Unit 2 Word-netTheory

Semantic Roles, Word Sense Disambiguation (WSD): Word-Net, Word-net Application in Query Expansion, Wiktionary, semantic relatedness, Measures of Word-Net Similarity, Similarity Measures. Resnick's work on Word-Net Similarity, Indian Language Word-nets and Multilingual Dictionaries, Multi- linguality, Metaphors, Co references

---

2.1 Semantic Roles

Semantic roles are the underlying relationships that exist between the elements of a sentence and their real-world referents. In other words, semantic roles describe the roles that different parts of a sentence play in relation to the actions or events that they describe.

Understanding semantic roles is important in natural language processing and computational linguistics, as it allows machines to better understand the meaning of a sentence and the relationships between its different elements.

Patient (also known as affected, undergoer)
Longacre 1983 defines a patient as the entity undergoing a change of state or location, or which is possessed, acquired or exchanged.
Larson 1984 defines the affected role as the thing that is affected by an event person or thing that undergoes a process, or person who experiences an event.
Examples
- The entity predicated with a state or location:
  - The door is open.
  - John is at home.
- The entity undergoing a change of state or location:

  - He opened the door.
  - The door swung open.
  - He threw the ball across the yard.
  - The ball rolled off the table.

Beneficiary
A beneficiary is the semantic role of a referent which is advantaged or disadvantaged by an event.
Example:
John helped Susan to buy her first car

Causer
Causer is the semantic role of the referent which instigates an event rather than actually doing it.
Example:
The rain destroyed the crops.

Experiencer
Experiencer is the semantic role of an entity (or referent) which receives, accepts, experiences, or undergoes the effect of an action.
Normally an experiencer is an entity that receives a sensory impression, or in some other way is the locus of some event or activity that involves neither volition nor a change of state.
Examples
- Lucretia saw the bicycle.
- It was Bill who smelled the bacon first.
- The explosion was heard by everyone.

Goal
Goal is the semantic role of the place to which something moves, or thing toward which an action is directed.
Examples

- John swam to the raft.
- He studied for the test.

Instrument
Instrument is the semantic role of an inanimate thing that an agent uses to implement an event. It is the stimulus or immediate physical cause of an event. Instrument words are usually nouns occurring in the noun phrase of a clause:
- Someone cut the bread with a knife.

Manner
is a semantic role that notes how the action, experience, or process of an event is carried out.
Example
- The girl walked to school slowly.

Measure
is a semantic role which notes the quantification of an event.
Example The new coat costs $70.

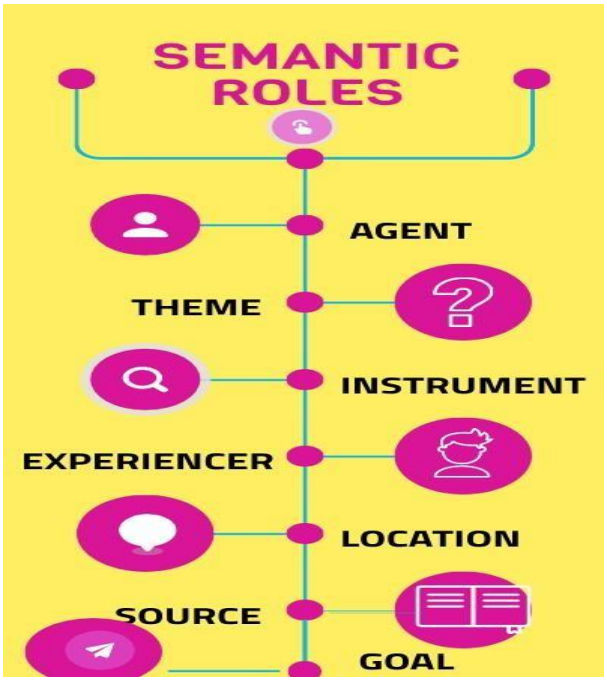
Time
is the semantic role of the temporal placement of an event.
Example
- The whistle will sound at noon.

Source
is the semantic role of the following referents:
The place of origin (with verbs of motion, locomotion, and propulsion)
The entity from which a physical sensation emanates (with verbs of sensation, attention, and speech)
The original owner in a transfer (with verbs of acquisition, transfer, and grab)
Examples
As the place of origin:
- John fell off the chair.(with a motion verb)
- The baby crawled from the kitchen to the door. (with a locomotion verb)
- John picked up the knife from the box. (with a propulsion verb)

As the entity from which a physical sensation emanates:
- John smelled the odor of onions. (with a sensation verb)
- The people watched the performance of
- he dancers. (with an attention verb
- The mother told her child a story. (with a speech verb) With speech verbs, the source is coreferential with the agent
As the original owner in a transfer:
- John obtained an application form from
- the office. (with an acquisition verb)
- John bought the book from Tom. (with a transfer verb)
- John grabbed the book from Tom. (with a grab verb)

********************************Semantic Roles*********************************
Agent :The ‘doer’ or instigator of the action denoted by the predicate.

Patient : The 'undergoer' of the action or event denoted by the predicate.
Theme : The entity that is moved by the action or event denoted by the predicate.
Experiencer : The living entity that experiences the action or event denoted by the predicate.
Goal : The location or entity in the direction of which something moves.
Benefactive : The entity that benefits from the action or event denoted by the predicate.
Source : The location or entity from which something moves Instrument: The medium by which the action or event denoted by the predicate is carried out.
Locative : The specification of the place where the action or event denoted by the predicate in situated.



2.2 Word Sense Disambiguation (WSD)

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. Lexical ambiguity, syntactic or semantic, is one of the very first problem that any NLP system faces. Part-of-speech (POS) taggers with high level of accuracy can solve Word's syntactic ambiguity. On the other hand, the problem of resolving semantic ambiguity is called WSD (word sense disambiguation). Resolving semantic ambiguity is harder than resolving syntactic ambiguity.

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.

2.2.1 Evaluation of WSD

The evaluation of WSD requires the following two inputs −

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 &minsu;

- **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.

2.2.2 Approaches and Methods to Word Sense Disambiguation (WSD)

Approaches and methods to WSD are classified according to the source of knowledge used in word disambiguation.

Let us now see the four conventional methods to WSD −

1. **Dictionary-based or Knowledge-based Methods**

As the name suggests, for disambiguation, these methods primarily rely on dictionaries, treasures and lexical knowledge base. They do not use corpora evidences for disambiguation. The Lesk method is the seminal dictionary-based method introduced by Michael Lesk in 1986. The Lesk definition, on which the Lesk algorithm is based is **"measure overlap between sense definitions for all words in context"**. However, in 2000, Kilgarriff and Rosensweig gave the simplified Lesk definition as **"measure overlap between sense definitions of word and current context"**, which further means identify the correct sense for one word at a time. Here the current context is the set of words in surrounding sentence or paragraph.

2. **Supervised Methods**

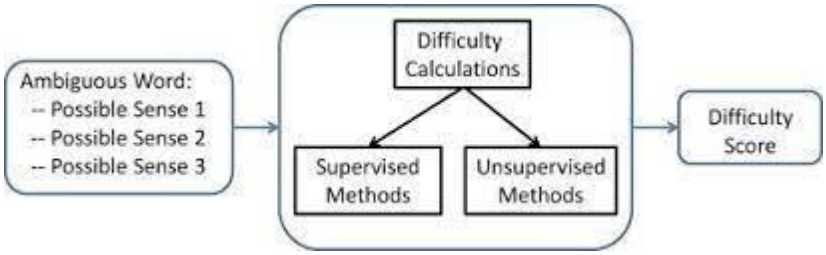
For disambiguation, machine learning methods make use of sense-annotated corpora to train. These methods assume that the context can provide enough evidence on its own to disambiguate the sense. In these methods, the words knowledge and reasoning are deemed unnecessary. The context is represented as a set of "features" of the words. It includes the information about the surrounding words also. Support vector machine and memory-based learning are the most successful supervised learning approaches to WSD. These methods rely on substantial amount of manually sense-tagged corpora, which is very expensive to create.

3. **Semi-supervised Methods**

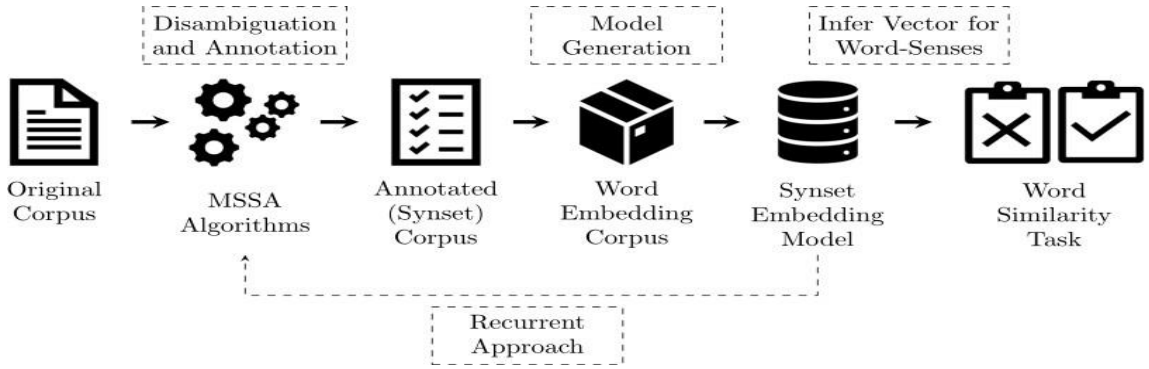
Due to the lack of training corpus, most of the word sense disambiguation algorithms use semi-supervised learning methods. It is because semi-supervised methods use both labelled as well as unlabeled data. These methods require very small amount of annotated text and large amount of plain unannotated text. The technique that is used by semisupervised methods is bootstrapping from seed data.

4. **Unsupervised Methods**

These methods assume that similar senses occur in similar context. That is why the senses can be induced from text by clustering word occurrences by using some measure of similarity of the context. This task is called word sense induction or discrimination. Unsupervised methods have great potential to overcome the knowledge acquisition bottleneck due to non-dependency on manual efforts.



Word Sense Disambiguation



Workflow for word sense disambiguation

For example, consider the sentence "I went to the bank to deposit my check." The word "bank" can refer to a financial institution, or it can refer to the edge of a river or other body ofwater. In this context, the correct sense of "bank" is the financial institution.

WSD algorithms use various techniques to identify the correct sense of a word in a given context. These techniques can include rule-based methods, statistical models, and machine learning algorithms. Some of the features that these algorithms may consider when disambiguating word senses include the surrounding words, the part of speech of the word in question, and the broader context of the sentence.

WSD is an important task in natural language processing, as it is necessary for many applications such as machine translation, text classification, and information retrieval. Accurately identifying the correct sense of a word in a given context can help improve the accuracy of these applications and make them more useful for end-users.

2.2.3 Applications of Word Sense Disambiguation (WSD)

Word sense disambiguation (WSD) is applied in almost every application of language technology.

Let us now see the scope of WSD −

1. **Machine Translation**

Machine translation or MT is the most obvious application of WSD. In MT, Lexical choice for the words that have distinct translations for different senses, is done by WSD. The senses in MT are represented as words in the target language. Most of the machine translation systems do not use explicit WSD module.

2. Information Retrieval (IR)

Information retrieval (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. The system basically assists users in finding the information they required but it does not explicitly return the answers of the questions. WSD is used to resolve the ambiguities of the queries provided to IR system. As like MT, current IR systems do not explicitly use WSD module and they rely on the concept that user would type enough context in the query to only retrieve relevant documents.

3. Text Mining and Information Extraction (IE)

In most of the applications, WSD is necessary to do accurate analysis of text. For example, WSD helps intelligent gathering system to do flagging of the correct words. For example, medical intelligent system might need flagging of "illegal drugs" rather than "medical drugs"

4. Lexicography

WSD and lexicography can work together in loop because modern lexicography is corpusbased. With lexicography, WSD provides rough empirical sense groupings as well as statistically significant contextual indicators of sense.

2.2.4 Difficulties in Word Sense Disambiguation (WSD)

Followings are some difficulties faced by word sense disambiguation (WSD) −

1. Differences between dictionaries

The major problem of WSD is to decide the sense of the word because different senses can be very closely related. Even different dictionaries and thesauruses can provide different divisions of words into senses.

2. Different algorithms for different applications

Another problem of WSD is that completely different algorithm might be needed for different applications. For example, in machine translation, it takes the form of target word selection; and in information retrieval, a sense inventory is not required.

3. 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.

4. Word-sense discreteness

Another difficulty in WSD is that words cannot be easily divided into discrete submeanings.

2.3 Word-Net

In the field of natural language processing, there are a variety of tasks such as automatic text classification, sentiment analysis, text summarization, etc. These tasks are partially based on the pattern of the sentence and the meaning of the words in a different context. The two different words may be similar with an amount of amplitude. For example, the words 'jog' and 'run', both of them are partially different and also partially similar to each other. To perform specific NLP-based tasks, it is required to understand the intuition of words in different positions and hold the similarity between the words as well. Here WordNET helps in solving the linguistic problems of the NLP models. WordNET is a lexical database of semantic relations between words in more than 200 languages. Which contains adjectives, adverbs, nouns, and verbs grouped differently into a set of

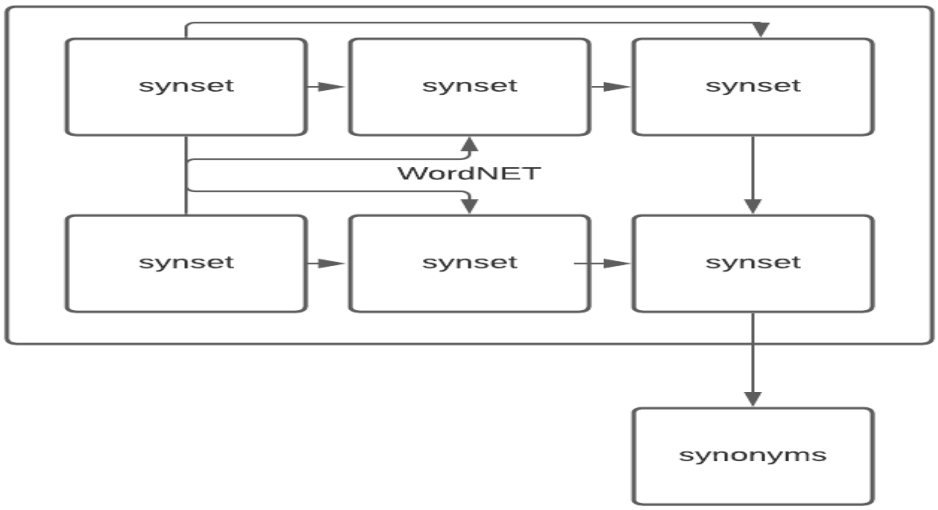
cognitive synonyms, where each word in the database is expressing its distinct concept. The cognitive synonyms which are called synsets are presented in the database with lexical and semantic relations. WordNET is publicly available for download and also we can test its network of related words and concepts.

The Distinction Between WordNET and Thesaurus

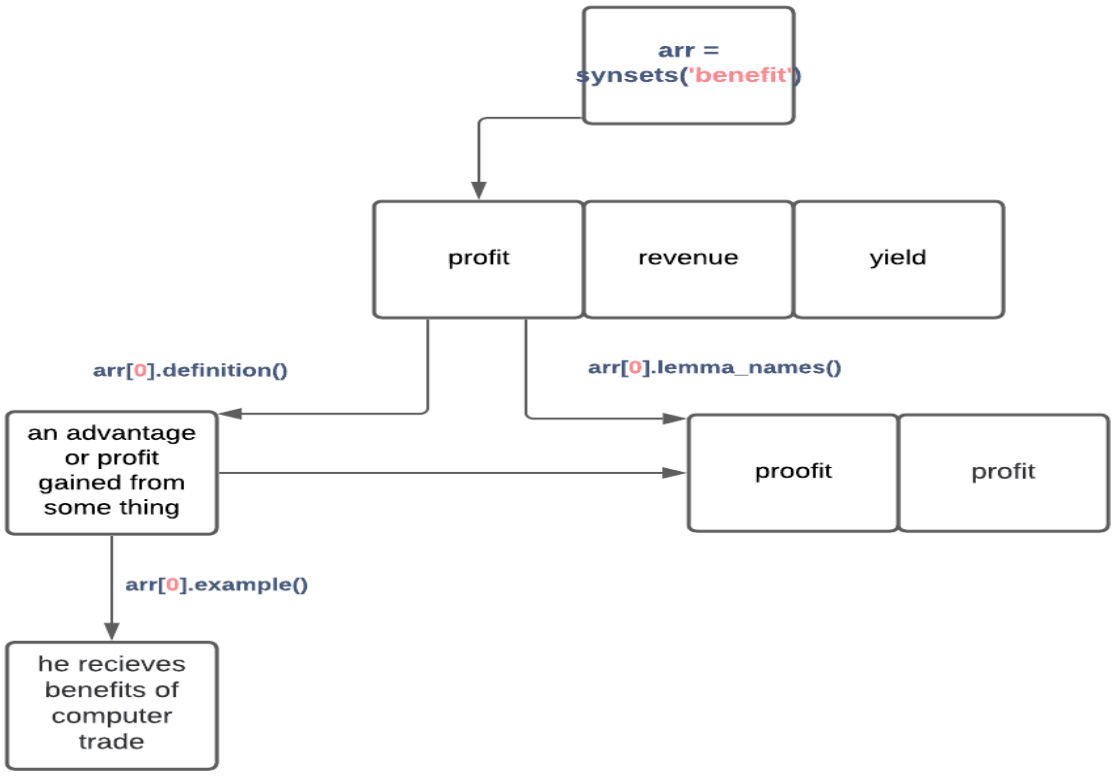
Where thesaurus is helping us in finding the synonyms and antonyms of the words the WordNET is helping us to do more than that. WordNET interlinks the specific sense of the words wherein thesaurus links words by their meaning only. In the WordNET the words are semantically disambiguated if they are in close proximity to each other. Thesaurus provides a level to the words in the network if the words have similar meaning but in the case of WordNET, we get levels of words according to their semantic relations which is a better way of grouping the words.

Structure of WordNET

The below image is a basic structure of the WordNET. The main concept of the relationship between the words in the WordNETs network is that the words are synonyms like sad and unhappy, benefit and profit. These words show the same concept of using them in similar contexts by interchanging them. These types of words are grouped into synsets which are unordered sets. Where synsets are linked together if they are having even small conceptual relations. Every synset in the network has its own brief definition and many of them are illustrated with the example of how to use them in a sentence. That definition and example part makes WordNET different from other



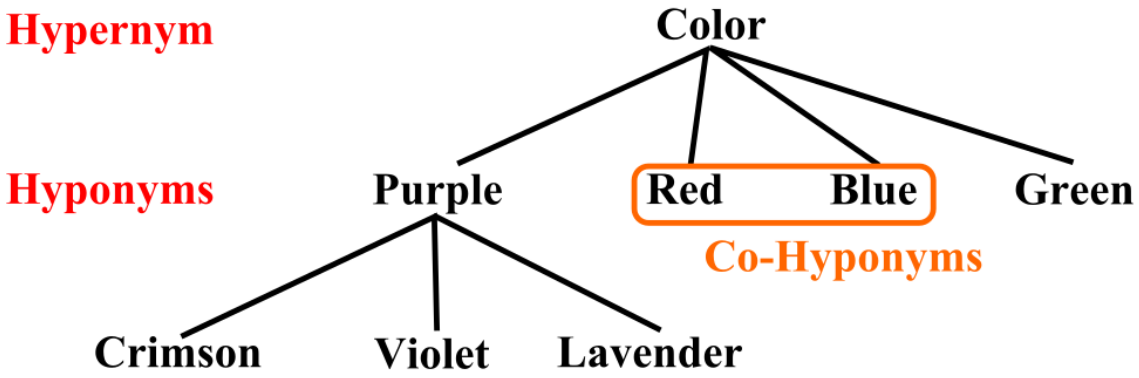
In the below picture we can see the structure of any synset where we are having synonyms of benefit in the array of synsets with the definition and the example of usage of benefit word. This synset is related to another synset word, where the words benefit and profit have exactly the same meaning.



Here we can see the structure of the wordnet and also how the synsets under the networks are interlinked because of the conceptual relation between the words.

Relations in the WordNET

1. **Hyponym:** In linguistics, a word with a broad meaning constitutes a category into which words with more specific meanings fall; a superordinate. For example, the colour is a hypernym of red. Where Hyponymy shows the relationship between a hypernym and a specific instance of a hyponym. A hyponym is a word or phrase whose semantic field is more specific than its hypernym. The semantic field of a hypernym, also known as a superordinate.



The above image is an example of the relationship between hyponyms and hypernym.

The reason for explaining these terms here is because in WordNET the most frequent relationships between synsets are based on these hyponym and hypernym relations. These are very beneficial in linking words like(paper, piece of paper). Saying more specifically with an example from the above picture like purple and violet, in WordNET the category colour includes purple which in turn includes violet. The root node of the hierarchy is the last point for every noun. In violet is a kind of purple and purple is a kind of colour then violet is a kind colour this is the hyponymy relation between the words which is transitive.

2. **Meronymy**: The wordnet hold follows the meronymy relation which defines the whole relationship between the synset for example a bike has two wheels handle and petrol tank. These components of a bike are inherited from their subordinates: if a bike has two wheels then a sports bike has wheels as well. In linguistics, we basically use this kind of relationship for adverbs which basically represents the characteristic of the noun. So the parts are inherited into a downward direction because all the bikes and types of bikes have two wheels, but not all kinds of automobiles consist of two wheels.
3. **Troponymy**: In linguistics, troponymy is the presence of a 'manner' relation between two lexemes. In WordNET Verbs describing events that necessarily and unidirectionally entail one another are linked: {buy}-{pay}, {succeed}-{try}, {show}-{see}, etc. basically the in the hierarchy verbs towards the bottom shows the manners are characterizing the events like communication-talk-whisper.
4. **Antonymy**: Adjective words under the WordNET arranged in the antonymy pairs like wet and dry, smile and cry. Each of these pairs of antonyms is linked with sets of semantic similar ones. The cry is linked to weep, shed tears, sob, wail etc. so that they all can be considered as the opposite of indirect antonyms of a smile.
5. **Cross – PoS Relations:** Most of the relations in the wordNET are in the same part of speech. On the basis of part of speech relations, we can divide WordNET into 4 types of 4 subnets one for each noun, verbs, adjective, and adverb. There are also some cross-PoS pointers available in the network which include a morphosemantic link that holds the words with the same meaning and shares a stem. For example, many pairs like (reader read) in which the noun of the pair has a semantic layer with respect to the verb have been specified.

Implementation of WordNET

For implement WordNET
Importing libraries:
import nltk
from nltk.corpus import wordnet

Downloading the wordnet:
nltk.download('wordnet')
Output:

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Unzipping corpora/wordnet.zip.
True
```

2.4 Word-net Application in Query Expansion

Query expansion is a technique used in information retrieval systems to improve the precision and recall of a search query by automatically expanding the query to include synonyms and related terms. WordNet is a useful resource for query expansion, as it provides a large numberof synonyms and related terms for English words.

Query Expansion is a most enhanced tool for improving the query statements and also improving the retrieval of relevant documents in match to the search.
Query expansion is a method to add more new terms along with the original query so as to a better retrieval performance.
The four different ways of query expansion is mentioned below:
• Manual (the user can choose the expansion terms).
• Interactive (system suggest the query expansion terms to the user to expand the query).
• Automatic (the entire process is invisible).
• Hybrid (mixture of more than one query expansion methods).

A) Automatic Query Expansion (AQE)

After the evaluation results obtained at the Text Retrieval Conference series (TREC) the AQE become popular and considered to be most common and preferred method. Most researchers had implemented this technique and brought noticeable improvements in IR. AQE is considered to be most promising technique to boost-up the retrieval effectiveness of document ranking. Even MySQL, Google Enterprise and Lucene provide AQE to users that can turned on or off.

B) Important Steps in Automatic Query Expansion

Based on the review, Automatic query expansion is categorized into four steps. The steps are briefed below.

1) Pre-processing of Raw Data

The data used for users query expansion is converted into a new effective format for further processing.
It involves the following steps:
a. Text extraction is done from documents like MS Word, HTML and PDF documents.
b. Tokenization (here individual words are extracted).
c. Stop word removal (the common words like prepositions and articles are removed).
d. Stemming of Word
e. Weighting of word (To understand the importance in every document, each word is assign with a score)

2) Query Term Features Generation and its Ranking

The query term is generated and ranked by the system for expansion features. The query term properties form a base for feature generation. With very few query term expansions features, the query is expanded. So ranking is very important. The input to this stage are user query and data. The output contains a set of expansion features additionally with their corresponding scores.

3) Selection of Query Term Features

The top most elements are considered for query expansion after completing the ranking of query term features. These elements are considered individually and not on their expansion feature's mutual dependencies among them

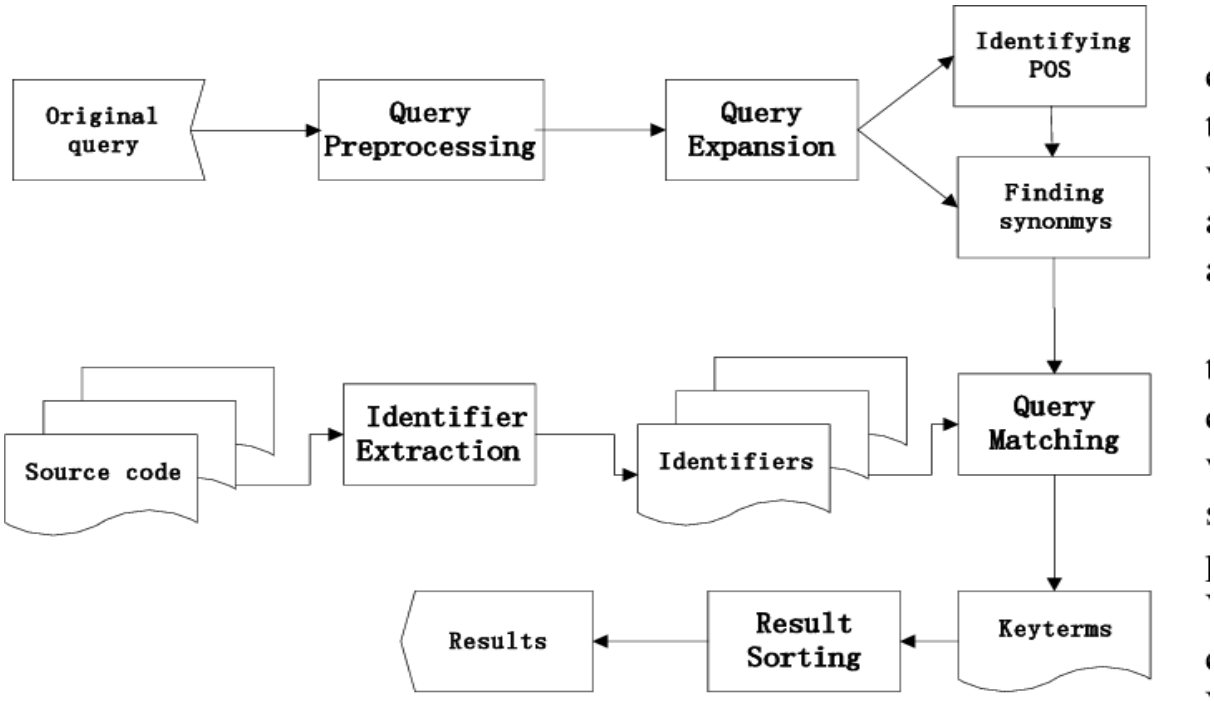
4) Query Formulation and Reformulation

The query formulation and reformulation is the final stage in query expansion. Here it is discussed on the submission of expanded query to information retrieval system in order to get effective results. Here a weight is assigned to each feature that describes the expanded query term reweighting. There are many application of AQE like Multimedia Information Retrieval, Question Answering, Information Filtering, Mobile Search, Cross-Language Information Retrieval, Expert Finding, Federated Search, Slot-based Document Retrieval, etc.

5) Categorization of Query Expansion Techniques

The five major classification of Automatic query expansion techniques are mentioned below:

1. Query based techniques
• Distribution difference based techniques
• Model based techniques
• Document summarization based techniques
2. Corpus based techniques
• Concept term based techniques
• Term clustering based techniques
3. Linguistic based techniques
• Stemming based techniques
• Ontology browsing based techniques
• Syntactic parsing based techniques
4. Web data based techniques
• Anchor Text based techniques
• Wikipedia based techniques
5. Search log data based techniques
• Related queries based techniques
• Exploiting query documents relationship based techniques



Query expansion via WordNet for effective code search

One common approach to using WordNet for query expansion is to identify the synonyms and related terms for each word in the original query, and then include these terms in the expanded query. For example, if the original query was "car", WordNet could be used to identify related terms such as "automobile", "vehicle", and "motorcar", which could then be included in the expanded query.

Another approach to using WordNet for query expansion is to use it to identify the semantic relationships between words in the query, and then use these relationships to identify additional terms that are related to the original query. For example, if the original query was "car", WordNet could be used to identify that "car" is a hyponym of "vehicle", and that "vehicle" is a hypernym of "car". This information could then be used to expand the query to include related terms such as "automobile", "truck", and "bus".

Overall, using WordNet for query expansion can help improve the accuracy and relevance of search results in information retrieval systems, and has been widely used in various applications such as web search engines, document retrieval systems, and question-answering systems.

2.5 Wiktionary

Wiktionary is a multilingual, web-based project to create a free content dictionary of all words in all languages. It is collaboratively edited via a wiki, and its name is a portmanteau of the words wiki and dictionary. It is available in 171 languages and in Simple English. Like its sister project Wikipedia, Wiktionary is run by the Wikimedia Foundation, and is written collaboratively by volunteers, dubbed "Wiktionarians". Its wiki software, MediaWiki, allows almost anyone with access to the website to create and edit entries. Because Wiktionary is not limited by print space considerations, most of Wiktionary's language editions provide definitions and translations of

words from many languages, and some editions offer additional information typically found in thesauri and lexicons. The English Wiktionary includes a thesaurus (formerly known as Wikisaurus) of synonyms of various words. Wiktionary data are frequently used in various natural language processing tasks.

Wiktionary has semi-structured data.Wiktionary lexicographic data can be converted to machine-readable format in order to be used in natural language processing tasks.

Wiktionary data mining is a complex task. There are the following difficulties: (1) the constant and frequent changes to data and schemata, (2) the heterogeneity in Wiktionary language edition schemata and (3) the human-centric nature of a wiki.

There are several parsers for different Wiktionary language editions:

- DBpedia Wiktionary: a subproject of DBpedia, the data are extracted from English, French, German and Russian wiktionaries; the data includes language, part of speech, definitions, semantic relations and translations. The declarative description of the page schema,regular expressions and finite state transducer are used in order to extract information.
- JWKTL (Java Wiktionary Library): provides access to English Wiktionary and German Wiktionary dumps via a Java Wiktionary API. The data includes language, part of speech, definitions, quotations, semantic relations, etymologies and translations. JWKTL is available for non-commercial use.
- wikokit: the parser of English Wiktionary and Russian Wiktionary.The parsed data includes language, part of speech, definitions, quotations, semantic relations and translations. This is a multi-licensed open-source software.
- Etymological entries have been parsed in the Etymological WordNet project.

The various natural language processing tasks were solved with the help of Wiktionary data:

- Rule-based machine translation between Dutch language and Afrikaans; data of English Wiktionary, Dutch Wiktionary and Wikipedia were used with the Apertium machine translation platform.
- Construction of machine-readable dictionary by the parser NULEX, which integrates open linguistic resources: English Wiktionary, WordNet, and VerbNet.The parser NULEX scrapes English Wiktionary for tense information (verbs), plural form and part of speech (nouns).

- Speech recognition and synthesis, where Wiktionary was used to automatically create pronunciation dictionaries. Word-pronunciation pairs were retrieved from 6 Wiktionary language editions (Czech, English, French, Spanish, Polish, and German). Pronunciations are in terms of the International Phonetic Alphabet. The ASR system based on English Wiktionary has the highest word error rate, where each third phoneme has to be changed.
- Ontology engineering and semantic network constructing.
- Ontology matching.
- Text simplification. Medero & Ostendorf assessed vocabulary difficulty (reading level detection) with the help of Wiktionary data. Properties of words extracted from Wiktionary entries (definition length and POS, sense, and translation counts) were investigated. Medero & Ostendorf expected that (1) very common words will be more likely to have multiple parts of speech, (2) common words to be more likely to have multiple senses, (3) common words will be more likely to have been translated into multiple languages. These features extracted from Wiktionary entries were useful in distinguishing word types that appear in Simple English Wikipedia articles from words that only appear in the Standard English comparable articles.
- Part-of-speech tagging. Li et al. (2012) built multilingual POS-taggers for eight resource-poor languages on the basis of English Wiktionary and Hidden Markov Models.
- Sentiment analysis.

2.6 Semantic relatedness

Semantic relatedness (SR) is defined as a measurement that quantitatively identifies some form of lexical or functional association between two words or concepts based on the contextual or semantic similarity of those two words regardless of their syntactical differences.
The need to determine semantic relatedness or its inverse, semantic distance, between two lexically expressed concepts is a problem that pervades much of natural language processing. Measures of relatedness or distance are used in such applications as word sense disambiguation, determining the structure of texts, text summarization and annotation, information extraction and retrieval, automatic indexing, lexical selection, and the automatic correction of word errors in text. It's important to note that semantic relatedness is a more general concept than similarity; similar entities are semantically related by virtue of their similarity (bank–trust company), but dissimilar entities may also be semantically related by lexical relationships such as meronymy (car–wheel) and antonymy (hot–cold), or just by any kind of functional relationship or frequent association (pencil–paper, penguin–Antarctica, rain–flood). Computational applications typically require relatedness rather than just similarity; for example, money and river are cues to the in-context meaning of bank that are just as good as trust company.

Lexical Resource–based Approaches to Measuring Semantic Relatedness

1. **Dictionary-based Approaches**

In this approach Longman Dictionary of Contemporary English (LDOCE) is converted into a network by creating a node for every headword and linking each node to the nodes for all the words used in its definition.
In this network, the similarity function simKF between words of the defining vocabulary is computed by means of spreading activation on this network. The function is extended to the rest of LDOCE by representing each word as a list $W = \{w1, \ldots, wr\}$ of the words in its definition; thus, for instance,

simKF(*linguistics*, *stylistics*)

= simKF(*{the, study, of, language, in, general, and, of, particular, languages, and, their, structure, and, grammar, and, history}, {the, study, of, style, in, written, or, spoken, language}*)

Built on this work to derive a **context-sensitive,** or **dynamic,** measure that takes into account the "associative direction" of a given word pair. For example, the context *{car, bus}* imposes the associative direction of vehicle (close words are then likely to include *taxi, railway, airplane*, etc.), whereas the context *{car, engine}* imposes the direction of components of car (*tire, seat, headlight*, etc.).

2. Approaches Based on Roget-structured Thesauri

Roget-structured thesauri, such as *Roget's Thesaurus* itself, the *Macquarie Thesaurus* , and others, group words in a structure based on **categories** within which there are several levels of finer clustering. The categories themselves are grouped into a number of broad, loosely defined classes. The user's main access is through the **index,** which contains category numbers along with **labels** representative of those categories for each word.
Methods of semantic distance that are based on Roget structured thesauri rely not only on the category structure but also on the index and on the **pointers** within categories that cross-reference other categories. In part as a consequence of this, typically no numerical value for semantic distance can be obtained: rather, algorithms using the thesaurus compute a distance implicitly and return a boolean value of 'close' or 'not close'.
Five types of semantic relations between words are there working with an abridged version of *Roget's Thesaurus*.
In their approach, two words were deemed to be related to one another, or semantically close, if their base forms satisfy any one of the following conditions:
1. They have a category in common in their index entries.
2. One has a category in its index entry that contains a pointer to a category of the other.
3. One is either a label in the other's index entry or is in a category of the other.
4. They are both contained in the same subcategory.
5. They both have categories in their index entries that point to a common category.

3. Approaches Using WordNet and Other Semantic Networks

Most of the methods use WordNet, a broad coverage lexical network of English words. Nouns, verbs, adjectives, and adverbs are each organized into networks of synonym sets (**synsets**) that each represent one underlying lexical concept and are interlinked with a variety of relations. A polysemous word will appear in one synset for each of its senses. Which can be further used for finding the semantic relatedness of that word. The noun network of WordNet was the first to be richly developed The backbone of the noun network is the subsumption hierarchy (hyponymy/ hypernymy), which accounts for close to 80% of the relations. At the top of the hierarchy are 11 abstract concepts, termed unique beginners, such as entity ('something having concrete existence; living or nonliving') and psychological feature ('a feature of the mental life of a living organism'). The maximum depth of the noun hierarchy is 16 nodes. The nine types of relations defined on the noun subnetwork, in addition to the synonymy relation that is implicit in each node are: the hyponymy (IS-A) relation and its inverse, hypernymy; six meronymic (PART-OF) relations — COMPONENT-OF, MEMBER-OF and SUBSTANCE-OF and their inverses; and antonymy, the COMPLEMENTOF relation.

4. Computing Taxonomic Path Length

A simple way to compute semantic relatedness in a taxonomy such as WordNet is to view it as a graph and identify relatedness with path length between the concepts:The shorter the path from one node to another, the more similar they are. Hirst and St-Onge adapted Morris and Hirst's (1991) semantic distance algorithm from Roget's Thesaurus to WordNet. They distinguished two strengths of semantic relations in WordNet. Two words are strongly related if one of the following holds:
1. They have a synset in common (for example, human and person).
2. They are associated with two different synsets that are connected by the antonymy relation (for example, precursor and successor).
3. One of the words is a compound (or a phrase) that includes the other and "there is any kind of link at all between a synset associated with each word" (for example, school and private school).
Two words are said to be in a medium-strong, or regular, relation if there exists an allowable path connecting a synset associated with each word.

5. **Scaling the Network**

In scaling ,we can create scales to measure communication constructs by first listing the key terms in the conceptual definition, then expanding the terms by looking up synonyms in dictionaries such as WordNet. By using this we can form the network and according to that it will be easy to find the semantic relatedness.

Relatedness measures quantify the degree to which two words are associated with each other (scissors-paper). Similarity is a subset of relatedness and quantifies how alike two concepts are based on their location within an is-a hierarchy (car-vehicle).

2.7 Measures of Word-Net Similarity

WordNet Similarity implements measures of similarity and relatedness that are all in some way based on the structure and content of WordNet. Measures of similarity use information found in an is– a hierarchy of concepts (or synsets), and quantify how much concept A is like (or is similar to) concept B. For example, such a measure might show that an automobile is more like a boat than it is a tree, due to the fact that automobile and boat share vehicle as an ancestor in the WordNet noun hierarchy. WordNet is particularly well suited for similarity measures, since it organizes nouns and verbs into hierarchies of is–a relations. In version 2.0, there are nine separate noun hierarchies that include 80,000 concepts, and 554 verb hierarchies that are made up of 13,500 concepts. Is–a relations in WordNet do not cross part of speech boundaries, so similarity measures are limited to making judgments between noun pairs (e.g., cat and dog) and verb pairs (e.g., run and walk). While WordNet also includes adjectives and adverbs, these are not organized into is–a hierarchies so similarity measures can not be applied.
However, concepts can be related in many ways beyond being similar to each other. For example, a wheel is a part of a car, night is the opposite of day, snow is made up of water, a knife is used to cut bread, and so forth. As such WordNet provides relations beyond is–a, including has–part, is–made–of, and is–an–attribute–of. In addition, each concept is defined by a short gloss that may include an example usage. All of this information can be brought to bear in creating measures of relatedness. As a result these measures tend to be more flexible, and allow for relatedness values to be assigned across parts of speech (e.g., the verb murder and the noun gun).

Measures of WordNet similarity refer to a variety of methods that are used to quantify the degree of similarity between two words or concepts in WordNet. These measures are importantfor a wide range of natural language processing applications such as information retrieval, text

classification, and word sense disambiguation.

There are several measures of WordNet similarity, including:

1. Path-based measures: These measures are based on the shortest path between two synsets in the WordNet hierarchy. The shorter the path, the more similar the two synsets are considered to be.
2. Information content measures: These measures take into account the frequency of a word or concept in a corpus of text, and use this information to estimate its semantic importance. Words with higher information content are considered to be more specific and less likely to be confused with other words, and thus are considered more similar.
3. Resnik's similarity measure: This measure is based on the information content of the lowest common subsumer of two synsets. The lowest common subsumer is the synset that is closest to both synsets in the WordNet hierarchy.
4. Lin's similarity measure: This measure is based on the information content of the common ancestor of two synsets. The common ancestor is the synset that is furthest from the root node of the WordNet hierarchy and is shared by both synsets.
5. Jiang-Conrath similarity measure: This measure is based on the information content of the lowest common subsumer of two synsets, as well as the information content of the two synsets themselves. It is designed to address some of the limitations of other measures, such as the fact that they do not take into account the specificity of the synsets being compared.

Overall, there are many different measures of WordNet similarity, each with its own strengths and weaknesses. The choice of measure will depend on the specific task and the characteristicsof the data being analyzed.

Using WordNet::Similarity WordNet::Similarity can be utilized via a command line interface provided by the utility program similarity.pl. This allows a user to run the measures interactively. In addition, there is a web interface that is based on this utility. WordNet::Similarity can also be embedded within Perl programs by including it as a module and calling its methods.

1. Command Line The utility similarity.pl allows a user to measure specific pairs of concepts when given in word#pos#sense form. For example, car#n#3 refers to the third WordNet noun sense of car. It also allows for the specification of all the possible senses associated with a word or word#pos combination. For example, in Figure 1, the first command requests the value of the lin measure of similarity for the second noun sense of car (railway car) and the first noun sense of bus (motor coach). The second command will return the score of the pair of concepts that have the highest similarity value for the nouns car and bus. In the third command, the –allsenses switch causes the similarity measurements of all the noun senses of car to be calculated relative to the first noun sense of bus.

```
> similarity.pl --type WordNet::Similarity::lin car#n#2 bus#n#1
car#n#2  bus#n#1  0.530371390319309        # railway car versus motor coach

> similarity.pl --type WordNet::Similarity::lin car#n bus#n
car#n#1  bus#n#1  0.618486790769613        # automobile versus motor coach

> similarity.pl --type WordNet::Similarity::lin --allsenses car#n bus#n#1
car#n#1  bus#n#1  0.618486790769613        # automobile versus motor coach
car#n#2  bus#n#1  0.530371390319309        # railway car versus motor coach
car#n#3  bus#n#1  0.208796988315133        # cable car versus motor coach
```

Figure 1: Command Line Interface

2. Programming Interface WordNet::Similarity is implemented with Perl's object oriented features. It uses the WordNet::QueryData package (Rennie, 2000) to create an object representing WordNet. There are a number of methods available that allow for the inclusion of existing measures in Perl source code, and also for the development of new measures. When an existing measure is to be used, an object of that measure must be created via the new() method. Then the getRelatedness() method can be called for a pair of word senses, and this will return the relatedness value. For example, the program in Figure 2 creates an object of the lin measure, and then finds the similarity between the first sense of the noun car (automobile) and the second sense of the noun bus (network bus).
3. WordNet::Similarity enables detailed tracing that shows a variety of diagnostic information specific to each of the different kinds of measures. For example, for the measures that rely on path lengths (lch, wup, path) the tracing shows all the paths found between the concepts. Tracing for the information content measures (res, lin, jcn) includes both the paths between concepts as well as the least common subsumer. Tracing for the hso measure shows the actual paths found through WordNet, while the tracing for lesk shows the gloss overlaps in WordNet found for the two concepts and their nearby relatives. The vector tracing shows the word vectors that are used to create the gloss vector of a concept.

```
#!/usr/bin/perl -w

use WordNet::QueryData;                          # use interface to WordNet
use WordNet::Similarity::lin;                    # use Lin measure

 $wnObj = new WordNet::QueryData;                  # create a WordNet object
$linObj = new WordNet::Similarity::lin($wnObj); # create a lin object

$value = $linObj -> getRelatedness ('car#n#1', 'bus#n#2');  # how similar?
```

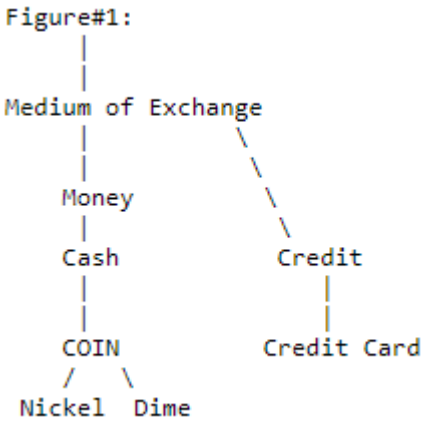
Figure 2: Programming Interface

2.8 Resnick's work on Word-Net Similarity

By using the standard argumentation of information theory by Ross [1976], information content of a concept is defined as the negative the log likelihood, -logp(c), where p(c) is the probability

of encountering such concept c. For example, in Figure#1, Money has a less information content than NICKEL as the probability of encountering the concept, p(Money) is much greater than encountering the probability of p(Nickel). Consequently, Nickel contains more information, i.e. more specific than Money as it is quantifiable, an exact form of money etc. Also, whenever the concept Nickel is used, the concept of Money is also used as it is the parent node of Nickel. As a result, Resnik concluded that the higher the terms in the taxonomy tree, the less information they contain but are encountered more frequently. E.g. the concept 'matter' would have a p(matter) of near 1 and a information content of near 0. Resnik also uses the following equation to define the similarity between two dis! tinct concepts [p450]:

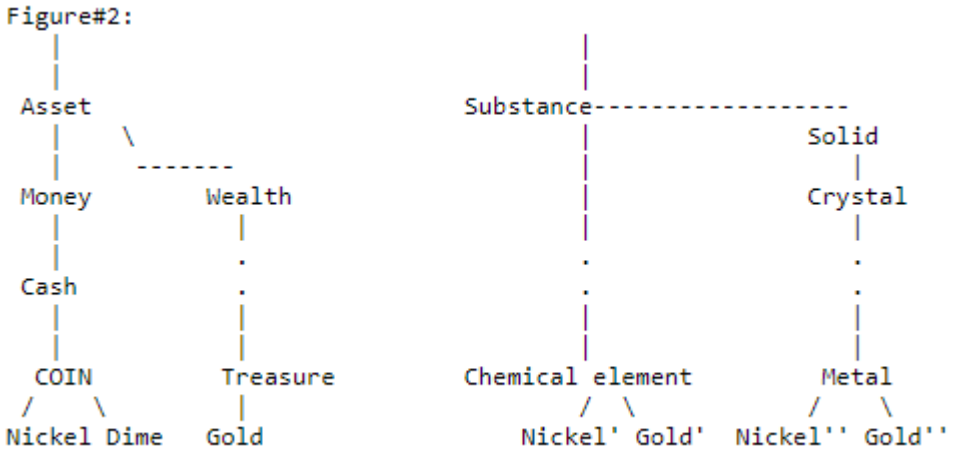
Sim(c1,c2) = max c in S(c1,c2) [-logp(c)] (2)



where c = concept and S(c1,c2) = set of concepts c, that subsume both c1 and c2 and -logp(c) = information it carries -- the information content. Then this equation simply means that we want a concept, with maximum information content, that subsumes both c1 and c2. Therefore in Figure#2, COIN is used for the similarity between concepts Nickel and Dime rather than Cash, since it is the one with more information content and also subsumes both Nickel and Dime. The equation for comparing two words (remember that a word can have multiple distinct concepts, e.g. orange) is required to consider all concepts that range over w1 and w2 as before,

Simword(w1,w2) = take max from all c1,c2 [sim(c1,c2)] (3)

Where c1 ranges over w1 and c2 ranges over w2. The formula then means we want to compare all concepts of w1 and w2, and take the maximum similar value Sim(c1,c2), where both w1 and w2 can be an instance of the maximum concept. For example, refer to Figure#2, for Nickel and Gold, the word Nickel has parents: COIN, Chemical element and metal; the word Gold has parents: Treasure, Chemical element and Metal. The maximum c for Nickel and Gold would be chosen to be ASSET, Chemical element or Metal. As in the real world, assume p(Asset) > p(Metal) > p(Chemical element), i.e. more general and less information content, Chemical element would be chosen as it's information content is the largest and its value, Sim(Nickel',Gold'), will be calculated using equation(2), and assigned to Simword(Gold,Nickel) for the two words Gold and Nickel.



2.9 Indian Language Word-nets

Indian Language WordNets The creation of IndoWordNet began in 2000 with Hindi WordNet. Due to the complex nature of Indian language families, and many other reasons such as morphological richness, gender information etc. it was decided that Hindi be used as a pivot for linking all the Indian Languages. Hindi shares many common features and borrowed concepts from ancient Indian languages like Sanskrit and is the most commonly spoken language in India. The expansion approach adopted for IndoWordNet creation is:

1. Creation of a Hindi synset with synonymous words.
2. Mapping of the synset with relations such as hypernymy and hyponymy etc.
3. Tagging of the synset with an ontological category.
4. Allotment of a unique synset ID to the concept described in the synset.
5. Creation of the same synset in the other Indian languages leading to an implicit linkage of relations, ontological categories.

Indian language Word-nets are lexical databases similar to WordNet that capture the lexical and semantic relationships among words in various Indian languages. These word-nets are important resources for natural language processing applications in Indian languages, which are diverse and complex, and pose significant challenges for natural language processing.

Several Indian language Word-nets have been developed over the years, each covering a specific language or group of languages. Some of the notable Indian language Word-nets include:

1. Hindi WordNet: This is a lexical database of Hindi words and their semantic relationships, developed at the Indian Institute of Technology Bombay.
2. Tamil WordNet: This is a lexical database of Tamil words and their semantic relationships, developed at the Indian Institute of Technology Madras.
3. Bengali WordNet: This is a lexical database of Bengali words and their semantic relationships, developed at Jadavpur University.
4. Marathi WordNet: This is a lexical database of Marathi words and their semantic relationships, developed at the Indian Institute of Technology Bombay.
5. Gujarati WordNet: This is a lexical database of Gujarati words and their semantic

relationships, developed at the Sardar Patel University.

Indian language Word-nets are valuable resources for various natural language processing tasks such as machine translation, information retrieval, and text classification in Indian languages. They help to capture the nuances and complexities of Indian languages and facilitate the development of effective natural language processing applications.

The large nationwide project of building Indian language wordnets was called the IndoWordNet project. IndoWordNet is a linked lexical knowledge base of wordnets of 18 scheduled languages of India, viz., Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Meitei, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu. The wordnets are getting created by using expansion approach from the Hindi WordNet.

Principles of wordnet construction

The wordnets follow the principles of minimality, coverage and replaceability for the synsets. That means, there should be at least a 'core' set of lexemes in the synset that uniquely give the concept represented by the synset (minimality), e.g., {house, family} standing for the concept of 'family' ("she is from a noble house"). Then the synset should cover ALL the words representing the concept in the language (coverage), e.g., the word 'ménage' will have to appear in the 'family' synset, albeit, towards the end of the synset, since its usage is rare. Finally, the words towards the beginning of the synset should be able to replace one another in reasonable amount of corpora (replaceability), e.g., 'house' and 'family' can replace each other in the sentence "she is from a noble house".

	Noun	Verb	Adjectives	Adverbs	Total
Assamese	9065	1676	3805	412	14958
Bengali	27281	2804	5815	445	36346
Bodo	8788	2296	4287	414	15785
Gujarati	26503	2805	5828	445	35599
Hindi	29807	3687	6336	541	40371
Kannada	12765	3119	5988	170	22042
Kashmiri	21041	2660	5365	400	29469
Konkani	23144	3000	5744	482	32370
Malayalam	20071	3311	6257	501	30140
Manipuri	10156	2021	3806	332	16351
Marathi	23271	3146	5269	539	32226
Nepali	6748	1477	3227	261	11713
Odiya	27216	2418	5273	377	35284
Punjabi	23255	2836	5830	443	32364
Sanskrit	32385	1246	4006	265	37907
Tamil	16312	2803	5827	477	25419
Telugu	12078	2795	5776	442	21091
Urdu	22990	2801	5786	443	34280

Table 1: Number of synsets in different wordnets

	Nouns		Verbs		Adjectives		Adverbs		Total
	D	H	D	H	D	H	D	H	
Assamese	7019	679	1300	36	2744	0	294	0	12072
Bengali	11049	7680	1824	99	3356	3	312	0	24323
Bodo	6940	603	1594	64	2854	1	293	0	12349
Gujarati	10910	7533	1825	99	3356	3	312	0	24038
Hindi	11584	8221	1988	212	3542	4	344	0	25895
Kannada	7806	1973	1921	154	3453	3	133	0	15443
Kashmiri	9363	6261	1767	100	3240	2	294	0	21027
Konkani	10545	6952	1888	128	3391	2	328	0	23234
Malayalam	9146	4754	1970	206	3525	4	340	0	19945
Manipuri	7192	823	1324	43	2712	0	244	0	12338
Marathi	9874	6556	1839	144	3092	0	333	0	21838
Nepali	5217	496	1114	42	2202	1	200	0	9272
Odiya	11039	7680	1679	66	3187	2	271	0	23924
Punjabi	10215	6382	1822	99	3355	3	312	0	22188
Sanskrit	8396	6470	1048	28	2873	2	241	0	19058
Tamil	8130	3066	1821	98	3353	3	312	0	16783
Telugu	6944	1843	1819	98	3350	0	312	0	14366
Urdu	10424	6816	1822	98	3356	3	313	0	22832

Table 2: Linkage Statistics for English to Indian Language WordNets. D stands for Direct links, and H stands for Hypernymy links

IndoWordNet is highly similar to EuroWordNet. However, the pivot language is Hindi which, of course, is linked to the English WordNet. Also typical Indian language phenomena like complex predicates and causative verbs are captured in IndoWordNet.

IndoWordNet is publicly browsable. The Indian language wordnet building efforts forming the subcomponents of IndoWordNet project are: North East WordNet project, Dravidian WordNet Project and Indradhanush project all of which are funded by the TDIL project.

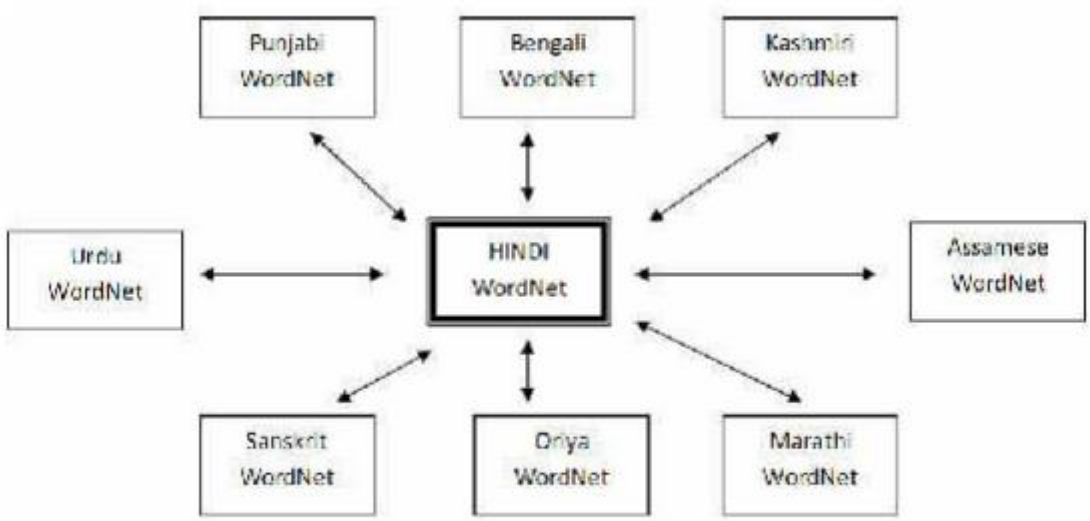


Figure 1: Linked Indo WordNet structure

Hindi WordNet Database Design

S. No.	Field Name	Purpose
01.	synset_id	Primary key: Uniquely identifies a concept/synset in the language
02.	concept_definition	The gloss / concept definition in a synset
03.	category_id	Foreign key from category table. Specifying if the concept is a noun, verb, adjective or adverb
04.	source_id	Foreign key from source table. Specifies the source from where the concept is taken
05.	Synset	inputting a set of synonyms representing that particular concept
06.	Example	examples of the given concept

2.10 Multilingual Natural Language Processing (NLP)

Multilingual NLP is a technology that integrates linguistics, artificial intelligence, and computer science to serve the purpose of processing and analyzing substantial amounts of natural human language in numerous settings.

How does multilingual NLP work?

There are many different forms of multilingual NLP, but in general, it enables computational software to understand the language of certain texts, along with contextual nuances. Multilingual NLP is also capable of obtaining specific data and delivering key insights. In short, multilingual NLP technology makes the impossible possible which is to process and analyze large amounts of data. Without it, this kind of task can probably only be executed by employing a very labor- and time-intensive approach.

Multilingual Dictionaries

Multilingual dictionaries are dictionaries that provide translations of words and phrases from one language to one or more other languages. These dictionaries are useful for individuals who need to communicate or work in multiple languages, such as language learners, travelers, and

professionals in international business.

Multilingual dictionaries can be printed or digital, and they can be available in various formats such as online dictionaries, mobile applications, or software programs. Some of the most popular multilingual dictionaries include:

1. Oxford Dictionaries: Oxford Dictionaries offer translations in various language pairs such as English to Spanish, French, German, Italian, and more.
2. Merriam-Webster Dictionary: This popular English language dictionary also provides translations of English words and phrases into Spanish.
3. Collins Dictionary: Collins Dictionary provides translations in multiple languages, including French, German, Italian, Spanish, and Portuguese.
4. Langenscheidt: Langenscheidt is a publisher of multilingual dictionaries for various language pairs, including German, English, Spanish, French, Italian, and Russian.
5. Google Translate: Google Translate is a popular online translation service that provides translations in over 100 languages.

Multilingual dictionaries are important resources for language learners, professionals, and individuals who need to communicate in multiple languages. They help to bridge the languagebarrier and promote cross-cultural communication and understanding.

Multilinguality

Multilinguality refers to the ability to use and communicate in more than one language. It is a valuable skill that allows individuals to communicate and work effectively in diverse cultural and linguistic contexts.

Multilinguality is becoming increasingly important in today's globalized world, where people are often required to communicate and collaborate with individuals from different countries and cultures. Multilinguality can facilitate cross-cultural communication and understanding, and it can open up new opportunities for personal and professional growth.

The benefits of multilinguality are numerous. They include:

1. Improved communication: Multilingual individuals are able to communicate with a wider range of people, which can improve their social and professional interactions.
2. Increased job opportunities: Many jobs require individuals to be able to communicate in multiple languages, and multilingual individuals have a competitive advantage in the job market.

Improved cognitive function: Multilingualism has been shown to improve cognitive

4. function, including memory, problem-solving, and decision-making skills.
5. Cultural awareness: Multilingual individuals are often more aware of cultural differences and are better equipped to navigate cross-cultural interactions.
6. Enhanced travel experiences: Multilingual individuals can communicate more effectively while traveling, which can enhance their travel experiences and allow them to connect more deeply with local cultures.

Overall, multilinguality is a valuable skill that can enhance personal and professional growth, improve communication and understanding, and open up new opportunities in today's globalized world.

What makes multilingual NLP difficult to scale?

One of the biggest obstacles preventing multilingual NLP from scaling quickly is relating to low availability of labelled data in low-resource languages.
Among the 7,100 languages that are spoken worldwide, each of them has its own linguistic rules and some languages simply work in different ways. For instance, there are undeniable similarities between Italian, French and Spanish, whilst on the other hand, these three languages are totally different from a specific Asian language group, that is Chinese, Japanese, and Korean which share some similar symbols and ideographs.
The outcome from this leads to the need to have various techniques to generate language models that can work with all these languages. In short, different languages often require different vector spaces, even if there are existing pre-trained language embeddings.
Even though pre-trained word embeddings in different languages exist, it is possible that all of them are in different vector spaces. This means that similar words can signify different vector representations, basically due to the natural characteristics of a certain language.
This is why scaling multilingual NLP applications can be challenging. They use large amounts of labelled data, process it, learn patterns, and generate prediction models. When building NLP on a text comprising different languages, it is best to consider multilingual NLP.
When we need to build NLP on a text containing different languages, we may look at multilingual word embeddings for NLP models that have the potential to scale effectively.

Solutions for tackling multilingual NLP challenges

1, Training specific non-English NLP models
The first suggested solution is to train an NLP model for a specific language. A well-known example would be a few new versions of Bidirectional Encoder Representations from Transformers (BERT) that have been trained in numerous languages.
However, the biggest problem with this approach is its low success rate of scaling. It takes lots of time and money to train a new model, let alone many models. NLP systems require various large models, hence the processes can be very expensive and time-consuming.
This technique also does not scale effectively in terms of inference. Using NLP in different languages means the business would have to sustain different models and provision several servers and GPUs. Again, this can be extremely costly for the business.
2, Leveraging multilingual models
The past years have seen that new emerging multilingual NLP models can be incredibly accurate, at times even more accurate than specific, dedicated non-English language models.

Whilst there are several high-quality pre-trained models for text classification, so far there has not been a multilingual model for text generation with impressive performance.
3, Utilizing translation
The last solution some businesses benefit from is to use translation. Companies can translate their non-English content to English, provide the NLP model with that English content, then translate the result back to the needed language.
As manual as it may sound, this solution has several advantages, including cost-effective workflow maintenance and easily supported worldwide languages.
Translation may not be suitable if your business is after quick results, as the overall workflow's response time must increase to include translating process.

2.11 Metaphor

Metaphors are figures of speech that describe one thing in terms of another, by asserting that one thing is another thing. They are used to make comparisons or to describe something abstract in more concrete terms.

Metaphors can be used in various contexts, such as literature, poetry, advertising, and everyday conversation. They can be powerful tools for expressing complex or abstract ideas, and for creating vivid and memorable images in the minds of the audience.

Some examples of common metaphors include:

1. "Life is a journey." This metaphor describes life in terms of a journey, with its ups and downs, twists and turns, and various milestones along the way.

2. "Love is a rose." This metaphor describes love in terms of a rose, with its beauty, delicacy, and potential to prick or hurt.
3. "The world is a stage." This metaphor describes the world in terms of a theatrical stage, with people playing various roles and performing their parts.
4. "Time is money." This metaphor describes time in terms of money, with time being a valuable resource that should be used wisely and efficiently.
5. "Her voice was music to my ears." This metaphor describes a pleasing sound in terms of music, with the sound having a beautiful and melodious quality.

Metaphors can be powerful tools for communicating complex or abstract ideas, and for creatingmemorable images and associations in the minds of the audience. They can help to make language more vivid and expressive, and can add depth and richness to writing and speech.

2.12 Coreference

Co-reference resolution is a natural language processing (NLP) task that involves identifying and linking all the expressions that refer to the same entity in a text. Co-reference is a common phenomenon in language, where different words or expressions refer to the same person, place,thing, or concept.

The goal of co-reference resolution is to identify all the co-referent expressions in a text and link them to the correct entity. This is a challenging task for computers because it requires themachine to have a deep understanding of the text and the ability to make connections betweendifferent expressions and concepts.

There are various approaches to co-reference resolution in NLP, including rule-based, statistical, and machine learning methods. Some of the most common techniques used in co- reference resolution include:

1. Mention-pair models: These models use statistical methods to identify and link all the expressions that refer to the same entity in a text. They rely on various features such asword frequency, syntactic structure, and contextual information to identify co-referentexpressions.
2. Entity-based models: These models use machine learning techniques to identify and link all the expressions that refer to the same entity in a text. They use entity-level features such as named entity recognition and coreference features to identify co- referent expressions.
3. Rule-based models: These models use a set of pre-defined rules to identify and link co-referent expressions in a text. They rely on linguistic and contextual rules to identify the relationships between different expressions and concepts in a text.

Co-reference resolution is an important task in NLP, as it can improve the accuracy and understanding of automated systems that deal with large amounts of text data. It is used in various applications such as machine translation, question answering, and text summarization,and it has the potential to improve the performance and efficiency of these systems.

****