quiz

- **address**
  - The stress can be on either syllable. Compare with transport, effect, outline

- **resent**
  - As a verb infinitive or as “re-sent” a letter

- **entrance**
  - As a noun or as a verb meaning to put someone in a trance

- **number**
  - As a noun but also as the comparative of the adjective “numb”
Ambiguity

- Not in computer languages (by design)!
- Or Lojban
- Noun–noun phrases: (XY)Z vs. X(YZ)
  - science fiction writer
  - customer service representative
  - state chess tournament
One Two Tree, by Noah Smith, Kevin Gimbel, and Jason Eisner

English has the wonderful feature that it lets you stick two nouns together into a **compound noun**, whose meaning derives in some idiosyncratic way from the meanings of its parts:

- **water fountain**: a fountain that *supplies* water
- **water ballet**: a ballet that *takes place in* water
- **water meter**: a device (called meter) that *measures* water
- **water barometer**: a barometer that *uses* water instead of mercury (to measure air pressure)
- **water biscuit**: a biscuit that is *made with* water
- **water glass**: a glass that is *meant to hold* water

Even more fun is that one of the two nouns in the compound noun could itself be a compound noun, as in the case of *ice cream soda*. But what's the recipe for that beverage? It depends. You make [[*ice cream* soda]] by dropping *ice cream* into *soda*, but you make [ice [[*cream soda*]]] by dropping *ice* into *cream soda*. 
The paragraph above used [square brackets] to distinguish two possible meanings of ice cream soda, one of them being the conventional meaning. Add brackets to each compound below to indicate whether the most likely meaning corresponds to [[X Y] Z] or [X [Y Z]].

a. ice cream soda
b. science fiction writer
c. customer service representative
d. state chess tournament
e. Mars Rover landing
f. plastic water cooler
g. typeface design report
R-2 Choose the most likely bracketing for the 4-word compound noun country song platinum album.

a. [country [song [platinum album]]]
b. [country [[song platinum] album]]
c. [[country song] [platinum album]]
d. [[country [song platinum]] album]
e. [[[country song] platinum] album]

R-3 Give a plausible definition of [[space mission] [[control freak] show]]. (If you must use compound nouns in your definition, define them too.)

R-4 Show the most likely bracketing for the 8-noun sequence below. As in the examples above, your bracketing must have the form [X Y], where each of X and Y is either a single-word noun or a compound noun (which must also be written as a bracketing [X Y] and so on.)

family    board    game    togetherness    effect    government    study    author
A computer program knows less about the world than you do, so it may have more trouble interpreting these sequences of nouns. How many bracketings must it choose among? Complete the following table by inserting the correct numbers for $f(5)$, $f(6)$, and $f(7)$.

<table>
<thead>
<tr>
<th>Number of words (call this $n$)</th>
<th>Number of bracketings (call this $f(n)$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

(Should be obvious)
(Should be obvious)
(As in R-1)
(As in R-2)
(As in R-3)
R-7  For each phrase below, list all possible bracketings. The definition of “bracketing” is the same as in R-4, except that now X is also allowed to be a single-word adjective.

a. roasted red potato pancake

b. crazy monkey cheap cider house
Solution

R-1

a. [[ice cream] soda]
b. [[science fiction] writer]
c. [[customer service] representative]
d. [state [chess tournament]]
e. [[Mars Rover] landing]
f. [plastic [water cooler]]
g. [[typeface design] report]

R-2

[[[country song] [platinum album]]]
R-3

Default answer: A drama about control freaks (i.e., freaks about control), performed during a space mission (i.e., a mission to space).

Many other answers are possible as long as each of the bracketings is correctly defined. Below are examples of correct answers for each bracketing:

For "[control freak]":
- a person who is obsessive about having things his way

For "[[control freak] show]":
- a show that is run by control freaks
- a show that contains control freaks (i.e., the actors are control freaks)
- a show that is designed for or intended for control freaks
- a display of behavior by a control freak

For "[space mission]":
- a mission into space
For "[[space mission] [[control freak] show]]":
- a control freak show that is broadcast to audiences on space missions
- a control freak show that is set on a space mission
- a control freak show that is about space missions
- a display of behavior by control freaks; the display is witnessed on a space mission

Some examples of incorrect answers:
- the show contains control freaks that are interested in space missions
- the show contains control freaks who run or are on a space mission

These are incorrect because they attach "space mission" to "control freak" instead of attaching it to "control freak show". For these answers to be correct, the bracketing would have to be "[[[space mission] [control freak]] show]".
R-4

[[[[family [board game]] [togetherness effect]] [government study]] author].

Although the following might also be defensible:
[[[[family [[board game] togetherness]] effect] [government study]] author]

Or even perhaps:
[[[[family [board game]] togetherness] effect] [government study]] author]
\( f(5) = 14 \)
\( f(6) = 42 \)
\( f(7) = 132 \)

There are \( f(5) \) bracketings of *togetherness effect government study author* — whatever \( f(5) \) turns out to be! Similarly there are \( f(3) \) \((= 2)\) bracketings of *family board game*. So you have to list \( f(3) \cdot f(5) \) bracketings that split the 8-word sequence into 3 words + 5 words like this. But the full list for \( f(8) \) must also consider other splits, such as 1 word + 7 words. The general principle is that

\[
 f(n) = \sum_{k=1}^{n-1} f(k) \cdot f(n - k) \quad \text{for any } n > 1
\]

You can therefore compute each line in the table from the previous lines. By the way, the resulting sequence of numbers is called the Catalan numbers: you can look it up.
a. There are only 3 bracketings (fewer than $f(4)=5$ because the rules from R-6 are “missing”):

- roasted [red [potato pancake]] - a roasted red pancake made of potatoes
- roasted [[red potato] pancake] - a roasted pancake made of red potatoes
- [[roasted [red potato]]] pancake - a pancake made of roasted red potatoes

Note that the 4th logical possibility, a red pancake made of roasted potatoes, is not consistent with this word order: you’d have to call it a *red roasted potato pancake*.

b. There are 7 bracketings (fewer than $f(5)=14$):

- [[crazy monkey] [[cheap cider] house]] - the house of crazy monkeys serves cider that is cheap
- [[crazy monkey] [cheap [cider house]]] - the house of crazy monkeys that serves cider is cheap
- [crazy [monkey [[cheap cider] house]]] - the crazy house of monkeys serves cider that is cheap
- [crazy [monkey [cheap [cider house]]]] - the crazy house of monkeys that serves cider is cheap
- [crazy [[monkey [cheap cider]] house]] - the crazy house serves cheap cider that’s for monkeys
- [[[crazy monkey] [cheap cider]] house] - the house serves cheap cider that’s for crazy monkeys
- [[[crazy [monkey [cheap cider]]] house]] - the house serves crazy, cheap cider that’s for monkeys

Again, there are logical possibilities that are not consistent with the word order, such as a house of monkeys that serves crazy, cheap cider.
NACLO Problem

• Fakepapershelfmaker, by Willie Costello
In English, we can combine two nouns to get a compound noun, such as in ‘mailbox’ or ‘sandcastle’. We can do this in Japanese as well, but just sticking the two words together isn’t enough. Instead, the words themselves undergo predictable changes:

\[
\begin{align*}
\text{ikebana} & \rightarrow \text{ike} + \text{hana} \\
\text{asagiri} & \rightarrow \text{asa} + \text{kiri} \\
\text{hoshizora} & \rightarrow \text{hoshi} + \text{sora}
\end{align*}
\]

‘flower arranging’  ‘morning fog’  ‘starry sky’

‘arrange’  ‘flower’  ‘morning’  ‘fog’  ‘star’  ‘sky’

Compound words can then be compounded again, creating compounds with three or more members. Study the diagrams below carefully. You’ll notice that the order in which the compound is built affects both the meaning and the final form of the word.

\[
\begin{align*}
\text{nuribashibako} & \rightarrow \text{hashibako} + \text{nuribashi} \\
\text{nurihashibako} & \rightarrow \text{nuri} + \text{hashi} + \text{hako}
\end{align*}
\]

‘lacquered box for chopsticks’  ‘box for lacquered chopsticks’

‘box for chopsticks’  ‘lacquered chopsticks’

‘lacquered’  ‘chopsticks’  ‘box’  ‘lacquered’  ‘chopsticks’  ‘box’
**F1.** The following is a list of several Japanese words with their English meanings. Use this word bank to write definitions of the Japanese compounds (a)-(f). Be very specific with how you phrase your definition. If your definition is ambiguous (has two meanings), it will not be counted.

<table>
<thead>
<tr>
<th>sakura</th>
<th>cherry blossom</th>
<th>kami</th>
<th>paper</th>
<th>nise</th>
<th>fake</th>
</tr>
</thead>
<tbody>
<tr>
<td>shiru</td>
<td>soup</td>
<td>tana</td>
<td>shelf</td>
<td>tsukuri</td>
<td>maker</td>
</tr>
<tr>
<td>iro</td>
<td>color(ed)</td>
<td>tanuki</td>
<td>raccoon</td>
<td>hako</td>
<td>box</td>
</tr>
</tbody>
</table>

(a) nisetanukijiru
(b) nisedanukijiru
(c) irogamibako
(d) irokamibako
(e) nisezakuradana
(f) nisesakuradana
F2. Match the following four-member Japanese compound words on the left with their English meanings on the right. (Some will require you to stretch your imagination a bit!) One of the Japanese words will correspond to two possible English meanings.

<table>
<thead>
<tr>
<th></th>
<th>(1) a fake (fraudulent) shelf-maker made of paper</th>
<th>(A) nisegamidanadzukuri</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2)</td>
<td>a maker of fake shelves for paper</td>
<td>(B) nisekamitanadzukuri</td>
</tr>
<tr>
<td>(3)</td>
<td>a fake (fraudulent) maker of shelves for paper</td>
<td>(C) nisegamitanadzukuri</td>
</tr>
<tr>
<td>(4)</td>
<td>a shelf-maker made of fake paper</td>
<td>(D) nisekamidanadzukuri</td>
</tr>
<tr>
<td>(5)</td>
<td>a maker of shelves for fake paper</td>
<td></td>
</tr>
</tbody>
</table>

F3. Explain your answers to F1 and F2 in the space provided below.
### Solution

**F1.** The following is a list of several Japanese words with their English meanings; use them to write definitions of the Japanese compounds.

<table>
<thead>
<tr>
<th>sakura</th>
<th>cherry blossom</th>
<th>kami</th>
<th>paper</th>
<th>nise</th>
<th>fake</th>
</tr>
</thead>
<tbody>
<tr>
<td>shiru</td>
<td>soup</td>
<td>tana</td>
<td>shelf</td>
<td>tsukuri</td>
<td>maker</td>
</tr>
<tr>
<td>iro</td>
<td>color(ed)</td>
<td>tanuki</td>
<td>raccoon</td>
<td>hako</td>
<td>box</td>
</tr>
</tbody>
</table>

(a) **nisetanukijiru**

fake soup made out of raccoons

(b) **nisedanukijiru**

soup made out of fake raccoons

(c) **irogamibako**

box for colored paper

(d) **irokamibako**

colored box for paper

(e) **nisezakuradana**

shelf for fake cherry blossoms

(f) **nisesakuradana**

fake shelf for cherry blossoms
F2. Match the following four-member Japanese compound words with their English meanings: one of the Japanese words has two possible meanings.

(1) a fake shelf-maker made of paper  B: nisekamitanadzukuri
(2) a maker of fake shelves for paper  D: nisekamidanadzukuri
(3) a fake maker of shelves for paper  D: nisekamidanadzukuri
(4) a shelf-maker made of fake paper  C: nisegamitanadzukuri
(5) a maker of shelves for fake paper  A: nisegamidanadzukuri
F3. Explain your answers.

When we compound two Japanese words, the first word modifies/describes the second. For example, adding hashi before hako makes a word meaning a box (hako) for chopsticks (hashi). As another example, adding nuri before hashi makes a word meaning chopsticks (hashi) that are lacquered (nuri).

Every simple (noncompound) word has two forms: the basic form, used when it occurs alone, and the variant form, sometimes used in compound words.

<table>
<thead>
<tr>
<th>Basic</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>hako</td>
<td>_bako</td>
</tr>
<tr>
<td>hana</td>
<td>_hana</td>
</tr>
<tr>
<td>hashi</td>
<td>_hashi</td>
</tr>
<tr>
<td>kami</td>
<td>_gami</td>
</tr>
<tr>
<td>kiri</td>
<td>_giri</td>
</tr>
<tr>
<td>sakura</td>
<td>_zakura</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basic</th>
<th>Variant</th>
</tr>
</thead>
<tbody>
<tr>
<td>shiru</td>
<td>_jiru</td>
</tr>
<tr>
<td>sora</td>
<td>_zora</td>
</tr>
<tr>
<td>tana</td>
<td>_dana</td>
</tr>
<tr>
<td>tanuki</td>
<td>_danuki</td>
</tr>
<tr>
<td>tsukuri</td>
<td>_dzukuri</td>
</tr>
</tbody>
</table>

The variant form has a different first letter, which depends on the first letter in the basic form. Specifically, we replace the initial h with b, initial k with g, initial s with z, initial sh with j, initial t with d, and initial ts with dz. As a side note, some letters do not require replacement, but they do not occur in the problem.
NACLO Problem Solutions

• One Two Tree

• Fakepapershelfmaker
Types of Ambiguity

• **Morphological:**
  – Joe is quite impossible. Joe is quite important.

• **Phonetic:**
  – Joe’s finger got number.

• **Part of speech:**
  – Joe won the first round.

• **Syntactic:**
  – Call Joe a taxi.

• **Prepositional phrase attachment:**
  – Joe ate pizza with a fork / with meatballs / with Samantha / with pleasure.

• **Sense:**
  – Joe took the bar exam.
Other Sources of Difficulty

- **Subjectivity:**
  - Joe believes that stocks will rise.

- **Cc attachment:**
  - Joe likes ripe apples and pears.

- **Negation:**
  - Joe likes his pizza with no cheese and tomatoes.

- **Referential:**
  - Joe yelled at Mike. He had broken the bike.
  - Joe yelled at Mike. He was angry at him.

- **Reflexive:**
  - John bought him a present.
  - John bought himself a present.

- **Ellipsis and parallelism:**
  - Joe gave Mike a beer and Jeremy a glass of wine.

- **Metonymy:**
  - Boston called and left a message for Joe.
Other Sources of Difficulties

• Non-standard, slang, and novel words and usages
  – A360, 7342.67, +1–646–555–2223
  – “spam” or “friend” as verbs
  – yolo, selfie, chillax – recently recognized as dictionary words
  – www.urbandictionary.com – (Parental Warning!)

• Inconsistencies
  – junior college, college junior
  – pet spray, pet llama

• Typoes and gramattical erorz 😊
  – reciept, John Hopkins, should of

• Parsing problems
  – Selbständigkeit (self-reliance)
  – cup holder
  – Federal Reserve Board Chairman
Other Sources of Difficulties

• Complex sentences
• Counterfactual sentences
• Humor and sarcasm
• Implicature/inference/world knowledge:
  – I was late because my car broke down.
  – Implies I have a car, I use the car to get to places, the car has wheels, etc.
  – What is not explicitly mentioned, what is world knowledge?
• Semantics vs. pragmatics
  – Do you know the time?
• Language is hard even for humans
  – Both first language and second language
The S&P 500 climbed 6.93, or 0.56 percent, to 1,243.72, its best close since June 12, 2001.

The Nasdaq gained 12.22, or 0.56 percent, to 2,198.44 for its best showing since June 8, 2001.

The DJIA rose 68.46, or 0.64 percent, to 10,705.55, its highest level since March 15.
The S&P 500 climbed 6.93, or 0.56 percent, to 1,243.72, its best close since June 12, 2001.
The Nasdaq gained 12.22, or 0.56 percent, to 2,198.44 for its best showing since June 8, 2001.
The DJIA rose 68.46, or 0.64 percent, to 10,705.55, its highest level since March 15.
NLP