Unit - III

Semantic passing:-

Introduction:-

* In language understanding is the identification of a meaning representation that is detailed enough to allow reasoning system to make deduction (the process of reaching a decision or answer by thinking about the known facts).
* Semantic passing is the task of converting a natural language utterance to a logical form: a machine understandable representation of its meaning. (Utterance - Spoken word)
* Semantics by its dictionary definition is the study of meaning, and passing is the examination of something in a minute way, that is, identifying and relating the pieces of information being passed. When these two concepts together, can call it as Semantic passing.
* Is the processing to identifies the meaning chunks contained in an information signal in an attempt to transform it into some data structures that can be manipulated by a computer to perform higher level tasks.
* Semantic passing can thus be understood as extracting the precise meaning of an utterance. Applications of Semantic passing include machine translation, question answering, ontology induction, automated reasoning, and code generation.

* An ontology (a branch of metaphysics concerned with the nature and relations of being (things, ideas and reality)) can be created that can capture the various granularities and aspects of meanings that are embodied in such a variety of applications.
* There are two compromise approaches have emerged in the natural language processing Semantic parsing.

In the first approach, a specific, rich meaning representation is created for a limited domain for use by applications

In the second approach, a related set of intermediate meaning representations is created, going from low-level anaysis to a mid level analysis.

* Semantic parsing is a process of segmentation and this parsing have different types are follows:

(1) Shallow Semantic parsing :- identifying entities in an utterance and labelling them with the roles. Shallow Semantic parsing is sometimes known as slot filling or frame Semantic parsing
* slot filling parsing is the largely an encoder-decoder model

(2) Deep Semantic parsing :- Is also known as compositional Semantic parsing, is concerned with producing precise meaning representations of utterances that significant compositionally.

(3) Representation languages :- Early Semantic parsers used highly domain specific meaning representation languages, like prolog, lambda calculus, lambda dependency based compositional Semantics (λ-DCS), SQL, Python and Alexa meaning representations.

* Datasets used for training statistical semantic parsing models are divided into two main classes based on application - those used for question answering via knowledge base queries, and those used for code generation.

(1) Question answering :- Semantic parsing is the Air Travel Information System (ATIS) dataset, which contains questions and commands about upcoming flights as well as corresponding SQL.

(2) Code generation :- It includes two trading card datasets that link the text that appears on cards to code that precisely represents one was constructed linking magic : The gathering card texts to Java snippets, the other by linking Hearthstone card texts to python snippets. (Snippet - Small piece of code extract)

* The task of producing the output of the first type is often called deep semantic parsing, and the task of producing the output of the second type is often called shallow semantic parsing.

* In semantic parsing, it has exactly two types of meaning representations : a domain dependent, deeper representation and a set of relatively general purpose, lowlevel and intermediate representations.

* Is the process of identifying meaning chunks contained in an information signal - to transform computers to perform higher level tasks.

* Semantic parser component includes - two kinds of Semantic Parsers: grammar - based and neural - based
The grammar - based parser consists of modules of lexicon recognizer, grammar compiler, scorer, reranker and trainer.
The neural - based components consists of an input layer, a processing layer, and an output layer.
* Semantic parsers can be learned automatically from sequences paired with their logical form.

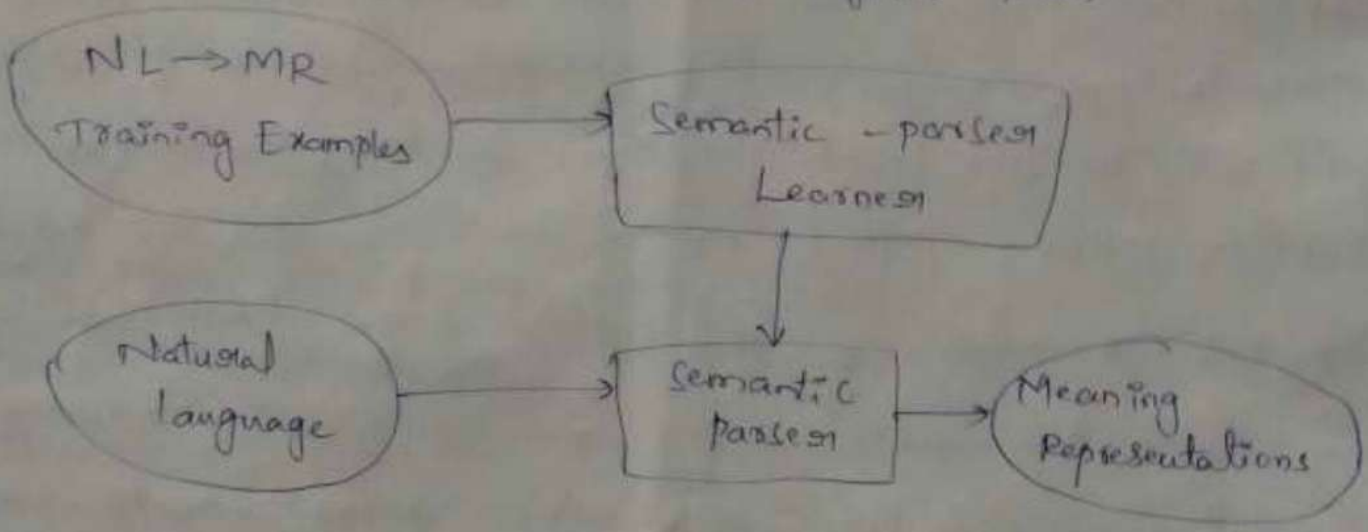


fig: Semantic parsing using Natural language.

* Semantic parsing maps Natural language sentences to completely formal computer language.
* Semantic parsers can be effectively learned from supervised corpora consisting of only sentences paired with their formal representations.

Semantic Interpretation :-

* Semantic Interpretation means adding semantics to a grammar.
* Semantic parsing can be considered as part of a larger process, semantic interpretation, which involves various components that together define a representation of text that can be fed into a computer to allow further computations manipulations and search, which are prerequisite for any language understanding system or application.
* Here follows the theory about structure of semantics.

Semantic theory should be able to:

i) Explain sentence having ambiguous meaning: The bill is large is ambiguous in the sense that is could represent money or the beak of a bird.

ii) Resolve the ambiguities of words in context. The bill is large but need not be paid, the theory should be able to disambiguate the monetary meaning of bill.

iii) Identify meaningless but syntactically well-formed sentence: colorless green ideas sleep furiously.

iv) Identify syntactically or transformationally unrelated paraphr of concept having the same semantic content.

→ "The Structure of Semantic Theory," Katz and Fodor put forward what were the properties a semantic theory should possess.

* Semantic Interpretation tags provide a practical way of handling the information that is presented in the grammar matching with the speech recognizer.
* for example - A grammar for arithmetic expressions, augmented with semantics.

Arithmetic expressions with grammar rules:

Exp(x) → Exp(x_1) operator (op) Exp(x_2) {x = Apply(op, x_1, x_2)}

Exp(x) → (Exp(x))

Exp(x) → Number(x)

Number(x) → Digit (x)

Number(x) → Number (x_1) Digit (x_2) {x = 10 × x_1 + x_2}

Digit (x) → x {0 ≤ x ≤ 9}

Operator (x) → x {x ∈ {+, −, ÷, ×}}

Here, Variable 'x' represents the Semantics of a constituent.

* Each rule is augmented with a variable indicating the Semantic interpretation of the phrase.
* Parse tree with Semantic interpretations for the string "(3 + (4 ÷ 2))"

Semantics of a digits, 3, 4, 2

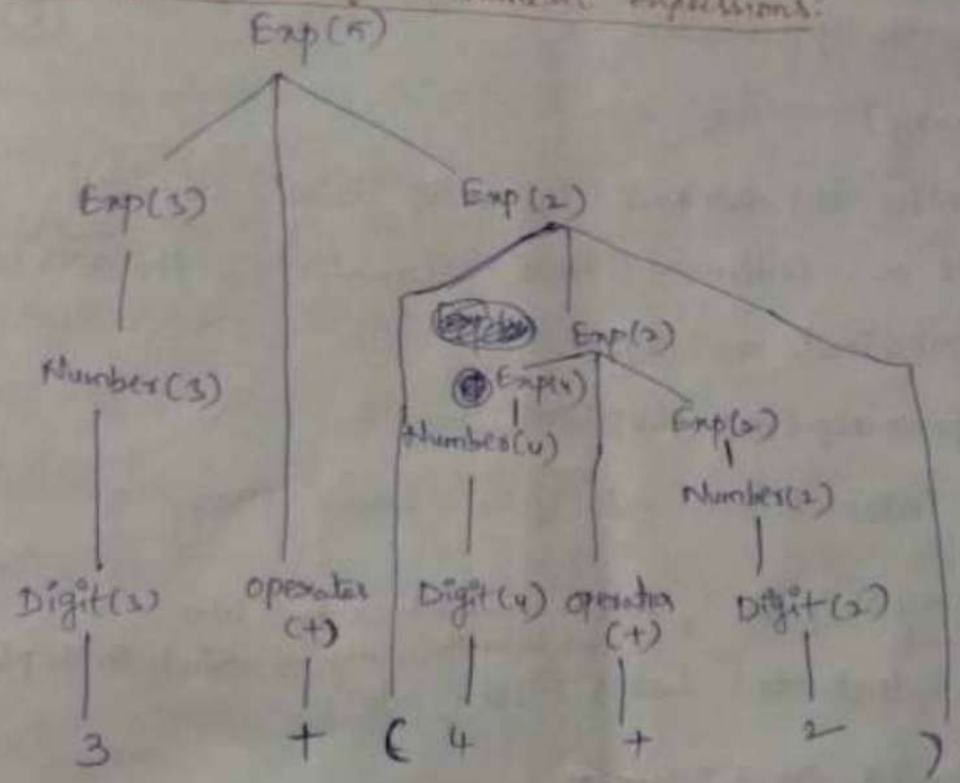
Semantics of an expressions 3 + (4 ÷ 2)

Here operators are +, ÷

applying to the Semantics of the phrases are

phrase 3, phrase 4, phrase 2

Top down approach using arithmetic expressions:



Semantics of English :-

Ex:- John Loves mary

The logical sentence Loves (John, mary)

The NP "John" should have its semantic interpretation - the logical term John.

The complicated part is the VP "Loves mary"

The semantic interpretation of this phrase is neither a logical term nor a complete logical sentence.

"Loves mary" is a predicate, when combined with a term that represents a person yields (John) a complete logical sentence.

↓
Hidden word

"Loves mary" as the predicate

λx Loves (x, mary)

NP with Semantics obj followed by a VP with Semantics predicate yields a Sentence whose Semantic is the result of applying predicate to object.

S (pred (obj)) → NP (obj) VP (pred).

The Semantic Interpretation of "John Loves mary" is

$(\lambda x$ Loves (x, mary)) (John), → x is substituted by John can substitutes any of object in the place x

which is equivalent to Loves (John, Mary).

VPs are represented as predicates

The verb "loves" is represented as $\lambda y \lambda x$ Loves (x, y),

When given the argument mary, returns the predicate

λx Loves (x, mary).

↑

In the place we can substitute another variable y

A grammar that can derive a semantic interpretation for "John Loves mary"

Name (John) → John

Name (Mary) → Mary

Verb ($\lambda y \lambda x$ Loves (x, y)) → Loves.

* λ (lambda) is a formal system in mathematical logic for expressing computation based on function abstraction and application using variable binding and substitution.

* In the following subsections for achieving a semantic representation

① Structural Ambiguity :- This is a sentence level phenomenon and essentially means transforming a sentence into its underlying syntactic representation.

* Because syntax and semantics have such a strong interaction, most theories of semantic interpretation refer to the underlying syntactic representation.

* Syntax has become the first stage of processing followed by various other stages in the process of semantic interpretation.

* Structural ambiguity arises when a sentence has more than one meaning due to the way words are arranged in that sentence.

- for example, the sentence

"Sara caught the butterfly by the tree"

is structurally ambiguous because it has two meanings:

i) Sara caught the butterfly while she was standing by the tree

ii) Sara caught that butterfly which was fluttering near the tree.

* One type of ambiguity is semantic ambiguity, also known as lexical ambiguity. Semantic ambiguity arises when a sentence has more than one meaning due to certain word in the sentence which itself has more than one meaning.

(2) Word Sense :-

* In any language, it is almost certainly the case that the same word type, or word lemma, is used in different contexts and with different morphological variants to represent different entities or concepts in the world.

* A word sense corresponds either neatly to a seme (the smallest possible unit of meaning) or a sememe (larger unit of meaning).

* "Sense" is a verb meaning "feel" [illegible], These are different sense words, ie, taste, sight, touch, smell and hearing.

* Word sense disambiguation, in natural language processing, 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, we use the word nail to represent a part of the human anatomy and also to represent the generally metalic object used to secure other objects.

* Humans are adept at identifying, through context, with sense of the word is intended (planned).

* For example, The presence of words such as hammer and hardware store in sentences 1 and 2,

and of clipped and manicure in sentences 3 and 4, enable humans to easily disambiguate the sense in which nail is used:

Ex: ① He nailed the loose arm of the chair with a hammer.
② He went to the
③ He bought a box of nails from the hardware store
④ He went to the beauty salon to get his nails clipped
⑤ He went to get a manicure. His nails had grown very long.

③ Entity and Event Resolution :-

* Any discourse inevitably consists of a set of entities participating in a series of explicit or implicit events over a period of time.

* The next important Component of Semantic interpretation is the identification of various entities that are sprinkled across the discourse using the same or different phrases.

* Reconciling what type of entity or event is being considered along with disambiguating various ways in which the same entity is referred to over a discourse, is critical to creating a semantic representation.

* Entity resolution is the process that resolves entities and detects relationships. This process incoming identity records in 3 phases: Recognize, resolve and relate.

* Event resolution means the occurrence of one or more of the following events.

* These are two different tasks have become popular over the years: named entity recognition and coreference resolution. These two tasks are implemented under information extraction.

* Named entity recognition also known as named entity identification, entity chunking, and entity extraction is the subtask of information extraction.

NER helps to easily identify the key elements in a text, like names of people, places, brands.

* Coreference resolution is the task of finding all expressions that refer to the same entity in a text.

Ex:- When he arrived home, John went to sleep.

Here, he is referred to John, and "he" came before "John" in sentence

* coreference resolution is amazing to find the complex relationship between the 'signifier' and the 'signified' ie, which expression or phrase refers to a particular entity in a text.

(4) predicate Argument Structure :- Once we have the word Senses, entities, and events identified, another level of Semantic structure comes into play: identifying the participants of the entities in these events.

* Resolving the argument structure of the predicates in a Sentence is where we identify which entities play what part in which event.

* An argument structure typically indicates the number of arguments a lexical item takes, their syntactic analysis, and their semantic relation.

* A word which functions are the verb does here, we call a predicate and words which function as the nouns do are called arguments. Here are some other predicates and arguments.

Selena	slept
argument	predicate
Tom	is tall
argument	predicate

→ Generally, this process can be defined as the identification of who did what to whom, when, where, why and how.

Ex:- Bell Atlantic Corp. said it will acquire one of Control Data Corp's Computer maintenance businesses.

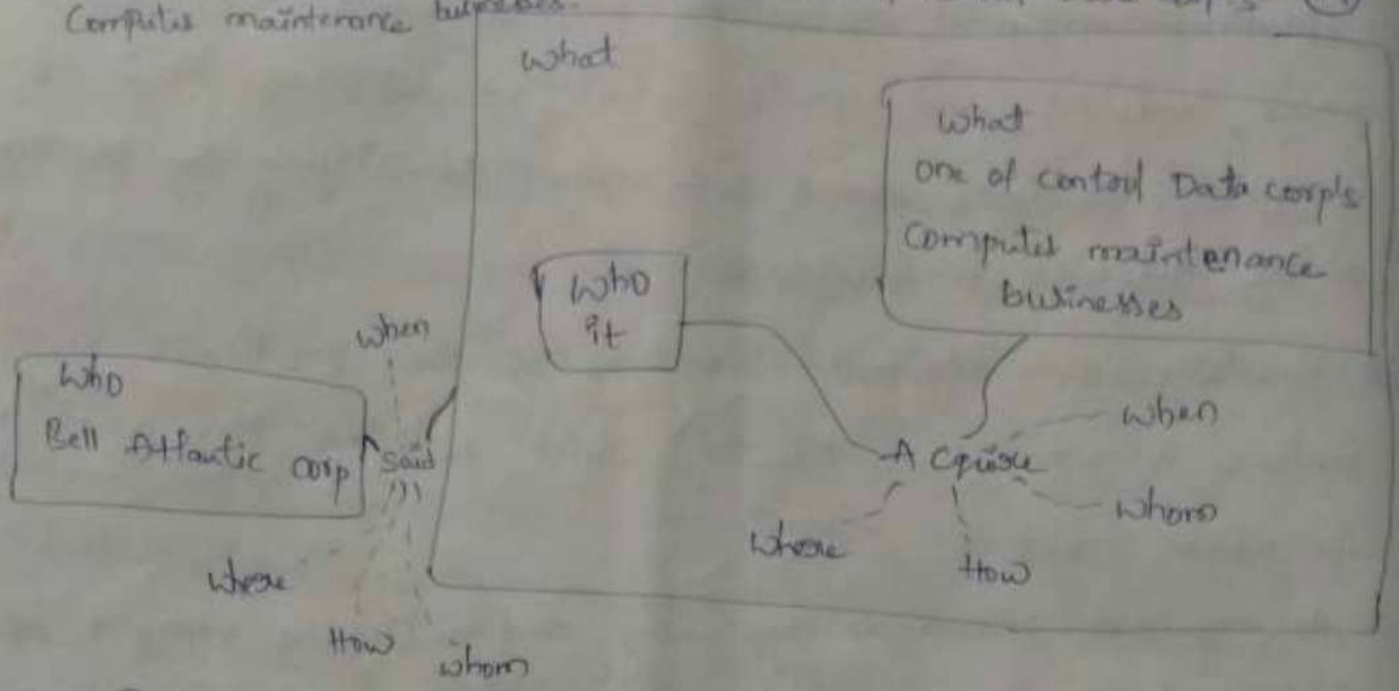


Fig:- A representation of who did what to whom, when, where, why and how.

(5) Meaning Representation:- The final process of the Semantic interpretation is to build a semantic representation or meaning representation that can be manipulated by algorithms to various application ends. This process is sometimes called the deep representation.

* It shows how to put together the building blocks of Semantic systems.

* Meaning representation is the act or action of representing. It aims to encode the semantic information into a real-valued representation, which will be utilized in further textual classification or matching tasks.

* The following two examples show sample sentences and their meaning representations for the Robocup and Geoquery domains.

Ex: If our player 2 has the ball, then position our player 5 in the midfield.

((bowner (player our 2)) (do (player our 5) (pos (midfield))))

Which rivers is the longest?

answer (x1, longest (x1, river (x1)))

System paradigms:-

* The main processing paradigms for NLP are search and classification. The search methods all derive from techniques for searching graphs that are common across computing, such as breadth-first and depth-first search.
* A linguistic paradigm is any kind of rational tabulation of linguistic forms, such as phrases, words, or phonemes, intended to illustrate contrasts and systematic variation.
* The system paradigm represents a dramatic theoretical shift from empirical laws and human rules approach for understanding communication.
* Paradigms are important because they define how we perceive reality and how we behave within it.

* It is important to get a perspective on the various primary dimensions on which the problem of semantic interpretation has been tackled

* The approaches generally fall into the following three categories:

① System architectures:- As the conceptual model that defines the structure, behavior, and more views of a system. There are different kinds of methods to find the solution to the ambiguity.

a) Knowledge based :- These systems use a predefined set of rules or a knowledge base to obtain a solution to a new problem.

b) Unsupervised :- These systems tend to require minimal human intervention to be functional by using existing resources that can be bootstrapped for a particular application or problem domain

c) Supervised :- These systems involve the manual annotation of some phenomena that appear in a sufficient quantity of data so that machine learning algorithms can be applied.

d) Semi supervised :- manual annotation is usually very expensive and does not yield enough data to completely capture a phenomenon

* In such instances, researches can automatically expand the dataset on which their models are trained either by employing machine generated output directly or by bootstrapping off an existing model by having humans correct its output.

(2) Scope:- extent or range of view, outlook, application, operation, effectiveness. There are different methods are follows:

a) Domain dependent:- These systems are specific to certain domains, such as airtravel reservations and simulated football coaching.

b) Domain independent:- These systems are general enough that the techniques can be applicable to multiple domains without little or no change.

(3) Coverage:- The extent to which something deals with something else. There are different methods are follows.

a) Shallow:- These systems tend to produce an intermediate representation that can then be converted to one that a machine can take it action on.

b) Deep:- These systems usually create a terminal representation that is directly consumed by a machine or application.

* The three major research paradigms associated with out disciplines are positivism, postpositivism and interpretivism.

Positivism is based on the assumptions that can be understood, identified.

Postpositivism is based on the observations that impacts on the conclusions.

* Interpretivism is based on the assumption that reality is subjective, multiple and constructed
* A paradigm is a way of organizing and conducting Sensory Information. It is helps to organize information and understan the world.

Word Sense :—

* Word Sense disambiguation is an important method of NLP by which the meaning of a word is determined, which is used in a particular content.
* In a compositional approach to semantics, where the meaning of the whole is composed on the meaning of parts, the small parts under consideration in textual discourse : either tokens as they appear in the text or their lemmatized forms.
* Attempts to solve this problem range from rules based and knowledge based to completely unsupervised, supervised and semi supervised learning methods.
* Very early systems were predominantly rule based or knowled based and used dictionary definitions of sense of words.
* Unsupervised word sense induction or disambiguation techniqu try to induce the senses of word as it appears in various corpora.

* most relevant Supervised approaches to word sense disambiguation, usually application independent - level of granularity. approaches are generating a ranking, or distribution of membership sense.
* Word sense ambiguities can be of 3 principle types: i) homonymy ii) polysemy iii) categorical ambiguity
* Homonymy:- defined as the words having same spelling or same form but having different and unrelated meaning
* Polysemy:- Is a Greek word, which means many signs, It has the same spelling but different and related meaning.
* Both polysemy and homonymy words have the same syntax or spelling. The main difference blw them is that in polysemy, the meanings of the words are related but in homonymy, the meanings of the words are not related.
* Categorical ambiguity:- The ambiguity is used to grammatical Category of noun, Verb.
* Categorical ambiguity can be resolved with syntactic information (part of speech), word senses have been annotated for each Part of Speech separately, the sense annotation has been done Per lemma.

* In the English language, more than 38% of words are polysemous, a term referring to a single word which may have multiple definitions or "senses"
* The word "set" for instance - which can be used as a noun, verb, and adjective, most polysemous words recorded in the English dictionary
* The human brain's neural networks perform semantic processing, storage, and retrieval in a way which makes us naturally gifted at word-sense disambiguation (WSD).
* Words with multiple senses pose a significant problem in NLP, a branch of AI which studies the ability of computers to interpret and "understand" the human language
* The most common problem is the difference between various dictionaries or text corpus. Different dictionaries have different meanings for words, which makes the sense of the words to be perceived different.
* for example word sense using in English language.

Word is "bark"

i) The word refers to the outer covering of the tree

ii) The other refers to the sound made by a dog

Systems :-

* Researchers have explored various system architectures to address the sense disambiguation problem
* It can classify these systems into four main categories:
  i) rules based or knowledge
  ii) Supervised
  iii) Unsupervised
  iv) Semi supervised
* The first generation of word sense disambiguation systems was primarily based on the dictionary sense definitions.
* Much of this information is historical and cannot readily be translated and made available for building systems today. But some of techniques and algorithms are still available.
* The simplest and oldest dictionary based sense disambiguation ~~systems was primarily based on dictionary~~ algorithm was introduced by lesk.
* The core of the algorithm is that the dictionary sense whose terms most closely overlap with the terms in the content.
* WSD having overlap Based Approaches
  * Require a machine Readable Dictionary (MRD).
  * Find the overlap between the features of different senses of an ambiguous word (sense bag) and the features of the words in its content (content bag).

* Sense bag contains multiple semantic words, where as content bag contains sentence complete words.

* Lesk's Algorithm for Simple WSD:

Sense Bag :- Contains the words in the definition of a candidate sense of the ambiguous word.

Content bag :- Contains the words in the definition of each sense of each content word.

* Ex:- on burning coal we get ash.

⇒ problem is the finding sense bag of "ash" word and also construct the content bag.

Steps :-

function SIMPLIFIED LESK (word, sentence)
returns best sense of word
best_sense ← most frequent sense for word
max_overlap ← 0
Content ← set of words in sentence
foreach sense in senses of word do
Signature ← set of words in the gloss and examples of sense
Overlap ← max_overlap then
max_overlap ← overlap
best_sense ← sense
end return (best_sense)

* The COMPUTER OVERLAP 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.
* Basic lesk algorithm implementation involves the following steps.
  i) Count the number of words in the neighborhood of the word and in the dictionary definition of that sense for each sense of the word being disambiguated.
  ii) The sense to be picked is the one with the greatest number of items in this count.
* Ex:- On burning coal we get ash
  So here is taking the definitions from wordnet.

Ash

Sense 1
Trees of the olive family with pinnate leaves, thin furrowed bark and gray

Sense 2
The solid residue left when combustible material is thoroughly burned or oxidized

Sense 3
To convert into ash

coal

Sense 1
A piece of glowing carbon or burnt wood

Sense 2
charcoal.

Sense 3
A black solid combustible substance formed by the partial decomposition of vegetable matter without free access to air, used as a fuel for

* In simple terms, lesk's algorithm counts the number of overlaps between all dictionary terms of a word, surrounding it, known as a "context window".
* The word with the highest number of overlaps, without including stop words (such as "the", "a", "and") it to be the word's "sense".

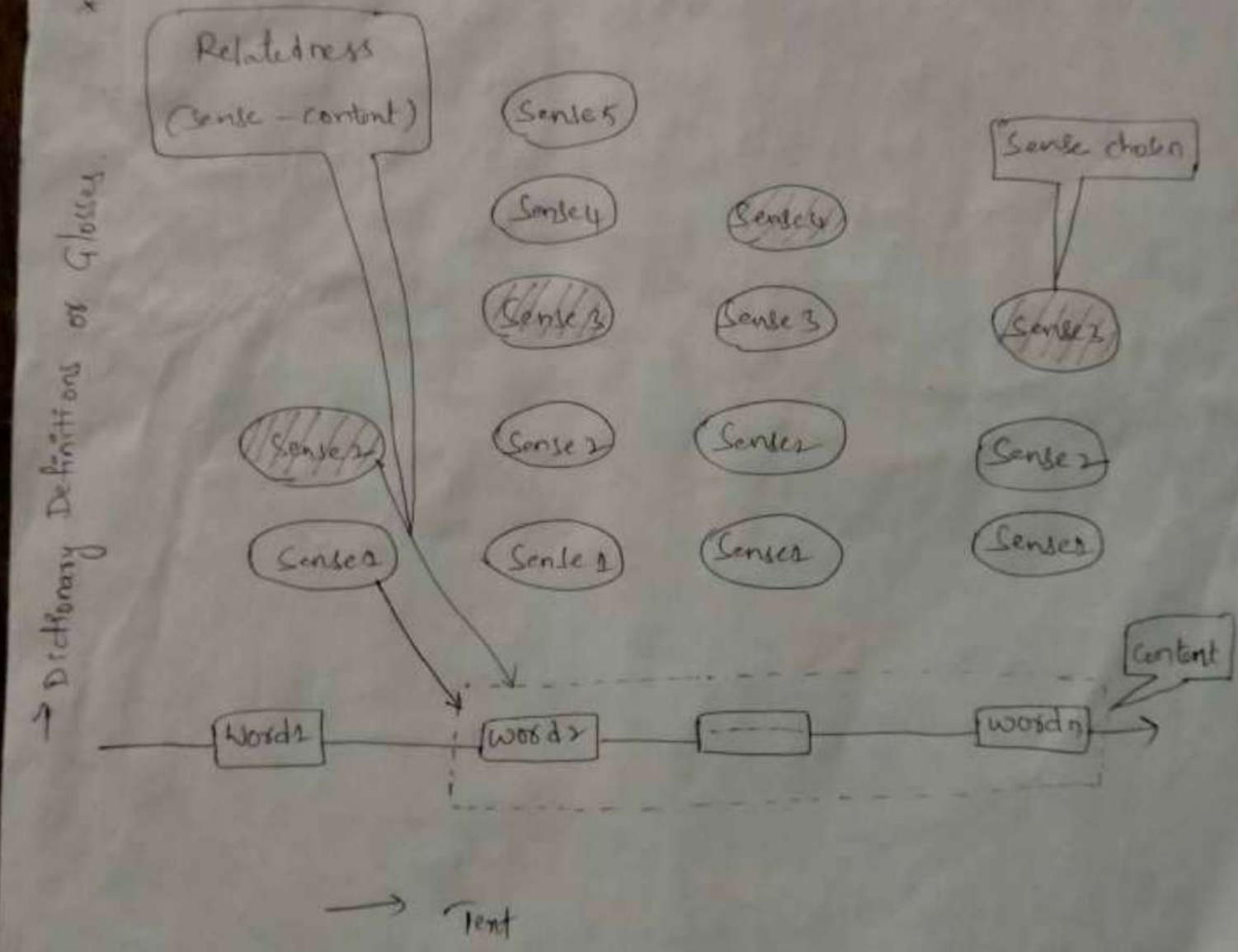


fig: Lesk Algorithm with word sense Disambiguation.

Software :-

* Nlp Software helps computers understand, interpret, and manipulate human language, Ex :- Speech and text through the use of artificial intelligence.
* Amazon Comprehend is an Nlp Service, integrated with the Amazon web Services infrastructure. use this API for Nlp tasks such as sentiment analysis, topic modeling, entity recognition.
* There are different Natural processing Tools through Saas (Software as Service) tools or open-source libraries.
* Saas tools are ready to use and powerful cloud based solutions that can be implemented with low or no code.
* Saas tools and Nlp libraries are shown below.

1) monkey learn :- Is a user friendly, Nlp powerd platform that helps to gain valuable insights from your text data.
To perform text analysis tasks such as sentiment analysis, topic classification or keyword extraction.

2) Aylien :- Is a Saas API that uses deep learning and Nlp to analyse large volumes of text based data, such as academic publications, real time content from news outlets and social media data.
It can use Nlp tasks like text summarization, article extraction, entity extraction and sentiment analysis.

3) IBM Watson :- Is a suite of AI services stored in the IBM cloud one of its key feature is NLU, which allows us to identify and extract keywords, categories, emotions, entities.

It adapted to many different activities to healthcare finance, trove of documents (a store of valuable)

4) Google cloud :- The Google cloud natural language API provides several pretrained models for sentiment analysis, content classification & entity extraction. It also offers AutoML Natural language. it allows to build customized ML models.

5) Amazon comprehend :- Is an NLP service, integrated with the Amazon web services infrastructure. This NLP tasks such as sentiment analysis, topic modeling, entity recognition.
It allows to perform advanced analysis of medical data using machine learning.

6) NLTK :- The natural language Toolkit (NLTK) with python is one of the leading tools in NLP model building. Focused on research and education in the NLP field, NLTK is bolstered by an active community, as well as a range of sample datasets, and resources include a python handbook.
NLTK provides plenty of components for NLP tasks, like tokenisation, tagging, stemming, parsing and classification.

7) Stanford core NLP :- It is a popular library built and maintained by the NLP community at Stanford university. It's written in Java. The Core NLP toolkit allows to perform a variety of NLP tasks, such as part-of-speech, tagging, tokenization, and named entity recognition.

main advantages are scalability and optimization for speed.

8) Text Blob :- It is a Python library that works as an extension of NLTK, allowing to perform much more intuitive and user-friendly interface.

It is the best choice for beginers, who want to tackle NLP tasks like Sentiment analysis, text classification, POS Tagging.

9) Spacy :- one of the newest open-source natural language processing with python libraries on our list is spacy. It is lightning-fast, easy to use, well documented, and designed to support large volumes of data, boasts a series of pretrained NLP models that make your job even easier.

Spacy keeps its menu short and serves up the best available option for each task successfully.

It is the best tool kit to prepare text for deep learning, and excels at extraction tasks. It is only available in English

10) Gensim:- Is a highly specialized python library that largely deals with topic modeling tasks using algorithms like Latent Dirichlet Allocation (LDA) [LDA is a type of topic modeling algorithm, it represents of a fixed number of topics, and given this number of topics learn the topic distribution that each document in a collection of documents has] LDA mainly used in classification problems where you have a categorical output variable.

It is also excellent at recognizing text similarities, indexing texts, and navigating different documents.