Unit-III

Semantic Analysis and Pragmatics: Introduction, Meaning Representation(85), Lexical Semantics(B0), Ambiguity, Word Sense Disambiguation(71), Discourse Processing: Introduction, Cohesion(78), Reference Resolution(A2), Discourse Coherence and Structure(A4).

INTRODUCTION:

Semantic Analysis is a subfield of Natural Language Processing (NLP) that attempts to understand the meaning of Natural Language. Understanding Natural Language might seem a straightforward process to us as humans. However, due to the vast complexity and subjectivity involved in human language, interpreting it is quite a complicated task for machines. Semantic Analysis of Natural Language captures the meaning of the given text while taking into account context, logical structuring of sentences and grammar roles.

Parts of Semantic Analysis

Semantic Analysis of Natural Language can be classified into two broad parts:

1. Lexical Semantic Analysis: Lexical Semantic Analysis involves understanding the meaning of each word of the text individually. It basically refers to fetching the dictionary meaning that a word in the text is deputed to carry.

2. Compositional Semantics Analysis: Although knowing the meaning of each word of the text is essential, it is not sufficient to completely understand the meaning of the text.

Semantic Analysis is a subfield of Natural Language Processing (NLP) that attempts to understand the meaning of Natural Language. Understanding Natural Language might seem a straightforward process to us as humans. However, due to the vast complexity and subjectivity involved in human language, interpreting it is quite a complicated task for machines. Semantic Analysis of Natural Language captures the meaning of the given text while taking into account context, logical structuring of sentences and grammar roles.

Parts of Semantic Analysis
Semantic Analysis of Natural Language can be classified into two broad parts:

1. Lexical Semantic Analysis: Lexical Semantic Analysis involves understanding the meaning of each word of the text individually. It basically refers to fetching the dictionary meaning that a word in the text is deputed to carry.

2. Compositional Semantics Analysis: Although knowing the meaning of each word of the text is essential, it is not sufficient to completely understand the meaning of the text.

For example, consider the following two sentences:

- **Sentence 1:** Students love Nlp.
- **Sentence 2:** Nlp loves Students.

Although both these sentences 1 and 2 use the same set of root words {student, love, Nlp}, they convey entirely different meanings.

Hence, under Compositional Semantics Analysis, we try to understand combinations of individual words from the meaning of the text.

Tasks involved in Semantic Analysis

In order to understand the meaning of a sentence, the following are the major processes involved in Semantic Analysis:

1. Word Sense Disambiguation
2. Relationship Extraction

Word Sense Disambiguation:

In Natural Language, the meaning of a word may vary as per its usage in sentences and the context of the text. Word Sense Disambiguation involves interpreting the meaning of a word based upon the context of its occurrence in a text.

For example, the word ‘Bark’ may mean ‘the sound made by a dog’ or ‘the outermost layer of a tree.’

Likewise, the word ‘rock’ may mean ‘*a stone*‘ or ‘*a genre of music*‘ – hence, the accurate meaning of the word is highly dependent upon its context and usage in the text.

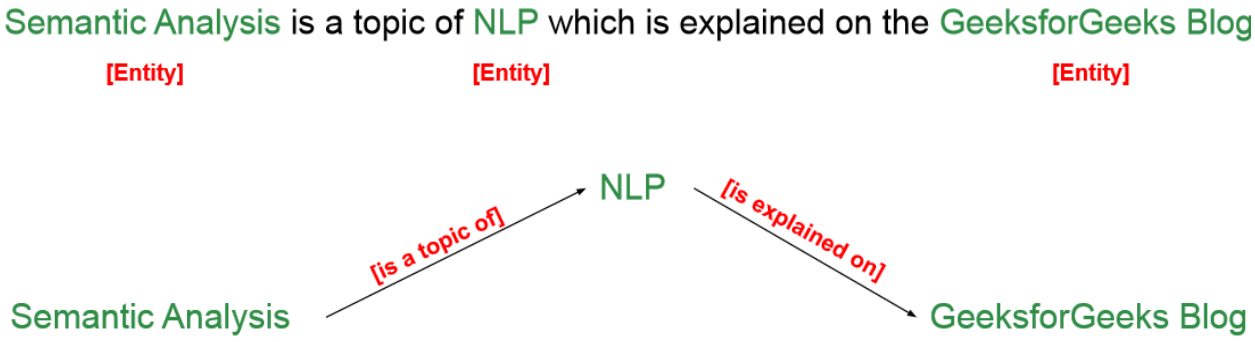
Thus, the ability of a machine to overcome the ambiguity involved in identifying the meaning of a word based on its usage and context is called Word Sense Disambiguation.

Relationship Extraction:

Another important task involved in Semantic Analysis is Relationship Extracting. It involves firstly identifying various entities present in the sentence and then extracting the relationships between those entities.

For example, consider the following sentence:

Semantic Analysis is a topic of NLP which is explained on the GeeksforGeeks blog. The entities involved in this text, along with their relationships, are shown below.



Lexical Semantics

Certainly! Lexical semantics is a subfield of linguistics that focuses on the study of meaning in words and the relationships between words. It plays a crucial role in understanding how words contribute to the overall meaning of sentences and texts. Here are some key aspects and concepts related to lexical semantics:

9. **Word Meaning:**

Denotation and Connotation: Lexical semantics examines both the denotative (literal) and connotative (emotional or cultural) meanings of words.

Sense and Reference: Words often have multiple senses (different meanings) and references (the actual entities or concepts they represent).

9. **Word Relations**:

Synonymy and Antonymy: Lexical semantics explores relationships between words, such as synonymy (similar meanings) and antonymy (opposite meanings).

Hyponymy and Hypernymy: Words can also be related in terms of hyponymy (specific to general) and hypernymy (general to specific) relationships.

9. **Semantic Fields:**

Lexical Fields: Words can be grouped into semantic fields based on shared characteristics or themes. Understanding these fields helps identify semantic relationships between words.

9. Polysemy and Homonymy:

Polysemy: A single word having multiple related meanings. Lexical semantics delves into how these meanings are connected.

Homonymy: Different words that happen to have the same form. Lexical semantics distinguishes between homonyms with unrelated meanings and those with a historical connection.

5. Word Sense Disambiguation:

Contextual Analysis: Lexical semantics aids in disambiguating word meanings based on the context in which they appear. This is crucial in natural language processing tasks.

6. Lexical Relations in Sentences:

Thematic Roles: Lexical semantics helps identify the roles that words play in a sentence, such as the agent, patient, or theme.

Selectional Restrictions: Certain words have restrictions on the types of arguments they can take, and lexical semantics helps in understanding these constraints.

7. Lexical Semantics in Computational Linguistics:

Word Embeddings: Techniques like word embeddings use distributional semantics to represent words based on their contextual usage.

Ontologies and Semantic Networks: Lexical semantics is employed in building ontologies and semantic networks, facilitating knowledge representation in artificial intelligence.

8. Philosophical Implications:

Philosophical Semantics: Lexical semantics contributes to philosophical discussions on language, meaning, and reality, such as debates on reference and truth conditions.

9. Recent Developments:

Cognitive Semantics: Integrating insights from cognitive science, lexical semantics explores how mental representations and cognitive processes influence word meaning.

Ambiguity and Uncertainty in Language :

Ambiguity, generally used in natural language processing, can be referred as the ability of being understood in more than one way. In simple terms, we can say that ambiguity is the capability of being understood in more than one way. Natural language is very ambiguous. NLP has the following types of ambiguities –

Lexical Ambiguity

The ambiguity of a single word is called lexical ambiguity. For example, treating the word silver as a noun, an adjective, or a verb.

Syntactic Ambiguity

This kind of ambiguity occurs when a sentence is parsed in different ways. For example, the sentence “The man saw the girl with the telescope”. It is ambiguous whether the man saw the girl carrying a telescope or he saw her through his telescope.

Semantic Ambiguity

This kind of ambiguity occurs when the meaning of the words themselves can be misinterpreted. In other words, semantic ambiguity happens when a sentence contains an ambiguous word or phrase. For example, the sentence “The car hit the pole while it was moving” is having semantic ambiguity because the interpretations can be “The car, while moving, hit the pole” and “The car hit the pole while the pole was moving”.

Anaphoric Ambiguity

This kind of ambiguity arises due to the use of anaphora entities in discourse. For example, the horse ran up the hill. It was very steep. It soon got tired. Here, the anaphoric reference of “it” in two situations cause ambiguity.

Pragmatic ambiguity

Such kind of ambiguity refers to the situation where the context of a phrase gives it multiple interpretations. In simple words, we can say that pragmatic ambiguity arises when the statement is not specific. For example, the sentence “I like you too” can have multiple interpretations like I like you (just like you like me), I like you (just like someone else dose)

NLP - Word Sense Disambiguation

We understand that words have different meanings based on the context of its usage in the sentence. If we talk about human languages, then they are ambiguous too because many words can be interpreted in multiple ways depending upon the context of their occurrence.

Word sense disambiguation, in natural language processing (NLP), may be defined as the ability to determine which meaning of word is activated by the use of word in a particular context. Lexical ambiguity, syntactic or semantic, is one of the very first problem that any NLP system faces. Part-of-speech (POS) taggers with high level of accuracy can solve Word's syntactic ambiguity. On the other hand, the problem of resolving semantic ambiguity is called WSD (word sense disambiguation). Resolving semantic ambiguity is harder than resolving syntactic ambiguity.

For example, consider the two examples of the distinct sense that exist for the word "bass" –

1. I can hear bass sound.
2. He likes to eat grilled bass.

The occurrence of the word bass clearly denotes the distinct meaning. In first sentence, it means frequency and in second, it means fish. Hence, if it would be disambiguated by WSD then the correct meaning to the above sentences can be assigned as follows –

1. I can hear bass/frequency sound.
2. He likes to eat grilled bass/fish.

Evaluation of WSD

The evaluation of WSD requires the following two inputs –

A Dictionary :The very first input for evaluation of WSD is dictionary, which is used to specify the senses to be disambiguated.

Test Corpus :Another input required by WSD is the high-annotated test corpus that has the target or correct-senses. The test corpora can be of two types

Lexical sample – This kind of corpora is used in the system, where it is required to disambiguate a small sample of words.

All-words – This kind of corpora is used in the system, where it is expected to disambiguate all the words in a piece of running text.

Approaches and Methods to Word Sense Disambiguation (WSD)

Approaches and methods to WSD are classified according to the source of knowledge used in word disambiguation.

Let us now see the four conventional methods to WSD –

Dictionary-based or Knowledge-based Methods

As the name suggests, for disambiguation, these methods primarily rely on dictionaries, treasures and lexical knowledge base. They do not use corpora evidences for disambiguation. The Lesk method is the seminal dictionary-based method introduced by Michael Lesk in 1986. The Lesk definition, on which the Lesk algorithm is based is "measure overlap between sense definitions for all words in context". However, in 2000, Kilgarriff and Rosensweig gave the simplified Lesk definition as "measure overlap between sense definitions of word and current context", which further means identify the correct sense for one word at a time. Here the current context is the set of words in surrounding sentence or paragraph.

Supervised Methods

For disambiguation, machine learning methods make use of sense-annotated corpora to train. These methods assume that the context can provide enough evidence on its own to disambiguate the sense. In these methods, the words

knowledge and reasoning are deemed unnecessary. The context is represented as a set of "features" of the words. It includes the information about the surrounding words also. Support vector machine and memory-based learning are the most successful supervised learning approaches to WSD. These methods rely on substantial amount of manually sense-tagged corpora, which is very expensive to create.

Semi-supervised Methods

Due to the lack of training corpus, most of the word sense disambiguation algorithms use semi-supervised learning methods. It is because semi-supervised methods use both labelled as well as unlabeled data. These methods require very small amount of annotated text and large amount of plain unannotated text. The technique that is used by semisupervised methods is bootstrapping from seed data.

Unsupervised Methods

These methods assume that similar senses occur in similar context. That is why the senses can be induced from text by clustering word occurrences by using some measure of similarity of the context. This task is called word sense induction or discrimination. Unsupervised methods have great potential to overcome the knowledge acquisition bottleneck due to non-dependency on manual efforts.

Applications of Word Sense Disambiguation (WSD)

Word sense disambiguation (WSD) is applied in almost every application of language technology.

Let us now see the scope of WSD –

Machine Translation

Machine translation or MT is the most obvious application of WSD. In MT, Lexical choice for the words that have distinct translations for different senses, is done by WSD. The senses in MT are represented as words in the target language. Most of the machine translation systems do not use explicit WSD module.

Information Retrieval (IR)

Information retrieval (IR) may be defined as a software program that deals with the organization, storage, retrieval and evaluation of information from document

repositories particularly textual information. The system basically assists users in finding the information they required but it does not explicitly return the answers of the questions. WSD is used to resolve the ambiguities of the queries provided to IR system. As like MT, current IR systems do not explicitly use WSD module and they rely on the concept that user would type enough context in the query to only retrieve relevant documents.

Text Mining and Information Extraction (IE)

In most of the applications, WSD is necessary to do accurate analysis of text. For example, WSD helps intelligent gathering system to do flagging of the correct words. For example, medical intelligent system might need flagging of "illegal drugs" rather than "medical drugs"

Lexicography

WSD and lexicography can work together in loop because modern lexicography is corpusbased. With lexicography, WSD provides rough empirical sense groupings as well as statistically significant contextual indicators of sense**.**

Difficulties in Word Sense Disambiguation (WSD)

Followings are some difficulties faced by word sense disambiguation (WSD) –

Differences between dictionaries

The major problem of WSD is to decide the sense of the word because different senses can be very closely related. Even different dictionaries and thesauruses can provide different divisions of words into senses.

Different algorithms for different applications

Another problem of WSD is that completely different algorithm might be needed for different applications. For example, in machine translation, it takes the form of target word selection; and in information retrieval, a sense inventory is not required.

Inter-judge variance

Another problem of WSD is that WSD systems are generally tested by having their results on a task compared against the task of human beings. This is called the problem of interjudge variance.

Word-sense discreteness

Another difficulty in WSD is that words cannot be easily divided into discrete submeanings.

Reference Resolution:

Reference resolution is a vital aspect of natural language processing (NLP) that addresses the challenge of identifying and linking words or phrases in a text to the entities or concepts they refer to. This process is crucial for achieving a deeper understanding of the context and meaning encoded in natural language.

In the complex structure of human language, words and phrases often act as references to previously mentioned or upcoming entities. Understanding these references is essential for accurately interpreting the content of a text and is a fundamental step toward enabling machines to comprehend language in a manner similar to humans.

Reference resolution can be broadly categorized into two types:
1. Anaphora resolution
2. Cataphora resolution.

1. Anaphora Resolution:
Anaphora refers to a situation where a word or phrase refers to something mentioned earlier in the text.
For **example**, in the sentence "John went to the store. He bought some groceries," the word "He" is an anaphor that refers to "John."
Anaphora resolution involves determining the antecedent (the word or phrase to which the anaphor refers).

2. Cataphora Resolution:

Cataphora is the opposite of anaphora. In cataphora, a pronoun or phrase refers to something mentioned later in the text.
For **example**, "After he arrived, John went to the store." Here, "he" is a cataphor referring to "John," who is mentioned later.
Cataphora resolution involves identifying the later mention to which the cataphor refers.

Deep learning approaches, particularly neural network models, have shown success in capturing complex relationships and contextual information for reference resolution. End-to-end models, such as those based on transformer architectures, have achieved state-of-the-art results in various NLP tasks, including coreference resolution.

Reference resolution is essential for applications like text summarization, question answering, and machine translation, where understanding the relationships between entities in a text is crucial for generating accurate and coherent outputs.

Reference Resolution

Interpretation of the sentences from any discourse is another important task and to achieve this we need to know who or what entity is being talked about. Here, interpretation reference is the key element. **Reference** may be defined as the linguistic expression to denote an entity or individual. For example, in the passage, Ram, the manager of ABC bank, saw his friend Shyam at a shop. He went to meet him, the linguistic expressions like Ram, His, He are reference.

On the same note, **reference resolution** may be defined as the task of determining what entities are referred to by which linguistic expression.

Terminology Used in Reference Resolution

We use the following terminologies in reference resolution −

Referring expression − The natural language expression that is used to perform reference is called a referring expression. For example, the passage used above is a referring expression.

Referent − It is the entity that is referred. For example, in the last given example Ram is a referent.

Corefer − When two expressions are used to refer to the same entity, they are called corefers. For example, **Ram** and **he** are corefers.

Antecedent − The term has the license to use another term. For example, **Ram** is the antecedent of the reference **he**.

Anaphora & Anaphoric − It may be defined as the reference to an entity that has been previously introduced into the sentence. And, the referring expression is called anaphoric.

Discourse model – The model that contains the representations of the entities that have been referred to in the discourse and the relationship they are engaged in.

Types of Referring Expressions

Let us now see the different types of referring expressions. The five types of referring expressions are described below –

Indefinite Noun Phrases

Such kind of reference represents the entities that are new to the hearer into the discourse context. For example – in the sentence Ram had gone around one day to bring him some food – some is an indefinite reference.

Definite Noun Phrases

Opposite to above, such kind of reference represents the entities that are not new or identifiable to the hearer into the discourse context. For example, in the sentence - I used to read The Times of India – The Times of India is a definite reference.

Pronouns

It is a form of definite reference. For example, Ram laughed as loud as he could. The word **he** represents pronoun referring expression.

Demonstratives

These demonstrate and behave differently than simple definite pronouns. For example, this and that are demonstrative pronouns.

Names

It is the simplest type of referring expression. It can be the name of a person, organization and location also. For example, in the above examples, Ram is the name-refereeing expression.

Reference Resolution Tasks

The two reference resolution tasks are described below.

Coreference Resolution

It is the task of finding referring expressions in a text that refer to the same entity. In simple words, it is the task of finding corefer expressions. A set of coreferring expressions are called coreference chain. For example - He, Chief Manager and His - these are referring expressions in the first passage given as example.

Constraint on Coreference Resolution

In English, the main problem for coreference resolution is the pronoun it. The reason behind this is that the pronoun it has many uses. For example, it can refer much like he and she. The pronoun it also refers to the things that do not refer to specific things. For example, It's raining. It is really good.

Pronominal Anaphora Resolution

Unlike the coreference resolution, pronominal anaphora resolution may be defined as the task of finding the antecedent for a single pronoun. For example, the pronoun is his and the task of pronominal anaphora resolution is to find the word Ram because Ram is the antecedent.

Applications of reference resolution :

Reference resolution plays a crucial role in various natural language processing (NLP) applications by enhancing the understanding of textual content.Here are some applications of reference resolution :

1.**Coreference Resolution**:

- Coreference resolution is a specific type of reference resolution that focuses on identifying when two or more expressions in a text refer to the same entity. This is essential for maintaining coherent and accurate representation of information in applications such as text summarization, information extraction, and question answering.

2. Text Summarization:

- Reference resolution is crucial in text summarization systems to ensure that pronouns and other references are correctly linked to their corresponding

antecedents. This helps in generating concise and coherent summaries by avoiding ambiguity in the relationships between entities.

3.**Question Answering**:

- In question answering systems, understanding references is essential for correctly interpreting and answering user queries. Effective resolution of references contributes to better comprehension of context and facilitates the extraction of relevant information from the text.

4.**Machine Translation**:

- Reference resolution aids in maintaining the consistency of translated content. Ensuring that pronouns and references are accurately translated and aligned with their counterparts in the source language is crucial for producing high-quality translations.

5.Information Extraction:

- Reference resolution is important in information extraction tasks where the goal is to identify and extract specific pieces of information from unstructured text. Accurate resolution of references helps in linking entities and events, improving the overall quality of extracted information.

Discourse Coherence and Structure

The most difficult problem of AI is to process the natural language by computers or in other words natural language processing is the most difficult problem of artificial intelligence. If we talk about the major problems in NLP, then one of the major problems in NLP is discourse processing – building theories and models of how utterances stick together to form **coherent discourse**. Actually, the language always consists of collocated, structured and coherent groups of sentences rather than isolated and unrelated sentences like movies. These coherent groups of sentences are referred to as discourse.

Concept of Coherence

Coherence and discourse structure are interconnected in many ways. Coherence, along with property of good text, is used to evaluate the output quality of natural language generation system. The question that arises here is what does it mean for a text to be coherent? Suppose we collected one sentence from every page of the newspaper, then will it be a discourse? Of-course, not. It is because these sentences do not exhibit coherence. The coherent discourse must possess the following properties –

Coherence relation between utterances

The discourse would be coherent if it has meaningful connections between its utterances. This property is called coherence relation. For example, some sort of explanation must be there to justify the connection between utterances.

Relationship between entities

Another property that makes a discourse coherent is that there must be a certain kind of relationship with the entities. Such kind of coherence is called entity-based coherence.

Discourse structure

An important question regarding discourse is what kind of structure the discourse must have. The answer to this question depends upon the segmentation we applied on discourse. Discourse segmentations may be defined as determining the types of structures for large discourse. It is quite difficult to implement discourse segmentation, but it is very important for **information retrieval, text summarization and information extraction** kind of applications.

Algorithms for Discourse Segmentation

In this section, we will learn about the algorithms for discourse segmentation. The algorithms are described below –

Unsupervised Discourse Segmentation

The class of unsupervised discourse segmentation is often represented as linear segmentation. We can understand the task of linear segmentation with the help of an example. In the example, there is a task of segmenting the text into multi-

paragraph units; the units represent the passage of the original text. These algorithms are dependent on cohesion that may be defined as the use of certain linguistic devices to tie the textual units together. On the other hand, lexicon cohesion is the cohesion that is indicated by the relationship between two or more words in two units like the use of synonyms.

Supervised Discourse Segmentation

The earlier method does not have any hand-labeled segment boundaries. On the other hand, supervised discourse segmentation needs to have boundary-labeled training data. It is very easy to acquire the same. In supervised discourse segmentation, discourse marker or cue words play an important role. Discourse marker or cue word is a word or phrase that functions to signal discourse structure. These discourse markers are domain-specific.

Text Coherence

Lexical repetition is a way to find the structure in a discourse, but it does not satisfy the requirement of being coherent discourse. To achieve the coherent discourse, we must focus on coherence relations in specific. As we know that coherence relation defines the possible connection between utterances in a discourse. Hebb has proposed such kind of relations as follows –

We are taking two terms $\mathbf{S_0}$ and $\mathbf{S_1}$ to represent the meaning of the two related sentences –

Result

It infers that the state asserted by term $\mathbf{S_0}$ could cause the state asserted by $\mathbf{S_1}$. For example, two statements show the relationship result: Ram was caught in the fire. His skin burned.

Explanation

It infers that the state asserted by $\mathbf{S_1}$ could cause the state asserted by $\mathbf{S_0}$. For example, two statements show the relationship – Ram fought with Shyam's friend. He was drunk.

Parallel

It infers p(a1,a2,…) from assertion of S_0 and p(b1,b2,…) from assertion S_1. Here ai and bi are similar for all i. For example, two statements are parallel – Ram wanted car. Shyam wanted money.

Elaboration

It infers the same proposition P from both the assertions – S_0 and S_1 For example, two statements show the relation elaboration: Ram was from Chandigarh. Shyam was from Kerala.

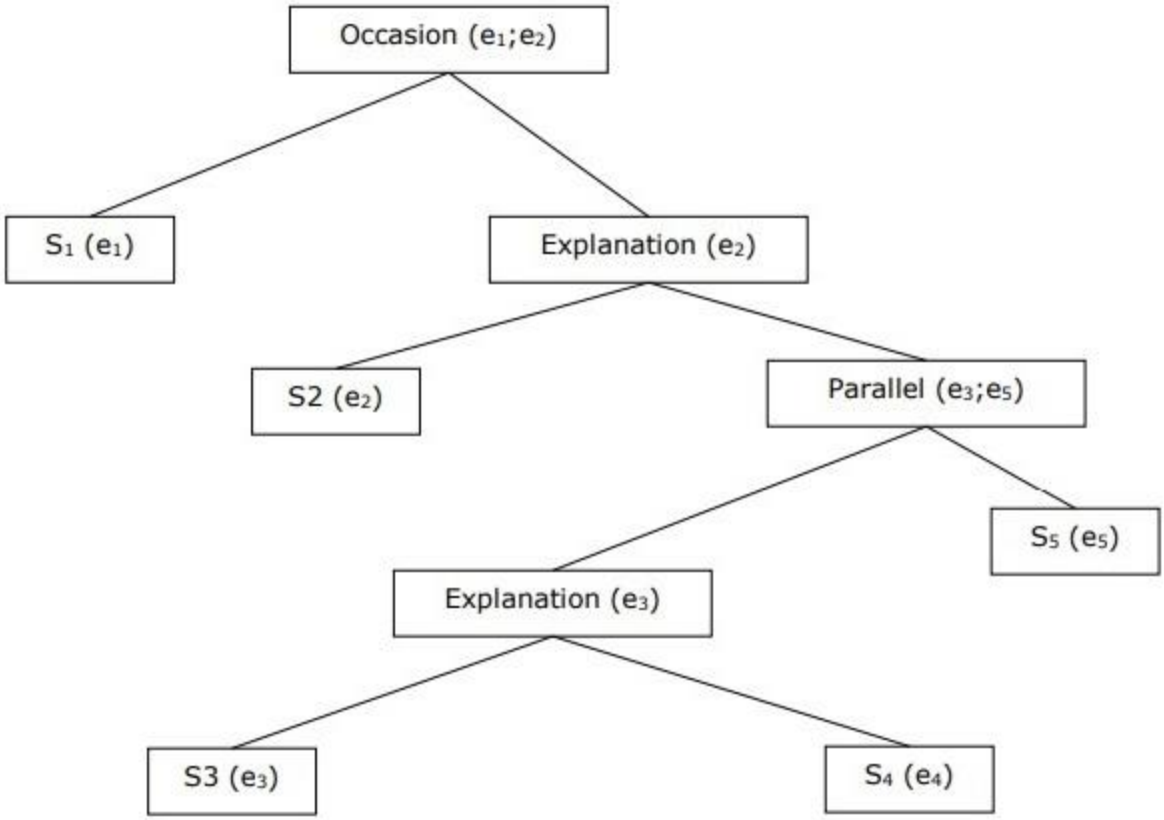
Occasion

It happens when a change of state can be inferred from the assertion of S_0, final state of which can be inferred from S_1 and vice-versa. For example, the two statements show the relation occasion: Ram picked up the book. He gave it to Shyam.

Building Hierarchical Discourse Structure

The coherence of entire discourse can also be considered by hierarchical structure between coherence relations. For example, the following passage can be represented as hierarchical structure –

- S_1 – Ram went to the bank to deposit money.
- S_2 – He then took a train to Shyam's cloth shop.
- S_3 – He wanted to buy some clothes.
- S_4 – He do not have new clothes for party.
- S_5 – He also wanted to talk to Shyam regarding his health



COHESION:

Cohesion in NLP generally refers to the measure of how well the parts of a text stick together or how closely related the words in a text are to each other. It is a concept used to assess the overall flow and unity of a piece of text. Cohesion is crucial for ensuring that a text is coherent and easy to understand.

There are various linguistic devices and techniques that contribute to cohesion in a text, including:

1. Reference: Using pronouns or other words to refer back to previously mentioned entities.

For example, "The cat sat on the mat. It purred softly."

2. Ellipsis: Omitting words that can be easily inferred from the context.

For example, "Mary likes coffee; John, tea."

3. Conjunctions: Using words like "and," "but," "however," etc., to connect ideas and show the relationship between different parts of the text.

4. Lexical Cohesion: Referring to the use of words with related meanings to link different parts of a text. This includes synonyms, antonyms, hypernyms, hyponyms, etc.

5. Repetition: Repeating words, phrases, or structures for emphasis or to reinforce a point.

6. Parallelism: Structuring sentences or phrases in a similar way to create a sense of balance and harmony.
