

Unit 5

Lexical Semantics

Lexeme, Lexicon, Senses, WordNet, WSD, Word Similarity

Natural Language Processing (NLP)
MDS 555

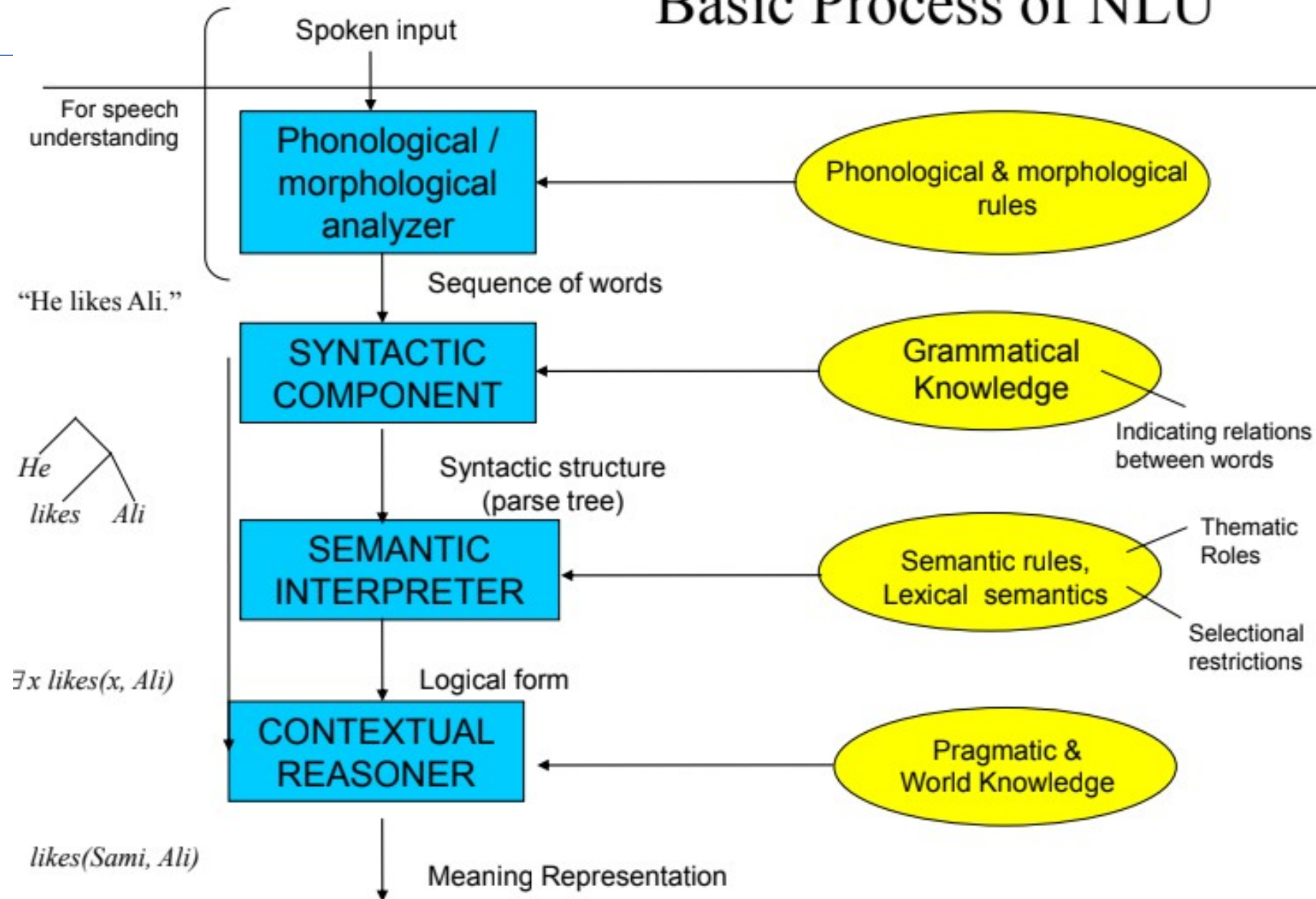


Objective

- Lexeme
- Lexicon
- Senses
- Lexical relations
- WordNet (Lexical Database)
- Word Sense Disambiguation (WSD)
- Word Similarity



Basic Process of NLU



Semantic Analysis

- **Syntactic Analysis** focuses on the **structure** and **grammar** of language
- The purpose of **semantic analysis** is to draw **exact meaning** or **dictionary meaning** from the text.
- The work of **semantic analyzer** is to check the text for meaningfulness.
- Semantic analysis is a process in natural language processing (NLP) that involves **understanding the meaning** of words, phrases, sentences, and larger text segments.



Senses / Meaning

- Traditionally, meaning in language has been studied from three perspectives
 - The meaning of **individual words**
 - The meaning of **individual sentences or utterances**
 - The meaning of a **text or discourse**
- Meaning is a notion in semantics classically defined as having two components:
 - **Reference**, anything in the referential realm denoted by a word or expression, and
 - **Sense**, the system of paradigmatic and syntagmatic relationships between a lexical unit and other lexical units in a language.



Paradigmatic and syntagmatic relationships

- Paradigmatic
 - base: **paradigm** – means pattern and
- Syntagmatic
 - base: **syntagma** – means specific order of word or elements
- Both syntagmatic relationship and paradigmatic relationship, are concepts in linguistics that **showcase the types of relationship, that exists between words and sound segments of a language.**
- They focus on **inclusivity** and **co-occurrence relationships**, that exist between words and sound segments in a language.



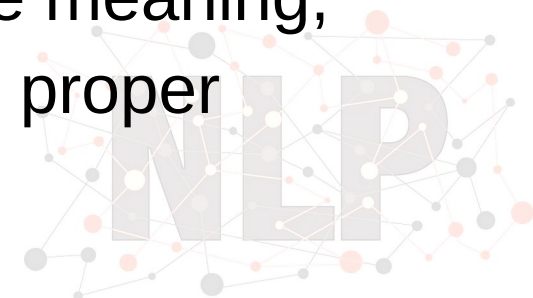
Paradigmatic and syntagmatic relationships

- Paradigmatic Relationships:
 - These are **associations between words** that can substitute for each other in the same context. They are based on **choice and contrast**.
 - Example: In the sentence "The cat is on the mat,"
 - the word "**cat**" could be replaced with "**dog**", "**bird**" or "**rabbit**"
 - These words **are in a paradigmatic relationship** because **they belong to the same category** (animals) and can replace each other without altering the grammatical structure of the sentence.



Paradigmatic and syntagmatic relationships

- Syntagmatic Relationships:
 - These are the relationships between words that **appear together in a sequence**. They are based on combination and order.
 - Example: In the same sentence, "**The cat is on the mat**"
 - the words "**the**", "**cat**", "**is**", "**on**", "**the**" and "**mat**" are in a syntagmatic relationship because they form a meaningful sentence when arranged in this specific order.
 - Changing the order to "**Mat the is on cat the**" disrupts the meaning, illustrating how syntagmatic relationships depend on the proper arrangement of elements.



Semantic Analysis

- It can be done in two parts
 - the **study of the meaning** of individual words
 - the individual words will be combined to **provide the meaning** of the sentences



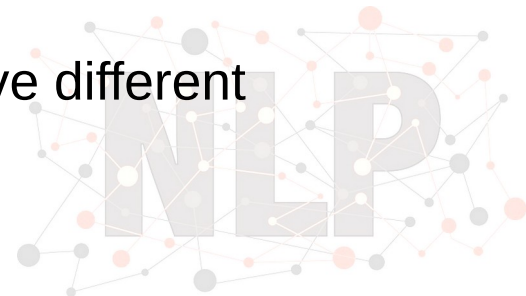
Lexeme

- A lexeme is the **basic unit of meaning** in a language, representing a word or a set of words with related forms.
 - For example, "run," "runs," "ran," and "running" => lexeme "run"
- It is an entry in the lexicon includes
 - An **orthographic** representation
 - A **phonological** form
 - A **symbolic** meaning representation



Lexeme

- An orthographic representation
 - **Visual representation** of words, including spelling, letter arrangement, and punctuation. It is concerned with how words look in written form.
 - The word "**color**" in American English vs. "**colour**" in British English.
 - The orthographic difference lies in the spelling ("color" vs. "colour"), though both words represent the same concept.
- A phonological form
 - Involves the **sound structure** of words, including pronunciation, stress, intonation, and the relationship between sounds in a language.
 - Example: The words "read" (present tense) and "read" (past tense).
 - These words are spelled the same (orthographically identical) but have different pronunciations:
 - **/ri:d/** (present) vs. **/rɛd/** (past), showing their phonological difference.



Lexeme

- A symbolic meaning representation
 - Refers to the **abstract** or representational meaning of a **word**, **symbol**, or **sign**.
 - It goes beyond the literal meaning to include what a word or symbol represents in a **broader cultural, social, or conceptual context**.
 - A common example of symbolic meaning is the word "heart."
 - **Literal Meaning:** A heart is an organ that pumps blood through the body.
 - **Symbolic Meaning:** A **heart** often symbolizes **love, affection, or emotion** in various cultures.
 - heart shape is commonly used to represent love
 - What about **<3** => **I <3 you**



Elements of Semantic Analysis

- Lexical Relationship: 8 Types of lexical relationships-

- 1) Synonymy

- 2) Antonymy

- 3) Hyponymy

- 4) Homonymy

- 5) Polysemy

- 6) Meronymy / Partonymy

- 7) Holonymy

- 8) Homophony



Synonymy

- Different ways of expressing related concepts
- It is the relation between two lexical items having **different forms but expressing the same or a close meaning**.
 - Examples are 'author/writer', 'fate/destiny', 'big/large'
- Synonyms are almost never truly substitutable-
 - Used in different contexts
 - Have different implications



Antonymy

- It is the relation between two lexical items having **symmetry between their semantic components relative to an axis**.
- The scope of antonymy is as follows –
 - Application of property or not – Example is ‘life/death’, ‘certitude/incertitude’
 - Application of scalable property – Example is ‘rich/poor’, ‘hot/cold’
 - Application of a usage – Example is ‘father/son’, ‘moon/sun’.



Hypernym and Hyponyms

- Is A relationship
 - It may be defined as the relationship between a generic term and instances of that generic term.
- The generic term is called hypernym
- The instances are called hyponyms
 - The word color is hypernym and
 - the color blue, yellow etc. are hyponyms



Homonymy

- Lexemes that share a form
 - Phonological, orthographic or both (Spelled and/or Pronounced the same)
- It may be defined as the words having same spelling or same form but having different and unrelated meaning.
 - For example, the word “**Bat**” is a homonymy word
 - Bat can be an implement to hit a ball
 - Bat is a nocturnal flying mammal



Polysemy

- It is a word or phrase with different but related sense
- We can say that polysemy has the same spelling but different and related meaning
 - For example, the word “**bank**” is a polysemy word having the following meanings
 - A **financial institution**
 - The **building** in which such an institution is located
 - A synonym for “to rely on”



Homonym vs Polysemy

- Homonyms: same word, different meaning
 - bank (river)
 - bank (financial)
- Polysemy: different senses of same word
 - That dog has floppy ears.
 - He has a good ear for jokes.
 - bank(financial) - the building, the institution, the notion of where money is stored



Meronymy and holonym

- The relationship where a word represents a whole of which another word is a part.
 - The smaller part is called the "**meronym**"
 - and the larger whole is called the "**holonym**"
 - Metaphore
 - The **White House** released new figures today
 - **Simhadarbar** is sent to villages.... => **Public Administration**
 - As per Officials of Simhadarbar => **Government**
 - **Page** is a meronym of **book**



Lexical Database

WordNet



Lexical Database

- A lexical database is an organized description of the lexemes of a language
 - It contains structured information about words
 - The main difference between lexical databases and dictionaries is that dictionaries aim to explain or translate words, while lexical databases are primarily developed for research purposes.



Lexical Database: WordNet

- WordNet is a lexical database for the English language that groups words into sets of synonyms called **synsets**.
- Each **synset** represents a distinct concept and includes various words that share a common meaning.
 - It was **created** by **cognitive scientists at Princeton University** and is widely used in natural language processing and linguistic research.
- Original Paper: <https://aclanthology.org/H94-1111.pdf>



WordNet

- Each entry is annotated with one or more **senses**
- Each sense provides a variety of information

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (frequency) {offset} <lexical filename> [lexical file number]
(gloss) "an example sentence"

Display options for word: word#sense number (sense key)

Noun

- (1){03730361} <noun.artifact>[06] [S:](#) (n) **mask#1 (mask%1:06:00::)** (a covering to disguise or conceal the face)
- (1){01051399} <noun.act>[04] [S:](#) (n) **mask#2 (mask%1:04:00::)** (activity that tries to conceal something) *"no mask could conceal his ignorance"; "they moved in under a mask of friendship"*
- {08270371} <noun.group>[14] [S:](#) (n) [masquerade#1 \(masquerade%1:14:00::\)](#), [masquerade party#1 \(masquerade party%1:14:00::\)](#), [masque#1 \(masque%1:14:00::\)](#), **mask#3 (mask%1:14:00::)** (a party of guests wearing costumes and masks)
- {03730526} <noun.artifact>[06] [S:](#) (n) **mask#4 (mask%1:06:01::)** (a protective covering worn over the face)

Verb

- (1){02152033} <verb.perception>[39] [S:](#) (v) [dissemble#2 \(dissemble%2:39:00::\)](#), [cloak#1 \(cloak%2:39:00::\)](#), **mask#1 (mask%2:39:00::)** (hide under a false appearance) *"He masked his disappointment"*
- (1){01361031} <verb.contact>[35] [S:](#) (v) **mask#2 (mask%2:35:00::)** (put a mask on or cover with a mask) *"Mask the children for Halloween"*
- {02163017} <verb.perception>[39] [S:](#) (v) [disguise#1 \(disguise%2:39:00::\)](#), **mask#3 (mask%2:39:01::)** (make unrecognizable) *"The herb masks the garlic taste"; "We disguised our faces before robbing the bank"*
- {01361558} <verb.contact>[35] [S:](#) (v) **mask#4 (mask%2:35:02::)** (cover with a sauce) *"mask the meat"*

WordNet Entries

- Senses contains
 - **Gloss**
 - A definition of the sense
 - List of Synonyms
 - Commonly referred to as a **synset**
 - Example sentence

WordNet Search - 3.1

- [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (frequency) {offset} <lexical filename > [lexical file number]
(gloss) "an example sentence"

Display options for word: word#sense number (sense key)

Noun

- (1){03730361} <noun.artifact>[06] [S:](#) (n) **mask#1 (mask%1:06:00::)** (a covering to disguise or conceal the face)
- (1){01051399} <noun.act>[04] [S:](#) (n) **mask#2 (mask%1:04:00::)** (activity that tries to conceal something) *"no mask could conceal his ignorance"; "they moved in under a mask of friendship"*
- {08270371} <noun.group>[14] [S:](#) (n) [masquerade#1 \(masquerade%1:14:00::\)](#), [masquerade party#1 \(masquerade party%1:14:00::\)](#), [masque#1 \(masque%1:14:00::\)](#), **mask#3 (mask%1:14:00::)** (a party of guests wearing costumes and masks)
- {03730526} <noun.artifact>[06] [S:](#) (n) **mask#4 (mask%1:06:01::)** (a protective covering worn over the face)

Verb

- (1){02152033} <verb.perception>[39] [S:](#) (v) [dissemble#2 \(dissemble%2:39:00::\)](#), [cloak#1 \(cloak%2:39:00::\)](#), **mask#1 (mask%2:39:00::)** (hide under a false appearance) *"He masked his disappointment"*
- (1){01361031} <verb.contact>[35] [S:](#) (v) **mask#2 (mask%2:35:00::)** (put a mask on or cover with a mask) *"Mask the children for Halloween"*
- {02163017} <verb.perception>[39] [S:](#) (v) [disguise#1 \(disguise%2:39:00::\)](#), **mask#3 (mask%2:39:01::)** (make unrecognizable) *"The herb masks the garlic taste"; "We disguised our faces before robbing the bank"*
- {01361558} <verb.contact>[35] [S:](#) (v) **mask#4 (mask%2:35:02::)** (cover with a sauce) *"mask the meat"*

Lexicographic Categories

- Super-senses
- Coarse-grained semantic category

Category	Example	Category	Example	Category	Example
ACT	service	GROUP	place	PLANT	tree
ANIMAL	dog	LOCATION	area	POSSESSION	price
ARTIFACT	car	MOTIVE	reason	PROCESS	process
ATTRIBUTE	quality	NATURAL EVENT	experience	QUANTITY	amount
BODY	hair	NATURAL OBJECT	flower	RELATION	portion
COGNITION	way	OTHER	stuff	SHAPE	square
COMMUNICATION	review	PERSON	people	STATE	pain
FEELING	discomfort	PHENOMENON	result	SUBSTANCE	oil
FOOD	food			TIME	day



Sense Relations

- Hypernym: Relation between a concept and its superordinate
 - Food is a hypernym of cake
- Hyponym: Relation between a concept and its subordinate
 - Corgi is a hyponym of dog
- Meronym: Relation between a part and its whole
 - Wheel is a meronym of car
- Holonym: Relation between a whole and its parts
 - Car is a holonym of wheel
- Antonym: Relation between two semantically opposite concepts
 - Leader is an antonym of follower

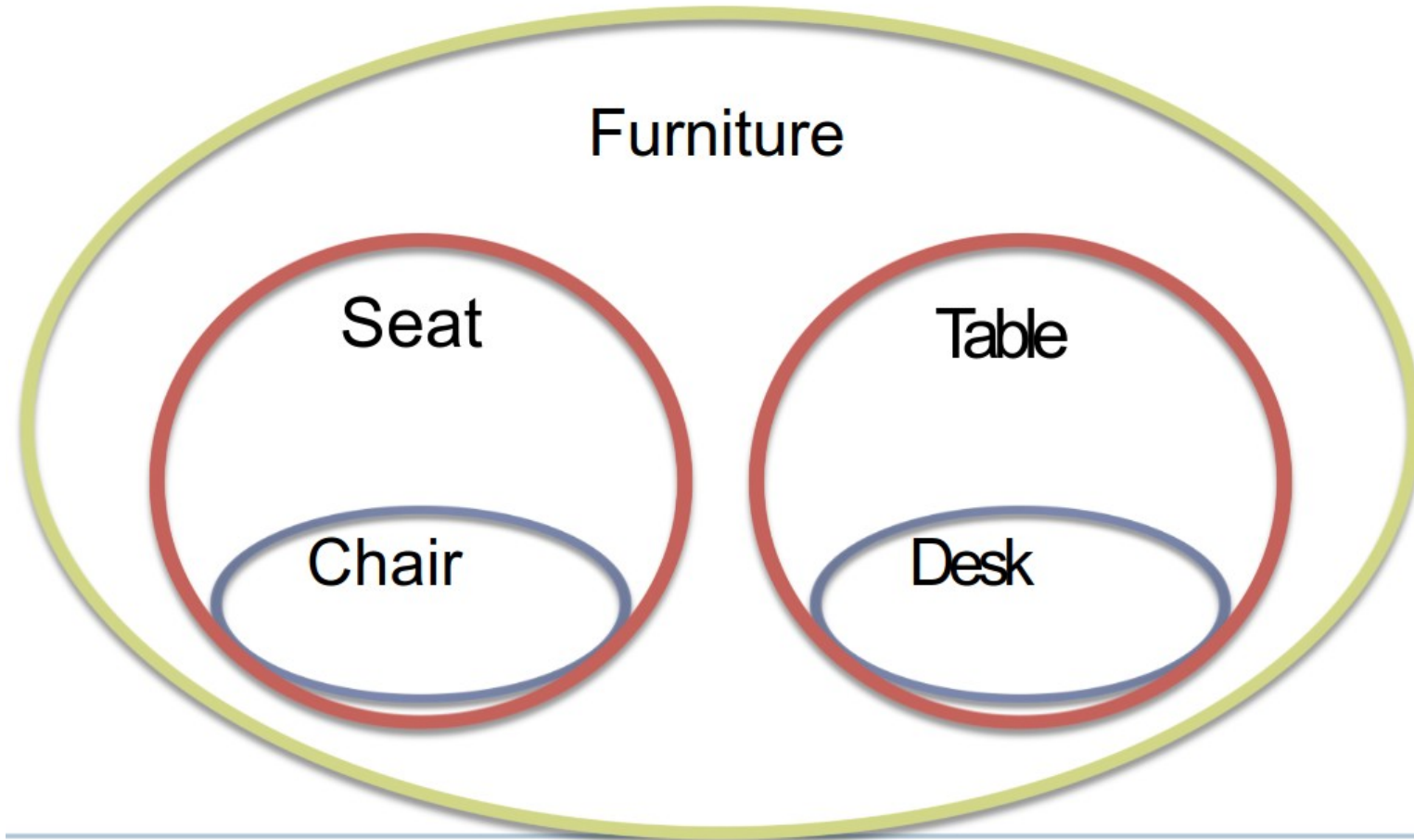


Sense Relations

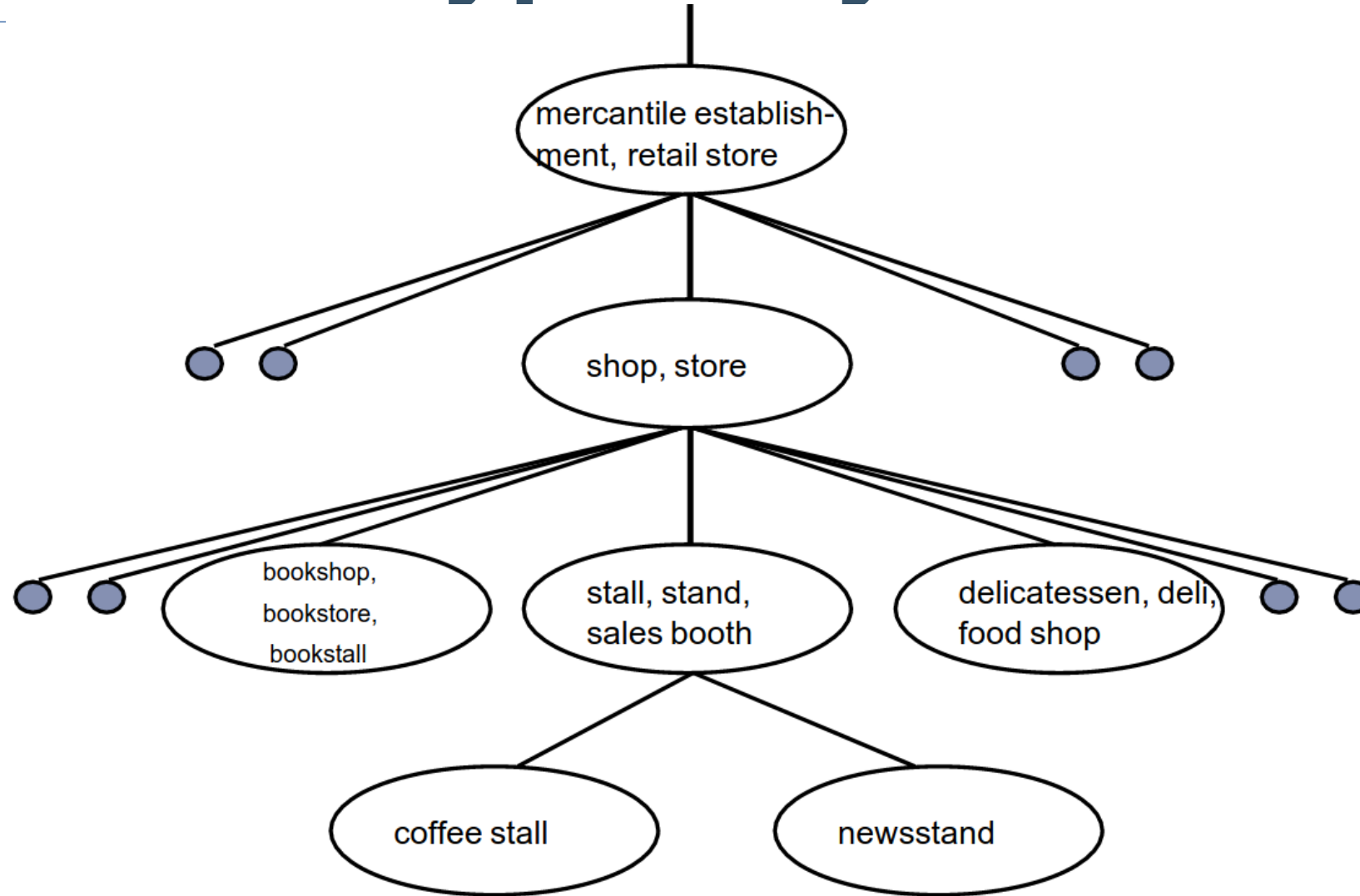
<i>Category</i>	<i>Relation</i>	<i>Type</i>	<i>Example</i>
Noun	Hypernym/hypo	Sem	dog IS A KIND OF animal
	Meronym	Sem	arm IS A PART OF body
Verb	Implication: Cause Precondition Troponym Inclusion Opposition	Sem Lex	to kill CAUSES to die to succeed ENTAILS DOING to try to limp IS ONE WAY TO walk snore ENTAILS DOING to sleep to die ANTONYM to be born
Adj	Antonym	Lex	hot ANTONYM cold
Adv	Derived adj	Lex	quickly DERIVED FROM quick
	Antonym	Lex	quickly ANTONYM slowly



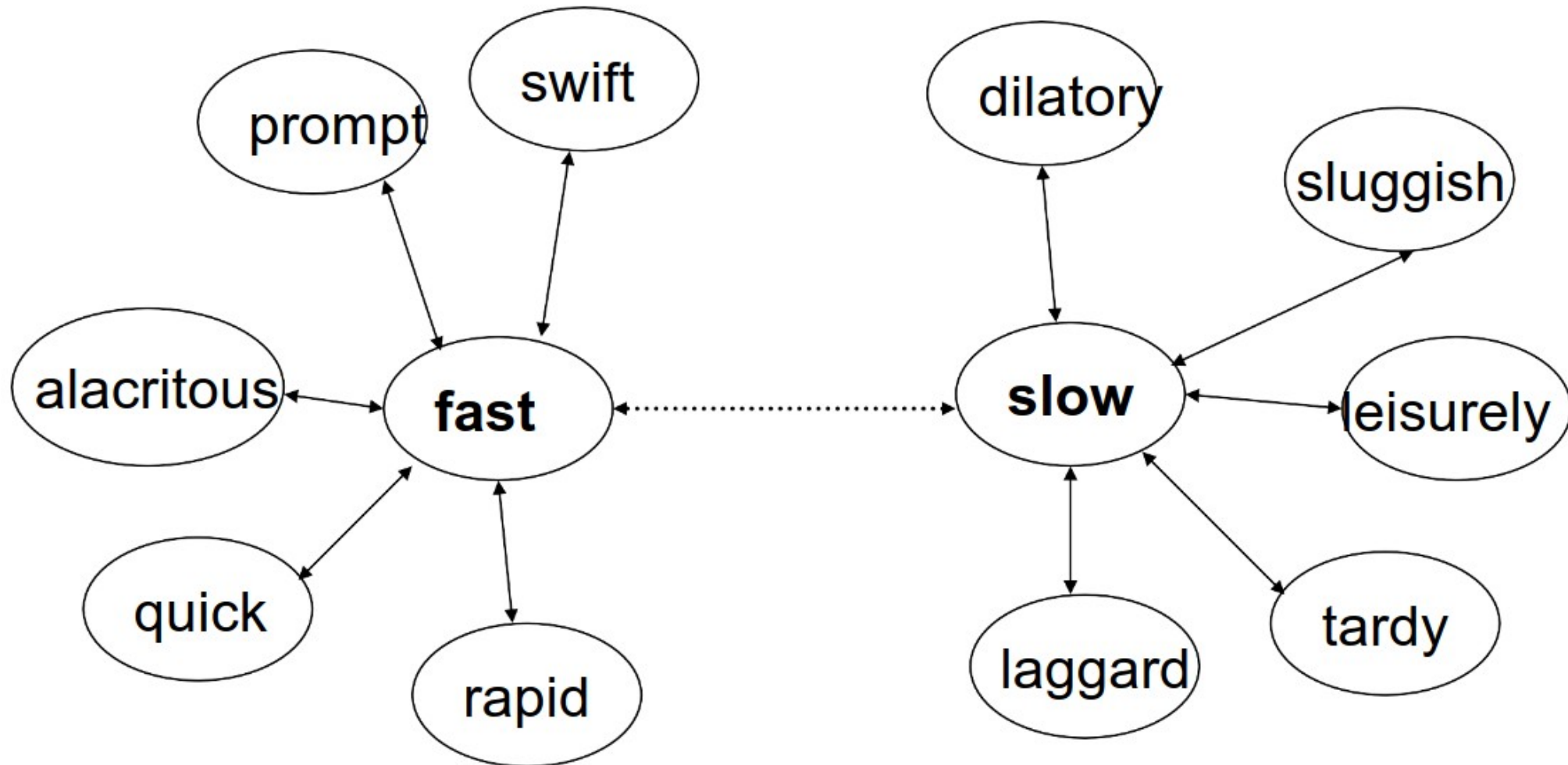
Nouns : Hyperonym



Nouns : Hyperonym

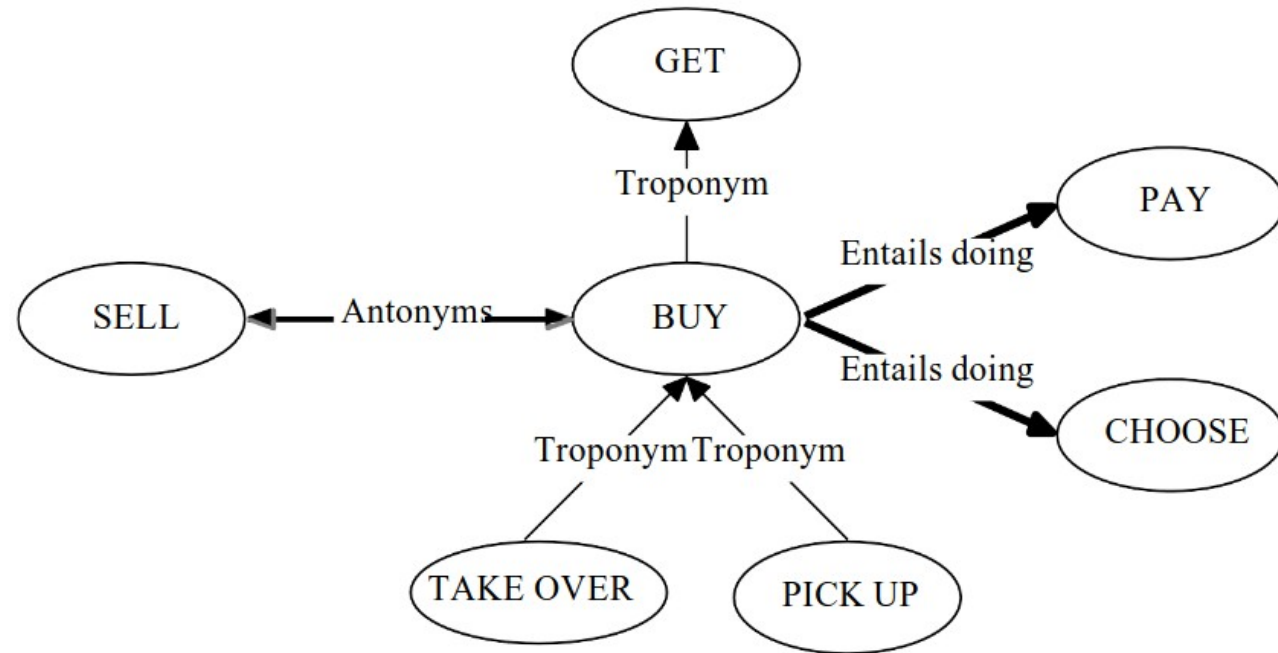


Adjectives (Antonymy, Similarity)



Semantic Network

- Meaning of a word = relations with other words
 - e.g.: to buy



WordNet for multiple languages

- EuroWordNet
 - Create synsets, create relations for every language
 - Then map synsets
- MultiWordNet
 - Create synsets for a new WordNet mapped to the English wordnet synsets (Princeton WordNet, PWN)
 - Importing the semantic relations the new wordnet



Word Sense Disambiguation (WSD)



Word sense disambiguation

- *Word sense disambiguation is the **problem of selecting a sense** for a word from a set of predefined possibilities*
 - Sense Inventory usually comes from a dictionary or thesaurus
 - Knowledge intensive methods
 - supervised learning



Computer vs Humans

- *Polysemy* – most words have many possible meanings.
 - "Bank" (the side of a river) and "bank" (a financial institution).
 - A computer program has no basis for knowing which one is appropriate, even if it is **obvious to a human**...
 - Ambiguity is **rarely a problem for humans** in their day to day communication, except in extreme cases...



Ambiguity for Computer

- The fisherman jumped off the **bank** and into the water.
- The **bank** down the street was robbed!
- Back in the day, we had an entire **bank** of computers devoted to this problem.
- The **bank** in that road is entirely too steep and is really dangerous.
- The plane took a **bank** to the left, and then headed off towards the mountains.



Ambiguity for Humans

- DRUNK GETS NINE YEARS IN VIOLIN CASE
- FARMER BILL DIES IN HOUSE
- STOLEN PAINTING FOUND BY TREE
- RED TAPE HOLDS UP NEW BRIDGE
- DEER KILL 300,000
- RESIDENTS CAN DROP OFF TREES
- INCLUDE CHILDREN WHEN BAKING COOKIES
- MINERS REFUSE TO WORK AFTER DEATH



Two variants of WSD task

- Lexical Sample task
 - Small pre-selected set of target words
 - And inventory of senses for each word
- All-words task
 - Every word in an entire text
 - A lexicon with senses for each word
 - Sort of like part-of-speech tagging



WSD - Applications

- Machine Translation
- Information Retrieval
- Text Summarization
- Question Answering
- Sentiment Analysis
- Text Classification
- Grammar and Style checking of writing
- Speech Context Identification



Approaches to WSD

- Dictionary-Based Approaches
- Supervised Machine Learning
- Unsupervised Machine Learning
- Knowledge-Based Approaches



WSD - Dictionary-Based Approaches

- These approaches **use dictionaries or lexical resources** like WordNet to look up word senses.
 - **WordNet**, for example, provides a **structured database** of words and their senses, along with relationships between these senses (e.g., hypernyms and hyponyms).
- Dictionary-based methods match the word in context to its sense in the dictionary.



WSD - Supervised Machine Learning

- In supervised approaches, a WSD model is **trained on labeled data** where words are tagged with their correct senses.
- Features derived from the context (surrounding words) are used to train classifiers like
 - decision trees,
 - support vector machines
 - neural networks to predict senses.



WSD - Unsupervised Machine Learning

- Unsupervised methods do not rely on labeled data.
- Instead, they use clustering or similarity-based techniques to group instances of a word into clusters representing different senses.
- Common approaches include **clustering algorithms** like
 - k-means
 - distributional similarity based on word co-occurrence.



WSD - Knowledge-Based Approaches

- These approaches incorporate external knowledge sources, such as **ontologies**, **semantic networks**, or **domain-specific databases**, to disambiguate word senses.
- They rely on semantic relationships and knowledge about the entities mentioned in the text.



WSD - Hybrid Approaches

- Some WSD systems combine multiple techniques,
 - such as using a **dictionary-based method** as a fallback when supervised machine learning models are uncertain.



Lesk Algorithm

- Michael E. Lesk introduced the Lesk algorithm in 1986 as a classic approach for word sense disambiguation
- The 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.



Lesk Algorithm

- Intuition: word overlap between context and dictionary entries
 - Unsupervised, but knowledge rich

The **bank** can guarantee deposits will eventually cover future tuition costs because it invests in adjustable-rate mortgage securities.

WordNet

bank ¹	Gloss:	a financial institution that accepts deposits and channels the money into lending activities
	Examples:	“he cashed a check at the bank”, “that bank holds the mortgage on my home”
bank ²	Gloss:	sloping land (especially the slope beside a body of water)
	Examples:	“they pulled the canoe up on the bank”, “he sat on the bank of the river and watched the currents”



Lesk Algorithm

- Simplest implementation:
 - Count overlapping content words between glosses and context
- Lots of variants:
 - Include the examples in dictionary definitions
 - Include hypernyms and hyponyms
 - Give more weight to larger overlaps (e.g., bigrams)
 - Give extra weight to infrequent words (e.g., idf weighting)
 - ...
- Works reasonably well!



Lesk Algorithm

function SIMPLIFIED LESK(*word*, *sentence*) **returns** best sense of *word*

best-sense \leftarrow most frequent sense for *word*

max-overlap \leftarrow 0

context \leftarrow set of words in *sentence*

for each *sense* **in** senses of *word* **do**

signature \leftarrow set of words in the gloss and examples of *sense*

overlap \leftarrow COMPUTEOVERLAP(*signature*, *context*)

if *overlap* > *max-overlap* **then**

max-overlap \leftarrow *overlap*

best-sense \leftarrow *sense*

end

return(*best-sense*)

Figure 20.3 The Simplified Lesk Algorithm. The COMPUTEOVERLAP function returns the number of words in common between two sets, ignoring function words or other words on a stop list. The original Lesk algorithm defines the *context* in a more complex way. The *Corpus Lesk* algorithm weights each overlapping word w by its $-\log P(w)$, and includes labeled training corpus data in the *signature*.



Word Similarity



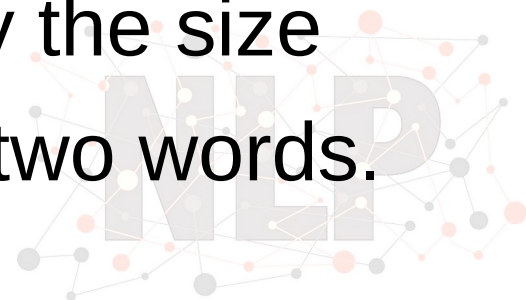
Word similarity

- Word similarity refers to the degree of resemblance or likeness between two words based on their meaning or semantic content.
- It quantifies how close or related two words are **in terms of their semantic interpretation**.
- Word similarity is a fundamental concept in natural language processing (NLP) and computational linguistics,
- It plays a crucial role in various NLP tasks
 - including information retrieval,
 - machine translation,
 - document clustering, and more.



Lexical Similarity

- Lexical similarity measure focuses on the surface form of words and consider factors such as spelling and character overlap. Common lexical similarity measures include:
 - **Jaccard Similarity**: It measures the size of the intersection of characters or n-grams divided by the size of the union of characters or n-grams between two words.



Lexical Similarity

- **Edit Distance** (Levenshtein Distance): It quantifies the minimum number of single-character edits (insertions, deletions, substitutions) required to transform one word into another.
- **Cosine Similarity**: It computes the cosine of the angle between two vectors representing the word frequencies in a document-term matrix. While it's often used for documents, it can also be applied to words.



Semantic Similarity

- It measure of how closely related two words are in terms of their meaning.
- It **quantifies** the degree to which two words share a similar concept or are used in similar contexts.
- Words that have a high semantic similarity typically share the same or closely related meanings, while those with low similarity are conceptually distant.



Semantic Similarity

- Synonymy
 - Words that are synonyms (e.g., "car" and "automobile") have a high semantic similarity because they essentially mean the same thing.
- Hypernymy/Hyponymy
 - Words in a hypernym-hyponym relationship (e.g., "dog" and "animal") have a measurable semantic similarity based on their hierarchical relationship in a taxonomy like WordNet.
- Antonymy
 - Words that are antonyms (e.g., "hot" and "cold") may have a lower semantic similarity because they represent opposite concepts.



Semantic Similarity: Measure

- Lexical Resources:
 - WordNet: Uses hierarchical relationships between words (synsets) to measure similarity.
 - Word Embeddings: Models like Word2Vec, GloVe, or BERT learn vector representations of words based on their usage in large corpora. The cosine similarity between these vectors represents semantic similarity.
- Path-Based Measures:
 - Leacock-Chodorow (LCH): Based on the shortest path between two words in a taxonomy.
 - Wu-Palmer (WUP): Considers the depth of the words and their lowest common ancestor in a taxonomy.



Semantic Similarity: Measure

- Information-Theoretic Measures:
 - Jiang-Conrath (JCN): Combines the information content of the words and their lowest common ancestor.
 - Lin Similarity: Uses the ratio of shared information content to the total information content



Semantic Similarity: Measure

- Information-Theoretic Measures:
 - Jiang-Conrath (JCN): Combines the information content of the words and their lowest common ancestor.
 - Lin Similarity: Uses the ratio of shared information content to the total information content



Information Content

- Information content is often derived from the probability of encountering a concept in a large text corpus. The Information Content IC can be computed as

$$IC(c) = -\log(P(c)) \qquad P(c) = \frac{\text{frequency of occurrences of } c}{\text{total frequency of all concepts in the corpus}}$$

- The concept "animal" is very broad and general, so it might appear frequently in a corpus, leading to a lower IC value.
- The concept "dog" is more specific and would appear less frequently, leading to a higher IC value.



Semantic Similarity: Uses

- Text Mining
 - Identifying similar terms or concepts in documents
- Information Retrieval
 - Enhancing search results by finding documents with semantically similar terms
- Natural Language Processing (NLP)
 - Word sense disambiguation, machine translation, and more



Study Materials

- Word Net: <https://wordnetcode.princeton.edu/5papers.pdf>
- <http://disi.unitn.it/~ldkr/ldkr2017/slides/11.KDI.WordNet.A.Closer.Look.pdf>
- <http://lintool.github.io/UMD-courses/CMSC723-2009-Fall/session11-slides.pdf>
- <https://web.stanford.edu/~jurafsky/slp3/slides/Chapter18.wsd.pdf>
-



Thank you

