UNIT 5: Discourse Processing

- Discourse processing in natural language processing (NLP) refers to the study and analysis of text beyond the level of individual sentences.
- It focuses on understanding the connections, relationships, and coherence between sentences and larger units of text, such as paragraphs and documents.
- Discourse processing aims to capture the overall meaning, structure, and flow of a piece of text, taking into account various linguistic and contextual factors.
- The main goal of discourse processing is to extract meaningful information and infer the intended meaning from a text, which can be useful in a variety of NLP applications, such as text summarization, sentiment analysis, question answering, and information extraction.
- It involves 4 subtasks, including:
    - Coherence and cohesion analysis
    - Discourse structure analysis
    - Rhetorical parsing
    - Discourse-level sentiment analysis

Coherence and cohesion analysis

- This involves examining the relationships between sentences and identifying how they connect to form a coherent and cohesive text.
- It includes identifying discourse markers (e.g., "however," "therefore"), and coreference resolution (e.g., identifying pronouns and their referents), and lexical cohesion (e.g. Identifying related words or concepts across sentences).
- Example: Text: "John loves hiking. He often goes to the mountains. The fresh air and beautiful scenery make him feel alive."

- Coherence and cohesion analysis identifies that "He" in the second sentence refers to "John" in the first sentence through coreference resolution.

2. Discourse structure analysis:

- This involves analyzing the hierarchical structure of a text to determine its organization and how different parts relate to each other. It includes identifying discourse relations (e.g., cause-effect, contrast, temporal) between sentences and determining the overall discourse structure (e.g., introduction, body, conclusion).
- Example: Text: "First, we will discuss the problem. Then, we will propose a solution. Finally, we will evaluate the results."
- Discourse structure analysis identifies the temporal relations between sentences and the overall structure of the text (sequence of events).

3. Rhetorical parsing:

- This involves identifying rhetorical devices and patterns used in a text, such as argumentation, persuasion, or narrative techniques.
- It helps in understanding the author's intent and the overall rhetorical structure of the text.
- Example:
- Text: "On the one hand, reducing taxes can stimulate economic growth. On the other hand, it can lead to a decrease in government revenue. Therefore, a careful balance is necessary."
- Rhetorical parsing identifies the use of contrastive rhetoric (on the one hand, on the other hand) to present different perspectives and the subsequent conclusion.

4.**Discourse-level sentiment analysis:**

- This involves determining the sentiment or attitude expressed at the discourse level, considering the overall context and flow of the text.
- It helps in understanding the overall sentiment of a document or identifying shifts in sentiment.
- Example:
- Text: "The movie started off slow, but it quickly picked up pace. However, the ending was disappointing."
- Discourse-level sentiment analysis takes into account the overall sentiment of the text and identifies a mixed sentiment (positive at the beginning, negative at the end).
- Discourse processing techniques often utilize machine learning algorithms, linguistic rules, and semantic representations to model and analyze the relationships between sentences.
- They help in capturing the global structure and coherence of text, enabling deeper understanding and interpretation.

Cohesion:

- In NLP, cohesion refers to the linguistic devices and techniques used to create a sense of unity and connectedness within a text.
- It involves the explicit and implicit relationships between words, phrases, and sentences that contribute to the overall coherence of the text.
- Cohesion ensures that the different parts of a text are linked together logically and smoothly, allowing readers or language processing systems to understand the flow of information.
- There are 5 types of cohesion that can be found in a text:
  - Reference Cohension
  - Substitution cohesion
  - Ellipsis cohesion
  - Lexical cohesion
  - Conjunction cohesion

1. Reference cohesion: It involves the use of words or expressions to refer back to previously mentioned entities or ideas. This can include pronouns, demonstratives, or definite/indefinite articles.

- Example: "John bought a new car. It is red and very fast."
- In this example, "It" refers back to the previously mentioned car, creating reference cohesion.

2. Substitution cohesion: It occurs when a word or phrase is substituted by another word or phrase to avoid repetition.

- Example: "John likes swimming, and Mary does too."
- In this example, "does too" substitutes the repetition of "likes swimming" in the second part of the sentence.

3. Ellipsis cohesion: It involves the omission of words or phrases that can be inferred from the context.

- Example: "John went to the store, and Mary to the library."
- In this example, the verb "went" is omitted in the second part of the sentence, but it can be inferred from the previous context.

4. Lexical cohesion: It is based on the use of related words or synonyms across sentences or paragraphs.

- Example: "The weather was hot. The sun was shining brightly. People were enjoying the beach."
- In this example, the words "weather," "sun," and "beach" are used to establish lexical cohesion.

5. Conjunction cohesion: It involves the use of conjunctions or connectors to link sentences or ideas together.

- Example: "I bought some groceries. In addition, I need to do laundry."
- In this example, the conjunction "In addition" establishes cohesion between the two sentences.

- Cohesion plays a crucial role in enhancing the clarity and understanding of a text.
- By creating connections between different parts of a text, cohesion helps readers or language processing systems comprehend the relationships and follow the flow of information smoothly.

Reference Resolution

- Reference resolution in natural language processing (NLP) refers to the process of identifying and connecting pronouns, definite/indefinite articles, or other referring expressions to their respective referents in the text.
- It involves determining what entity or concept a pronoun or reference refers to in order to establish a coherent and cohesive understanding of the text.
- Reference resolution is a challenging task because it requires understanding the context and identifying the correct antecedent or referent for a given expression.
- The resolution can be explicit, where the referent is mentioned explicitly in the text, or implicit, where it relies on contextual information and background knowledge.
- **Pronoun reference resolution:** "John saw a dog in the park. It was chasing a ball."
- In this example, the pronoun "It" refers to the previouslymentioned noun "dog." Reference resolution identifies the antecedent and connects the pronoun to its referent.
- **Definite article reference resolution:** "I bought a book. The book is very interesting."
- Here, the definite article "The" refers back to the noun "book" mentioned earlier in the text. Reference resolution connects the definite article to its referent.
- **Indefinite article reference resolution**: "I saw a car accident. An ambulance arrived at the scene quickly."

- In this example, the indefinite article "An" introduces a new entity, an ambulance, which is the referent of the article.
- **Demonstrative reference resolution**: "This is a beautiful painting. That one is even more impressive."
- Here, the demonstratives "This" and "That" establish reference to different paintings. Reference resolution identifies the specific referents based on the spatial or contextual information.
- **Coreference resolution**: "John met Mary. He gave her a gift."
- In this example, coreference resolution connects the pronouns "He" and "her" to their respective antecedents, "John" and "Mary," to establish the reference between them.
- Reference resolution is a critical component in various NLP applications, such as question answering, summarization, machine translation, and information extraction.
- It helps in understanding the relationships between entities and concepts in a text and enables the construction of a coherent and meaningful representation of the information.

Discourse Coherence and Structure

- Discourse coherence and structure in NLP refer to the organization, flow, and logical connections between sentences and larger units of text.
- It focuses on understanding how individual sentences relate to each other and contribute to the overall meaning and structure of the discourse.
- Discourse coherence ensures that the text is coherent, understandable, and connected, while discourse structure deals with the arrangement and organization of different parts of the text
- **Discourse Coherence**: Discourse coherence involves the relationships and connections between sentences, ensuring that they form a cohesive and meaningful text.
- Example: ***"I went to the grocery store. Bought some fruits and vegetables. The cashier was friendly."***
- In this example, coherence is established through the use of explicit semantic connections.

- The second sentence can be understood as an elaboration of the first sentence, describing the action that occurred at the grocery store.
- The third sentence introduces a new piece of information related to the grocery store visit.
- **Discourse Structure**: Discourse structure refers to the overall organization and arrangement of sentences, paragraphs, or sections within a text.
- It captures the logical and hierarchical relationships between different parts of the text.
- Example:
- Text: ***"Introduction: In this paper, we will discuss the importance of renewable energy.***
- Body: First, we will examine the environmental benefits. Then, we will explore the economic advantages. Finally, we will address the challenges.
- **Conclusion:** In conclusion, renewable energy holds great potential for a sustainable future."
- In this example, the text exhibits a clear discourse structure.
- It starts with an introduction, followed by the body that consists of three distinct sections (environmental benefits, economic advantages, challenges), and ends with a conclusion.
- The discourse structure helps readers to navigate and comprehend the content effectively.
- Discourse coherence and structure can also involve explicit linguistic devices that establish connections between sentences.
- Example: ***"John loves hiking. As a result, he spends most of his weekends exploring different trails."***
- In this example, the use of the discourse marker "As a result" establishes a causal relationship between the two sentences.
- It indicates that John's love for hiking leads to him spending most of his weekends exploring trails.
- Discourse coherence and structure are important for understanding and generating coherent and meaningful text.

Language Modeling

- Introduction
- n-Gram Models
- Language Model Evaluation
- Parameter Estimation
- Language Model Adaptation
- Types of Language Models
- Language Specific Modeling Problems
- Multilingual and Cross-lingual Language Modeling

- Statistical Language Model is a model that specifies the a priori probability of a particular word sequence in the language of interest.
- Given an alphabet or inventory of units ∑ and a sequence W= w1w2…..wt ∑* a language model can be used to compute the ϵ probability of W based on parameters previously estimated from a training set
- The inventory ∑ is the list of unique words encountered in the training data.
- Selecting the units over which a language model should be defined is a difficult problem particularly in languages other than English.

Introduction

- A language model is combined with other model or models that hypothesize possible word sequences.
- In speech recognition a speech recognizer combines acoustic model scores with language model scores to decode spoken word sequences from an acoustic signal.
- Language models have also become a standard tool in information retrieval, authorship identification, and document classification

n-Gram Models

N-gram models are a type of probabilistic language model used in NLP to predict the next item in a sequence, typically words in a sentence.

- An N-gram is a contiguous sequence of n items from a given sample of text or speech.

- The "n" in N-gram represents the number of items in the sequence:
  - Unigram (1-gram): A single word.
  - Bigram (2-gram): A sequence of two words.
  - Trigram (3-gram): A sequence of three words.
  - 4-gram, 5-gram, etc.: Longer sequences of words.
- n-gram models work by calculating the probability of a word based on the occurrence of the previous (n-1) words.
- The probability of a word sequence can be estimated using the chain rule of probability, which in the context of an N-gram model is simplified by making the Markov assumption that the probability of a word depends only on the previous (n-1) words.
-
  -