21CSE356T

NATURAL LANGUAGE PROCESSING

UNIT-3

Word2Vec

CBOW

Skip-gram and GloVe

Discourse Segmentation

Text Coherence

Discourse Structure

Reference Resolution

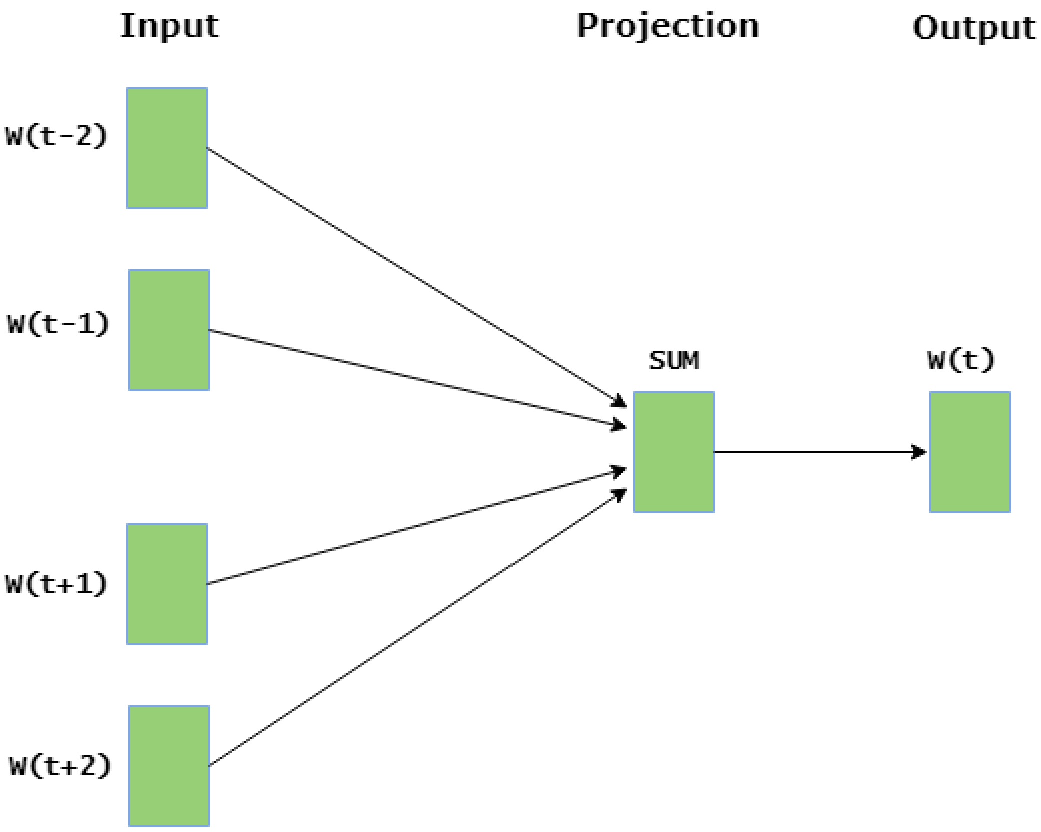
Pronominal Anaphora Resolution

Coreference Resolution

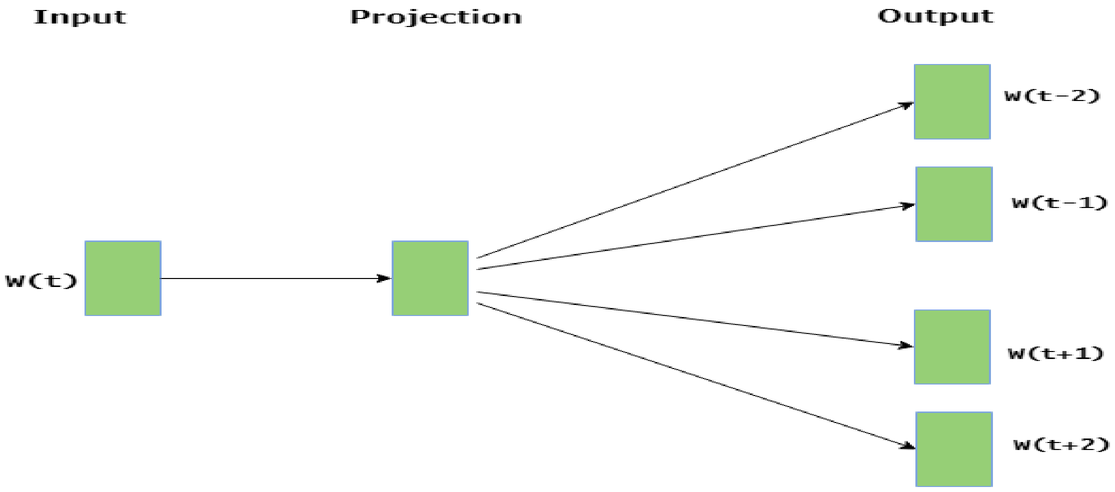
Word2Vec

Word2Vec is a widely used method in natural language processing (NLP) that allows words to be represented as vectors in a continuous vector space. Word2Vec is an effort to map words to high-dimensional vectors to capture the semantic relationships between words, developed by researchers at Google. Words with similar meanings should have similar vector representations, according to the main principle of Word2Vec. Word2Vec utilizes two architectures

- **CBOW** Continuous Bag of Words): The CBOW model predicts the current word given context words within a specific window. The input layer contains the context words and the output layer contains the current word. The hidden layer contains the dimensions we want to represent the current word present at the output layer.



Skip Gram : Skip gram predicts the surrounding context words within specific window given current word. The input layer contains the current word and the output layer contains the context words. The hidden layer contains the number of dimensions in which we want to represent current word present at the input layer.

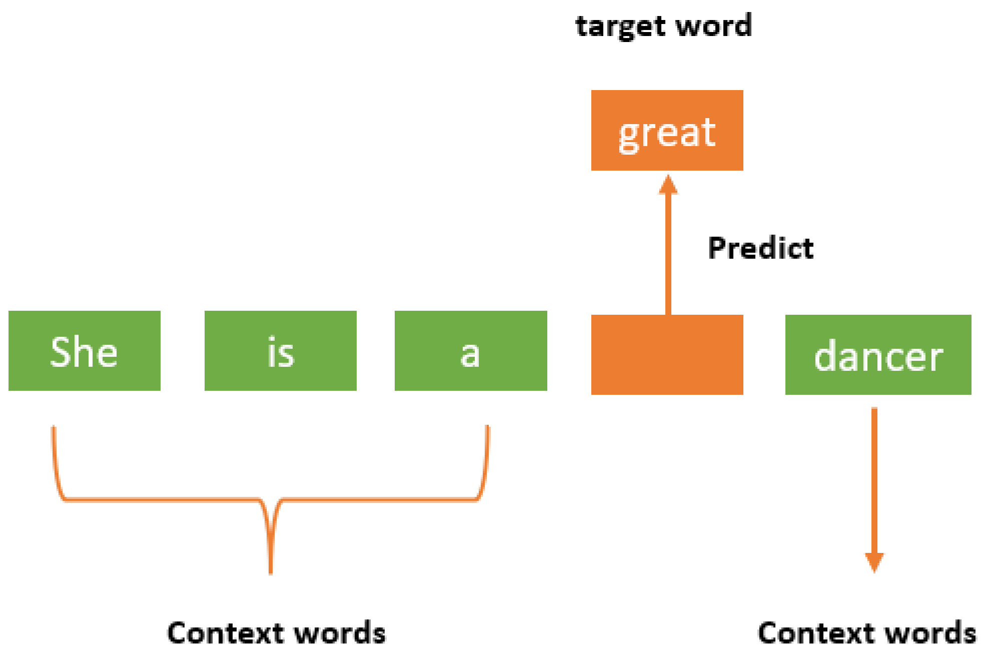


Need of Word2Vec

In natural language processing (NLP), Word2Vec is a popular and significant method for representing words as vectors in a continuous vector space. Word2Vec has become popular and is utilized in many different NLP applications for several reasons:

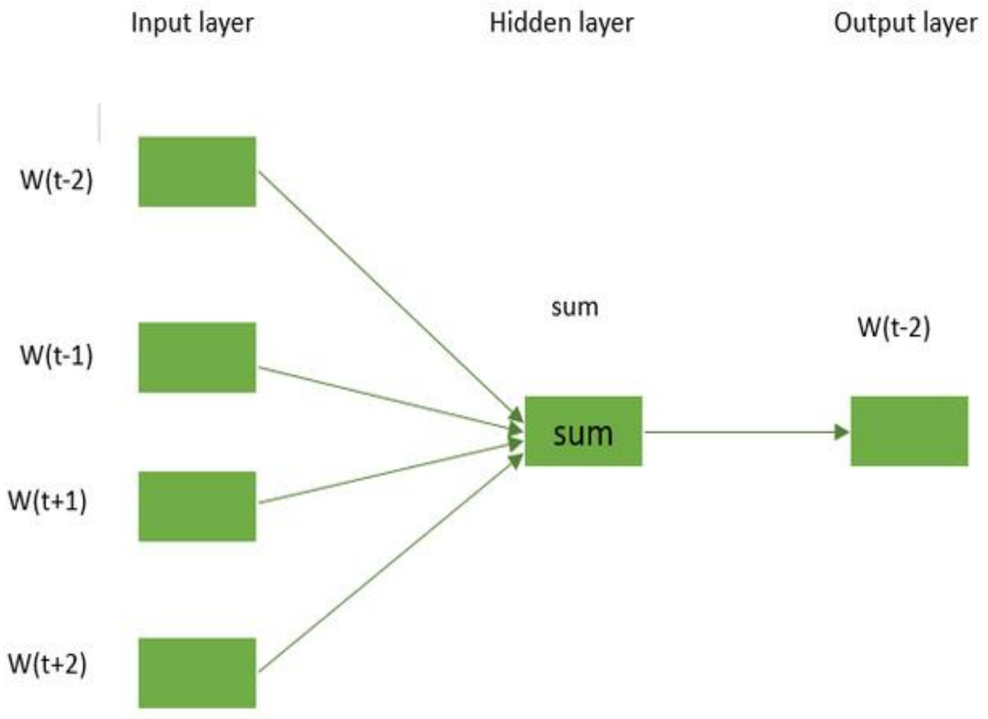
- **Semantic Representations:** Word2Vec records the connections between words semantically. Words are represented in the vector space so that similar words are near to one another. This enables the model to interpret words according to their context within a particular corpus.
- **Distributional Semantics:** The foundation of Word2Vec is the distributional hypothesis, which holds that words with similar meanings are more likely to occur in similar contexts. Word2Vec generates vector representations that reflect semantic similarities by learning from the distributional patterns of words in a large corpus.
- **Vector Arithmetic:** Word2Vec generates vector representations that have intriguing algebraic characteristics. Vector arithmetic, for instance, can be used to record word relationships. One well-known example is that the vector representation of “queen” could resemble the vector representation of “king” less “man” plus “woman.”
- **Efficiency:** Word2Vec’s high computational efficiency makes training on big datasets possible. Learning high-dimensional vector representations for a large vocabulary requires this efficiency.
- **Transfer Learning:** A variety of natural language processing tasks can be initiated with pre-trained Word2Vec models. Time and resources can be saved by fine-tuning the embeddings discovered on a sizable dataset for particular uses.
- **Applications:** Word2Vec embeddings have shown promise in a number of natural language processing (NLP) applications, such as machine translation, text classification, sentiment analysis, and information retrieval. These applications are successful in part because of their capacity to capture semantic relationships.
- **Continuous Bag of Words (CBOW)** is a popular natural language processing technique used to generate word embeddings. Word

embeddings are important for many NLP tasks because they capture semantic and syntactic relationships between words in a language. CBOW is a neural network-based algorithm that predicts a target word given its surrounding context words. It is a type of "unsupervised" learning, meaning that it can learn from unlabeled data, and it is often used to pre-train word embeddings that can be used for various NLP tasks such as sentiment analysis, text classification, and machine translation.



CBOW model

The CBOW model uses the target word around the context word in order to predict it. Consider the above example "She is a great dancer." The CBOW model converts this phrase into pairs of context words and target words. The word pairings would appear like this ([she, a], is), ([is, great], a) ([a, dancer], great) having window size=2.



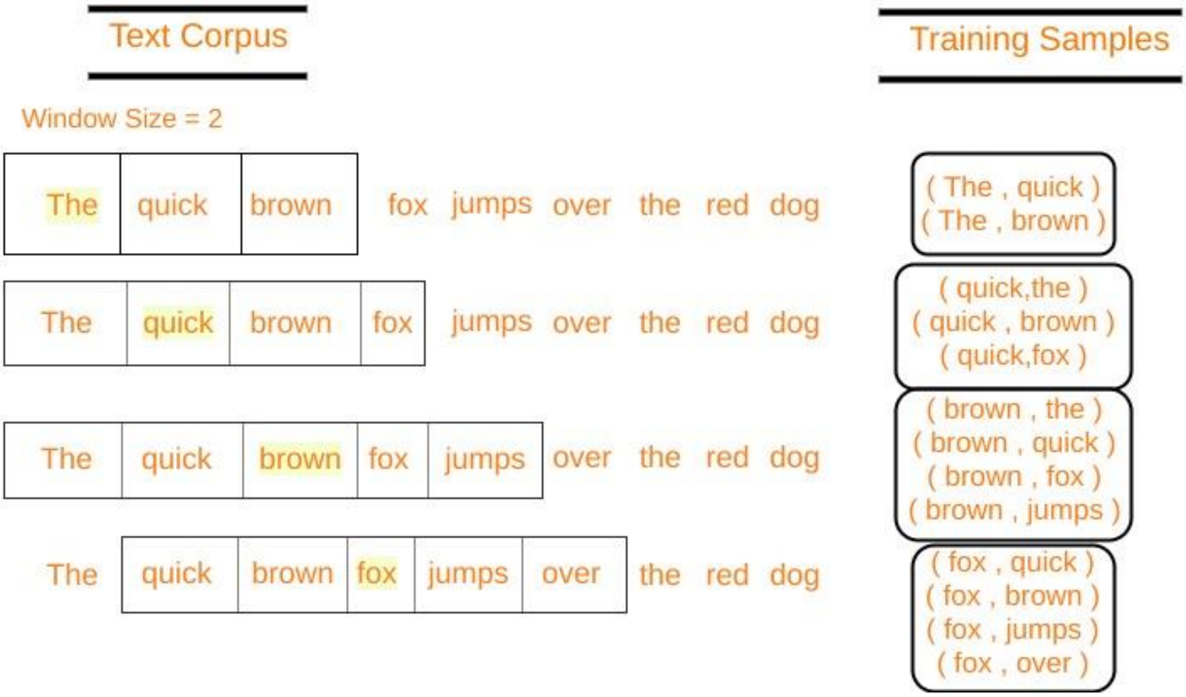
The model considers the context words and tries to predict the target term. The four 1∗W input vectors will be passed to the input layer if have four words as context words are used to predict one target word. The hidden layer will receive the input vectors and then multiply them by a W∗N matrix. The 1∗N output from the hidden layer finally enters the sum layer, where the vectors are element-wise summed before a final activation is carried out and the output is obtained from the output layer.

Difference between Bag-of-Words (BoW) model and the Continuous Bag-of-Words (CBOW)

Bag-of-Words (BoW) model	**Bag-of-Words (CBOW)**
It represents text as a collection of words and their frequency in a given document or corpus	It does not consider the order or context in which the words appear, and therefore, it may not capture the full meaning of the text
The BoW model is simple and easy to implement, but it has limitations in capturing the meaning of language	the CBOW model is a neural network-based approach that captures the context of words. It learns to predict the target word

Skip-Gram

skip-gram architecture of word2vec, the input is the center word and the predictions are the context words. Consider an array of words W, if W(i) is the input (center word), then W(i-2), W(i-1), W(i+1), and W(i+2) are the context words if the sliding window size is 2.



	Skip-gram	CBOW (Continuous Bag of Words)
Context and Target	the model **predicts the context words** given a target word	the model **predicts the target word** given the context words.
Window Size	Skip-gram typically **uses a larger window size,** which determines the maximum distance between the target word and its context words. This **allows Skip-gram to capture broader semantic relationships between words.**	CBOW **uses a smaller window size,** focusing on the immediate context words, and thus tends to be more **efficient in terms of computational resources.**
Training Efficiency	Skip-gram, with its larger window size and individual context word predictions, requires **more training iterations** and can be **computationally expensive.**	**CBOW is generally faster to train compared to Skip-gram** since it aggregates the context word vectors to predict the target word
Rare Words	**Skip-gram performs better with rare words** since it generates more training examples for infrequent words.	**CBOW, on the other hand, may struggle with rare words** as they might not appear frequently in the context window.
Semantic Relationships	Skip-gram is **known to capture more fine-grained semantic relationships between words** due to its focus on predicting the context words given a target word**. It tends to produce better word embeddings** for tasks like word analogy and word similarity	**CBOW,** while generally faster and computationally efficient, **might lose some of the finer semantic details captured by Skip-gram.**

Discourse segmentation in Natural Language Processing (NLP) refers to the task of dividing a text or discourse (e.g., a speech or a written document) into coherent units, such as sentences, paragraphs, or segments that are meaningful and logically connected. The goal is to understand the structure of the discourse and how different parts of the text relate to each other.

Discourse segmentation helps in a variety of NLP tasks, including summarization, machine translation, information retrieval, and question answering, by enabling systems to better understand the organization of text at a higher level than just sentence-by-sentence or word-by-word.

key components

Basic Units of Discourse Segmentation

Sentences: Dividing the text into individual sentences.

Topics: Identifying segments of text that cover a specific topic.

Discourse Units: Groupings of related sentences or phrases that form a coherent unit of thought.

Levels of Discourse:

Microstructure: Focuses on sentence-level relations and how they connect.

Macrostructure: Concerns the overall structure of a document, such as the hierarchical organization of topics and sub-topics.

3. Challenges in Discourse Segmentation:

Ambiguity: The same sentence can have different meanings in different contexts, and identifying the right segmentation depends on understanding these ambiguities.

Coherence: Proper segmentation relies on identifying connections between ideas and events in the discourse, which is often subtle.

Contextual Understanding: Some connections, like causal relationships or contrasts, depend on understanding the broader context.

Methods for Discourse Segmentation:

Rule-Based Approaches: These rely on predefined linguistic rules (such as punctuation marks, connectives, and discourse markers) to segment text.

Machine Learning-Based Approaches: Supervised models use labeled data to learn how to segment discourse effectively. Features might include sentence length, punctuation patterns, and lexical cohesion.

Deep Learning-Based Approaches: Modern NLP models (like Transformer-based models) are used to perform discourse segmentation by learning complex patterns and relationships from large corpora.

Applications:

Text Summarization: Breaking down a document into its core segments helps summarize it effectively.

Dialogue Systems: In conversational AI, discourse segmentation can help understand turns and topic shifts in dialogue.

Information Retrieval: It improves the ability to retrieve contextually relevant documents by segmenting and understanding their structure.

Machine Translation: Better discourse segmentation leads to more accurate translations by maintaining the coherence of the original text.

Text Coherence in Natural Language Processing (NLP) refers to the logical flow and consistency of ideas within a text. It is what makes a piece of writing or speech seem cohesive and easy to follow for a reader or listener. Coherent text ensures that sentences and ideas connect in a way that feels natural and that the text as a whole conveys a unified message.

Coherence goes beyond grammatical correctness (syntax) and involves the semantic and logical relationships between sentences and paragraphs. It is one of the key elements that separate random collections of words or sentences from meaningful, well-structured discourse.

Keys of Text Coherence

I. **Cohesion**
II. **Coherence:**

Cohesion refers to the grammatical and lexical connections that bind sentences together (like pronouns, conjunctions, or lexical repetition).

Coherence is more abstract. It involves the overall meaning, flow of ideas, and how well the sentences relate to each other at the semantic level.

Types of Coherence:

Local Coherence: Ensures that individual sentences or discourse units are logically connected to each other in the immediate context. This involves things like maintaining topic consistency or logical flow within a paragraph.

Global Coherence: Deals with the larger structure of the text, including how different parts of the text work together to express a unified theme or message across the entire discourse.

Lexical Cohesion

Repetition: Repeating key terms or phrases to maintain topic consistency (e.g., repeating "climate change" throughout an article).

Synonymy: Using synonyms or related terms to avoid redundancy while maintaining the same topic (e.g., using "environment" instead of "nature").

Collocation: Using words that commonly occur together (e.g., "global warming" and "greenhouse gases").

Approaches to Text

Rule-Based Approaches:

These approaches rely on predefined rules to ensure coherence. For example, rules might be applied to check that topics are consistent or that discourse markers are used properly.

Statistical Models:

Statistical models use data-driven approaches, often drawing from large corpora to determine common patterns of coherence in texts. For instance, coherence can be measured based on how often certain discourse markers or words appear together.

Machine Learning Approaches:

Supervised learning can be used to train models to recognize and score the coherence of a text. Features might include sentence-level coherence, word similarity, topic modeling, and discourse structure.

Deep Learning Approaches:

Modern techniques, particularly with Transformers (e.g., BERT, GPT), can capture long-range dependencies and understand contextual relationships across larger portions of text. These models can be used for text generation, summarization, and coherence checking by learning from large datasets.

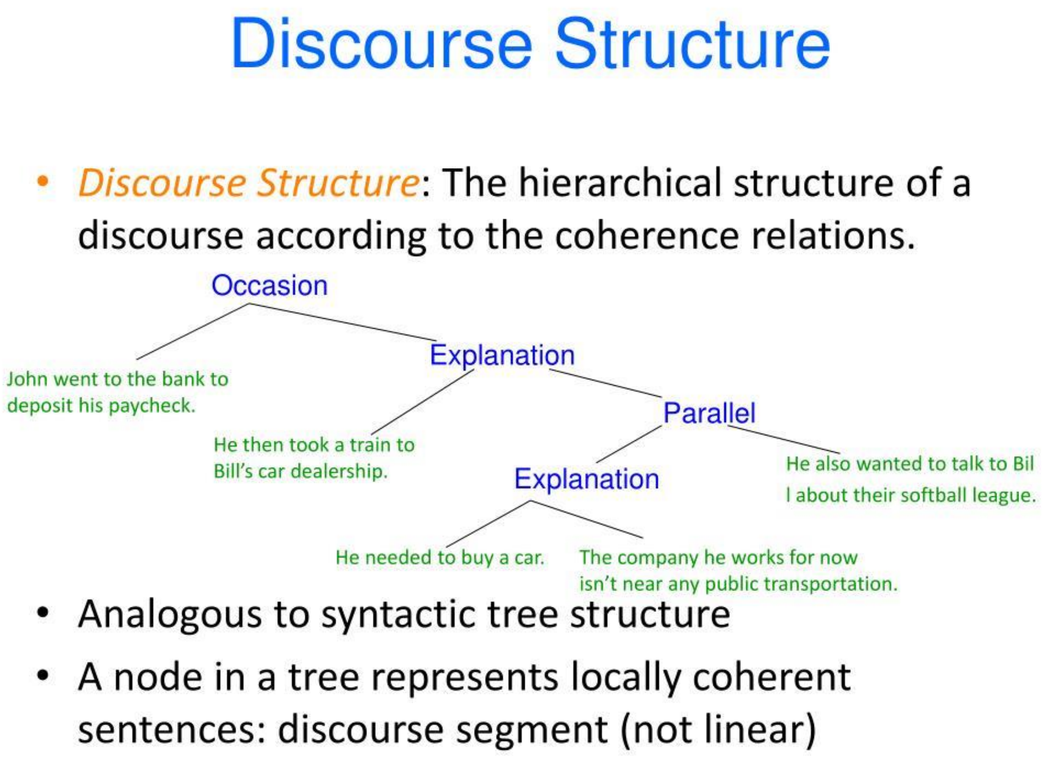
Applications

Text Summarization:

A coherent summary must capture the main points of a document while maintaining logical flow and connection between ideas. Coherence is essential to ensure the summary is readable and meaningful.

Machine Translation:

Ensuring coherence is crucial in machine translation to produce translations that make sense within the context of the entire sentence or passage, rather than being a mere word-for-word translation.



Discourse Structure

- *Discourse Structure*: The hierarchical structure of a discourse according to the coherence relations.

Occasion

John went to the bank to deposit his paycheck.

Explanation

He then took a train to Bill's car dealership.

Parallel

Explanation

He also wanted to talk to Bil l about their softball league.

He needed to buy a car.

The company he works for now isn't near any public transportation.

- Analogous to syntactic tree structure
- A node in a tree represents locally coherent sentences: discourse segment (not linear)

Discourse Structures

1. **Segmentation and Boundary Detection**:

    Discourse segmentation involves breaking a text into smaller, coherent segments (discourse units). It is closely related to identifying boundaries between different discourse units, such as sentences, paragraphs, or thematic segments.

2. **Discourse Tree Structures**:

    A **discourse tree** represents the hierarchical structure of a discourse. It is usually a tree-like structure where each node represents a discourse unit, and the edges between nodes represent the discourse relations.

    One well-known model for discourse structure is the **Rhetorical Structure Theory** (RST), which posits that text can be broken down into units connected by rhetorical relations (such as **elaboration**, **contrast**, and **cause-effect**).

3. **Rhetorical Structure Theory (RST)**:

    RST is a framework used to analyze the discourse structure of a text. It involves representing the text as a tree, where nodes correspond to text segments, and edges represent rhetorical relations (e.g., **contrast**, **elaboration**).

    In RST, a text can be analyzed for how smaller units (e.g., sentences) are connected into larger units, reflecting the structure of the argument or narrative.

4. **Discourse Representation Theory (DRT)**:

    DRT focuses on the interpretation of discourse in a logical form, aiming to explain how reference and meaning flow through a text.

    It models the process of discourse comprehension by maintaining a representation of the discourse's context and updating it as new information (from the text) is introduced.

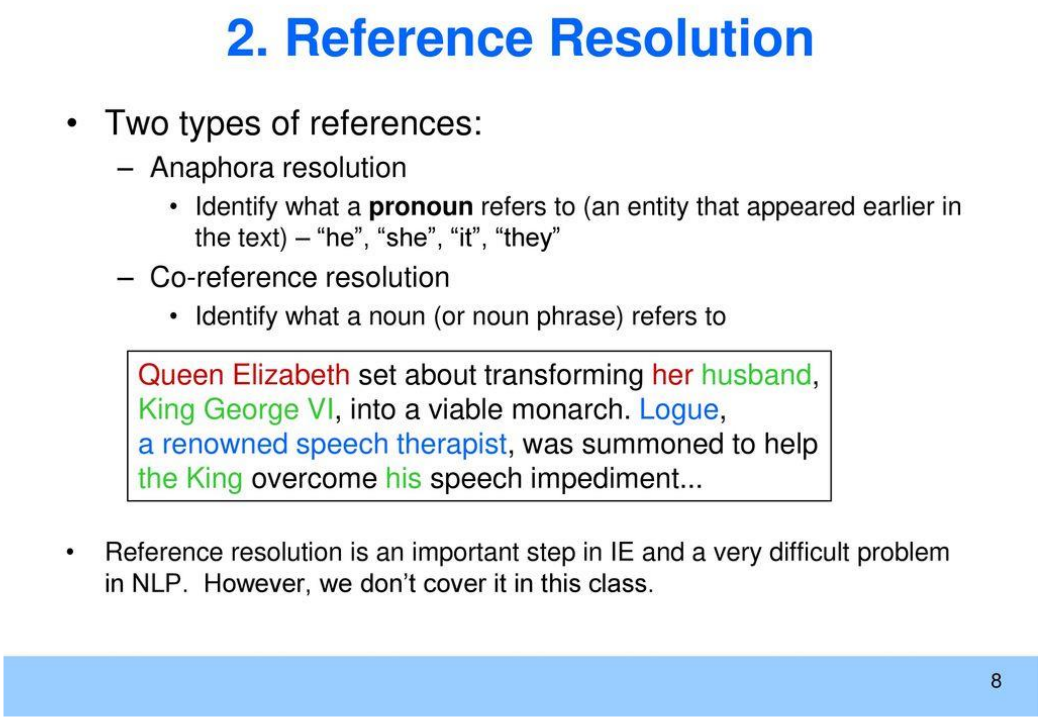
Types of Discourse Structures:

1. **Narrative Structure**:
    - A narrative structure typically involves events or actions that occur in a sequence, often with causal relationships between them. For example, in a story, one event leads to another (cause-effect), and characters or objects may be introduced and developed over time.
2. **Argumentative Structure**:
    - Argumentative texts often involve claims supported by reasons or evidence. The discourse structure in these texts reflects the logical progression of an argument, where evidence supports claims, and counterarguments may be addressed.
3. **Expository Structure**:

    Expository texts, such as instructional or informational texts, often follow a clear structure of introducing a topic, explaining it in detail, and then concluding or summarizing the information.

4. **Conversational Structure**:

discourse structure involves turns of speech between participants. The structure also depends on discourse markers like greetings, interruptions, questions, and responses.



2. Reference Resolution

- Two types of references:
  - Anaphora resolution
    - Identify what a **pronoun** refers to (an entity that appeared earlier in the text) – “he”, “she”, “it”, “they”
  - Co-reference resolution
    - Identify what a noun (or noun phrase) refers to

Queen Elizabeth set about transforming her husband, King George VI, into a viable monarch. Logue, a renowned speech therapist, was summoned to help the King overcome his speech impediment...

- Reference resolution is an important step in IE and a very difficult problem in NLP. However, we don’t cover it in this class.

8

Reference Resolution in Natural Language Processing (NLP) refers to the task of determining what an expression (typically a pronoun or a noun phrase) refers to in a text. This is essential for understanding the relationships between different parts of a sentence or across sentences. The goal is to link expressions (like "he," "she," "it," or "the president") to their proper referents (such as a specific person, object, or concept) within the discourse.

Reference resolution involves two main components:

1. **Anaphora Resolution**: Resolving references to previously mentioned entities within the discourse, such as resolving pronouns (e.g., "he," "she," "it").
2. **Cataphora Resolution**: Resolving references that point forward to an entity mentioned later in the discourse (less common than anaphora).

Reference resolution is a key part of understanding coherence and context in a text and is necessary for tasks like machine translation, text summarization, and question answering.

Key Types

1. **Pronouns**

   Pronouns are common referring expressions that substitute for nouns. Examples include "he," "she," "it," "they," "this," "that," and "who."

   For example

   "John went to the store. He bought some milk."

   "He" refers to "John."

2. **Definite Noun Phrases**

   Definite noun phrases (e.g., "the book," "the president," "the car") often refer to specific entities that are assumed to be known to the listener or reader.

   For example:

   "I saw a dog outside. The dog was barking."

   "The dog" refers back to "a dog."

3. **Ellipsis**

   Ellipsis is the omission of a noun or phrase that is implied from the context.

   For example:

   "I need to buy a car, but I don’t have enough money for [a car]."

   The implied noun "a car" is resolved based on the context.

4. **Demonstratives**

Demonstratives (e.g., "this," "that," "these," "those") are used to refer to specific entities, and their reference often depends on context.

For example:

- "I like this book," where "this" refers to a specific book in the context.

5. **Named Entities**

Named entities like people, organizations, locations, etc., need to be tracked across a document or conversation.

For example:

"Barack Obama was elected president. He served two terms."

"He" refers to "Barack Obama."

Challenges

1. **Ambiguity**:

Pronouns and other referring expressions can sometimes have multiple possible antecedents, leading to ambiguity. For example, in the sentence:

- "John told David that he should study more."
- It’s unclear whether "he" refers to John or David.

2. **Long-Distance Dependencies**:

The referent of an expression can be far away in the text, making it challenging to resolve. For example, in a long passage, a pronoun might refer to an entity mentioned much earlier in the text.

3. **Multiple Antecedents**:

Sometimes, a referring expression could be associated with more than one potential antecedent. For example:

"Alice and Bob were at the party. She left early."

It’s unclear whether "she" refers to Alice or Bob.

4. **Complex Sentence Structures**:

   In complex sentences with embedded clauses or nested structures, reference resolution can become difficult due to the syntactic complexity of determining which noun a pronoun refers to.

5. **Context and World Knowledge**:

   Reference resolution often requires external knowledge beyond the text, such as understanding the relationships between entities or resolving temporal and spatial references.

   Pronominal Anaphora Resolution in Natural Language Processing (NLP) is a specific subtask of **anaphora resolution**, where the goal is to identify the antecedent (i.e., the noun or noun phrase) to which a **pronoun** (the anaphor) refers. Pronominal anaphora resolution focuses on resolving references made by pronouns, which are linguistic elements used to replace nouns or noun phrases, like "he," "she," "it," "they," "this," "that," and others.

   This process is crucial for understanding and generating coherent text, as pronouns often appear throughout a text, and without identifying their correct antecedents, machines would struggle to understand or produce meaningful sentences.

   Key Components of Pronominal Anaphora Resolution

1. **Anaphor**: The pronoun or other referring expression that requires resolution (e.g., "he," "she," "it," "they").
2. **Antecedent**: The noun or noun phrase to which the pronoun refers (e.g., "John," "Mary," "the car," "the dog").
3. **Coreference**: The broader task that involves linking all expressions in a text that refer to the same entity. Pronominal anaphora resolution is a subset of coreference resolution, as it focuses on resolving references made by pronouns.

Challenges in Pronominal Anaphora Resolution

1. **Ambiguity**:
   - A single pronoun might refer to multiple potential antecedents. For example:
     - "John and Bill went to the park. He played soccer."
     - "He" could refer to either John or Bill.
2. **Long-Distance Dependencies**:

   The antecedent and the anaphor can be far apart in the text, making it difficult to identify the correct reference.

     - "The dog was barking loudly. John, who was passing by, noticed it."
     - "It" refers to "the dog," even though "John" appears in between.
3. **Gender and Number Agreement**:

   Pronouns often agree in gender and number with their antecedents, but this might not always be clear. For instance:

     - "The teacher gave a lecture. She was very clear."
     - "She" refers to "The teacher," and they share the same gender (feminine singular).
4. **Syntactic and Semantic Factors**:

   Syntactic proximity (how close the antecedent is) and semantic compatibility (the meaning of the antecedent and the pronoun) are crucial. For example, a pronoun typically refers to the most recent noun phrase in a sentence unless other semantic cues suggest otherwise.
5. **Contextual Understanding**:

Resolving pronouns often requires understanding the broader context, including world knowledge or common sense. For example, in the sentence:

"He took a break after working hard."

"He" is likely referring to a male individual, but the specific identity can depend on context, such as previous knowledge of who is being discussed.

Coreference Resolution in Natural Language Processing (NLP) is the task of determining which words or phrases in a text refer to the same entity. This process involves identifying all mentions of the same real-world entity throughout a document or discourse. For example, in a sentence like:

- "John went to the store. He bought some milk."

"John" and "He" are coreferential, meaning they refer to the same entity, John. Coreference resolution helps machines understand how different parts of the text are connected and ensures that relationships between entities are properly captured.

Coreference resolution is fundamental for understanding context, maintaining consistency, and improving the coherence of texts in various NLP applications, such as machine translation, information extraction, question answering, and summarization.

Key Concepts in Coreference Resolution:

1. **Mention**:

A mention refers to a word or phrase in the text that represents an entity. Mentions can be noun phrases, pronouns, or even named entities.

For example, in the sentence "Barack Obama was elected president. He served two terms," "Barack Obama" and "He" are both mentions that refer to the same entity.

2. **Coreferential Mentions**:

Coreferential mentions are words or phrases that refer to the same entity.

In the sentence: "Alice bought a car. It is red," "Alice" and "It" are coreferential.

3. **Anaphora and Cataphora**:

   Anaphora: A form of reference where a pronoun or noun phrase refers to a previously mentioned entity. Example: "John went to the park. He played soccer." "He" refers to "John."

   Cataphora: A reference where the pronoun or noun phrase appears before its antecedent. Example: "She went to the store. Mary was happy." "She" refers to "Mary."

4. **Coreference Chain**:

   A **coreference chain** is a group of mentions that all refer to the same entity. For example, in the text:

   "John went to the store. He bought some milk."

   "John" and "He" form a coreference chain.

Challenges in Coreference Resolution:

1. **Ambiguity**:

   A single mention can have multiple possible antecedents (referents). For example:

   "The teacher told the student to study. She was helpful."

   "She" could refer to either the teacher or the student.

2. **Pronoun Resolution**:

   Determining which noun a pronoun (e.g., "he," "she," "it," "they") refers to is often challenging, especially when there are multiple entities in the discourse.

3. **Long-Distance Coreference**:

Coreferential mentions may be far apart in a document, making it difficult to identify links between them. For example:

"John went to the park. After a while, he met his friend."

"John" and "he" refer to the same person, but they are separated by other information.

4. **Definite Noun Phrases**:

Noun phrases like "the president," "the car," and "the book" can refer to specific entities, but resolving them requires understanding the context in which they appear. For example:

"The president of the company met with the CEO. He was very experienced."

"He" likely refers to "the president," not the CEO, based on prior context.

5. **Semantic Understanding**:

Coreference resolution often requires deep semantic understanding, as a pronoun might refer to something semantically related, even if it's not the closest noun. For example:

"The computer broke down. It was very frustrating."

"It" refers to "the computer," but it's based on the semantic relationship rather than syntactic proximity.

6. **Complex Sentence Structures**:

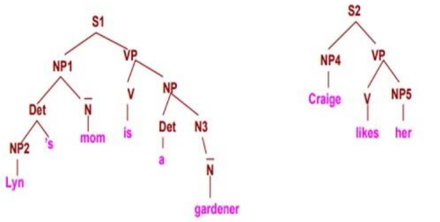
In complex sentences, resolving coreference can be difficult due to syntactic complexity. For example:

"John told Mary that he had bought a new car."

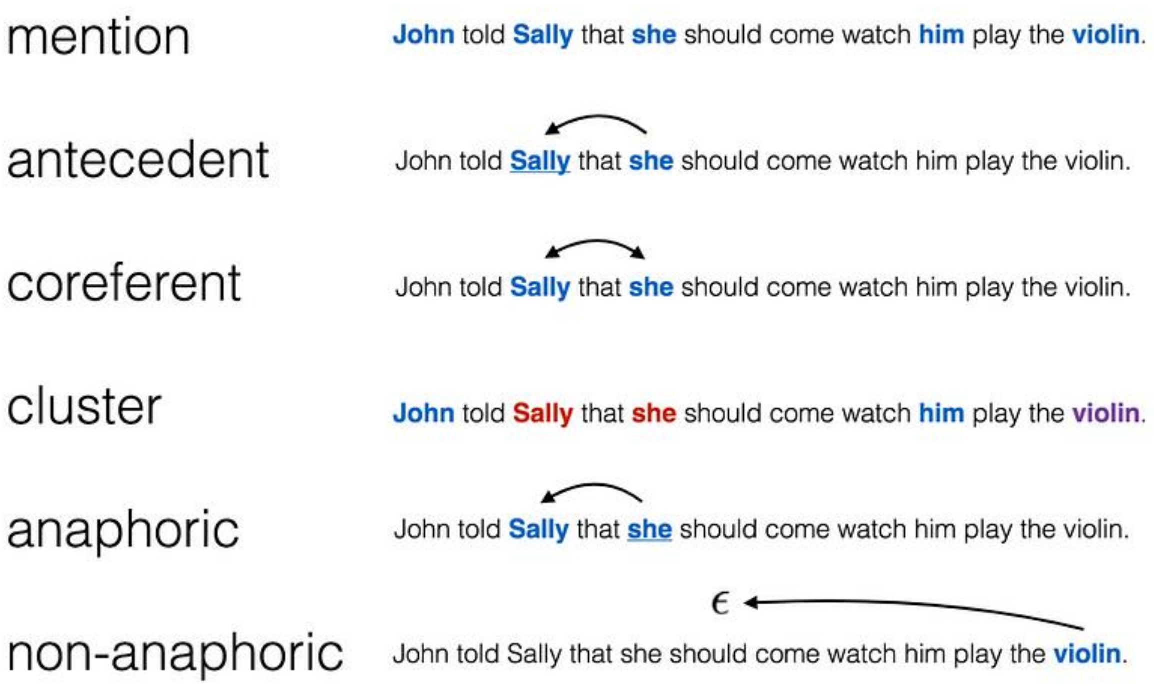
It may not be immediately clear if "he" refers to John or Mary, depending on sentence structure and discourse context.

Hobb’s Algorithm

- Simple syntax-based algorithm that on syntactic parser .
- Searches syntactic trees of current and preceding sentences in breadth-first, left-to-right manner.
- Stops when it finds matching NP.



- Start search at NP5 in S2.
- Reject NP4 as no NP node between it and X (S2).
- What would have happened if the subject was *Craige's mom*?
- Move to S1. NP1 is first NP we encounter, so finish.
- Result: *Lyn's mom*



mention John told Sally that she should come watch him play the violin.

antecedent John told Sally that she should come watch him play the violin.

coreferent John told Sally that she should come watch him play the violin.

cluster John told Sally that she should come watch him play the violin.

anaphoric John told Sally that she should come watch him play the violin.

non-anaphoric ϵ John told Sally that she should come watch him play the violin.