21CSE356T
NATURAL LANGUAGE PROCESSING
UNIT-3
SEMANTIC AND DISCOURSE ANALYSIS

- **Representing Meaning**
- **Lexical Semantics**
- **Word Senses**
  - **Relation between Senses**
- **Word Sense Disambiguation**
- **Word Embeddings**
  - **Word2Vec**
    - **CBOW**
    - **Skip-gram**
    - **GloVe**
- **Discourse Segmentation**
- **Text Coherence**
- **Discourse Structure**
- **Reference Resolution**
- **Pronominal Anaphora Resolution**
- **Coreference Resolution**

SEMANTIC AND DISCOURSE ANALYSIS

Semantic analysis focuses on understanding the meaning of words, phrases, sentences, and texts in context. It handles tasks like word sense disambiguation, semantic role labeling, and text entailment.

Key Tasks in Semantic Analysis

1. **Word Sense Disambiguation (WSD):** Determining the correct meaning of a word based on context.
   - Example: "bank" can mean a financial institution or a riverbank.
2. **Semantic Role Labeling (SRL):** Identifying the roles of words in a sentence, such as agents, objects, and instruments.
   - Example: "John [Agent] bought a book [Object] for Mary [Beneficiary]."
3. **Named Entity Recognition (NER):** Extracting entities like names, dates, locations, etc., from text.
   - Example: "Barack Obama was born in Hawaii" → Barack Obama [PERSON], Hawaii [LOCATION]
4. **Coreference Resolution:** Resolving references to the same entity in a text.
   - Example: "Mary dropped her phone. She picked it up." → "She" = Mary, "it" = phone.
5. **Sentiment Analysis:** Analyzing the sentiment or emotion expressed in a piece of text (positive, negative, neutral).
6. **Semantic Similarity and Paraphrasing:** Measuring how similar two texts are in meaning.
   - Example: "The cat is on the mat" ≈ "A cat sits on a mat."
7. **Textual Entailment:** Determining if one sentence logically follows from another.
   - Example: Premise: "All cats are animals." Hypothesis: "A cat is an animal." (True)

Approaches to Semantic Analysis

- **Lexical Semantics:** Studies the meaning of words and their relationships (e.g., synonyms, antonyms, hypernyms).
- **Distributional Semantics:** Uses statistical methods to learn word meanings based on their distribution in large corpora (e.g., word embeddings like Word2Vec, GloVe, BERT).
- **Compositional Semantics:** Studies how word meanings combine to form the meaning of phrases and sentences.

Discourse analysis examines how sentences and text form a cohesive, coherent whole. It involves understanding relationships between sentences, discourse structures, and conversational dynamics.

Key Tasks in Discourse Analysis

1. **Coherence and Cohesion:** Understanding how sentences and phrases connect logically.
   - Example: "I bought a car. It is red." ("It" refers to "car.")
2. **Discourse Parsing:** Identifying the structure of a text and the relations between its parts.
   - Example: Identifying contrast, cause-effect, or elaboration between sentences.
3. **Rhetorical Structure Theory (RST):** Analyzing the organization of text using rhetorical relations.
   - Example: "Because John was late, he missed the train." (Cause-Effect)
4. **Anaphora and Cataphora Resolution:** Resolving backward ("anaphora") or forward ("cataphora") references in text.
   - Anaphora: "Jane loves her dog. She plays with it daily."
   - Cataphora: "When she arrived, Maria was exhausted."
5. **Dialog Modeling:** Understanding and generating context-aware responses in conversations.
   - Example: Building chatbots or virtual assistants.
6. **Topic Segmentation:** Dividing text or discourse into segments based on topic shifts.
   - Example: Separating sections of a news article by subject.

Approaches to Discourse Analysis

- **Rule-Based Approaches:** Using predefined linguistic rules to identify discourse relations.
- **Machine Learning Models:** Training classifiers on labeled data for tasks like coreference resolution or discourse parsing.
- **Deep Learning Models:** Neural networks (e.g., transformers like BERT, GPT) are highly effective for understanding context and maintaining coherence.

REPRESENTING MEANING

The frameworks that are used to specify the syntax and semantics of these representations will be called **meaning representation languages**. In the representational approach, we take **linguistic inputs** and **construct meaning representations.** The process such representations are created and assigned to linguistic inputs is called ***semantic analysis***.

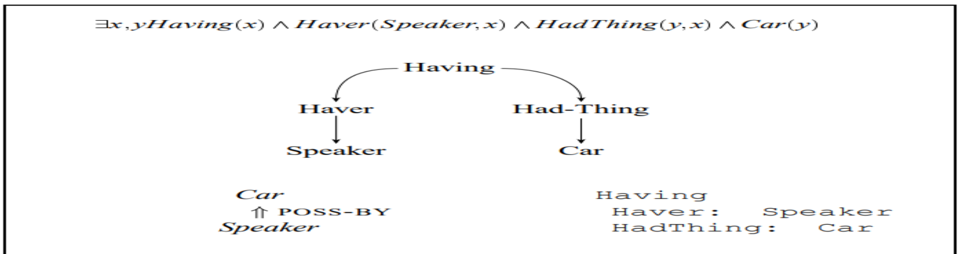


Figure 14.1 A list of symbols, two directed graphs, and a record structure: a sampler of meaning representations for *I have a car*.

Computational Desiderata for Representation

- **Verifiability:** Let us begin by considering the following simple question.

Does Maharani serve vegetarian food?

The meaning underlying the proposition: Maharani serves vegetarian food.

$$Serves(Maharani, VegetarianFood)$$

If the system finds a representation matching the input proposition in its knowledge base, it can return an **affirmative answer**. Otherwise, **it must either say No**, if its knowledge of local restaurants is complete, or say that it does not know if there is reason to believe that its knowledge is incomplete. This notion is known as **verifiability**.

- **Unambiguous Representations:** ***Vagueness does not give rise to multiple representations***.

I want to eat Italian food.

- **Canonical Form:** ***Inputs that mean the same thing should have the same meaning representation is known as the doctrine of canonical form***. The process of choosing the right sense in context is called **word sense disambiguation** or word sense tagging by analogy to **part-of-speech tagging.**
- **Inference and Variables: Inference** to refer generically to a system's ability to draw valid conclusions based on the meaning representation of inputs and its store of background knowledge. It must be possible for the system to draw **conclusions** about the truth of propositions that are not explicitly represented in the knowledge base.

Now consider the following somewhat more complex request.

I'd like to find a restaurant where I can get vegetarian food.

We can gloss a representation containing such variables as follows.

Serves (x; VegetarianFood)

Matching such a proposition succeeds only if the variable x can be replaced by some known object in the knowledge base in such a way that the entire proposition will then match.

- **Expressiveness:** To have a single meaning representation language that could adequately represent the meaning of any sensible natural language utterance.

LEXICAL SEMANTICS

The lexicon has a highly systematic structure ***that governs what words can mean, and how they can be used.*** This structure consists of relations among words and their meanings, as well as the ***internal structure of individual words***. The study of this systematic, meaning related, structure is called Lexical Semantics.

A **lexeme**, an individual entry in the lexicon. A lexeme should be thought of as a pairing of a particular orthographic and phonological form with some form of symbolic meaning representation. The **lexicon** is therefore a finite list made up of lexemes. The term sense to refer to a lexeme's meaning component.

Relations among Lexemes and their Senses

- **Homonymy**: Homonymy is defined as a relation that holds between words that have the same form with unrelated meanings.

  (16.1) Instead, a *bank* can hold the investments in a custodial account in the client's name.

  (16.2) But as agriculture burgeons on the east *bank*, the river will shrink even more.

  - **Citation-forms** are the orthographic-forms that are used to alphabetically index words in a dictionary, which in English correspond to what we have been calling the **root** form of a word.

    - ***Lexemes with the same orthographic form with unrelated meanings are called homographs.***
    - In **spelling correction**, homophones can lead to **real-word spelling errors, or malapropism**s, as when lexemes such as *weather* and *whether* are interchanged.
    - In speech recognition, homophones such as *to*, *two* and *too* cause obvious problems.
    - Finally, **text-to-speech** systems are vulnerable to homographs with distinct pronunciations.
- **Polysemy**: The phenomenon of a single lexeme with multiple related meanings is known as polysemy.
    - There are two criteria that are typically invoked to determine whether or not the meanings of two lexemes are related or not: **the history, or etymology**, of the lexemes in question, and how the words are conceived of by native speakers.
    - The issue of discovering the proper set of senses for a given lexeme is distinct from the process of determining which sense of a lexeme is being used in a given example. This latter task is called **word sense disambiguation**, or word sense tagging by analogy to **part-of-speech tagging**.
- **Synonymy**: Different lexemes with the same meaning.
    - **Substitutability**: Two SUBSTITUTABILITY lexemes will be considered synonyms if they can substitute for one another in a sentence without changing either the meaning or the acceptability of the sentence.
- **Hyponymy**: Pairings where one lexeme denotes a subclass of the other. For example, the relationship between car and vehicle is one of hyponymy. Example: **car is a hyponym of vehicle, and vehicle is hypernym of car**. The term **ontology** usually refers to an analysis of some domain, or microworld, into a set of distinct objects.

WORD SENSE

Words are ambiguous: the same word can be used to mean different things. The words ‘mouse’ or ‘bank’ are polysemous (**Greek ‘having many senses’, poly- ‘many’ + sema, ‘sign, mark’**). ***A sense (or word sense) is a discrete representation of one aspect of the meaning of a word***.

Example:

mouse^1 : .... a mouse controlling a computer system in 1968.

mouse^2 : .... a quiet animal like a mouse

bank^1 : ...a bank can hold the investments in a custodial account ...

bank^2 : ...as agriculture burgeons on the east bank, the river ...

One is based on the fact that dictionaries give textual definitions for each sense called **glosses**.

Glosses are not a formal meaning representation; they are just written for people. Consider the following fragments from the definitions of *right*, *left, red,* and *blood* from the **American Heritage Dictionary**

right	*adj.* located nearer the right hand esp. being on the right when facing the same direction as the observer.
left	*adj.* located nearer to this side of the body than the right.
red	*n.* the color of blood or a ruby.
blood	*n.* the red liquid that circulates in the heart, arteries and veins of animals.

- **How many senses do words have?**

Dictionaries and thesauruses give discrete lists of senses.

- **Relations Between Senses**

Synonymy: couch/sofa vomit/throw up filbert/hazelnut car/automobile

Antonymy: long/short big/little fast/slow cold/hot dark/light rise/fall up/down in/out

Hyponym: car is a hyponym of vehicle, dog is a hyponym of animal, mango is a hyponym of fruit.

Hypernym: car, and animal is a hypernym of dog.
Superordinate: The word superordinate is often used instead of hypernym.

Superordinate	vehicle	fruit	furniture	mammal
Subordinate	car	mango	chair	dog

IS-A hierarchy: Another name for the hypernym/hyponym structure is the IS-A hierarchy, in which we say A IS-A B, or B subsumes A.
Meronymy: the part-whole relation. A leg is part of a chair; a wheel is part of a car.
Structured Polysemy: BUILDING ↔ ORGANIZATION
Metonymy: *the White House* to refer to the administration whose office is in the White House.

AUTHOR (*Jane Austen wrote Emma*)	↔	WORKS OF AUTHOR (*I really love Jane Austen*)
FRUITTREE (*Plums have beautiful blossoms*)	↔	FRUIT (*I ate a preserved plum yesterday*)

WORD SENSE DISAMBIGUATION

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

- I can hear bass sound.
- 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 −

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

Evaluation of WSD

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)

1. **Selection Restriction-based Disambiguation:**
   Rule-to-Rule approach
   Blocks the formation of representations with selectional restriction violations
   Dishes + stir/ fry = food sense
   Dishes + wash = artifact sense
2. **Robust Word Sense Disambiguation:**
   Robust and Stand alone systems

Feature selection, Feature vector

Train classifier to assign words to senses

a) Dictionary-based or Knowledge-based Methods: These methods primarily rely on dictionaries, treasures and lexical knowledge base. They do not use corpora evidences for disambiguation. Lesk definition, on which the ***Lesk algorithm*** is based is "measure overlap between sense definitions for all words in context".

b) 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.

c) Semi-supervised Methods: In semi-supervised methods, we use both labelled as well as unlabelled data. These methods require very small amount of annotated text and large amount of plain unannotated text.

d) Unsupervised Methods: Unsupervised methods have great potential to overcome the knowledge acquisition bottleneck due to non-dependency on manual efforts.

Applications of Word Sense Disambiguation (WSD)

- **Machine Translation**: In MT, Lexical choice for the words that have distinct translations for different senses, is done by WSD.
- **Information Retrieval (IR):** 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.
- **Text Mining and Information Extraction (IE):** 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 corpus based.

Difficulties in Word Sense Disambiguation (WSD)

- **Differences between dictionaries**: 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.
- **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.

WORD EMBEDDINGS

In natural language processing (NLP), word embedding is a term used for ***the representation of words for text analysis, typically in the form of a real-valued vector*** that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning. Word embeddings can be obtained using a set of language modeling and feature learning techniques where words or phrases from the vocabulary are mapped to vectors of real numbers.

Features: Anything that relates words to one another. Eg: Age, Sports, Fitness, Employed etc. Each word vector has values corresponding to these features.

Goal of Word Embeddings

To reduce dimensionality

To use a word to predict the words around it

Inter word semantics must be captured

How are Word Embeddings used?

They are used as input to machine learning models.

Take the words —-> Give their numeric representation —-> Use in training or inference

To represent or visualize any underlying patterns of usage in the corpus that was used to train them.

Approaches to get Word Embeddings:

1) Word2Vec:

- In Word2Vec every word is assigned a vector. We start with either a random vector or **one-hot vector**.
- Word2Vec consists of models for generating word embedding. These models are shallow two-layer neural networks having one input layer, one hidden layer, and one output layer.
- Words with similar meanings have closer vector representations.
  - o Captures analogies: King - Man + Woman ≈ Queen
  - o Can be used for NLP tasks like sentiment analysis, text classification, and recommendation systems.

One-Hot vector: A representation where only one bit in a vector is 1. If there are 500 words in the corpus then the vector length will be 500. After assigning vectors to each word, we take a window size and iterate through the entire corpus. While we do this there are two **neural embedding methods** which are used.

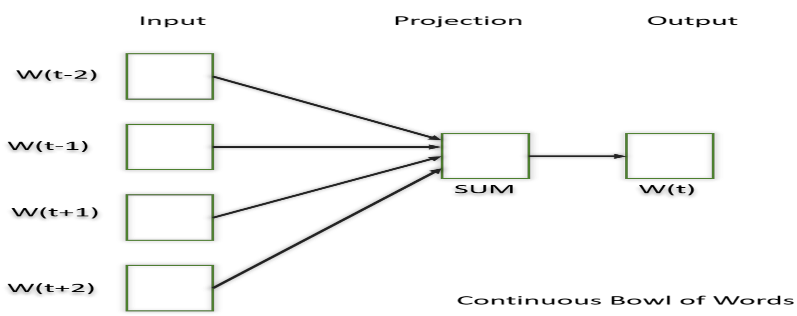
Word2Vec utilizes two architectures:

1.1) Continuous Bowl of Words (CBOW)

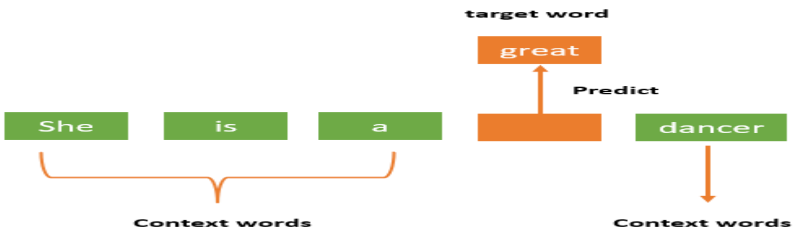
In this model what we do is we try to fit the neighbouring words in the window to the central word.

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.

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.



Example:



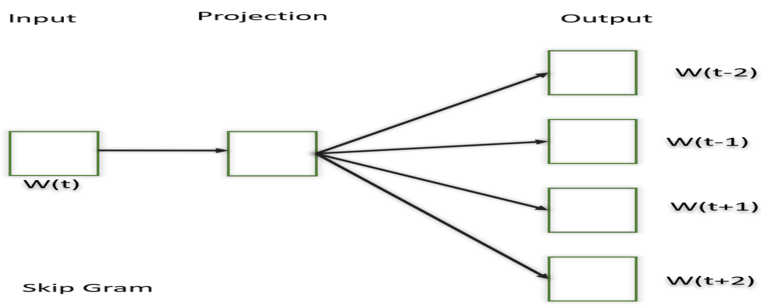
The CBOW model uses the target word around the context word in order to predict it. Consider the above example "**She is a great dancer**." The CBOW model converts this phrase into pairs of context words and target words. The word pairings would appear like this **([she, a], is), ([is, great], a) ([a,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

1.2) Skip Gram

In this model, we try to make the central word closer to the neighbouring words. It is the complete opposite of the CBOW model.

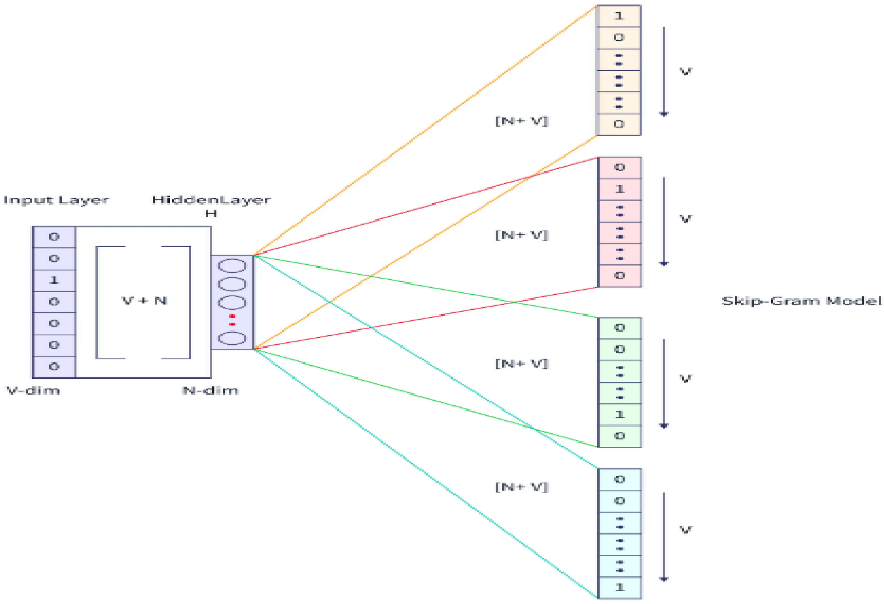
Skip gram predicts the surrounding context words within specific window given current word.

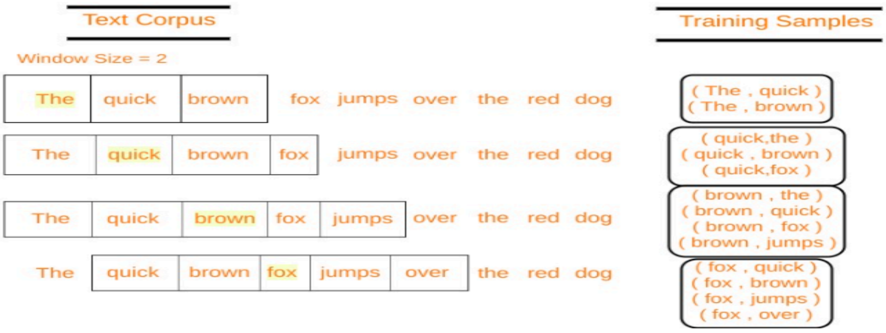
The **input** layer contains the current word and the output layer contains the context words. The **hidden** layer contains the number of dimensions in which we want to represent current word present at the input layer.



Let's say you have the sentence: **The dog fetched the ball.**

If you are trying to train a skip-gram model for the word "dog", the goal of the model is to predict the context words "the" and "fetched" given the input word "dog". So, the training data for the model would be pairs of the form ***(input word = "dog", context word = "the"), (input word = "dog", context word = "fetched")***.





2) GloVe (Global Vectors for Word Representation):

In this method, we take the corpus and iterate through it and get the co-occurrence of each word with other words in the corpus. We get a co-occurrence matrix through this. *The words which occur next to each other get a value of 1, if they are one word apart then 1/2, if two words apart then 1/3 and so on.* Let us take an example to understand how the matrix is created. We have a small corpus:

Corpus:

It is a nice evening.
Good Evening!
Is it a nice evening?

	It	Is	a	nice	evening	good
It	0					
Is	1+1	0				
A	1/2+1	1+1/2	0			
Nice	1/3+1/2	1/2+1/3	1+1	0		
evening	1/4+1/3	1/3+1/4	1/2+1/2	1+1	0	
Good	0	0	0	0	1	0

Initially, the vectors for each word are assigned randomly. Then we take two pairs of vectors and see how close they are to each other in space. ***If they occur together more often or have a higher value in the co-occurrence matrix and are far apart in space then they are brought close to each other. If they are close to each other but are rarely or not frequently used together then they are moved further apart in space.***

Benefits of using Word Embeddings:

- It is much faster to train than hand build models like WordNet (which uses ***graph embeddings***)
- Almost all modern NLP applications start with an embedding layer
- It Stores an approximation of meaning

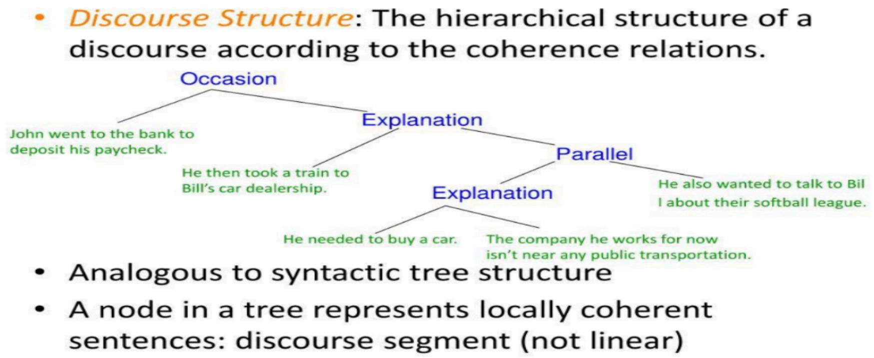
Drawbacks of Word Embeddings:

- It can be memory intensive
- It is corpus dependent. Any underlying bias will have an effect on your model
- It cannot distinguish between homophones. Eg: brake/break, cell/sell, weather/whether etc.

DISCOURSE SEGMENTATION

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.



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.

- **Separating a document into a linear sequence of subtopics:** Information retrieval, for example, for automatically ***segmenting a TV news broadcast or a long news story into a sequence of stories*** so as to find a relevant story, or for text summarization algorithms which need to make sure that different segments of the document are summarized correctly, or for information extraction algorithms which tend to extract information from inside a single discourse segment.

- **Unsupervised Discourse Segmentation:** Cohesion is the use of certain linguistic devices to link or tie together textual units. Lexical cohesion is cohesion indicated by relations between words in the two units, such as use of an identical word a synonym and hypernym.

- **Supervised Discourse Segmentation:** For the task of **paragraph segmentation,** it is trivial to find labeled training data from the web (marked with <p>) or other sources. A **discourse marker** is a word or phrase that functions to signal discourse structure.

- **Evaluating Discourse Segmentation**

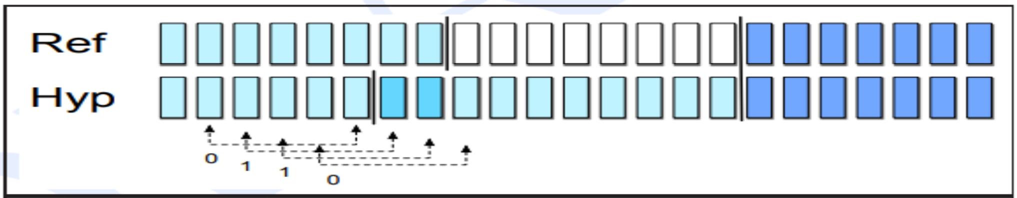


Figure 21.2 The WindowDiff algorithm, showing the moving window sliding over the hypothesis string, and the computation of $|r_i - h_i|$ at four positions. After Pevzner and Hearst (2002).

TEXT COHERENCE

Anaphoric expressions have often been called **cohesive devices**, since the coreference relations they establish serve to ‘tie’ different parts of a discourse together, thus making it cohesive.

- **The Phenomenon**

(18.71) John hid Bill’s car keys. He was drunk.

(18.72) ?? John hid Bill’s car keys. He likes spinach.

While most people find passage (18.71) to be rather unremarkable, they find passage (18.72) to be odd. Like passage (18.71), the sentences that make up passage (18.72) are well formed and readily interpretable.

The possible connections between utterances in a discourse can be specified as a set of coherence relations. A few such relations, proposed COHERENCE RELATIONS by Hobbs (1979a), are given below. The terms S0 and S1 represent the meanings of the two sentences being related.

Result: Infer that the state or event asserted by S_0 causes or could cause the state or event asserted by S_1.

(18.73) John bought an Acura. His father went ballistic.

Explanation: Infer that the state or event asserted by S_1 causes or could cause the state or event asserted by S_0.

(18.74) John hid Bill’s car keys. He was drunk.

Parallel: Infer $p(a_1, a_2, \ldots)$ from the assertion of S_0 and $p(b_1, b_2, \ldots)$ from the assertion of S_1, where a_i and b_i are similar, for all i.

(18.75) John bought an Acura. Bill leased a BMW.

Elaboration: Infer the same proposition P from the assertions of S_0 and S_1.

(18.76) John bought an Acura this weekend. He purchased a beautiful new Integra for 20 thousand dollars at Bill’s dealership on Saturday afternoon.

Occasion: A change of state can be inferred from the assertion of S_0, whose final state can be inferred from S_1, or a change of state can be inferred from the assertion of S_1, whose initial state can be inferred from S_0.

(18.77) John bought an Acura. He drove to the ballgame.

An Inference Based Resolution Algorithm

- **Modus ponens:**

$$\frac{\begin{array}{c}\alpha \Rightarrow \beta \\ \alpha\end{array}}{\beta}$$

An example of modus ponens is the following:

$$\frac{\begin{array}{c}\text{All Acuras are fast.} \\ \text{John's car is an Acura.}\end{array}}{\text{John's car is fast.}}$$

Deduction is a form of **sound inference**: if the premises are true, then the conclusion must be true.

- **Abductive Inference:**

$$\frac{\begin{array}{c}\alpha \Rightarrow \beta \\ \beta\end{array}}{\alpha}$$

Whereas deduction runs an implication relation forward, abduction runs it backward, reasoning from an effect to a potential cause. An example of abduction is the following:

$$\frac{\begin{array}{c}\text{All Acuras are fast.} \\ \text{John's car is fast.}\end{array}}{\text{John's car is an Acura.}}$$

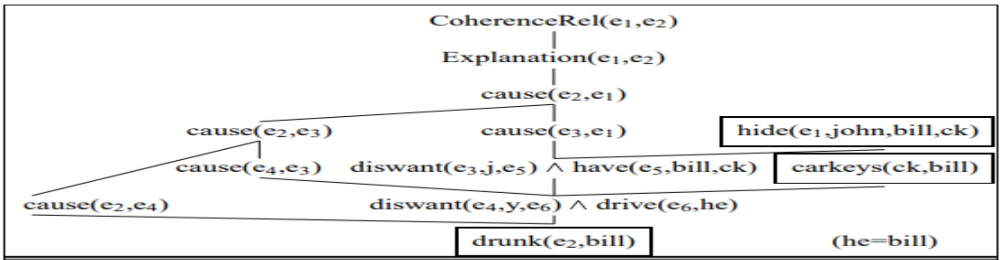


Figure 18.9 Establishing the coherence of passage (18.71).

- **Coherence and Coreference**

This approach provides an explanation for why the pronoun in passage (18.71) is most naturally interpreted as referring to Bill, but the pronoun in passage (18.96) is most naturally interpreted as referring to John.

(18.96) John lost Bill's car keys. He was drunk.

Establishing the coherence of passage (18.96) under Explanation requires an axiom that says that being drunk could cause someone to lose something. Because such an axiom will dictate that the person who is drunk must be the same as the person losing something, the free variable representing the pronoun will become bound to John. The only lexico-syntactic difference between passages (18.96) and (18.71), however, is the verb of the first sentence. The grammatical positions of the pronoun and potential antecedent noun phrases are the same in both cases, so syntactically-based preferences do not distinguish between these.

- **Discourse Connectives**

Sometimes a speaker will include a specific cue, called a **connective**, that serves to constrain the set of coherence relations that can hold between two or more utterances. For example, the connective because indicates the Explanation relationship explicitly, as in passage (18.97).

(18.97) John hid Bill's car keys because he was drunk.

The meaning of because can be represented as cause (e_2 ; e_1), which would play a similar role in the proof as the cause predicate that was introduced abductively via axiom (18.79).

- **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").

REFERENCE RESOLUTION

- A natural language expression used to perform reference is called a **referring expression**, and the entity that is referred to is called the **referent**. Thus, *John* and he in passage (18.1) are *referring expressions, and John is REFERENT.*

- Two referring expressions that are used to refer to the same entity are said to **corefer**, thus *John and he* **corefer** in passage (18.1).
- We call John the **antecedent** of he.
- Reference to an entity that has been previously introduced into the discourse is called **anaphora**, and the referring expression used is said to be **anaphoric**.
- Depending on the operative discourse context, you might DISCOURSE CONTEXT say *it, this, that, this car, that car, the car, the Acura, the Integra, or my friend's car*, among many other possibilities.
- For instance, you cannot simply say it or the Acura if the hearer has no prior knowledge of your friend's car, it has not been mentioned before, and it is not in the immediate surroundings of the discourse participants (i.e., the situational context of the discourse).
- A subset of these beliefs that has a special status form the hearer's mental model of the ongoing discourse, which we call a **discourse mode**l.
- When a referent is first mentioned in a discourse, we say that a representation for it is **evoked** into the model. Upon subsequent mention, this representation is **accessed** from the model.

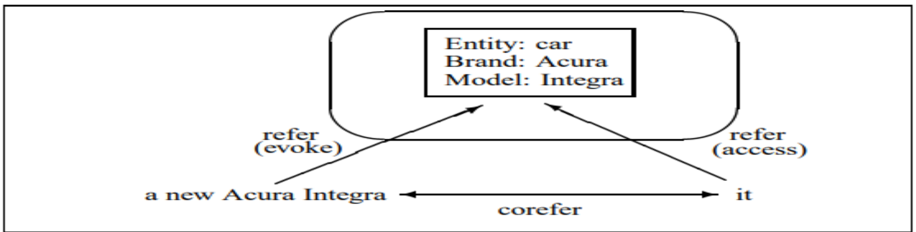


Figure 18.1 Reference operations and relationships. **THIS PICTURE TO BE REPLACED.**

Reference Phenomena

<u>Types of Referring Expression</u>

1. **Indefinite Noun Phrases**

Indefinite reference introduces entities that are new to the hearer into the discourse context. The most common form of indefinite reference is marked with the determiner a (or an), as in (18.5), but it can also be marked by a quantifier such as some (18.6) or even the determiner this (18.7).

(18.5) I saw an Acura Integra today.

(18.6) Some Acura Integras were being unloaded at the local dealership today.

(18.7) I saw this awesome Acura Integra today.

Such noun phrases evoke a representation for a new entity that satisfies the given description into the discourse model.

2. **Definite Noun Phrases**

Definite reference is used to refer to an entity that is identifiable to the hearer, either because it has already been mentioned in the discourse context (and thus is represented in the discourse model), it is contained in the hearer's set of beliefs about the world, or the uniqueness of the object is implied by the description itself. The case in which the referent is identifiable from discourse context is shown in (18.9).

(18.9) I saw an Acura Integra today. The Integra was white and needed to be washed.

Examples in which the referent is either identifiable from the hearer's set of beliefs or is inherently unique are shown in (18.10) and (18.11) respectively.

3. **Pronouns**

Another form of definite reference is pronominalization, illustrated in example (18.12).

(18.12) I saw an Acura Integra today. It was white and needed to be washed.
The constraints on using pronominal reference are stronger than for full definite noun phrases, requiring that the referent have a high degree of activation or **salience** in the discourse model. Pronouns usually (but not always) refer to entities that were introduced no further than one or two sentences back in the ongoing discourse, whereas definite noun phrases can often refer further back. Pronouns can also participate in **cataphora**, in which they are mentioned before their referents are, as in example (18.14).
(18.14) Before he bought it, John checked over the Integra very carefully.
Here, the pronouns he and it both occur before their referents are introduced.
Pronouns also appear in quantified contexts in which they are considered to be bound, as in example (18.15).
(18.15) Every woman bought her Acura at the local dealership.
Under the relevant reading, *her* does not refer to some woman in context, but instead behaves like a variable bound to the quantified expression *every woman*.

4. Demonstratives

Demonstrative pronouns, like ***this*** and ***that***, behave somewhat differently that simple definite ronouns like it. They can appear either alone or as determiners, for instance, ***this*** Acura, ***that*** Acura. The choice between two demonstratives is generally associated with some notion of spatial proximity: ***this*** indicating ***closeness*** and ***that signaling distance***.

5. One Anaphora

One-anaphora, exemplified in (18.18), blends properties of definite and indefinite reference.
(18.18) I saw no less than 6 Acura Integras today. Now I want one.
This use of one can be roughly paraphrased by one of them, in which them refers to a plural referent (or generic one, as in the case of (18.18), see below), and one selects a member from this set. Thus, *one* may evoke a new entity into the discourse model, but it is necessarily dependent on an existing referent for the description of this new entity.

<u>Three Types of Referents that Complicate the Reference Resolution Problem</u>

1. Inferrables

For instance, in some cases a referring expression does not refer to an entity that has been explicitly evoked in the text, but instead one that is inferentially related to an evoked entity. Such referents are called ***inferrables***. Consider the expressions a door and the engine in sentence
Eg: I almost bought an Acura Integra today, but a door had a dent and the engine seemed noisy.
The indefinite noun phrase a door would normally introduce a new door into the discourse context, but in this case the hearer is to infer something more: that it is not just any door, but one of the doors of the Integra.

2. Discontinuous Sets

In some cases, references using plural referring expressions like they and them (see page 672) refer to sets of entities that are evoked together, for instance, using another plural expression (their Acuras) or a conjoined noun phrase (John and Mary):
(18.23) John and Mary love their Acuras. They drive them all the time.

3. Generics

Making the reference problem even more complicated is the existence of generic reference. Consider example (18.25).
(18.25) I saw no less than 6 Acura Integras today. They are the coolest cars

PRONOMINAL ANAPHORA RESOLUTION

Pronominal Anaphora Resolution refers to the process in natural language processing (NLP) where a computer system identifies the antecedent (the noun phrase that a pronoun refers to) of a pronoun within a sentence or text, essentially figuring out "**who**" or "**what**" a pronoun is referring back to based on the surrounding context.

Preferences in Pronoun Interpretation

- **Recency**

The pronoun it is more likely to refer to the Legend than the Integra.
(18.49) John has an Integra. Bill has a Legend. Mary likes to drive it.

- **Grammatical Role**

(18.50) John went to the Acura dealership with Bill. He bought an Integra. [he = John]
(18.51) Bill went to the Acura dealership with John. He bought an Integra. [he = Bill]
(18.52) John and Bill went to the Acura dealership. He bought an Integra. [he = ??].

- **Repeated Mention**

(18.53) John needed a car to get to his new job. He decided that he wanted something sporty. Bill went to the Acura dealership with him. He bought an Integra. [he = John]

- **Parallelism**

There are also strong preferences that appear to be induced by parallelism effects.

(18.54) Mary went with Sue to the Acura dealership. Sally went with her to the Mazda dealership. [her = Sue]

- **Verb Semantics**

Certain verbs appear to place a semantically-oriented emphasis on one of their argument positions, which can have the effect of biasing the manner in which subsequent pronouns are interpreted.

(18.56) John telephoned Bill. He lost the pamphlet on Acuras.
(18.57) John criticized Bill. He lost the pamphlet on Acuras.

These examples differ only in the verb used in the first sentence, yet the subject pronoun in passage (18.56) is typically resolved to John, whereas the pronoun in passage (18.57) is resolved to Bill. Some researchers have

An Algorithm for Pronoun Resolution

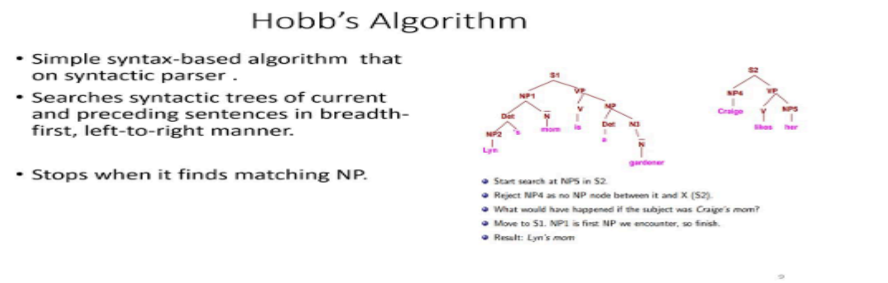
First, when a noun phrase that evokes a new entity is encountered, a representation for it must be added to the discourse model and a degree of salience (which we call a salience value) computed for it. The salience value is calculated as the sum of the weights assigned by a set of salience factors.

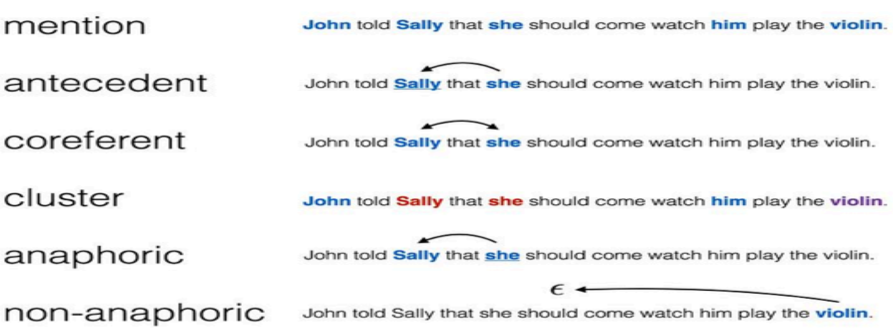
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative (direct object) emphasis	50
Indirect object and oblique complement emphasis	40
Non-adverbial emphasis	50
Head noun emphasis	80

Figure 18.5 Salience factors in Lappin and Leass's system.

Encoding a grammatical role preference scheme using the following hierarchy:
subject > existential predicate nominal > object > indirect object or oblique > demarcated adverbial PP

The steps taken to resolve a pronoun are as follows:
1. Collect the potential referents (up to four sentences back).
2. Remove potential referents that do not agree in number or gender with the pronoun.
3. Remove potential referents that do not pass intrasentential syntactic coreference constraints .
4. Compute the total salience value of the referent by adding any applicable values to the existing salience value previously computed during the discourse model update step.
5. Select the referent with the highest salience value. In the case of ties, select the closest referent in terms of string position (computed without bias to direction).





- ## Key points about pronominal anaphora resolution:
  - **Anaphora**: "Anaphora" is a linguistic term where a word or phrase refers back to a previously mentioned concept, and "pronominal" means that the referring word is a pronoun (like "he", "she", "it").
  - **Antecedent**: The noun phrase that the pronoun is referring back to is called the "antecedent".
    Example:
    "The dog chased the cat. It ran away." In this sentence, "It" is the pronoun, and "the cat" is the antecedent.

- ## Challenges in pronominal anaphora resolution:
  - **Ambiguity:** Sometimes, a pronoun could refer to multiple possible antecedents in a sentence, making it difficult to determine the correct reference.

- **Discourse context**: Understanding the wider context of a conversation or text is crucial to accurately resolve pronouns, as the relevant antecedent might be several sentences back.
- **Gender and number agreement**: Matching the gender and number of the pronoun to its antecedent is important for accurate resolution.

- **How computers perform pronominal anaphora resolution:**
  - **Rule-based approaches:** These systems use linguistic rules based on factors like proximity, grammatical structure, and semantic features to identify the most likely antecedent.
  - **Machine learning models:** Modern approaches often use statistical models trained on large amounts of text data to learn patterns in pronoun usage and identify the most probable antecedent for a given pronoun.

- **Importance of pronominal anaphora resolution:**
  - **Natural language understanding:** By correctly identifying the referents of pronouns, NLP systems can better understand the meaning of a text.
  - **Question answering systems:** Accurate pronoun resolution is crucial for systems that need to interpret questions and provide relevant answers.
  - **Text summarization:** Understanding pronoun references helps in creating concise summaries of longer texts

COREFERENCE RESOLUTION

- A natural language expression used to perform reference is called a referring expression, and the entity that is referred to is called the referent. Thus, John and he in passage (18.1) are referring expressions, and John is REFERENT.

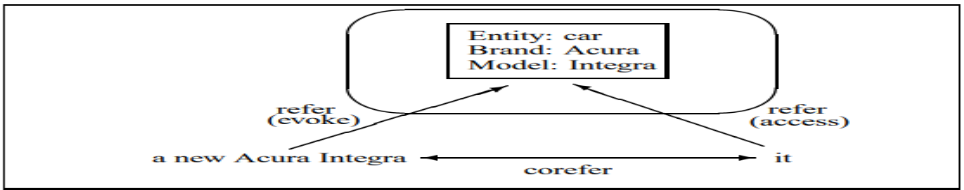


Figure 18.1 Reference operations and relationships. **THIS PICTURE TO BE REPLACED.**

Terminology Used 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.

Reference Resolution Tasks

● **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.

o **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.

Syntactic and Semantic Constraints on Coreference

- **Number Agreement:** Referring expressions and their referents must agree in number; for English, this means distinguishing between singular and plural references.

Singular	Plural	Unspecified
she, her, he, him, his, it	we, us, they, them	you

Figure 18.2 Number agreement in the English pronominal system.

- **Person and Case Agreement:** English distinguishes between three forms of person: first, second, and third.

	First	Second	Third
Nominative	I, we	you	he, she, they
Accusative	me, us	you	him, her, them
Genitive	my, our	your	his, her, their

Figure 18.3 Person and case agreement in the English pronominal system/

- **Gender Agreement**

masculine	feminine	nonpersonal
he, him, his	she, her	it

Figure 18.4 Gender agreement in the English pronominal system.

- **Syntactic Constraints**

Reference relations may also be constrained by the syntactic relationships between a referential expression and a possible antecedent noun phrase when both occur in the same sentence. For instance, the pronouns in all of the following sentences are subject to the constraints indicated in brackets.

(18.36) John bought himself a new Acura. [himself=John]
(18.37) John bought him a new Acura. [him$\neq$John]
(18.38) John said that Bill bought him a new Acura. [him$\neq$Bill]
(18.39) John said that Bill bought himself a new Acura. [himself=Bill]
(18.40) He said that he bought John a new Acura. [He$\neq$John;he$\neq$John]

English pronouns such as himself, herself, and themselves are called ***reflexives***. Oversimplifying the situation considerably, a reflexive corefers with the subject of the most immediate clause that contains it (ex. 18.36), whereas a nonreflexive cannot corefer with this subject (ex. 18.37).

- **Selectional Restrictions**

The selectional restrictions that a verb places on its arguments may be responsible for eliminating referents, as in example (18.45).

(18.45) John parked his Acura in the garage. He had driven it around for hours.

There are two possible referents for it, the Acura and the garage. The verb drive, however, requires that its direct object denote something that can be driven, such as a car, truck, or bus, but not a garage.