

Midterm Exam CS 7337 NLP

Paritosh Rai



CS 7337 – Natural Language Processing



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: See instructors note; submission should be in PDF or Word DOCX file format.

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

Homonymy and polysemy are similar concepts, and both refers to words or phrases having multiple meanings. However, they are little different.

**Homonymy:** Homonymy are two or more words have the same spelling and same pronunciation but a totally different meaning and origins.

Example:

My answers were *right*.

Turn *right* from the intersection.

Right in above sentences have same spelling, very different mean and not related to each other. One is related to correctness of answer and other is related to direction of turn.

**Polysemy:** Polysemy refers to words or phrases with different but related meaning.

Example:

I am reading *newspaper*.

He sued the *newspaper*.

In above example, newspaper have same spelling, different meaning, but are related to each other. One related to actual newspaper where and in second sentence newspaper refers to newspaper’s office or company.

|  |  |
| --- | --- |
| Polysemy | Homonymy |
| Different but related meaning | Different and unrelated Meaning |
| Can understand if know meaning of one word | Can not guess the other meaning as other word is not related |

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

NLP = NLU + NLG

**NLU: Natural Language Understanding**

NLU is part of NLP that can take text or speech understands the meaning. NLU leverages the rules and structure of language and process/ analyze the information. NLU understand the text or speech and extracts the useful information for downstream task.

Example: Reading the review from the blog and extracting positive and negative reviews.

**NLG: Natural Language Generation**

NLG takes the given data, analyzes, and generates narratives in conversational language. It is the ability to generate natural language based on information processed in previous stages. It serves two purposes for content developer, automation of content generation and delivery in appropriate format.

Example: Auto generation of twitter response or auto generate response on social media sites.

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 → 2

1. [10 pts] Show parse tree(s) for the expression 2 + 2 \* 2

E => E + E

=> E + E \* E (as E=>E\*E)

=> I + I \* I (as E=> I)

=> 2 + 2 \* 2 (as I=> 2)

Following are two ways parse trees that show two derivations of the string 2+ 2\*2:

* Precedence to Multiplication
* Precedence to Addition





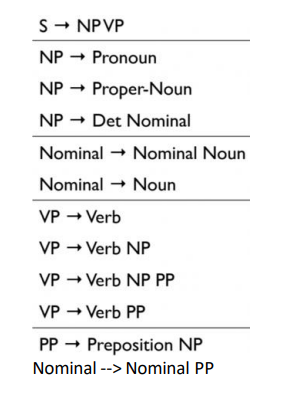
b. [10 pts] Describe any interesting observations in your answer to a.

Given any derivation of a string, it is possible to construct a parse tree that shows each of the steps in that derivation. However, two different derivations can give rise to the same parse tree since the parse tree does not show the order in which production rules are applied.

In case of Precedence of Multiplication resulting, value will be 6 and in case of precedence of Addition resulting value will be 8. So how rules are applied can impact the outcome. That is one of the reasons that different functions or library may give different outcome when effort is made to extract the PoS of same text or find right interpretation of text.

Q3. Consider the following grammar and sentence:

Nominal --> Nominal PP Sentence: I booked a flight from LA



Sentence: I booked a flight from LA

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

Sentence: I booked a flight from LA

The above sentence can be interpreted in two different ways:

First Interpretation: I booked the flight from (when I was in) LA

Second Interpretation: I booked the flight (, and my flight was starting) from LA.

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

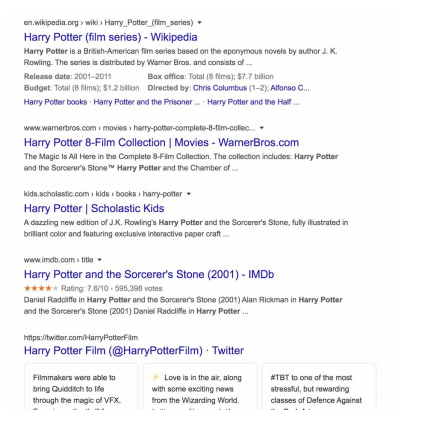
In First Interpretation Proposition Phrase (PP) is part of Verb Phrase (VP)



In Second Interpretation Proposition Phrase (PP) is part of Noun Phrase (NP)



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



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

1. [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.

Query or search on google is very generic. So, it looks for books, movie, social media, reviews, etc. Google maintains a vast library of the index. When a search is triggered in google, it instantly finds the things matching the search. It presents the results in the most helpful ways to find the most relevant links. It tries to add variety to outcomes in absences of guidance.

In our case of harry potter, it gave us information about book franchises, film, book reviews, and few others. Based on our pick of these links, it helps Google be intelligent to get the next step closer to what we are looking for and keep adjusting the query. Giving us multiple options in multiple fields allows google to streamline their next series of search results and modify the subsequent outcome in the right and focused direction.

Following are few factors that influence our outcome of search results:

Browsing/ Search History: Indicate interest of user’s interest and background.

Links clicked from first/previous outcome: Indicate what the user is looking for **now**, for example, gift, move, etc.

News: What is famous in the News today or at that moment (say there is a new Harry Potter store opened in your town)

Location: Something special happening in that location, say if the search is trigger outside move theater, then google can guess searcher is looking for a review of a movie or any deal on ticket price, or URL is banned in the country, or Harry Potter books are available in the local language.

User Information: Google keeps track of its users. As we search, we are logged into a Google account that gives google a lot of information about user demography, interest, etc., to align the search output based on user interest.

Get a more specific search: Sometimes, adding one or two more words helps google to get the outcome more quickly or closer to the user’s desire. Adding word like book to the search (harry potter book) will limit the search to the books and remove movies and other outcomes from the search outcome.

1. [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>

“[WordNet Search - 3.1 (princeton.edu)](http://wordnetweb.princeton.edu/perl/webwn?s=bank&sub=Search+WordNet&o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&h=)””

Top two sense on Bank in Wordnet are:

* [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#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"*

First one is related to bank of the river and other one is the financial institution. NLP uses Word Sense Disambiguation (WSD) to identify which meaning is used in ambiguous situation i.e., when word have multiple meaning like Polysemy (word have multiple senses). There are various ways to solve this problem, including lexical and dictionary-based methods and supervised and unsupervised machine learning methods.

Most popular used in nltk is Lexical based or **Lesk algorithm**. The basic principle behind this algorithm is to leverage dictionary or vocabulary definitions for a word we want to disambiguate in a body of text and compare the words in these definitions with a section of text surrounding our word of interest. Here we will be using the WordNet definitions for words (instead of a dictionary). The main objective is to return the synset with the maximum number of overlapping words or terms between the context sentence and the different definitions from each synset for the word we target for disambiguation. Following steps are followed:

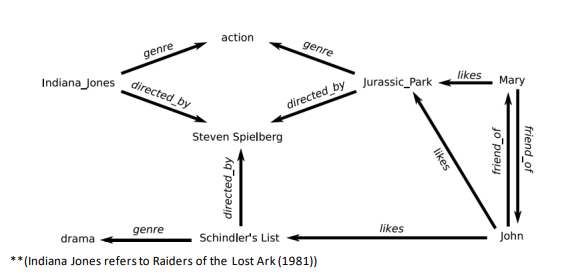
* Capture all the senses of the word from wordnet
* Compare each sense with the given text or sentence.
* Count the number of overlapping words between sense and given sentence
* Arrange the senses in the descending order of number of overlaps
* Use the sense with highest number of matching/overlap.

In our example

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

There is no common word with wordnet definitions on first defination. However, with second defination we see at least two words overlap. Therefore, second meaning makes more sense in given scenario.

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



1. [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”

Customer searches are many times motivated by recommendation of tools. It appears to impact customers' search intent, although these individual queries were not entered in succession. However, tools maintain the history of searches and the history of other customers who have done similar searches. The tool might be leveraging the knowledge of the customer and correlation of searches of other customers of giving the recommendation:

|  |  |  |
| --- | --- | --- |
| **Search** | **Customers Intend** | **Streaming service** |
| **Drama** | To watch Drama Movies (genre = “drama”) | One of the Drama movie given by tool was Schindler’s List. |
| **Jurassic Park** | Customer intention was to get information on movie Jurassic Park serise, genres, directors, actors and customer movie preferences. | John who likes dram movie like Schindler List also liked Jurassic Park. Based on previous search online tool might have recommended the Jurassic Park that might have triggered the search for customer. |
| **Indiana Jones: Raiders of the lost ark** | Customer search is very specific to a single move “Indiana Jones: Raiders of the lost ark”. He/ She is looking for information on movie, genres, director, actors, and customer movie preference. | The history of customer search show Jurassic Park serise of movies. Indiana Jones also has some director “Steven Spielberg”. |
| **Steven Spielberg** | Customer is looking for movies directed by Steven Spielberg (director = ‘Steven Spielberg) | Looking for Steven Spielberg makes perfect sense as Indiana Jones movie and Jurassic Park movies were directed by Steven Spielberg. |

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?

Based on Customer search (of “Indiana Jones”) and watching different move (“Jurassic Park”), I get the following insight about the customer:

Genre: The customer has a liking for action movies. As Indiana Jones and Jurassic Park, both are action movies.

Director: Customer likes movies directed by Steven Spielberg.

Like: Customer has similar liking as John and Mary.

Sequence: As Indian Jones has a series of movies; Jurassic Park also has a series of films. It can be an indicator that customer likes sequence movies. We can make a general observation or data point that if customers like one sequence of the movie. They may enjoy another sequence or series of movies as well.

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.

If the movie searched by customer “Indiana Jones: Raiders of the Lost Ark” is not available in the region (US, EU, Asia), I will work on a tool to give out by correlating the user preferences of past searches, wish lists, and choices they have made. I will look into option of relaxing the query to include other Indiana Jones movies, action movies, and other movies directed by Steven Spielberg.

I will take the following steps to build the experience from a technical design perspective:

Movie Database: Build a massive corpus of the movie with the movie name, director, year of release, genre, geo-restrictions, producers, media company, summary, actors, movie ratings, movie reviews…

User/Customer Database: Create a database that will keep growing and enhancing as customers make their choices and provide input. Some of the fields will be name, user id, avatar, location, age (or age group), gender (optional) etc. will be basic user information. The database will keep building as customers make choices.

Building correlation and Patterns: Based on choices made by the customer, new patterns and correlation will be derived. These patterns will allow us to recommend that customers and users watch the latest movies or what next, they would like to watch.

User Input: Build a front-end system to capture user input and searches.

Process Input: Create the right back-end system to capture the inputs like rating and reviews and take the feedback and process it to derive the proper outcomes.

* Capture the user input/search
* Process the input: normalize, tokenize, lemmatize, Disambiguation, spell correction, etc.

Custom Query Creation: User input, past correlation of data related to movies and user choices, and present input (current query + data correlation with other users + market trends + gender preferences + age group preferences + geo-restriction…) custom query is created.

Readjustment of Query/ Exception: Ensure to readjust the question to cast a broader net if the original custom query gives null results or a minimal outcome. For example, in the above ask, due to geo-restrictions, we cannot get “Indian Jones: Raiders of the Lost Ark,” the tool should be able to relax the query and just look for, say “Indian Jones.”

Sorting the Outcome: It is critical to sort the outcome based on query and user preferences

Display the Result: Display the output that is easy to understand and read.