WORD SENSE DISAMBIGUATION USING GRAPH THEORY

SUBMITTED BY

CHHAVI SHARMA
DTU/2K14/IT/016

AYUSH AGGARWAL
DTU/2K14/IT/013

Introduction

WHAT IS WSD?

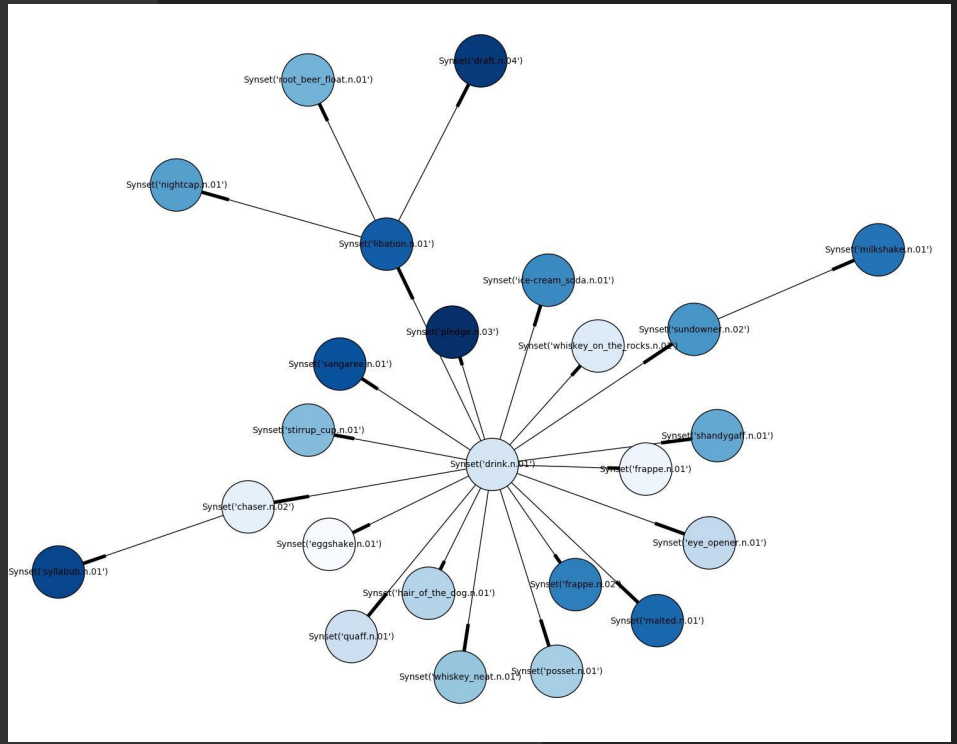
Word Sense Disambiguation is the process of identifying the sense of a polysemic word.

It is the task of identifying the intended meaning of the word in the given context out of the all the meanings possible of the word.

WORDNET

WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. WordNet is also freely and publicly available for download.

WordNet is available from http://wordnet.princeton.edu



GRAPH BASED WSD

Graph based algorithm works on the concept of identifying the most important node

Word meanings are plotted as nodes and their relations as edges

The Algorithm

Our WSD Algorithm exploits the relationships between two synsets such as synonymy , hyponymy and hypernymy to plot the graph and determine the important node

Code is available at https://github.com/ayush1252/WordSenseDisambgiuation

Step 1:- A query Q is entered in the form of an English sentence.

Step 2:- **Tokens** , the most important word phrases are extracted using **Rake Keyword Extractor** algorithm

Step 3:- Pick two words from the set of generated keywords and generate a subgraph G using **Wordnet**

To create a subgraph G

Initialise the Graph G to null

Create a set of synsets of both the keywords and add all those nodes to the graph and call them source and destination sets

Initialise a set of words/nodes that have been discovered yet

For every element x in syset of source:

Generate 3 sets of relations- **hypernymy** , **hyponymy** and **closure**

DFS the three subsets

(i) If the DFS length is greater than K , return

(ii) If the DFS discovered a node that was in the destination set , then

- Add all the nodes of the current path to the graph
- Add all the edges between all possible pair of the nodes
- Add all the nodes in the current path to the destination set

else follow through the DFS by repeating the process

Step 4:- For each Sub Graph G , calculate the **Local Centrality Measures** such as

- Degree Centrality
- Betweenness
- Hits
- Page Rank
- Closeness

Pick out the best source and destination context using all the above mentioned methods and pick out the ones

Step 5:- For each Sub Graph G , calculate the **Global Centrality Measures** such as

- Entropy
- Compactness
- Edge Density

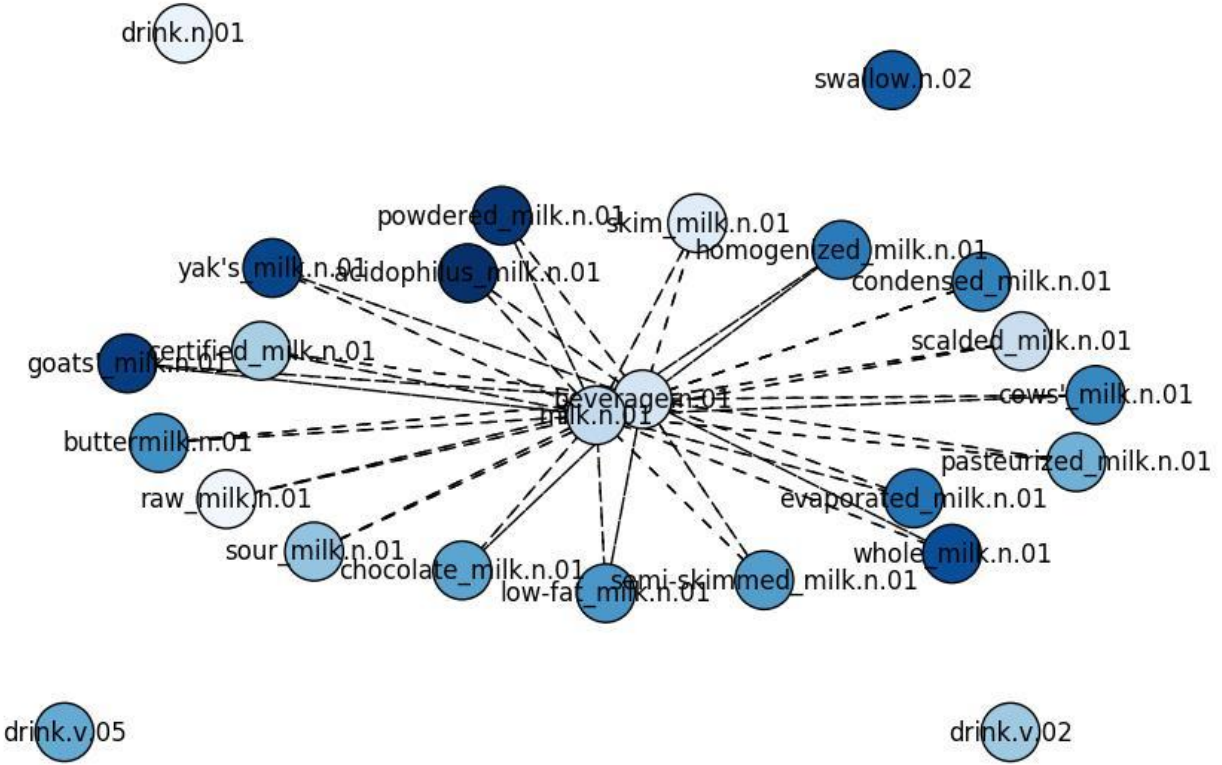
Use the global measures to weightage the chosen source and destination context and find out the final two source and destination synsets

WSD Algorithm

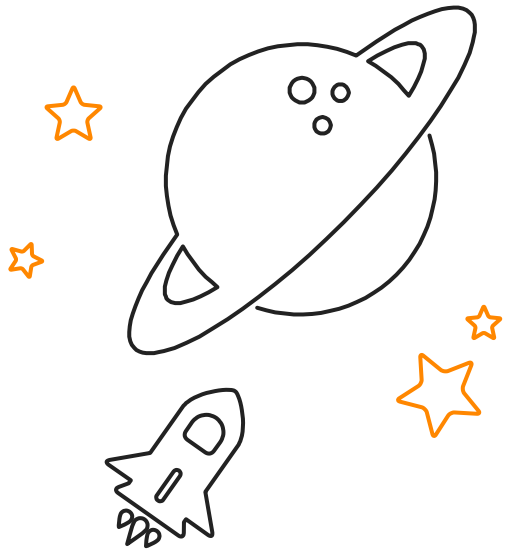
Our Algorithm takes input a sentence whose sense is to be disambiguate. It then uses a keyword extraction algorithm to extract content words only

- Generate sets of three relations , hyponymy meronymy and closure for the senses generated
- We perform a Depth First Search for all the words in those sets. Every time we encounter a node belong to destination along the path , we add the path to the senses

Graph Extracted from the algorithm after running on the keywords **drink** and **milk**



CENTRALITY MEASURES



Local and global measures, which determine the degree of relevance of a vertex v in graph G and the influence of a node over the network.

Local Measures

Degree Centrality

It is the simplest way to determine a vertex importance by its degree. The degree of a vertex refers to the number of edges incident on that vertex.

Betweenness Centrality

The betweenness centrality of a node ’v’ is the ratio of the shortest paths from one node to another node that are passing from ’v’ and the number of shortest paths between two nodes.

Local Measures

Closeness

Closeness or Key Player Problem(KPP) considers the importance of a vertex by its relative closeness with all the other vertices

PageRank

PageRank is one of the popular algorithms to rank the nodes or find the importance of a node in a network.

HITS

Two values are determined for a node, i.e. authority (a(v)) and hub value (h(v)).

Global Measures

Graph Entropy

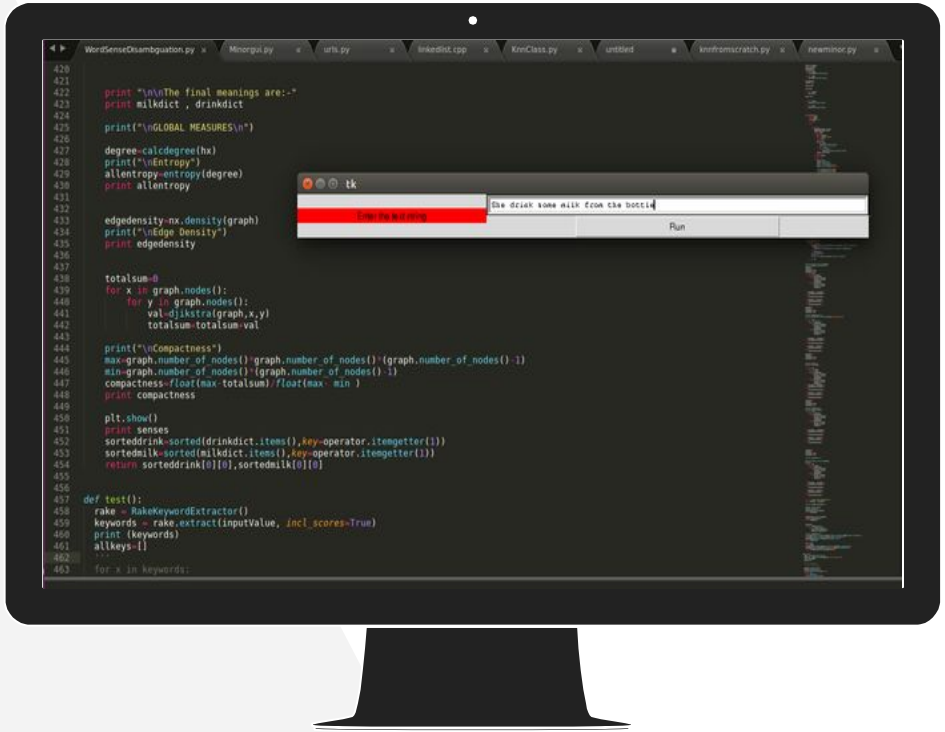
In a graph, high entropy indicates that many vertices are equally important, whereas low entropy indicates that only a few vertices are relevant.

Compactness

This measure represents the extent of cross referencing in a graph. When it is high, each vertex can be easily reached from other vertices.

Edge Density

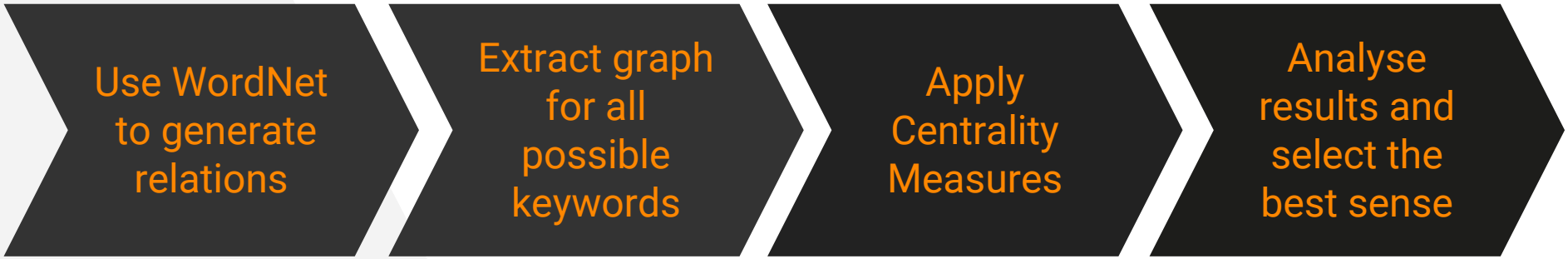
Edge density is calculated as the ratio of edges in a graph to the number of edges of a complete graph with |V | vertices.



Our Project

Python implementation of the above mentioned algorithm inculcating the nltk wordnet corpus along with networkX library to extract and analyze generate graph.

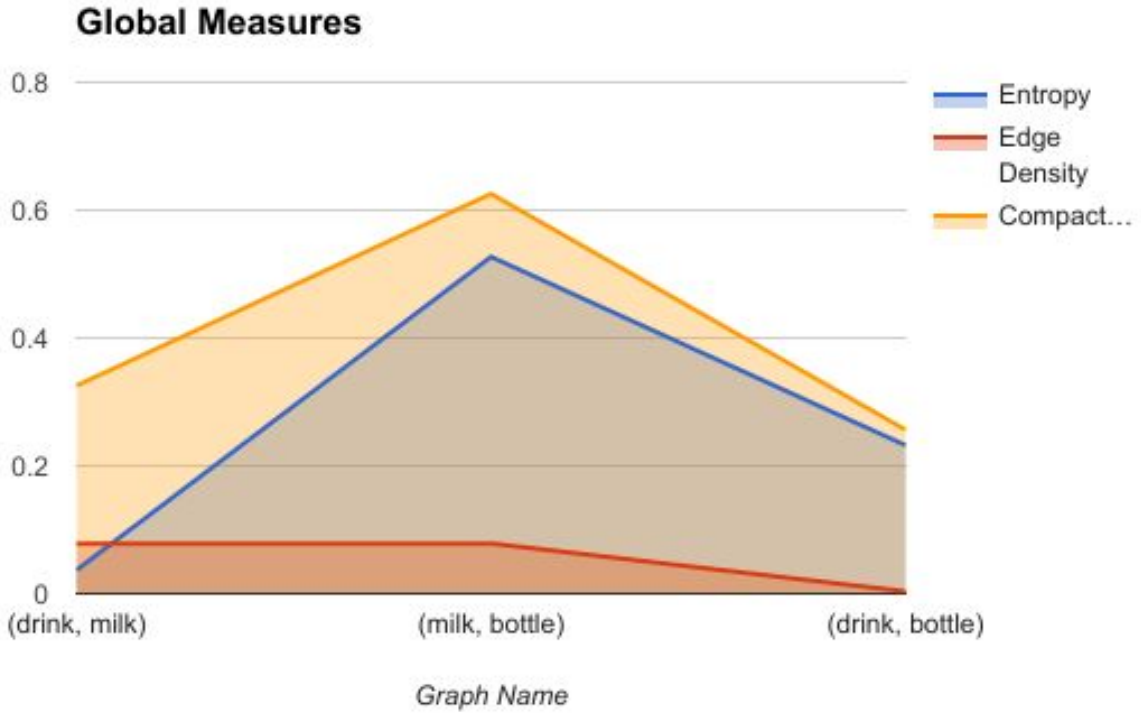
At a glance


Use WordNet to generate relations
Extract graph for all possible keywords
Apply Centrality Measures
Analyse results and select the best sense

Experimental Analysis

Global Centrality Measures for 'She drink some milk from the bottle'

Graph Name	Entropy	Edge Density	Compactness
(drink, milk)	**0.0371**	**0.0785**	**0.3255**
(milk, bottle)	**0.5270**	**0.0785**	**0.6255**
(drink, bottle)	**0.2321**	**0.0035**	**0.2565**


Global Measures
0.8
0.6
0.4
0.2
0
(drink, milk)
(milk, bottle)
(drink, bottle)
Entropy
Edge Density
Compact…
Graph Name

Experimental Analysis

Graph based algorithm provide a very fast method as compared to the sense-annotated supervised learning method.

Closeness and Degree Centrality were discovered as the most influential local methods while Entropy and Compactness were the global measures that distinguished the different graphs clearly

THANK YOU!