Module 4: Semantic Analysis notes . . .

Module 4: Semantic Analysis

Semantic Analysis – Introduction, meaning ... corpus study, Study of various dictionar... relations among lexemes & their se... homonymy, Polysemy, Synonymy, Hyponymy, Semantic Ambiguity. Word sense disambiguation, Lesk's Algorithm, Supervised (Naive Bayes, decision list. Introduction to semi-supervised method (Yarrowsky) and unsupervised (Hyperlex).

→ Semantics involves figuring out the meaning of the linguistic input [construct meaning representations] and process language to produce common sense knowledge about the world [extract data and construct models of the world] This module focuses on —

→ The study of meaning of the individual words [Lexical Semantics]

→ The study of how individual words combine to give meaning to a sentence. (or larger units)

Q) Where is semantic Analysis used? —

Ans) Semantic analysis is used in extracting important information from achieving human level accuracy in computers. It is used in tools like machine translations, chatbots, search engines and text analysis.

— Semantic Analysis is a subfield of NLP and machine learning. It tries to clear the context of any sentence/text and extract the emotions inherit in any sentence.

Steps to be carried out in semantic Analysis are: —

1) Segmentation I: Identify the clause boundaries and word boundaries.
2) Classification I: Determine the parts of speech.
3) Segmentation II: Identify constituents.
4) Classification II: Determine the syntactical categories for constituent

Meaning Representation:—

→ Building blocks of a semantic system :—

1) Entities: It represents the individual eg> particular person, location etc.
2) Concepts: This represents the general category of the individual such as person, nation etc.
3) Relations: Here relation between the entities and concepts is represented.
4) Predicates: It represents the verb structures.

Approaches to meaning representations:

1) FOPL [First order predicate logic]
2) semantic nets
3) Frames
4) conceptual dependency (CD)
5) Rule based architecture
6) case grammar.

Need of meaning representations:

1) Linking of linguistic elements to non linguistic elements
2) Representing variety at lexical level.
3) It can be used for reasoning.

→ Lexical semantics:

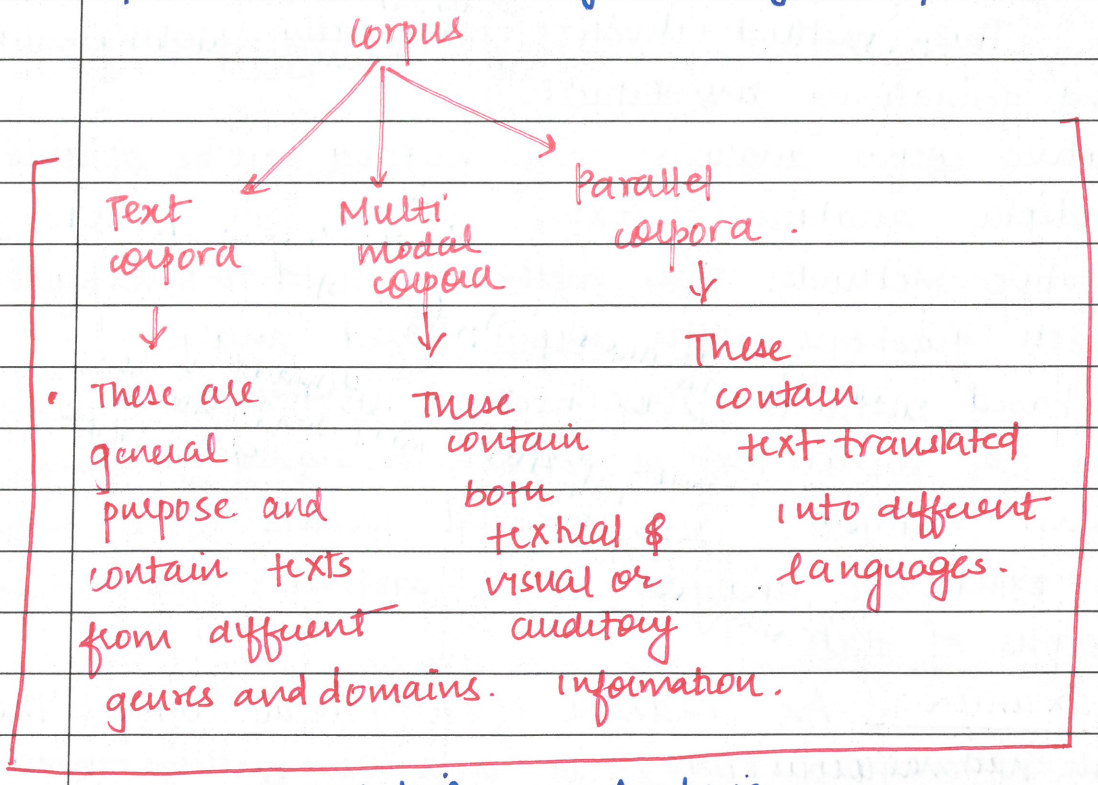
Lexical semantics is a part of semantic analysis. It studies the meaning of individual words that includes words, subwords, affixies (sub units), compound words and phrases.

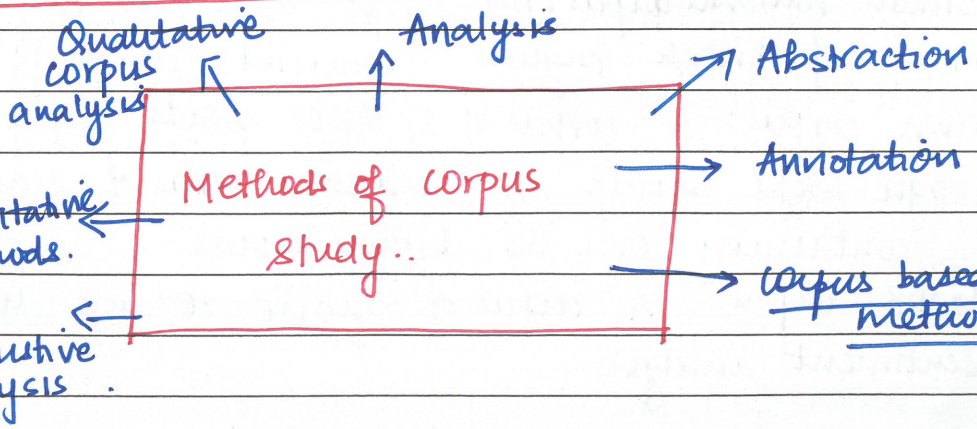
→ All words, subwords etc are collectively called as lexical items. Thus lexical semantics is the relationship between lexical items, meanings of sentences, and syntax of sentences.

Lexical semantics (linguistic study of systematic, meaning related structure of words or lexemes [the minimial unit and lexicon). This involves meanings of component words.

→ compositional semantics: This involves how words combine to form larger meanings.

→ corpus study [corpus study is corpus linguistics and it is rapidly growing it uses statistical analysis of large collections of written or spoken data to investigate linguistic phenomena.]





Methods of corpus study:—

Annotation: 1. Annotation consist of applications of scheme to texts.
2. Annotation may include structural make up, part of speech tagging, parsing and numerous other representations applications.

Abstraction: This method involves the translation (mapping) of tums in this scheme to tums in a theoretically motivated model or dataset.

Analysis: This method involves statistically probing, manipulating and generalizing the dataset.

Qualitative corpus analysis: This method can be of assistance for indepth analysis of texts.

Quantitative methods: This method is used to investigate research questions. with corpus based analysis.

Corpus based methods: This method involves the large scale study of written text, or spoken or signed utterances.

Exhaustive Analysis: This method involves the rigourous and exhaustive analyses of a particular feature in a corpus of texts.

Corpus Example: An example of a general corpus is the British national corpus (BNC). It is a '100' million-word corpus of British English compiled in the 1990's.

- Brown corpus: A corpus of 5,00,000 words compiled in 1961.
- Google books corpus: A massive corpus of scanned books containing over 155 billion words.
- IMDB corpus: A corpus of movie reviews useful for sentiment analysis.

Language dictionary - [Wordnet, Babelnet etc].

- A dictionary is listing of lexemes from a lexicon of one or more specific languages. They are arranged alphabetically.
- They include information on definitions, usage, etymologies translation etc.
- It is a lexicographical reference that shows interrelationships among the data.
- A clear distinction is made between general and specialized dictionaries. Specialized dictionaries includes words in a specialized fields rather than a complete range of words in the language.
- General dictionary are semasiological ie mapping word to definition while specialized dictionary are supposed to be onomasiological.
- They first identify concepts and then establish the terms used to designate them.
- The other dictionaries include are (1) Bilingual (translation) dictionaries, dictionaries of synonyms and rhyming dictionaries.
- The word dictionary is used to refer a general purpose (monolingual) dictionary.

There is also difference between (1) Prescriptive and descriptive dictionaries.

Prescriptive	Descriptive.
Reflects the correct use of language.	Reflects the recorded actual use.

Types of dictionaries:

In a general dictionary, each word may have multiple meanings. Some dictionaries include each separate meaning in the order of separate usage of words while others list definitions.

Specialised dictionary:- Specialised dictionary is also referred to as a Technical dictionary. It focuses on a specific subject field.

- Lexicographers categorise dictionaries into three types:-
- A multi-field dictionary: It covers several subject fields eg Business dictionary.
- A single field dictionary covers one subject (eg law) and
- A sub field dictionary - It covers a more specialised field (eg constitutional law).
- The 23 language Inter active Terminology for Europe is a multi field dictionary. The American national Biography is a single field. The African American national Biography project is a sub field dictionary.
- Another variant is the Glossary, an alphabetical list of defined terms in a specialised field, such as medicine (medical dictionary).

Defining Dictionaries:-

- A defining dictionary, provides a core glossary of the simplest meanings of the simplest concepts.

Historical dictionary:

A historical dictionary is a specific kind of ~~specific~~ descriptive dictionary It describes the development of words and senses over the time, using original source material to support its conclusions.

Dictionaries for natural language processing: These are built to be used for computer programs. These dictionaries are published online, & these are used by computer programs.

- Since most of the dictionaries control machine translations or cross lingual information retrieval (CLIR), the content is usually multilingual and usually of large size.
- To formalize the exchange and merge the dictionaries an ISO standard (Lexical makeup framework) LMF has been defined and used among the industrial and academic community.

→ **Babelnet dictionary:** ① Babelnet is a multilingual lexicalised semantic network. Babelnet was automatically created by linking most popular computational lexicon of English language, wordnet. The integration is done by using an automatic mapping and by filling in lexical gaps in resource poor languages by using statistical machine translation. The result is an encyclopaedic dictionary. This provides concepts and named entities that are lexicalised in many languages and connected with large number of semantic relations.

- Additional lexicalisations and definitions are added by linking to free-license word nets. Similar to wordnet, babelnet group-words in different languages are set into sets of synonyms, called as Babel synsets.
- Babelnet provides short definitions (called as glosses) in many languages are taken from wordnet.
- Babelnet stable realese: Babelnet 5.0/ February 2021.

Opuating system: Virtuoso universal server.

Type: Multilingual encyclopedic dictionary linked data.

website: Babelnet.org.

Statistics of Babelnet:

Babelnet (Version 5.0) covers 500 languages. It contains 26 million synsets and around 1.4 billion word senses.

Each babelnet synsets contains 2 synonyms per language ie word senses per average.

Version 5.0 also associates around 51 million images with Babel synsets and provides lemon RDF encoding of resource available via SPARQL endpoint 2.67 million synsets are assigned domain labels.

Applications

→ Babelnet has been shown to enable multilingual natural language processing applications.

→ The lexicalised knowledge available in Babelnet has been shown to obtain state of art results in:—

1) Semantic relatedness.
2) Multilingual word sense disambiguation.
3) Multilingual word sense disambiguation and entity linking
4) Video games with purpose.

Q. Explain Relations among lexemes and their senses? (V.V.Imp) MU23, MU24.

Ans) One way to approach lexical semantics is to study the relationship among lexemes. Semantics of a lexeme can be understood by analysing the relationship of lexemes with other lexemes. The important elements of semantic analysis; are as follows:—

1) Homonymy.
2) Hyponymy
3) Polysemy.
4) Synonymy
5) Antonymy.

Subject:Natural Language Processing

1 **Homonymy**: The first relationship that we discuss is homonymy. It is the simplest relationship that exists between the lexemes. Homonyms are the words that have the same form but have different, unrelated meanings.

A classical example of homonymy is Bank (river bank or financial institution).

A related idea is that homophones refer to words that are pronounced in the same way but different meaning or spelling of both.

for example: Bat is a homonym as it is used to hit the ball and also is flying mammal.

2 Hyponymy: The hypernym is a word with a general sense. The word automobile is a hypernym for a car and truck. The hyponym is a word with the most specific meaning. It is defined as relationship between the generic term and instances of that generic term. The generic term is called as hypernym and its instances are called as hyponyms.

For example: The word colour is a hypernym and the colour red, green etc are hyponyms.

for example: Parrot is a hyponym of bird.

3 **Polysemy**: Many words have more than one meaning of sense. Unlike homonyms, polysems are words with related meanings This linguistic phenomenon is called as polysemy or lexical ambiguity. Words that have several senses are ambiguous are polysemous. For example word 'chair' can refer to the piece of furniture or the act of presiding over a discussion etc. For example stream (flowing water) and stream (flow of data or video content).

* Difference between: Polysemy and Homonymy.

Homonymy involves words with same spellings or pronunciation but unrelated meanings. The meanings are not connected. However in polysemy words with multiple related meanings share a

a common core or a root meaning.

— Homonymy is often a matter of historical coincidence but polysemy reflects a word semantic development.

— Homonymy can lead to confusion but polysemy allows for a nuanced and context-dependent meanings, enriching language expression and communication.

4 Synonymy: The word synonym defines the relationship between diffuent words that have similar meaning. A simple way to decide whether two words are synonymous is to check whether they both can substitute each other ie check for substitutability Two words are synonym in a context if they can be substituted for each other without changing the meaning of the sentence. Examples are author/writer, 'fate/destiny'.

5 Antonymy: It is a relation between two lexical items possessing symmetry between their semantic components relative to an axis.

The scope of antonymy is as follows:-

1) Application of a property: Example is life/death, certitude/incertitude.

2) Application of scalable property: Example is rich/poor, hot/cold.

3) Application of usage: Example is father/son, moon/sun.

Module 4: Semantic Analysis

– Write short note on Wordnet?

Ans Wordnet is a large lexical database for English language.
– Wordnet consist of three databases - one for noun, one for verbs, and one for adjectives and adverbs.
– Inspired by psycholinguistic theories it was developed and maintained in a cognitive science lab, princeton university, under the direction of George A. Miller.
– Information is organised into sets of synonymous words called as synsets, each representing one base concept. These synsets are linked to each other by means of lexical and semantic relations.
– Lexical relations occur between word forms and semantic relations between word meanings. These relations include synonymy, hypernymy, homonymy etc.
– Wordnet lists all the senses of the word, each sense belonging to a different synset. Wordnet sense-entries consist of set of synonyms and a gloss.
– A gloss consist of dictionary-style definition and examples demonstrating the use of synset in a sentence.
– Nouns and verbs are organized into hierarchies based on hypernymy/hyponymy whereas adjectives are organized into clusters based on antonym pairs.

Applications of wordnet:—

1) Word sense disambiguation: Wordnet combines features of a number of ~~features~~ other resources commonly in the disambiguation work. It offers sense definition of words, identifies synsets of synonyms, defines a number of semantic relations

2) Automatic Query Expansion: Wordnet semantic relations can be used to expand queries so that the search of document is not confined to the pattern matching of the query terms but also covers synonyms.

3) Document structuring and categorization: The semantic information extracted from wordnet and wordnet conceptual representation of knowledge, have been used for text categorization.

4) Document summarization: Wordnet has been found useful in text summarization. The approach presented by Barzilay and Elhadad utilizes information from wordnet to compute lexical chains.

→ Word Sense Disambiguation (WSD):— (Imp).

The task of word sense disambiguation is to examine word tokens in a context and specify which sense of the word is being used.

Word sense Disambiguation (WSD): It is a challenge in natural language processing it is the process of figuring out the sense (meaning) of a word, which is activated by the use of the word in the current context.

WSD is a natural classification problem: given a word and its possible senses as defined by dictionary, classify the occurence of the word in a context into one or more of its senses classes.

The features of the context such as neighbouring words provide the evidence for classification. for example:

Little John was looking for his toy box. Finally he found it, The box was in the pen, John was very happy.

Wordnet lists 5 senses for the pen

pen – a writing instrument.

pen – a enclosure for livestock

playpen, pen – a portable enclosure in which baby can play.

pen – female swan

penitentiary, pen – a correctional institution for those convicted of major crimes.

Methods:

The approaches to WSD are as follows:–

1) Dictionary and knowledge based methods: These rely on dictionaries, thesauri

2) Supervised methods: These make use of sense-annotated corpora.

3) Semisupervised methods: These make use of secondary source of knowledge such as small annotated corpus as seed data in a bootstrapping process.

4) Unsupervised methods. These forsake external data in favour of working directly with unannotated raw corpora. Word sense discrimn is another name for these methods.

→ Dictionary and knowledge based approaches

The first dictionary based approach is the Lesk method (Lesk 1986)

It is predicated on the idea that the words employed in text have relationship to one another, by identifying the sense of definitions senses with the biggest word overlap in their definitions (two or more) words can be resolved.

– The usage of general word sense relatedness and computing the semantic similarity of each pair of word senses based on the specific lexical knowledge base such as wordnet. Graph based

techniques can also be used.
Selectional preferences (also known as selectional) constraints might be helpful. For instance; one can clarify that the phrase I am cooking bass (ie a musical instrument) does not refer to the activity of cooking.

(2) Supervised Techniques:

- The foundation of supervised approaches is the idea that the context might provide a sufficient justification for word disambiguation (hence would knowledge and reasoning might/are deemed unnecessary)
- Almost all machine learning algorithms now in use, along with related methods such as feature selection, parameter optimization and ensemble learning.

(3) Semi-supervised or minimally-supervised methods:

The bootstrapping method begins with tiny amount of seed data for each word, such as a few reliable decision rules or manually marked training instances (eg play in the context of bass)

- Any supervised approach can be used to train an initial classifier using the seeds. This classifier is then assigned to the corpus untagged section to extract a bigger training set, which contains the classifications with the highest degree of confidence.
- The procedure iterates until the entire corpus is eaten or a predetermined maximum number of iterations is reached.

(3) Unsupervised Techniques:—

- Unsupervised learning is one of the greatest challenge for WSD learners.
- The underlying assumption that similar senses occur in the similar contexts and thus senses are often induced from the text

by clustering word occurences by using some measure of similarity of context. Then the new occurences of the word are often classified into the closest induced cluster/senses.

WSD Evaluation: ① The evaluation of WSD systems requires a test corpus hand annotated with the target or correct senses and assumes that the corpus can be constructed.

② Two main peformance measures are used:—

- Precision: fraction of system assignments made that are correct.
- Recall: The fraction of total word instances correctly assigned by the system.

There are two kinds of test corpora:—

- Lexical sample: The occurence of a small sample of target words need to be disambiguated
- All words: all words in a piece of running text need to be disambiguated.

- Applications of WSD: WSD is applied in almost every application of NLP, following are the applications of WSD:—

1) Machine Translation: ① Machine translation is the most common application of WSD, in machine translation lexical choice of words have diffuent translations for diffuent senses is done by WSD.

2) Text mining and information extraction: In most of the text mining and Information extraction, WSD is necessary for accurate analysis of text. WSD helps in flagging of correct words.

3) Information retrieval: IR is defined as software that deals with organizn/ storage and retrieval and evaluation of information from document. WSD is used to solve ambiguities of the queries.

Lesk Algorithm:

The Lesk algorithm is based on the assumption that the words in a given neighbourhood will tend to share a common topic.

Lesk algorithm is used to compare the dictionary definition of the ambiguous word with the terms contained in the neighbourhood.

The core idea of the Lesk Algorithm is to choose the sense of a word whose definition has the overlap with the context in which the word appears. The context is typically defined as the words surrounding the target word.

Detailed Explanation of Lesk Algorithm:—

Basic steps:

1) **Identify the target word:** Determine the word that needs disambiguation in a given sentence or a context.
2) **Retrieve the possible senses:** Look up all possible senses (definitions) of the target word from a lexical resource such as dictionary or wordnet. Each sense will have a corresponding gloss (definition) and possibly example sentences.
3) **Extract Glosses:** For each possible sense of the target word, extract the glosses and any example sentences that might provide additional context.
4) **Determine the context:** Define the context by identifying the surrounding words in the sentence where the target word appears. The context can be the entire sentence or a window of words around the target word.
5) **Compute the overlap:** For each sense of the target word compute the overlap between the words in the gloss and the words in the context. This can be done by counting the number of shared words.

6) **Select sense:** choose the sense with the highest overlap score as the most appropriate sense for the target word in the context.

Example in detail:—
consider the word 'Bank' in the sentence "He sat on the bank of the river"

(1) Identify the target word:—
Target word: Bank.

(2) Retrieve the possible senses:—
Sense 1: "A financial institution where people deposit and withdraw money"
Sense 2: "The side of the river or a stream"

(3) Extract Glosses:
Gloss 1: "A financial institution where people deposit and withdraw money".
Gloss 2: "The side of a river or stream".

(4) Determine context:
Context words: ["He", "sat", "on", "the", "bank", "of", "the", "river"]

(5) Compute overlap:
Overlap for Sense 1: 0 (No common words between gloss 1 and context).
Overlap for sense 2: 1 (common word: "river").

(6) Select sense:
Sense 2 ("The side of a river or a stream") has the highest overlap score so it is chosen with the correct sense.

Yarowsky Algorithm (Imp)

The Yarowsky algorithm is developed by David Yarowsky it is a well known algorithm for word sense disambiguation (WSD) and it falls under the category of semisupervised learning. It leverages both labelled and unlabelled data to iteratively improve the accuracy of sense disambiguation. This algorithm is based on two key principles: one sense per collocation and one sense per discourse.

Key Principles:

1. One sense pa collocation: ① This principle states that a given word tends to have a consistent sense when it appears in the same collocation (specific lexical or syntactical context)

2. One sense pa discourse ① This principle states that a given word tends to maintain the same sense throughout a particular discourse or document.

Steps of Yarowsky Algorithm:

1. Initialization: ① start with the small set of seed examples where the correct sense of the target word is known. These seeds are ~~known~~ used to train classifier.

2. Bootstrapping: ① Use the initial classifier to label the sense of target word in the large corpus of unlabelled text.

② Identify new instances of the target word where the classifier is confident about the prediction.

③ Add these newly labelled instances to the training set and update the classifier.

3. **Iterative refinement:** Repeat the bootstrapping process iteratively, classify, label confidently predicted instances, retrain the classifier and refine the model. Each iteration uses the updated model to improve the accuracy of sense disambiguation

4. **Convergence:** The process continues until the classifier's performance stabilizes meaning that adding more labelled instances does not significantly change the model.

Example: consider the word plant, which can be either a factory or a living organism.

1. Initialization: seed examples for "plant" as a factory: "The new plant will produce cars".
   seed examples for "plant" as living organism: "She watered the plant every day.

2. Bootstrapping: ① Use these seeds to train an initial classifier
   ② Apply the classifier to a large corpus and identify the sentences where the classifier confidently predicts the sense of "plant".

3. Iterative Refinement: ① Add confident predictions to the training set. ② Retrain the classifier with the expanded training set. ③ Repeat the classification and labeling process.

4. convergence: Continue until the classifier's performance converges.

Advantages and Limitations:—

Advantages:

- Effiient use of a small amount of labelled data to bootstrap learning.
- Exploits large amounts of unlabelled data to improve pefomance.
- Incorporates robust linguistic principles (collocation & discourse consistency).

Limitations:

- Initial seed examples must be accurate and representative.
- Pefomance depends on the quality and representativeness of the seeds.
- May require several iterations to converge.

- Applications:

The Yarowsky algorithm has been widely used in various NLP tasks:

- Word sense disambiguation.
- Name entity recognition.
- Part of speech tagging.
- Machine Translation.

Conclusion: The Yarowsky algorithm is a seminal approach in semi supervised learning for word sense disambiguation. By iteratively bootstrapping from a small set of labelled example and leveraging large amounts of unlabelled data.

Naive Bayes classifier: [Supervised word disambiguation Algorithm]

Naive Bayes is a probabilistic classifier based on Bayes theorem with the "naive" assumption that the features (words) are independent given the class (sense).

Steps:

1. **Training:**

Calculate the prior probabilities: Compute the prior probabilities of each class (sense) from the training data.

$$P(s_i) = \frac{\text{Number of instances of sense } s_i}{\text{Total number of instances.}}$$

Calculate Likelihood: Compute the likelihood of each feature (word) given the class.

$$P(w_j | s_i) = \frac{\text{Count of } w_j \text{ in instances of } s_i + 1}{\text{Total number of words in } s_i + V}$$

where $V \rightarrow$ vocabulary size.

2. **Classification:** For a given context (set of words), compute the posterior probability for each class using Bayes theorem.

$$P(s_j | w_1, w_2 \ldots w_n) \propto P(s_i) \prod_{j=1}^{n} P(w_j | s_j)$$

Choose a class with the highest posterior probability.

Example: Consider the word 'plant' with two senses living organism and factory. Given a sentence "she bought a new plant for her garden," we calculate

→ P(factory| She bought new plant for her garden) $\propto$ P(factory) π_j P(wj|factory)

→ P(living organism | She bought a new plant for her garden) $\propto$ P(living organism) π_j P(wj| living organism)

- Compare the probability of both we find that P(living organism| She bought a new plant for her garden) $\gg$ P(factory| She bought a new plant for her garden)).

Hyperlex algorithm for (unsupervised word sense disambiguation)

The Hyperlex algorithm is an unsupervised word sense disambiguation based on distributional hypothesis, which suggests that words appearing in the similar contexts tend to have similar meanings. Hyperlex was proposed by Peter Vossen in 2002, it clusters word contexts to discover different senses of word without relying on labelled training data.

Hyperlex Algorithm overviews:—

1) Context Extraction: Extracts contexts in which the ambiguous word appears from a large corpus.
2) Feature Representation: Represent each context as a vector of features such as co-occuring words, parts of speech etc
3) Similarity computation: Computes similarities between context vectors using a distance metric like cosine similarity.
4) Clustering: Cluster the context vectors into a groups where each cluster represents a different sense of word.

5) Sense Assignment: Assign each occurrence of ambigious word to sense corresponding to the cluster it belongs to.

Important Questions:—

1] Explain WSD (word sense disambiguation) in detail?.
2] Explain Lesk Algorithm?
3] Write short note on 1) Babelnet.
4] Explain Yarowsky algorithm?
5] Explain homonymy, hyponymy, polysemy, synonymy with examples?. [V.V. imp].
6] Write short note on 1) Wordnet.

—×—×—×—×—×—×—×—

Subject:Natural Language Processing