Q9. Discourse structure and algorithms for segmentation

- Discourse in NLP is nothing but a coherent group of sentences

- Discourse analysis means extracting the meaning out of the corpus or text, and it helps train the NLP model better.

- Coherence is ~~[illegible]~~ used to evaluate the quality of a natural language generation system. A coherent discourse must possess the following properties-

→ Coherence relation between utterances-
The discourse would be coherent if it has ~~[illegible]~~ meaningful ~~relation~~ connections between it's utterances. This property is called coherence relations.

→ Relationship between entities-
Another property that makes discourse coherent is that there must be a certain kind of relationship between entities, and this is called entity-based coherence.

☆ Discourse Structure →

Discourse segmentation can be defined as determining the types of structures for a large discourse. Segmentation is difficult to implement, but is very important as it is used in fields like

- Information Retrieval
- Text summarization
- Information Extraction etc...

☆ Algorithms for Discourse Segmentation →

1. Unsupervised Discourse Segmentation -

- Often represented as linear segmentation

- eg. the task of segmenting the text into multi-paragraph units, where each unit represents a passage of the original text.

- These algorithms are dependant on cohesion - the use of certain linguistic devices to tie textual units together.

  Lexicon cohesion - cohesion indicated by the relationship between two or more words in two units.

2. Supervised Discourse Segmentation -

- The earlier method does not have any hand labelled segment boundaries, but supervised discourse segmentation needs to have boundary-labelled training data.

- discourse markers or cue words play a very important role in signalling discourse structure.

To achieve coherent discourse, we must focus on coherent relations in specific. Coherence relations define meaningful connections between utterances.
Hebb has proposed such kind of relations as follows →

Taking S_0 and S_1 to represent the meaning of two related sentences →

- ~~Result~~ Explanation-

It infers that the state asserted by S_1 could cause the state asserted by S_0

eg. Ram fought with Sham's friend. He was drunk.

- Result -

The state asserted by S_0 could cause the state asserted by S_1

eg. Rahul is late. He will be punished.

- Parallel-

The assertion from the statement S_0, ie. $p(a1, a2 \ldots)$ and the assertion from the statement S_2, ie. $p(b1, b2 \ldots)$ are infered.
a_i and b_i are similar for all i.

eg. Ram wants car. Sham wants money.

- Elaboration-

It infers the same proposition P from both S_0 and S_1

eg. Ram is from Delhi. Sham is from Mumbai.

- Occasion -

Occasion takes place when a change in state is infered from the assertion S_0 and the final state is infered from the assertion S_1.

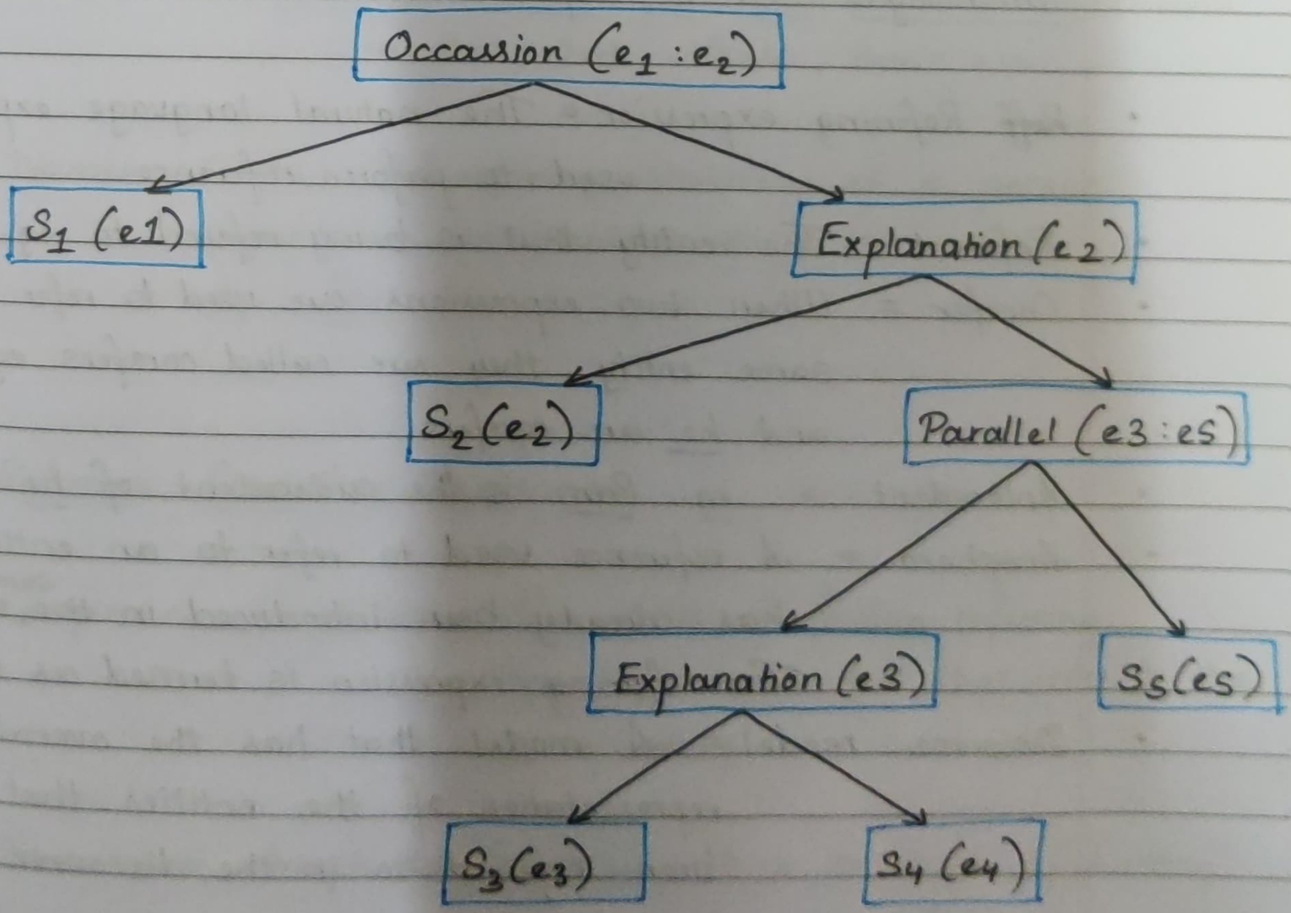
eg. Ram picked up the book. He gave it to Sham.

☆ Building Hierarchical Discourse Structure →

The coherence of the discourse can also be considered by a hierarchical structure of the coherence relations.

eg. The following passage can be represented as a hierarchical structure.

S_1 : Rahul went to the bank to deposit money
S_2 : He then went to Rohan's shop
S_3 : He wanted to buy a phone
S_4 : He did not have a phone
S_5 : He also wanted to $\overset{u}{\text{by}}$ a laptop from Rohan's shop



✯ Reference Resolution →

→ Interpretation of the sentences from any discourse is an important task, and for this we must know who or what entity is being talked about.

→ Reference is a linguistic expression used to denote an entity or individual.

→ eg. Ram saw his friend sham at the shop. He went to meet him.

Here, Ram, his and he are reference to the same individual.

Terminologies used in reference resolution -

- ~~Reff~~ Referring expression = The natural language expression used to perform reference
- Referent = The entity that is being refered to eg. Ram
- Corefer = When two expressions are used to refer to the same entity, they are called corefers. eg. Ram and he are corefers
- Antecedent = eg. Ram is the antecedent of he.
- Anaphora = A reference used to refer to an entity that has already been introduced in the same sentence. The referring expression is termed as anaphoric.
- Discourse model = A model that has the overall representation of the entities that have been referred to in the discourse.

Types of Referring Expressions -

1. Indefinite Noun Phrases - represent an entity that is new to the discourse hearer.
   eg. Rahul is doing **some** work.

2. Definite Noun Phrase - represents an entity that is not new, or is identifiable.
   eg. Rahul reads **the Times of India**

3. Pronouns - A form of definite reference
   eg. Rahul learned as much as **he** could

4. Demonstratives - They are also used to demonstrate nouns, but behave differently than simple pronouns
   eg. this, that, these, those etc.

5. Names - Can be ~~reff~~ referring expressions of a person, organisation or location etc.
   eg. **Ram** is a boy.

Reference Resolution Tasks →

- Coreference Resolution : The task of finding referring expressions in a text that refer to the same entity.

- Pronominal Anaphora Resolution : It is the task of finding the antecedent for a single pronoun.