ALMA MATER STUDIORUM

UNIVERSITÀ DI BOLOGNA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ARTIFICIAL INTELLIGENCE

Graph-Based Keyword Extraction from Scientific Paper Abstracts using Word Embeddings

Author:
Dinno Koluh

Supervisor:
Prof. Paolo Torroni

Co-Supervisor
Dr. Federico Ruggeri

Bologna
16. December 2023.

Problem statement

- Problem of keyword extraction
  - Important words in text
  - Inherently a ranking problem
- Application to scientific paper abstracts
- Why model the problem as a graph?
  - Well-established model
  - Model text as a graph
  - Ranking algorithms
- The role of word embeddings

NLP Pipeline

- Text preprocessing
- Broken down into blocks

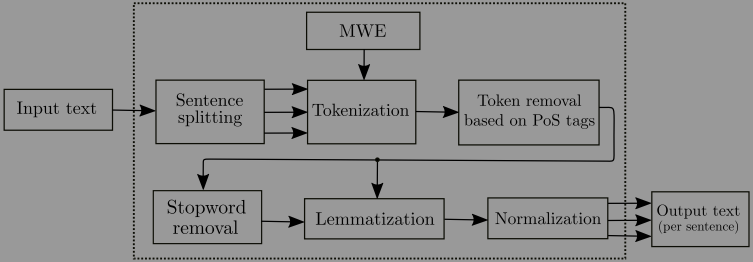
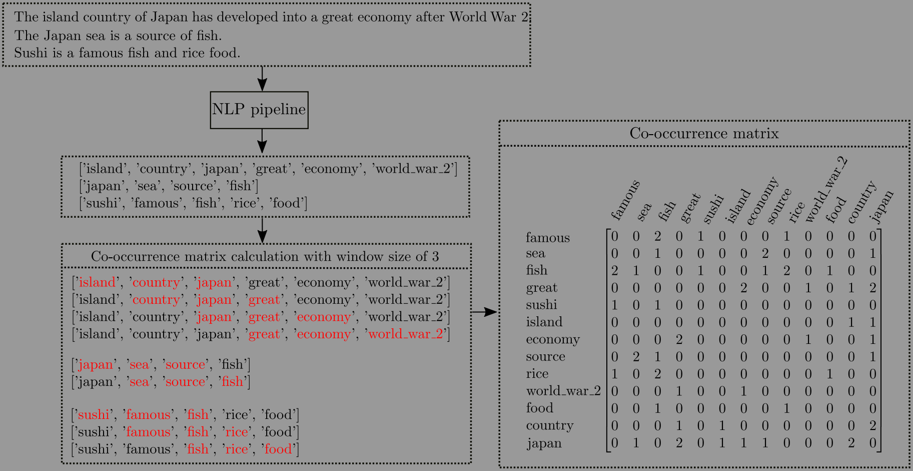


Figure 1: The NLP pipeline

Graph construction

- Distributional hypothesis



The island country of Japan has developed into a great economy after World War 2
The Japan sea is a source of fish.
Sushi is a famous fish and rice food.

NLP pipeline

['island', 'country', 'japan', 'great', 'economy', 'world_war_2']
['japan', 'sea', 'source', 'fish']
['sushi', 'famous', 'fish', 'rice', 'food']

Co-occurrence matrix calculation with window size of 3

['island', 'country', 'japan', 'great', 'economy', 'world_war_2']
['island', 'country', 'japan', 'great', 'economy', 'world_war_2']
['island', 'country', 'japan', 'great', 'economy', 'world_war_2']
['island', 'country', 'japan', 'great', 'economy', 'world_war_2']

['japan', 'sea', 'source', 'fish']
['japan', 'sea', 'source', 'fish']

['sushi', 'famous', 'fish', 'rice', 'food']
['sushi', 'famous', 'fish', 'rice', 'food']
['sushi', 'famous', 'fish', 'rice', 'food']

Co-occurrence matrix

$$
\begin{array}{l|ccccccccccccc}
 & \text{famous} & \text{sea} & \text{fish} & \text{great} & \text{sushi} & \text{island} & \text{economy} & \text{source} & \text{rice} & \text{world\_war\_2} & \text{food} & \text{country} & \text{japan} \\
\hline
\text{famous} & 0 & 0 & 2 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{sea} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 1 \\
\text{fish} & 2 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 0 & 1 & 0 & 0 \\
\text{great} & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 1 & 0 & 1 & 2 \\
\text{sushi} & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{island} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
\text{economy} & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
\text{source} & 0 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\text{rice} & 1 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
\text{world\_war\_2} & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{food} & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
\text{country} & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
\text{japan} & 0 & 1 & 0 & 2 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 2 & 0
\end{array}
$$

Figure 2: Co-occurrence matrix construction

Graph construction

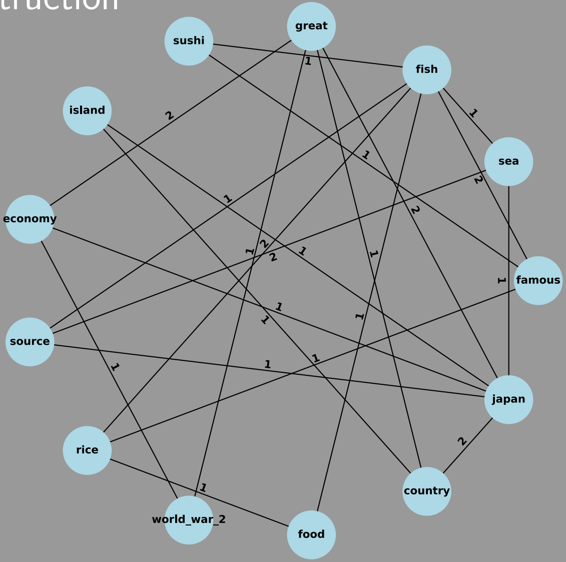


Figure 3: Graph representation of the co-occurrence matrix

Ranking algorithms

- Degree centrality (based on node degree)
- Closeness centrality (based on distance to other nodes)
- Betweenness centrality (based on number of shortest paths that pass through the node i.e. information flow)
- Eigenvector centrality (based on direct and neighbour connections)
- PageRank algorithm (Google's web page ranking algorithm applied to text)

Ranking algorithms

- Application on examples sentences:
- TABLE

Ranking algorithms

- Full pipeline for keyword extraction

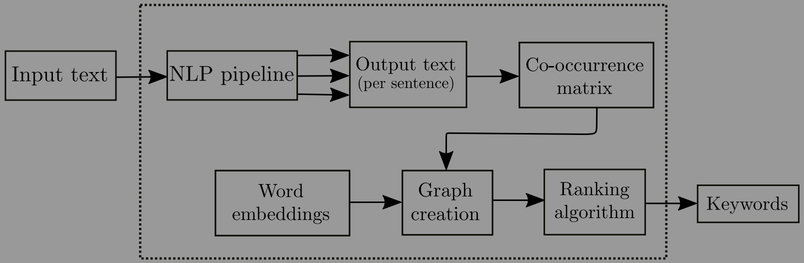


Figure 4: Full pipeline of keyword extraction

Experiments and Results

- Dataset: 5000 computer science paper abstracts with human assigned keywords, scraped from IEEE Xplore digital library
- Standard evaluation metrics (precision, recall, F-score)
- Important to address
    - Human assigned keywords are subjective and can be generated
    - Expected precision, recall and F-score from available literature in the range of $10 - 40\%$
    - Selecting the top n keywords calculated by the model, where n is the number of true keywords from the dataset

Experiments and Results

- TABLE

Q&A