# Knowledge-Based Categorization of Scientific Articles for Similarity Predictions

Nolwenn Bernard[1] Jonathan Weber[2] Germain Forestier[2] Michel Hassenforder[2]
Bastien Latard[1,2]

[1]MDPI AG, Basel, Switzerland

[2]Université de Haute-Alsace, IRIMAS, Mulhouse, France

25-27 August 2020

# Table of Contents

# Table of Contents

# Motivation

- Explosion of digital scientific articles, more than 3 million in 2018
  - Processing is time consuming for researchers

*https://www.the-scientist.com/the-nutshell/data-overflow-compromising-science-34785*

- Find relevant articles

# Table of Contents

# Proposed Approach


Figure 1: Global Workflow

# Categorization - Original approach

BabelNet[1]

- Multilingual encyclopedic dictionary and semantic network
- Integration of semantic lexicons and collaborative databases

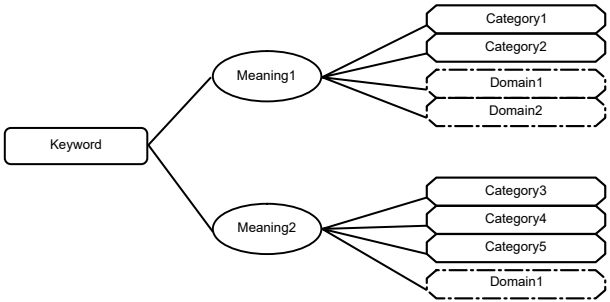


Figure 2: BabelNet architecture

[1] Roberto Navigli and Simone Paolo Ponzetto. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network". In: *Artificial Intelligence* 193 (2012), pp. 217–250.

## Categorization

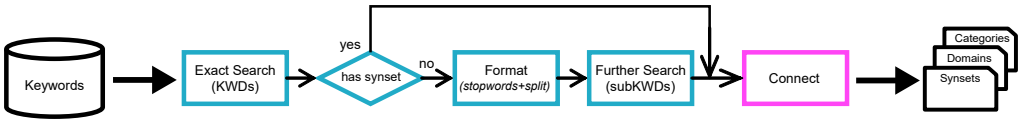


Figure 3: Simplified workflow of the categorization stage

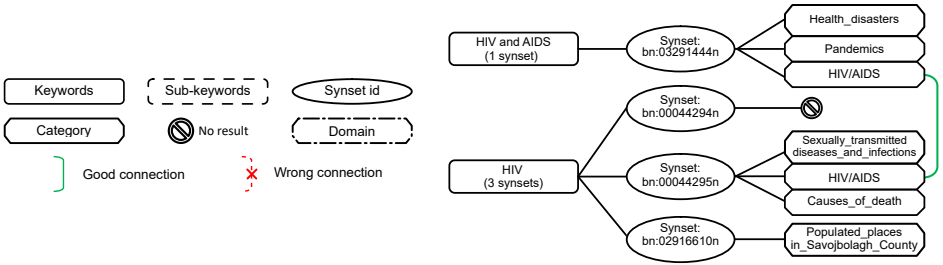


Figure 4: Simple connection by categories

## Categorization

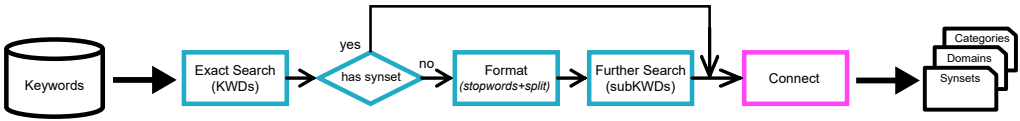


Figure 3: Simplified workflow of the categorization stage

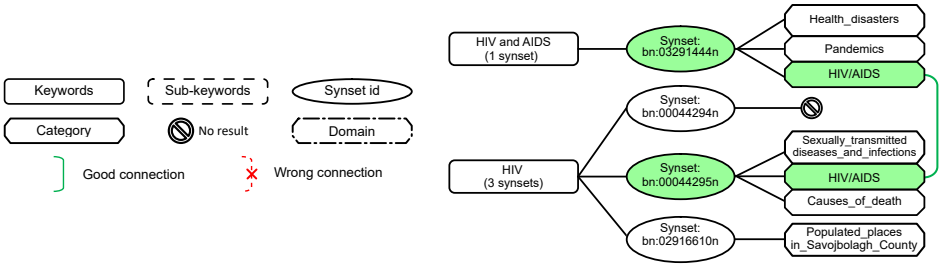


Figure 4: Simple connection by categories

# Categorization - Proposed Approach

- Assign all categories from connected synsets
- Synsets connections by domains

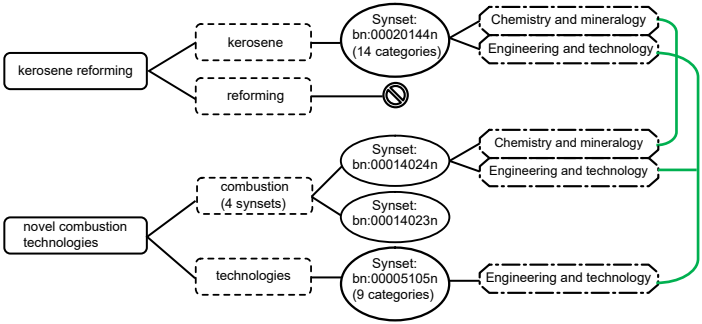


Figure 5: Domains connection

# Categorization - Proposed Approach

- Assign all categories from connected synsets
- Synsets connections by domains

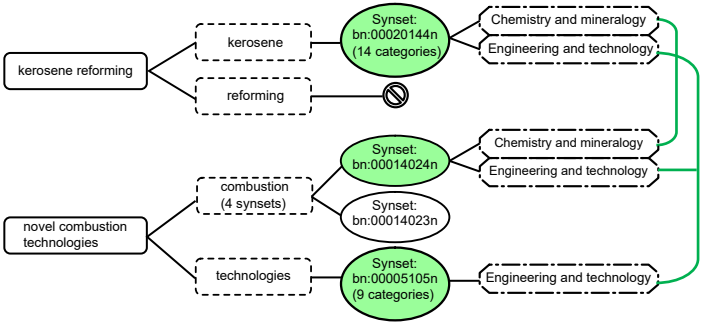


Figure 5: Domains connection

# Proposed Approach


- Data augmentation: retrieve semantic neighbors sharing at least one category with the article
- Similarity metric[2]: based on weighted Jaccard indexes
  - Keyword intersection
  - Keyword-Neighbor intersection
  - Neighbor intersection

[2] Bastien Latard. "Scientific Search Engines: From the Categorization to the Information Retrieval". PhD thesis. Université de Haute-Alsace, 2019.

# Table of Contents

# Experiment Results

- Web of Science Dataset WOS-46985[3]: 46,985 articles divided in 7 domains and 134 categories
- Evaluation using article's domain
- w2v-cos approach computes cosine similarity using article's mean vector

[3] Kamran Kowsari et al. "Hdltex: Hierarchical deep learning for text classification". In: *ICMLA*. IEEE, 2017, pp. 364–371.

# Experiment Results

| Approach | $\overline{P@k}$(%) | Article coverage (%) |
|---|---|---|
| w2v-cos | **19** | 99.9 |
| Original | 25 | 46.8 |
| Proposed | **49** | 88.2 |

Average for $k$ between 1 and 100

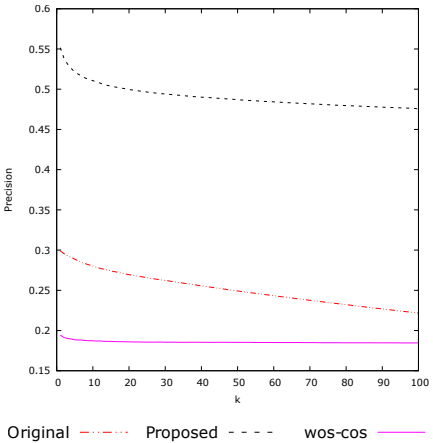


Figure 6: P@k

# Experiment Results

| Approach | $\overline{P@k}$(%) | Article coverage (%) |
|---|---|---|
| w2v-cos | 19 | 99.9 |
| Original | **25** | **46.8** |
| Proposed | **49** | **88.2** |

Average for $k$ between 1 and 100

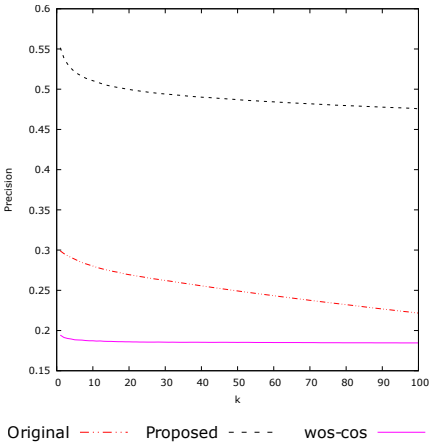


Figure 6: P@k

# Table of Contents

# Conclusion

- The results show that the proposed approach can compete with probabilistic methods
- The proposed approach achieves 88.2% of coverage against 48.6% for the original approach
- The average $P@k$ for the proposed approach is 49%
- Future works: study of reproducibility and comparison with other probabilistic methods to support first assumption

**Thank you for your attention**