21CSE356T

NATURAL LANGUAGE PROCESSING

UNIT-3

- Representing Meaning
- Lexical Semantics
- Word Senses
- Relation between Senses
- Word Sense Disambiguation
- Word Embeddings

Semantic Discourse Analysis

Semantic Discourse Analysis (SDA) in NLP is a method used to analyze the meaning and relationships between sentences and utterances within a text or conversation. It goes beyond analyzing individual words or sentences in isolation and focuses on understanding how meaning is constructed and conveyed across a larger discourse.

Here's a breakdown of key aspects:

- **Focus on Meaning:**
  - SDA emphasizes the semantic relationships between different parts of a text. This includes understanding how sentences connect to each other logically, how they build upon previous information, and how they contribute to the overall meaning of the discourse.
  - It looks at how meaning is created through the interaction of words, phrases, and sentences within a specific context.
- **Discourse Level Analysis:**
  - Unlike traditional semantic analysis that focuses on isolated sentences, SDA examines the structure and coherence of larger units of text, such as paragraphs, conversations, or entire documents.
  - It considers how discourse elements, such as pronouns, conjunctions, and discourse markers, contribute to the overall flow and meaning of the text.
- **Key Concepts and Techniques:**
  - **Coherence and Cohesion:** SDA analyzes how sentences and utterances are linked together to create a coherent and cohesive text. Cohesion refers to the explicit linguistic links between sentences, while coherence refers to the underlying logical connections.
  - **Reference Resolution:** Identifying how pronouns and other referring expressions relate to their antecedents in the text.
  - **Discourse Relations:** Identifying the semantic relationships between discourse segments, such as cause-effect, contrast, elaboration, and explanation. Rhetorical Structure Theory is a common framework used to identify those relationships.
  - **Topic Modeling and Topic Segmentation:** Identifying the main topics discussed in a discourse and dividing the discourse into segments based on topic changes.
  - **Dialogue Analysis:** Analyzing the structure and meaning of conversations, including turn-taking, speech acts, and dialogue acts.

    - **Sentiment Analysis within context:** understanding how sentiment changes and is affected by the surrounding conversation.
- **Applications:**
    - **Question Answering:** Understanding the context of a question and finding the relevant information in a large text.
    - **Summarization:** Generating concise summaries of long texts by identifying the key semantic relationships.
    - **Dialogue Systems:** Building more natural and engaging chatbots and virtual assistants.
    - **Information Extraction:** Extracting specific information from texts by understanding the semantic relationships between different parts of the text.
    - **Sentiment Analysis:** more accurate sentiment analysis by understanding the context of the text.
    - **Machine Translation:** Improving the accuracy of machine translation by considering the semantic relationships between sentences in the source and target languages.

Representing Meaning:Semantic Discourse Analysis within NLP, we're essentially addressing how we capture and structure the complex web of relationships and interpretations that arise within a text. It's about moving beyond surface-level word analysis to a deeper understanding of what the language is conveying. Here's a breakdown of key ways meaning is represented:

1. Semantic Networks and Graphs:

- These structures represent concepts as nodes and the relationships between them as edges.
- In discourse analysis, these can be used to visualize how different topics and entities connect throughout a text.
- For example, a graph could show how a character in a story relates to other characters, locations, and events.

2. Logical Formalisms:

- Systems like first-order logic can be used to represent the meaning of sentences in a precise and unambiguous way.
- This is particularly useful for tasks like question answering, where the system needs to reason about the information in a text.

- These formalisms allow for the representation of things like:
    - Relationships between entities (e.g., "John is the brother of Mary").
    - Quantifiers (e.g., "All dogs bark").
    - Logical connectives (e.g., "and," "or," "not").

3. Discourse Representation Structures (DRS):

- DRS is a specific formalism designed to capture the meaning of discourse, including phenomena like pronouns, temporal relationships, and discourse relations.
- It builds a representation of the discourse as it unfolds, keeping track of the entities and events that have been mentioned.

4. Vector Space Models:

- Techniques like word embeddings (e.g., Word2Vec, GloVe) and sentence embeddings (e.g., Sentence-BERT) represent words and sentences as vectors in a high-dimensional space.
- The idea is that semantically similar words or sentences will be located close to each other in this space.
- This allows for tasks like:
    - Finding semantically related documents.
    - Measuring the similarity between sentences.

5. Rhetorical Structure Theory (RST):

- RST focuses on identifying the rhetorical relationships between segments of a text, such as:
    - Cause-effect.
    - Elaboration.
    - Contrast.
- By representing these relationships, we can gain a better understanding of the overall structure and purpose of the discourse.

Key Considerations:

- **Context:** Meaning is highly dependent on context. Representations must account for the surrounding text, the speaker's intentions, and the background knowledge of the participants.
- **Ambiguity:** Natural language is inherently ambiguous. Representations must be able to handle multiple possible interpretations.

- **Dynamic Meaning:** The meaning of a discourse evolves as it unfolds. Representations must be able to capture this dynamic aspect.

Lexical semantics: Lexical semantics in NLP focuses on the study of word meaning, examining how words relate to concepts and to each other within a language. It delves into the internal structure of words, analyzing their sense relations like synonymy, antonymy, hyponymy, and meronymy, which reveal how words are organized in a semantic network. This area of NLP is crucial for tasks such as word sense disambiguation, where the correct meaning of a word in a specific context must be determined, and for information retrieval, where understanding the semantic similarity between words is essential for accurate search results. By analyzing lexical resources like WordNet and employing techniques like word embeddings, NLP systems can better understand the nuanced meanings of words and their roles in constructing larger linguistic structures, ultimately enhancing the system's ability to process and understand natural language.

The study of lexical semantics concerns:

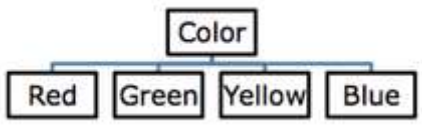
- the classification and decomposition of lexical items
- the differences and similarities in lexical semantic structure cross-linguistically
- the relationship of lexical meaning to sentence meaning and syntax.

Lexical relations

Hyponymy and hypernymy

Hyponymy and hypernymy refer to a relationship between a general term and the more specific terms that fall under the category of the general term.

For example, the colors *red*, *green*, *blue* and *yellow* are hyponyms. They fall under the general term of *color*, which is the hypernym.



Taxonomy showing the hypernym "color"
Color (hypernym) → red, green, yellow, blue (hyponyms)

Hyponyms and hypernyms can be described by using a taxonomy, as seen in the example.

Synonym

Synonym refers to words that are pronounced and spelled differently but contain the same meaning.

Happy, joyful, glad[6]

Antonym

Antonym refers to words that are related by having the opposite meanings to each other. There are three types of antonyms: graded antonyms, complementary antonyms, and relational antonyms.

Sleep, awake[6]
long, short

Homonymy

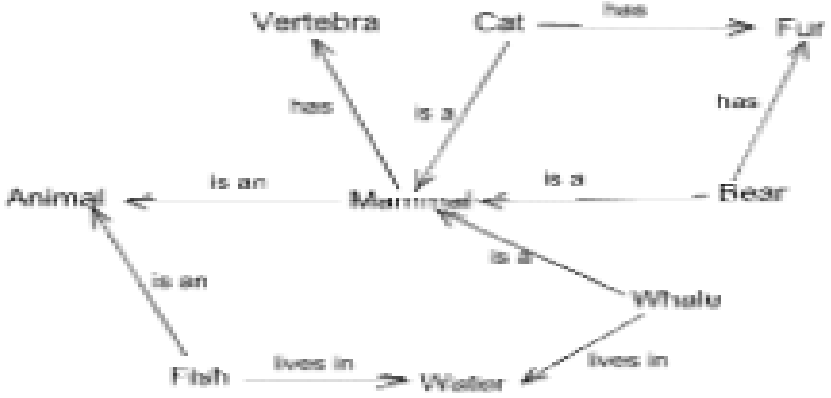
Homonymy refers to the relationship between words that are spelled or pronounced the same way but hold different meanings.

bank (of river)
bank (financial institution)

Polysemy

Polysemy refers to a word having two or more related meanings.

bright (shining)
bright (intelligent)



An example of a semantic network

Semantic networks

Lexical semantics also explores whether the meaning of a lexical unit is established by looking at its neighbourhood in the semantic network,[7] (words it occurs with in natural sentences), or whether the meaning is already locally contained in the lexical unit.

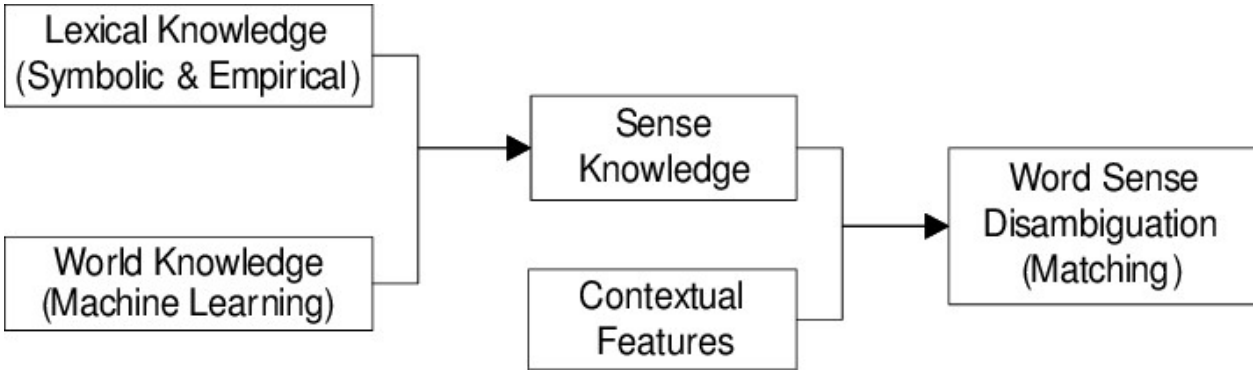
In English, WordNet is an example of a semantic network. It contains English words that are grouped into synsets. Some semantic relations between these synsets are meronymy, hyponymy, synonymy, and antonymy.

Word Senses

In Natural Language Processing (NLP), "word senses" refer to the different meanings or interpretations that a single word can have in various contexts. Words can be polysemous, meaning they have multiple related meanings, or homonymous, where they have entirely different meanings that are not related.

For example, the word "bank" can refer to:

1. A financial institution (e.g., "I need to go to the bank to withdraw cash.").
2. The side of a river (e.g., "The fisherman sat on the bank of the river.").



Relation between Senses

The relation between word senses is a critical aspect of Word Sense Disambiguation (WSD) in Natural Language Processing (NLP). Here are some key points regarding these relationships:

1. **Polysemy and Homonymy**: Words can have multiple senses due to polysemy (related meanings) or homonymy (unrelated meanings). Understanding these relationships helps in disambiguating the intended meaning based on context.

2. **Graph-based Approaches**: Many modern WSD techniques utilize graph-based methods, where word senses are represented as nodes in a graph. The edges represent relationships between these senses, allowing algorithms to leverage these connections to determine the most appropriate sense in a given context.

3. **Contextual Influence**: The meaning of a word often depends on its surrounding words (context). By analyzing the relationships between senses and their contexts, models can better predict which sense is intended. This is particularly important in languages like Telugu, where the nuances of word senses can significantly affect meaning.

4. **Hybrid Models**: Some approaches combine supervised and unsupervised learning to capture the relationships between different senses more effectively. These models can learn from labeled data while also leveraging the structure of the language to improve disambiguation.

5. **Deep Learning**: Recent advancements in deep learning have allowed for more sophisticated models that can understand the relationships between word senses in a more nuanced way. These models can capture complex patterns in data, improving the accuracy of sense disambiguation.

6. **Research and Literature**: There is a wealth of research on the relationships between word senses, including systematic analyses and surveys that explore various methodologies and their effectiveness. This body of work highlights the ongoing development in understanding and modeling these relationships.

Word Sense Disambiguation

Word sense disambiguation (WSD) in Natural Language Processing (NLP) is the problem of identifying which "sense" (meaning) of a word is activated by the use of the word in a particular context or scenario. In people, this appears to be a largely unconscious process. The challenge of correctly identifying words in NLP systems is common, and determining the specific usage of a word in a sentence has many applications. The application of Word Sense Disambiguation involves the area of Information Retrieval, Question Answering systems, Chat-bots, etc.

Word Sense Disambiguation (WSD) is a subtask of Natural Language Processing that deals with the problem of identifying the correct sense of a word in context. Many words in natural language have multiple meanings, and WSD aims to

disambiguate the correct sense of a word in a particular context. For example, the word “bank” can have different meanings in the sentences “I deposited money in the bank” and “The boat went down the river bank”.

WSD is a challenging task because it requires understanding the context in which the word is used and the different senses in which the word can be used. Some common approaches to WSD include:

1. Supervised learning: This involves training a machine learning model on a dataset of annotated examples, where each example contains a target word and its sense in a particular context. The model then learns to predict the correct sense of the target word in new contexts.
2. Unsupervised learning: This involves clustering words that appear in similar contexts together, and then assigning senses to the resulting clusters. This approach does not require annotated data, but it is less accurate than supervised learning.
3. Knowledge-based: This involves using a knowledge base, such as a dictionary or ontology, to map words to their different senses. This approach relies on the availability and accuracy of the knowledge base.
4. Hybrid: This involves combining multiple approaches, such as supervised and knowledge-based methods, to improve accuracy.

WSD has many practical applications, including machine translation, information retrieval, and text-to-speech systems. Improvements in WSD can lead to more accurate and efficient natural language processing systems.

Word Sense Disambiguation (WSD) is a subfield of Natural Language Processing (NLP) that deals with determining the intended meaning of a word in a given context. It is the process of identifying the correct sense of a word from a set of possible senses, based on the context in which the word appears. WSD is important for natural language understanding and machine translation, as it can improve the accuracy of these tasks by providing more accurate word meanings. Some common approaches to WSD include using WordNet, supervised machine learning, and unsupervised methods such as clustering.
The noun ‘star’ has eight different meanings or senses. An idea can be mapped to each sense of the word. For example,

- *“He always wanted to be a Bollywood star.” The word ‘star’ can be described as “A famous and good singer, performer, sports player, actor, personality, etc.”*
- *“The Milky Way galaxy contains between 200 and 400 billion stars”. In this, the word star means “a big ball of burning gas in space that we view as a point of light in the night sky.”*

Difficulties in Word Sense Disambiguation

There are some difficulties faced by Word Sense Disambiguation (WSD).

- **Different Text-Corpus or Dictionary:** One issue with word sense disambiguation is determining what the senses are because different dictionaries and thesauruses divide words into distinct senses. Some academics have proposed employing a specific lexicon and its set of senses to address this problem. In general, however, research findings based on broad sense distinctions have outperformed those based on limited ones. The majority of researchers are still working on fine-grained WSD.
- **PoS Tagging:** Part-of-speech tagging and sense tagging have been shown to be very tightly coupled in any real test, with each potentially constraining the other. Both disambiguating and tagging with words are involved in WSM part-of-speech tagging. However, algorithms designed for one do not always work well for the other, owing to the fact that a word's part of speech is mostly decided by the one to three words immediately adjacent to it, whereas a word's sense can be determined by words further away.

Sense Inventories for Word Sense Disambiguation

Sense Inventories are the collection of abbreviations and acronyms with their possible senses. Some of the examples used in Word Sense Disambiguation are:

- **Princeton WordNet:** is a vast lexicographic database of English and other languages that is manually curated. For WSD, this is the de facto standard inventory. Its well-organized Synsets, or clusters of contextual synonyms, are nodes in a network.
- **BabelNet:** is a multilingual dictionary that covers both lexicographic and encyclopedic terminology. It was created by semi-automatically mapping numerous resources, including WordNet, multilingual versions of WordNet, and Wikipedia.
- **Wiktionary:** a collaborative project aimed at creating a dictionary for each language separately, is another inventory that has recently gained popularity.

Approaches for Word Sense Disambiguation

There are many approaches to Word Sense Disambiguation. The three main approaches are given below:

1. Supervised: The assumption behind supervised approaches is that the context can supply enough evidence to disambiguate words on its own (hence, world knowledge and reasoning are deemed unnecessary).

Supervised methods for Word Sense Disambiguation (WSD) involve training a model using a labeled dataset of word senses. The model is then used to

disambiguate the sense of a target word in new text. Some common techniques used in supervised WSD include:

1. Decision list: A decision list is a set of rules that are used to assign a sense to a target word based on the context in which it appears.
2. Neural Network: Neural networks such as feedforward networks, recurrent neural networks, and transformer networks are used to model the context-sense relationship.
3. Support Vector Machines: SVM is a supervised machine learning algorithm used for classification and regression analysis.
4. Naive Bayes: Naive Bayes is a probabilistic algorithm that uses Bayes' theorem to classify text into predefined categories.
5. Decision Trees: Decision Trees are a flowchart-like structure in which an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.

Random Forest: Random Forest is an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes.

- **Supervised WSD Exploiting Glosses:** Textual definitions are a prominent source of information in sense inventories (also known as glosses). Definitions, which follow the format of traditional dictionaries, are a quick and easy way to clarify sense distinctions
- **Purely Data-Driven WSD:** In this case, a token tagger is a popular baseline model that generates a probability distribution over all senses in the vocabulary for each word in a context.
- **Supervised WSD Exploiting Other Knowledge**: Additional sources of knowledge, both internal and external to the knowledge base, are also beneficial to WSD models. Some researchers use BabelNet translations to fine-tune the output of any WSD system by comparing the output senses' translations to the target's translations provided by an NMT system.

2. Unsupervised: The underlying assumption is that similar senses occur in similar contexts, and thus senses can be induced from the text by clustering word occurrences using some measure of similarity of context. Using fixed-size dense vectors (word embeddings) to represent words in context has become one of the most fundamental blocks in several NLP systems. Traditional word embedding approaches can still be utilized to improve WSD, despite the fact that they conflate words with many meanings into a single vector representation. Lexical databases (e.g., WordNet, ConceptNet, BabelNet) can also help unsupervised

systems map words and their senses as dictionaries, in addition to word embedding techniques.

3. Knowledge-Based: It is built on the idea that words used in a text are related to one another, and that this relationship can be seen in the definitions of the words and their meanings. The pair of dictionary senses having the highest word overlap in their dictionary meanings are used to disambiguate two (or more) words. Lesk Algorithm is the classical algorithm based on Knowledge-Based WSD. Lesk algorithm assumes that words in a given "neighborhood" (a portion of text) will have a similar theme. The dictionary definition of an uncertain word is compared to the terms in its neighborhood in a simplified version of the Lesk algorithm.