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 (nodes and edges)
  - Ranking algorithms
  - Unsupervised
- The role of word embeddings
- Goal of thesis:
  - Verify results from available literature
  - Get insights in the area of information retrieval

NLP Pipeline

- Text preprocessing
- Broken down into blocks

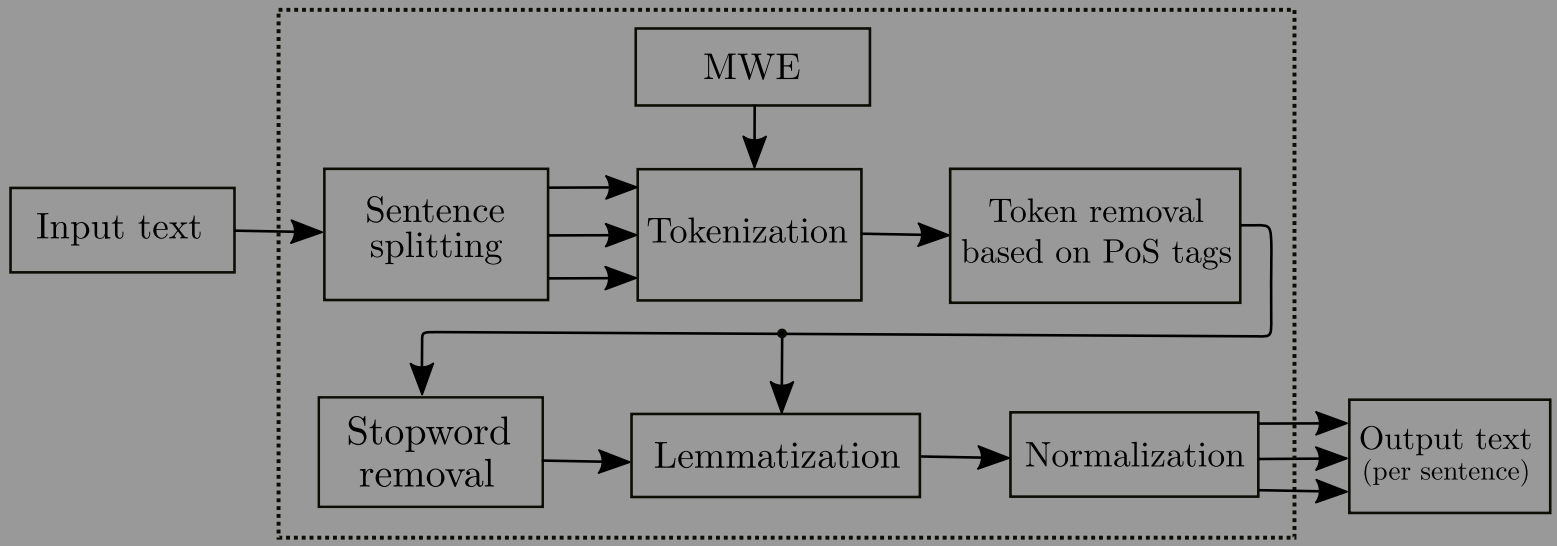

Figure 1: The NLP pipeline.

Graph construction

- Distributional hypothesis

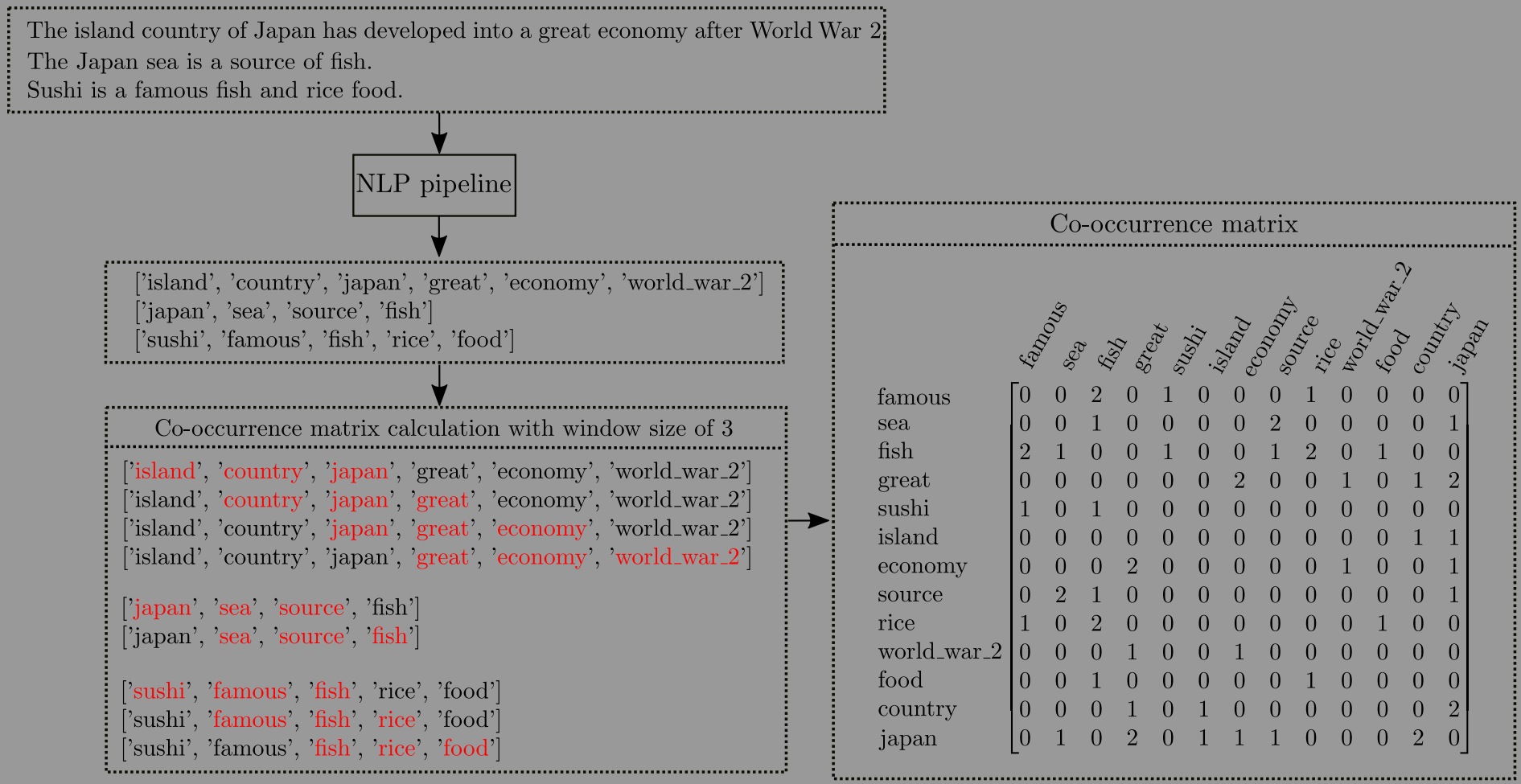

Figure 2: Co-occurrence matrix construction.

Graph construction

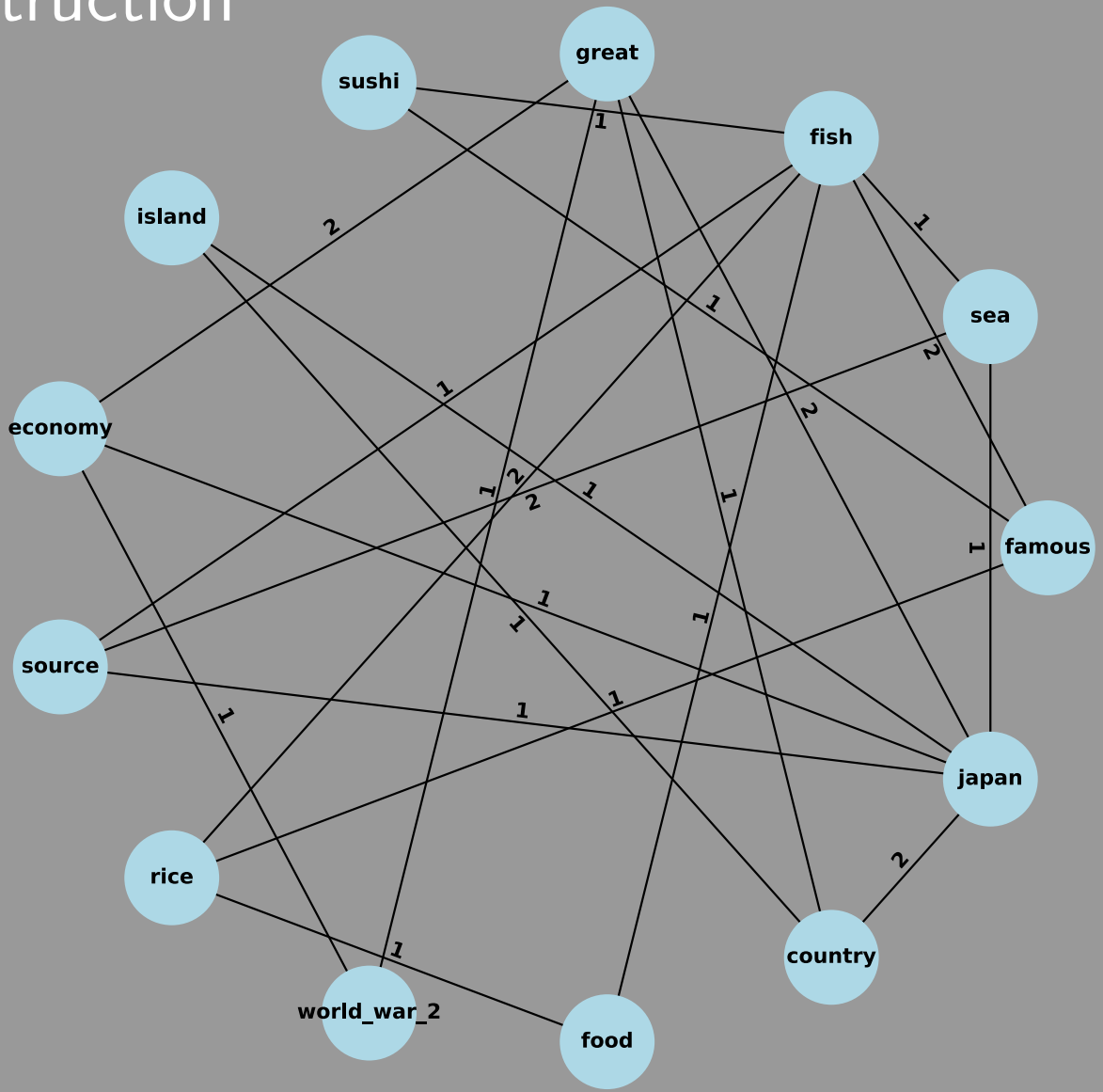

Figure 3: Graph representation of the co-occurrence matrix.

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

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.

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

Table 1: Ranking values.

Degree Centrality		Closeness Centrality		Betweenness Centrality		Eigenvector Centrality		PageRank	
Word	Ranking	Word	Ranking	Word	Ranking	Word	Ranking	Word	Ranking
japan	0.5	japan	0.462	japan	0.566	japan	0.515	fish	0.148
fish	0.5	source	0.444	fish	0.505	great	0.431	japan	0.148
great	0.333	sea	0.444	source	0.227	country	0.33	great	0.111
country	0.25	fish	0.429	sea	0.227	economy	0.309	famous	0.074
famous	0.25	island	0.353	economy	0.129	fish	0.279	rice	0.074
rice	0.25	economy	0.353	food	0.068	source	0.27	source	0.074
source	0.25	food	0.333	sushi	0.068	sea	0.27	sea	0.074
sea	0.25	sushi	0.333	island	0.064	island	0.171	country	0.074
economy	0.25	great	0.3	great	0.03	famous	0.164	economy	0.074
food	0.167	world_war_2	0.293	country	0.023	rice	0.164	food	0.037
sushi	0.167	country	0.293	famous	0.015	world_war_2	0.15	sushi	0.037
island	0.167	famous	0.273	rice	0.015	food	0.09	island	0.037
world_war_2	0.167	rice	0.273	world_war_2	0.011	sushi	0.09	world_war_2	0.037

Ranking algorithms

- Full pipeline for keyword extraction

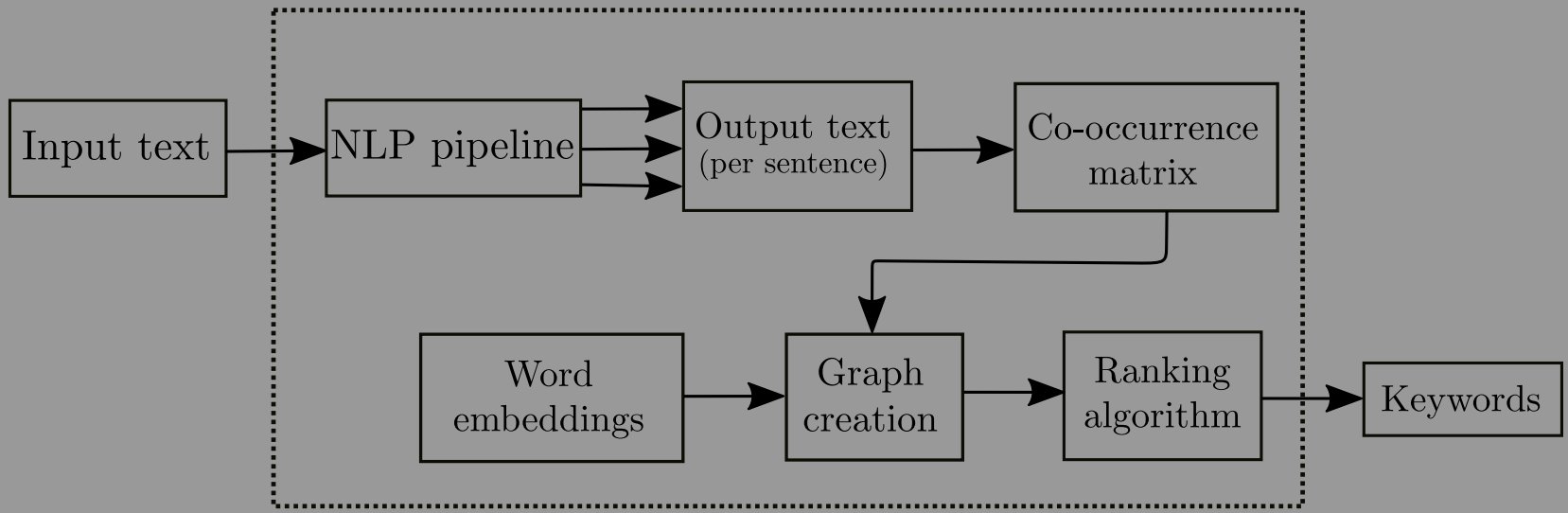

Figure 4: Full pipeline of keyword extraction

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 human-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 2: Keyword extraction results for precision recall and F-score for different window sizes.

Window Size	Ranking Algorithm	Precision (%)	Recall (%)	F-score (%)
3	Degree Centrality	29.37	33.19	31.16
	Closeness Centrality	25.33	28.62	26.87
	Betweenness Centrality	25.83	29.17	27.39
	Eigenvector Centrality	27.99	31.64	29.70
	PageRank	**29.60**	**33.45**	**31.40**
5	Degree Centrality	29.38	33.20	31.17
	Closeness Centrality	24.93	28.16	26.45
	Betweenness Centrality	25.66	28.98	27.21
	Eigenvector Centrality	27.93	31.57	29.63
	PageRank	29.37	33.19	31.16
10	Degree Centrality	28.55	32.26	30.28
	Closeness Centrality	24.03	27.13	25.48
	Betweenness Centrality	25.31	28.58	26.84
	Eigenvector Centrality	24.34	27.49	25.81
	PageRank	28.57	32.28	30.31

Q&A