

**Name : Sarvagya Singh**

**SAPID : 60009200030**

**BATCH : K1**

**Experiment No 7**

**Aim: Implement Lesk's algorithm for word sense disambiguation**

**Theory: -**

In Natural Language Processing (NLP), word sense disambiguation (WSD) is the challenge of determining which “sense” (meaning) of a word is activated by its use in a specific context, a process that appears to be mostly unconscious in individuals.

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.**

**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.

**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.

**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 semi supervised methods is bootstrapping from seed data.

**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.

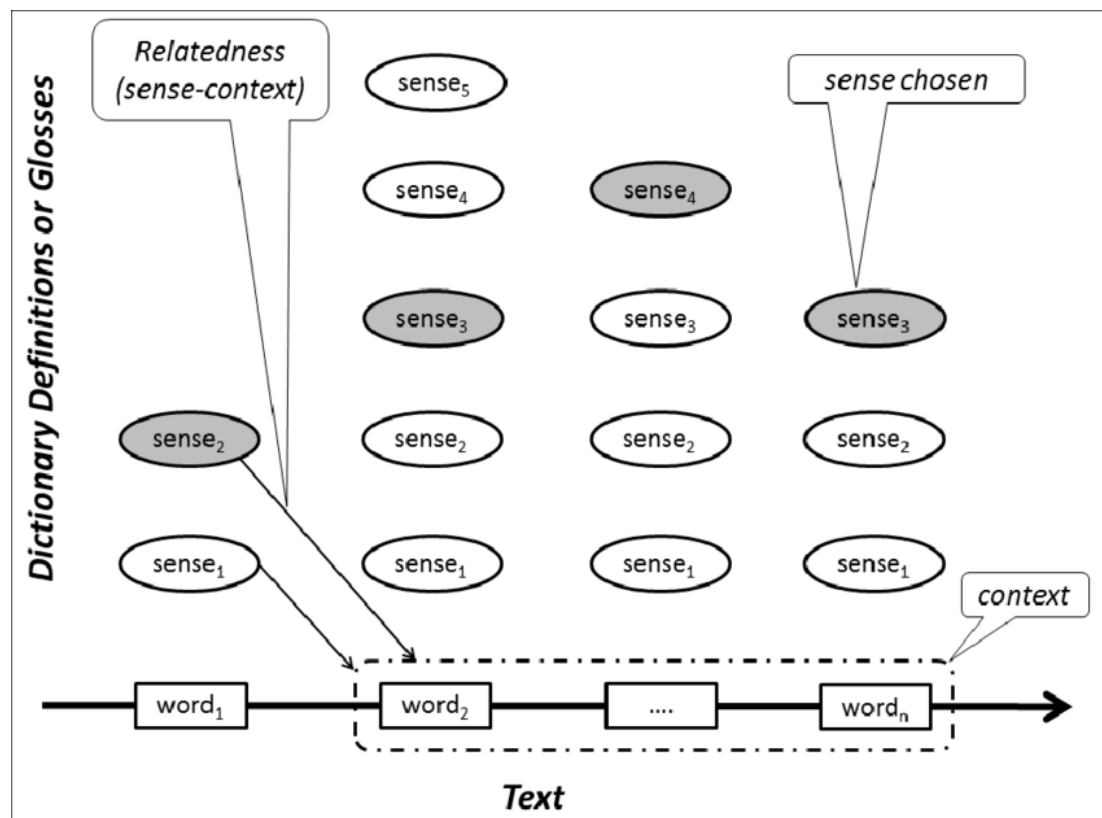
**Lesk Algorithm:-Dictionary based approach**

Lesk Algorithm is a way of Word Sense Disambiguation. The Lesk algorithm is a dictionary-based approach that is considered seminal. It is founded on the idea that words used in a text are related to one another, and that this relationship can be seen in the definitions of the words and their meanings. The pair of dictionary senses having the highest word overlap in their dictionary meanings are used to disambiguate two (or more) terms. Michael E. Lesk introduced the Lesk algorithm in 1986 as a classic approach for word sense disambiguation in Natural Language Processing. The Lesk algorithm assumes that words in a given “neighborhood” (a portion of text) will have a similar theme. The dictionary definition of an uncertain word is compared to the terms in its neighborhood in a simplified version of the Lesk algorithm.

Basic Lesk Algorithm implementation involves the following steps:

- 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.
- The sense to be picked is the one with the greatest number of items in this count.

Basically, the context is chosen from meaning of the nearest words. Following is the simplified pictorial representation of the same...



### Advantages and Disadvantages

There are numerous advantages to Lesk’s algorithm, the primary being that its simplicity makes it easy to implement, applicable in a variety of different contexts, and thus easily generalizable. Lesk notes that the algorithm does not depend on global information, meaning that since the same word could be referenced many times throughout a text but change each time, the meaning of a word is only derived from the collection of immediate supporting words in its context window, rather than from the entire text itself. Despite its simplicity and

power, the biggest drawback to Lesk's original algorithm is its performance — its accuracy was proposed by Lesk to be only around 50–70%, and has been shown to be much lower when experimentally validated against sense-tagged texts (Viveros-Jiménez, 2013). The algorithm also notably suffers from low recall, in that it cannot provide a corresponding contextual definition for many words simply because there is either no overlap to be found between dictionary definitions at all, or that several definitions have the same number of overlaps. Furthermore, Lesk leaves several questions unanswered.

**Lab Experiment to be performed in this session: -**

Step 1: Import Library's

Step2: Perform Tokenization

Step3: Take the ambiguous word and find the semantic of a word within various different context.

# CL-LAB-7

Name: Sarvagya Singh

SAP: 60009200030

Div/Batch: K/K1

## Lesk Algorithm

In [ ]:

```
import nltk
from nltk.wsd import lesk
from nltk.tokenize import word_tokenize
nltk.download('all')
```

```
[nltk_data] Downloading collection 'all'
[nltk_data] |
[nltk_data] | Downloading package abc to /root/nltk_data...
[nltk_data] | Unzipping corpora/abc.zip.
[nltk_data] | Downloading package alpino to /root/nltk_data...
[nltk_data] | Unzipping corpora/alpino.zip.
[nltk_data] | Downloading package averaged_perceptron_tagger to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping taggers/averaged_perceptron_tagger.zip.
[nltk_data] | Downloading package averaged_perceptron_tagger_ru to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping
[nltk_data] | taggers/averaged_perceptron_tagger_ru.zip.
[nltk_data] | Downloading package basque_grammars to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping grammars/basque_grammars.zip.
[nltk_data] | Downloading package bcp47 to /root/nltk_data...
[nltk_data] | Downloading package biocreative_ppi to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/biocreative_ppi.zip.
[nltk_data] | Downloading package bllip_wsj_no_aux to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping models/bllip_wsj_no_aux.zip.
[nltk_data] | Downloading package book_grammars to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping grammars/book_grammars.zip.
[nltk_data] | Downloading package brown to /root/nltk_data...
[nltk_data] | Unzipping corpora/brown.zip.
[nltk_data] | Downloading package brown_tei to /root/nltk_data...
[nltk_data] | Unzipping corpora/brown_tei.zip.
[nltk_data] | Downloading package cess_cat to /root/nltk_data...
[nltk_data] | Unzipping corpora/cess_cat.zip.
[nltk_data] | Downloading package cess_esp to /root/nltk_data...
[nltk_data] | Unzipping corpora/cess_esp.zip.
[nltk_data] | Downloading package chat80 to /root/nltk_data...
[nltk_data] | Unzipping corpora/chat80.zip.
[nltk_data] | Downloading package city_database to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/city_database.zip.
[nltk_data] | Downloading package cmudict to /root/nltk_data...
[nltk_data] | Unzipping corpora/cmudict.zip.
[nltk_data] | Downloading package comparative_sentences to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/comparative_sentences.zip.
[nltk_data] | Downloading package comtrans to /root/nltk_data...
[nltk_data] | Downloading package conll2000 to /root/nltk_data...
[nltk_data] | Unzipping corpora/conll2000.zip.
[nltk_data] | Downloading package conll2002 to /root/nltk_data...
[nltk_data] | Unzipping corpora/conll2002.zip.
[nltk_data] | Downloading package conll2007 to /root/nltk_data...
```

```
[nltk_data] | Downloading package crubadan to /root/nltk_data...
[nltk_data] | Unzipping corpora/crubadan.zip.
[nltk_data] | Downloading package dependency_treebank to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/dependency_treebank.zip.
[nltk_data] | Downloading package dolch to /root/nltk_data...
[nltk_data] | Unzipping corpora/dolch.zip.
[nltk_data] | Downloading package europarl_raw to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/europarl_raw.zip.
[nltk_data] | Downloading package extended_omw to
[nltk_data] | /root/nltk_data...
[nltk_data] | Downloading package floresta to /root/nltk_data...
[nltk_data] | Unzipping corpora/floresta.zip.
[nltk_data] | Downloading package framenet_v15 to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/framenet_v15.zip.
[nltk_data] | Downloading package framenet_v17 to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/framenet_v17.zip.
[nltk_data] | Downloading package gazetteers to /root/nltk_data...
[nltk_data] | Unzipping corpora/gazetteers.zip.
[nltk_data] | Downloading package genesis to /root/nltk_data...
[nltk_data] | Unzipping corpora/genesis.zip.
[nltk_data] | Downloading package gutenber to /root/nltk_data...
[nltk_data] | Unzipping corpora/gutenberg.zip.
[nltk_data] | Downloading package ieer to /root/nltk_data...
[nltk_data] | Unzipping corpora/ieer.zip.
[nltk_data] | Downloading package inaugural to /root/nltk_data...
[nltk_data] | Unzipping corpora/inaugural.zip.
[nltk_data] | Downloading package indian to /root/nltk_data...
[nltk_data] | Unzipping corpora/indian.zip.
[nltk_data] | Downloading package jeita to /root/nltk_data...
[nltk_data] | Downloading package kimmo to /root/nltk_data...
[nltk_data] | Unzipping corpora/kimmo.zip.
[nltk_data] | Downloading package knbc to /root/nltk_data...
[nltk_data] | Downloading package large_grammars to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping grammars/large_grammars.zip.
[nltk_data] | Downloading package lin_thesaurus to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/lin_thesaurus.zip.
[nltk_data] | Downloading package mac_morpho to /root/nltk_data...
[nltk_data] | Unzipping corpora/mac_morpho.zip.
[nltk_data] | Downloading package machado to /root/nltk_data...
[nltk_data] | Downloading package masc_tagged to /root/nltk_data...
[nltk_data] | Downloading package maxent_ne_chunker to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping chunkers/maxent_ne_chunker.zip.
[nltk_data] | Downloading package maxent_treebank_pos_tagger to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping taggers/maxent_treebank_pos_tagger.zip.
[nltk_data] | Downloading package moses_sample to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping models/moses_sample.zip.
[nltk_data] | Downloading package movie_reviews to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/movie_reviews.zip.
[nltk_data] | Downloading package mte_teip5 to /root/nltk_data...
[nltk_data] | Unzipping corpora/mte_teip5.zip.
[nltk_data] | Downloading package mwa_ppdb to /root/nltk_data...
[nltk_data] | Unzipping misc/mwa_ppdb.zip.
[nltk_data] | Downloading package names to /root/nltk_data...
[nltk_data] | Unzipping corpora/names.zip.
[nltk_data] | Downloading package nombank.1.0 to /root/nltk_data...
[nltk_data] | Downloading package nonbreaking_prefixes to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/nonbreaking_prefixes.zip.
[nltk_data] | Downloading package nps_chat to /root/nltk_data...
[nltk_data] | Unzipping corpora/nps_chat.zip.
[nltk_data] | Downloading package omw to /root/nltk_data...
[nltk_data] | Downloading package omw-1.4 to /root/nltk_data...
```

```
[nltk_data] | Downloading package opinion_lexicon to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/opinion_lexicon.zip.
[nltk_data] | Downloading package panlex_swadesh to
[nltk_data] | /root/nltk_data...
[nltk_data] | Downloading package paradigms to /root/nltk_data...
[nltk_data] | Unzipping corpora/paradigms.zip.
[nltk_data] | Downloading package pe08 to /root/nltk_data...
[nltk_data] | Unzipping corpora/pe08.zip.
[nltk_data] | Downloading package perluniprops to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping misc/perluniprops.zip.
[nltk_data] | Downloading package pil to /root/nltk_data...
[nltk_data] | Unzipping corpora/pil.zip.
[nltk_data] | Downloading package pll196x to /root/nltk_data...
[nltk_data] | Unzipping corpora/pll196x.zip.
[nltk_data] | Downloading package porter_test to /root/nltk_data...
[nltk_data] | Unzipping stemmers/porter_test.zip.
[nltk_data] | Downloading package ppattach to /root/nltk_data...
[nltk_data] | Unzipping corpora/ppattach.zip.
[nltk_data] | Downloading package problem_reports to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/problem_reports.zip.
[nltk_data] | Downloading package product_reviews_1 to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/product_reviews_1.zip.
[nltk_data] | Downloading package product_reviews_2 to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/product_reviews_2.zip.
[nltk_data] | Downloading package propbank to /root/nltk_data...
[nltk_data] | Downloading package pros_cons to /root/nltk_data...
[nltk_data] | Unzipping corpora/pros_cons.zip.
[nltk_data] | Downloading package ptb to /root/nltk_data...
[nltk_data] | Unzipping corpora/ptb.zip.
[nltk_data] | Downloading package punkt to /root/nltk_data...
[nltk_data] | Unzipping tokenizers/punkt.zip.
[nltk_data] | Downloading package qc to /root/nltk_data...
[nltk_data] | Unzipping corpora/qc.zip.
[nltk_data] | Downloading package reuters to /root/nltk_data...
[nltk_data] | Downloading package rslp to /root/nltk_data...
[nltk_data] | Unzipping stemmers/rslp.zip.
[nltk_data] | Downloading package rte to /root/nltk_data...
[nltk_data] | Unzipping corpora/rte.zip.
[nltk_data] | Downloading package sample_grammars to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping grammars/sample_grammars.zip.
[nltk_data] | Downloading package semcor to /root/nltk_data...
[nltk_data] | Downloading package senseval to /root/nltk_data...
[nltk_data] | Unzipping corpora/senseval.zip.
[nltk_data] | Downloading package sentence_polarity to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/sentence_polarity.zip.
[nltk_data] | Downloading package sentiwordnet to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/sentiwordnet.zip.
[nltk_data] | Downloading package shakespeare to /root/nltk_data...
[nltk_data] | Unzipping corpora/shakespeare.zip.
[nltk_data] | Downloading package sinica_treebank to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/sinica_treebank.zip.
[nltk_data] | Downloading package smultron to /root/nltk_data...
[nltk_data] | Unzipping corpora/smultron.zip.
[nltk_data] | Downloading package snowball_data to
[nltk_data] | /root/nltk_data...
[nltk_data] | Downloading package spanish_grammars to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping grammars/spanish_grammars.zip.
[nltk_data] | Downloading package state_union to /root/nltk_data...
[nltk_data] | Unzipping corpora/state_union.zip.
[nltk_data] | Downloading package stopwords to /root/nltk_data...
[nltk_data] | Unzipping corpora/stopwords.zip.
[nltk_data] | Downloading package subiectivity to
```

```

[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/subjectivity.zip.
[nltk_data] | Downloading package swadesh to /root/nltk_data...
[nltk_data] | Unzipping corpora/swadesh.zip.
[nltk_data] | Downloading package switchboard to /root/nltk_data...
[nltk_data] | Unzipping corpora/switchboard.zip.
[nltk_data] | Downloading package tagsets to /root/nltk_data...
[nltk_data] | Unzipping help/tagsets.zip.
[nltk_data] | Downloading package timit to /root/nltk_data...
[nltk_data] | Unzipping corpora/timit.zip.
[nltk_data] | Downloading package toolbox to /root/nltk_data...
[nltk_data] | Unzipping corpora/toolbox.zip.
[nltk_data] | Downloading package treebank to /root/nltk_data...
[nltk_data] | Unzipping corpora/treebank.zip.
[nltk_data] | Downloading package twitter_samples to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/twitter_samples.zip.
[nltk_data] | Downloading package udhr to /root/nltk_data...
[nltk_data] | Unzipping corpora/udhr.zip.
[nltk_data] | Downloading package udhr2 to /root/nltk_data...
[nltk_data] | Unzipping corpora/udhr2.zip.
[nltk_data] | Downloading package unicode_samples to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping corpora/unicode_samples.zip.
[nltk_data] | Downloading package universal_tagset to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping taggers/universal_tagset.zip.
[nltk_data] | Downloading package universal_treebanks_v20 to
[nltk_data] | /root/nltk_data...
[nltk_data] | Downloading package vader_lexicon to
[nltk_data] | /root/nltk_data...
[nltk_data] | Downloading package verbnet to /root/nltk_data...
[nltk_data] | Unzipping corpora/verbnet.zip.
[nltk_data] | Downloading package verbnet3 to /root/nltk_data...
[nltk_data] | Unzipping corpora/verbnet3.zip.
[nltk_data] | Downloading package webtext to /root/nltk_data...
[nltk_data] | Unzipping corpora/webtext.zip.
[nltk_data] | Downloading package wmt15_eval to /root/nltk_data...
[nltk_data] | Unzipping models/wmt15_eval.zip.
[nltk_data] | Downloading package word2vec_sample to
[nltk_data] | /root/nltk_data...
[nltk_data] | Unzipping models/word2vec_sample.zip.
[nltk_data] | Downloading package wordnet to /root/nltk_data...
[nltk_data] | Downloading package wordnet2021 to /root/nltk_data...
[nltk_data] | Downloading package wordnet2022 to /root/nltk_data...
[nltk_data] | Unzipping corpora/wordnet2022.zip.
[nltk_data] | Downloading package wordnet31 to /root/nltk_data...
[nltk_data] | Downloading package wordnet_ic to /root/nltk_data...
[nltk_data] | Unzipping corpora/wordnet_ic.zip.
[nltk_data] | Downloading package words to /root/nltk_data...
[nltk_data] | Unzipping corpora/words.zip.
[nltk_data] | Downloading package ycoe to /root/nltk_data...
[nltk_data] | Unzipping corpora/ycoe.zip.
[nltk_data] |
[nltk_data] Done downloading collection all

```

Out[ ]:

True

In [ ]:

```

def get_semantic(seq, key_word):

    # Tokenization of the sequence
    temp = word_tokenize(seq)

    # Retrieving the definition
    # of the tokens
    temp = lesk(temp, key_word)
    return temp.definition()

```

In [ ]:

```
keyword = 'book'
seq1 = 'I love reading books on coding.'
seq2 = 'The table was already booked by someone else.'

print(get_semantic(seq1, keyword))
print(get_semantic(seq2, keyword))
```

a number of sheets (ticket or stamps etc.) bound together on one edge  
arrange for and reserve (something for someone else) in advance

In [ ]:

```
keyword = 'book'
seq1 = 'I love reading lovbooks on coding.'
seq2 = 'The table was already booked by someone else.'

print(get_semantic(seq1, keyword))
print(get_semantic(seq2, keyword))
```

In [ ]:

```
keyword = 'jam'
seq1 = 'My mother prepares very yummy jam.'
seq2 = 'Signal jammers are the reason for no signal.'

print(get_semantic(seq1, keyword))
print(get_semantic(seq2, keyword))
```

press tightly together or cram  
deliberate radiation or reflection of electromagnetic energy for the purpose of disrupting  
enemy use of electronic devices or systems

In [ ]: