**CS 7337 – Natural Language Processing**

**Midterm Exam**

**Instructions:** Clarity of answers is more important than length of answers. Although not required (unless indicated otherwise), feel free to use graphs, charts, visuals, et al in your answers if you feel these artifacts can help support your answers. There are no bonus points for using these artifacts.

**Due date:** 1 hour before Live Session 10, in PDF or Word DOCX file format.

**Q1. a.** **[5 pts]** **Define homonymy and polysemy and give an example of each.**

**Solution:**

**Homonymy** 🡺 Relation between words with identical forms but different meanings.

Examples: (Source: [Wordnet Link](http://wordnetweb.princeton.edu/perl/webwn?s=address&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=), [Homonymy](https://www.thoughtco.com/homonymy-words-and-meanings-1690839))

**Address (n)**: The place where a person or organization can be found or communicated with.

**Sentence:** Please deliver the mail to my home address.

**Address (v)**: Speak to (give a speech to)

**Sentence:** The President addressed the large gathering.

**Polysemy** 🡺 The coexistence of many possible meanings for a word or phrase.

Examples: (Source: [Wordnet Link](http://wordnetweb.princeton.edu/perl/webwn?s=key&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=000000000000000000000), [Polysemy](http://esl.fis.edu/teachers/support/vocabPoly.htm))

**Key (n):** metal device shaped in such a way that when it is inserted into the appropriate lock the lock's mechanism can be rotated

**Sentence:** This key does not fit the lock.

**Key (adj):** Cardinal or Fundamental

**Sentence:** The key issue was not addressed in our HOA meeting.

**b**. **[5 pts]** **Define NLU and NLG and give an example of each.**

**Solution:**

Natural Language Understanding (NLU) – Get the machine to produce a useful representation of some inputted natural language.

Examples: Named Entity Recognition, Sentiment Analysis, Query Understanding

Natural Language Generation (NLG) – Get the machine to produce usable, natural language output based on structured data.

Examples: Conversational agents (Siri/Alexa), Chatbots, Document summarization

**Q2.** **You are given the following grammar for expressions:**

E  I I  a

E  E + E I  b

E  E \* E I  0

E  (E) I  1

1. **[10 pts]** **Show parse tree(s) for the expression 1 + 1 \* 1**

**Solution:**

|  |  |
| --- | --- |
|  |  |
| **Left Derivation Tree** (Source: [YouTube](https://www.youtube.com/watch?v=QUvkwHECm1A)) |  |
| **Right Derivation Tree** |  |

**b.** **[10 pts]** **Describe any interesting observations in your answer to a**.

**Solution:**

The expression can be parsed 2 different ways. The addition can be done before or after the multiplication. This could lead to the parser to have 2 very different semantics.

The ambiguity could have been avoided by using parenthesis to specify the order of execution.

**Q3. Consider the following grammar and sentence:**

**A screenshot of a cell phone

Description automatically generated**

**Sentence**: *I booked a flight from LA*

**a.** **[10 pts]** **In what way is this sentence ambiguous? Describe different interpretations of this sentence.**

**Solution:**

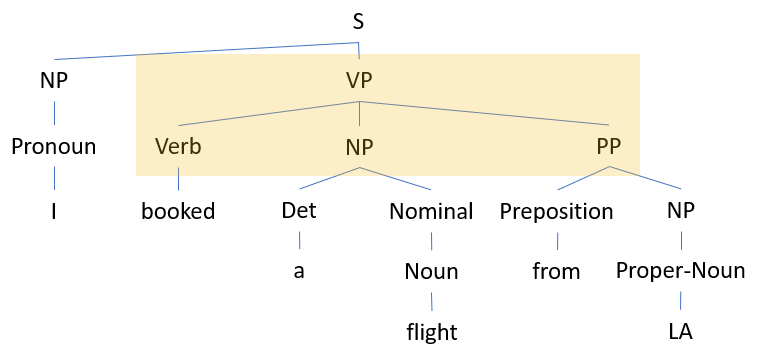
The above sentence can be interpreted as:

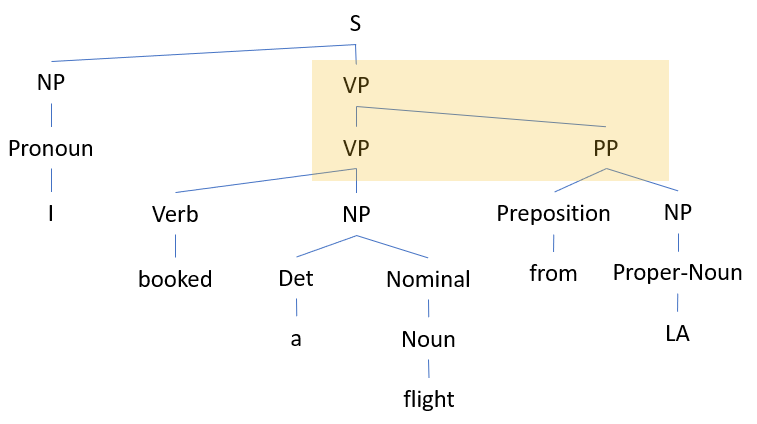
* I was in LA when I booked the flight.
* I booked a flight which starts from LA.

**b**. **[10 pts]** **Show the parse trees for this sentence and where the ambiguity manifests in the parse trees.**

**Solution:** (Source: <http://erg.delph-in.net/logon>)

The ambiguous part of the parse tree is highlighted.





**Q4. The image below shows Google search results for the query “harry potter”**

A screenshot of a social media post

Description automatically generated

As the results show, the query could represent any of the seven books in the harry potter franchise, any of the film adaptations of the books, a theme park, or a ride, an audiobook, cartoons, et al.

**a. [10 pts] Discuss why google shows a mix of such results and what factors can influence the search results for this query that will be presented to you.**

**Solution:**

When Google crawls the web, it only records the page code. (Source: [BERT](https://www.blog.google/products/search/search-language-understanding-bert/), [How google algorithm works](https://www.wordtracker.com/academy/google/how-it-works/how-google-algorithm-works)) The crawler hops from link-link to discover new pages. Google displays a mix of results due to the below reasons:

* Bidirectional Encoder Representations from Transformers (BERT) – Models process words in relation to all the other words in a sentence. These models consider the context of the word by inspecting the words that come before and after it.
* Indexing – The search results are indexed since searching the index is far quicker than searching the entire database each time.
* Ranking Algorithms – set of weighted metrics that determine the page rank for display on the results page.
* Relevance – based on placement of keywords, such as page titles and anchor text.
* Authority – based on PageRank (Link to a page with lot of votes) and Relevance (link with a relevant anchor)
* Trust – Age of content/domain, bad neighbors, duration to domain registration expiration, etc.
* Usability – down-weight a page with ads, page speed, etc. to determine the usability of the search results.
* Results type and Personalization – Device type, localized searches, mixed media content, previous search history determine the placement of results on the search page.

**b. [15 pts] Consider the following sentence**:

*The* ***bank*** *can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.*

The word **bank** has multiple senses. Use Wordnet to show the top two sense, glossaries and examples for **bank** and describe (at a high level) how you can use this information to find the proper sense for this word in a sentence.

Wordnet link: <http://wordnetweb.princeton.edu/perl/webwn>

**Solution:**

Source: [pywsd](https://github.com/alvations/pywsd), [Wordnet](http://wordnetweb.princeton.edu/perl/webwn?c=0&sub=Change&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&i=-1&h=000000000000000000&s=bank)

The top 2 noun senses of the word “**bank**“ are:

### Noun

* [S:](http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=bank&i=0&h=000000000000000000" \l "c) (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
* [S:](http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=bank&i=1&h=000000000000000000#c) (n) [depository financial institution](http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=depository+financial+institution), **bank**, [banking concern](http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=banking+concern), [banking company](http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=banking+company) (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*

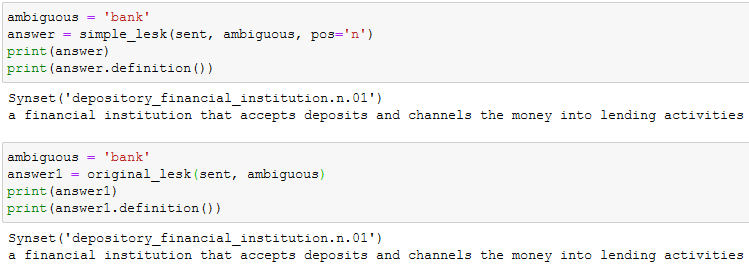
I’ve used pywsd to perform word sense disambiguation on the given sentence. The simple\_lesk and original\_lesk provided the same results but the usage was slightly different.

Wordnet is a lexical database for the English Language, created by Princeton and is part of the nltk corpus. It focuses on the relationship between words along with their definitions.

**Code:**

from pywsd.lesk import simple\_lesk, original\_lesk

sent = "The bank can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.".lower()



**Q5.** **You are building an online moving streaming service which enables looking up information on movies, genres, directors, actors and customer movie preferences.**

A close up of a device

Description automatically generated

\*\*(Indiana Jones refers to Raiders of the Lost Ark (1981))

**a. [10 pts] What is the customers intent (i.e. what are they looking for) with the following queries? (these are individual queries, not queries entered in succession)**

*“Drama”, “Jurassic Park”, “Indiana Jones: Raiders of the lost ark”, “Steven Spielberg”*

**Solution:**

Ontologies represent complex relationships between entities. In the above search scenario, the intent of the customer may be varied. In an ontological search, the result may be different from the starting point. It can be used as a relational search mechanism. I’ll dive deeper into the concept in the below explanation.

In the first scenario, the customer is looking to watch a movie of Drama genre. In the above ontology diagram, it would return Schindler’s List. However, the customer may decide that since only John likes the movie but not Mary, they may decide to watch other movies from the same director and settle on watching Jurassic Park since both Mary and John like the movie.

In the second scenario, the customer may have watched Jurassic Park and would like to watch something similar. He may start their search with a movie they liked in the past and move (Jurassic Park) on to watch a similar movie. Since Indiana Jones was directed by Steven Spielberg and its genre is Action, they may settle on watching Indiana Jones: Raiders of the lost ark.

The other scenarios are similar. The customers start with something that they are familiar with (Director, genre or movie) and proceed to find something with related to that entity.

**b. [5 pts] A customer searches for “Indiana Jones” but clicks on and watches “Jurassic Park” – what insights can you get from this customer action?**

**Solution:**

The customer may have already watched Indiana Jones and is looking for something similar. He may decide to watch Jurassic Park because it is directed by Spielberg and is of the same genre. As mentioned in the previous answer, customers may base their movie search based on their established preference and expand to include other movies based on similar genre, director, actors, etc.

**c. [10 pts] The customer searches for “Indiana Jones: Raiders of the lost Ark” but it’s not available in their region (US, EU, Asia). What search results would you show the customer? Discuss how you would build that experience from a technical design perspective.**

**Solution:**

If the customer searches for Indiana Jones: Raiders of the lost Ark, but it is not available in the region, we would have to show movies that are similar to the search query and is available in the region.

From a design perspective, we could add an additional relationship that shows whether the movie “is available” in a region. This will help the search engine return movies that are available in the query region.