Supplementary information: Concepts & texts in the practice of life science. The case of “signaling”

Wiktor Rorot

October 18, 2024

Data & methods

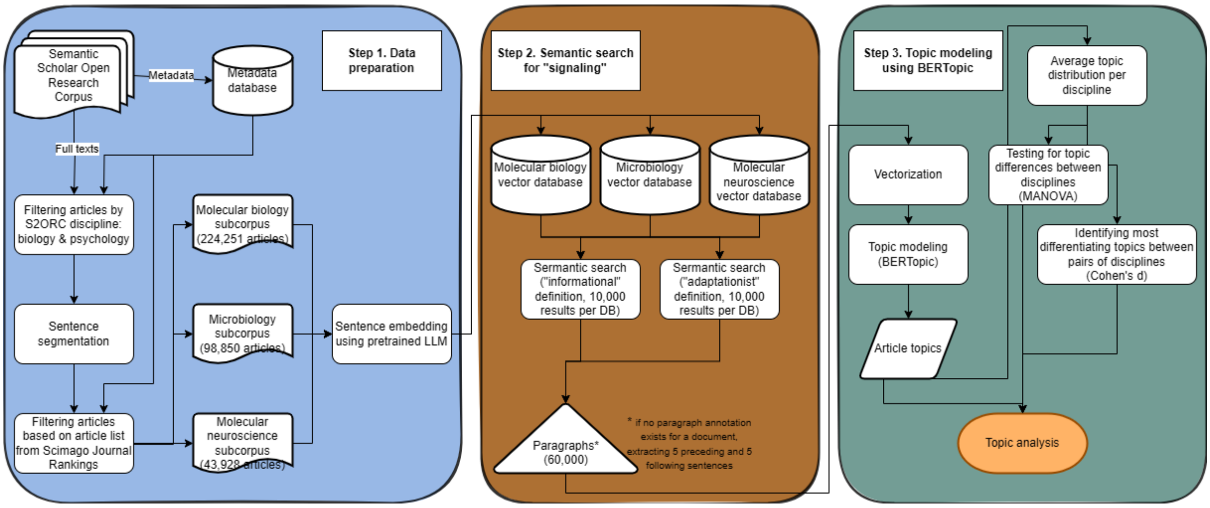
Step 1. Data preparation
Semantic Scholar Open Research Corpus
Metadata
Metadata database
Full texts
Filtering articles by S2ORC discipline: biology & psychology
Sentence segmentation
Filtering articles based on article list from Scimago Journal Rankings
Molecular biology subcorpus (224,251 articles)
Microbiology subcorpus (98,850 articles)
Molecular neuroscience subcorpus (43,928 articles)
Sentence embedding using pretrained LLM
Step 2. Semantic search for "signaling"
Molecular biology vector database
Microbiology vector database
Molecular neuroscience vector database
Semantic search ("informational" definition, 10,000 results per DB)
Semantic search ("adaptationist" definition, 10,000 results per DB)
Paragraphs* (60,000)
* if no paragraph annotation exists for a document, extracting 5 preceding and 5 following sentences
Step 3. Topic modeling using BERTopic
Vectorization
Topic modeling (BERTopic)
Article topics
Average topic distribution per discipline
Testing for topic differences between disciplines (MANOVA)
Identifying most differentiating topics between pairs of disciplines (Cohen's d)
Topic analysis

Figure 1: Figure S1. Flowchart presenting methodology of the study.

Dataset is based on the Semantic Scholar Open Research Corpus [S2ORC, release from August 2024; Lo et al. (2020)]. For identifying scientific fields, journal lists from Scimago Journal Ranking (2023 edition) were used for each discipline. For text processing, Python and the package spaCy (Honnibal and Montani 2017) were used. For sentence embedding, the SentenceTransformers package with pretrained `multi-qa-MiniLM-L6-cos-v1` model (Reimers and Gurevych 2019). For semantic search, a Milvus database (Wang et al. 2021) using the sentence embeddings was created and queried. For each dataset and definition (“adaptationist” or “semantic”) 10,000 semantically closest sentences were identified. For each sentence, using the metadata in S2ORC, I have found the corresponding paragraph. If no paragraphs could be found (not all articles in S2ORC have paragraph annotations), 5 preceding and 5 following sentences were used as an approximation. The resulting corpus of 59,100 paragraphs was vectorized and embedded using the SentenceTransformers package with pretrained `all-MiniLM-L6-v2` model. Paragraph vectors were used to trained a BERTopic model (Grootendorst 2022). The topic representation was fine-tuned by removing most frequent [based on a stop word list provided with the scikit-learn Python package; Pedregosa et al. (2011)] and least frequent terms, and selecting only nouns, adjectives, or noun and adjective pairs. No topic reduction was performed resulting in automatic detection of 163 topics. Outliers (documents not classified to any topic) were removed using the appropriate BERTopic function, by classifying them acccording to the most frequent topic within the document. The model was evaluated using standard coherence scores, $C_v = 0.677157502559966$ and $U_{mass} = -3.175586425117176$ (Mimno et al. 2011). The topic distributions within disciplines were analyzed using methodology proposed by (Lawley et al. 2023), using MANOVA [from the Statsmodels package; Seabold and Perktold (2010); isometric log-ratio transformation using the Scikit-bio package] to test if topic distributions are identical between groups and η^2 to evaluate effect size. For the purpose of topic distribution analysis, only topic distributions for paragraphs from articles that have unambiguous disciplinary membership

Table 1: Table S1. $N = 10$ most distinctive topics for each pair (by Cohen's d). Mi-Ne - Microbiology vs. Molecular Neuroscience, Mi-Mo - Microbiology vs. Molecular Biology, Ne-Mo - Molecular Neuroscience vs. Molecular Biology

Mi-Ne	Mi-Mo	Ne-Mo
2 : 0.163487	2 : 0.152754	139 0.195651
12: 0.144260	139: 0.146577	0 0.136896
0 : 0.143777	77: 0.075825	12 0.129947
82: 0.103929	23: 0.073181	82 0.101314
71: 0.086895	122 : 0.072909	14 0.091880
91: 0.082464	32: 0.071220	77 0.082301
15: 0.081965	46: 0.069545	10 0.077460
65: 0.078874	93: 0.062745	41 0.075498
14: 0.078318	9 : 0.062039	80 0.072106
23: 0.078028	75: 0.061762	127 0.070788

Table 2: Table S2. $N = 10$ least distinctive topics for each pair (by Cohen's d). Mi-Ne - Microbiology vs. Molecular Neuroscience, Mi-Mo - Microbiology vs. Molecular Biology, Ne-Mo - Molecular Neuroscience vs. Molecular Biology

Mi-Ne	Mi-Mo	Ne-Mo
57: 0.000060	106: 0.000035	74: 0.000185
27: 0.000068	51: 0.000036	8 : 0.000255
31: 0.000112	160: 0.000162	78: 0.000266
75: 0.000370	132: 0.000166	137: 0.000418
42: 0.000374	37: 0.000528	162: 0.000515
70: 0.000387	136: 0.000667	145: 0.000900
67: 0.000526	24: 0.001239	33: 0.001109
49: 0.000623	42: 0.001278	102: 0.001496
25: 0.000663	82: 0.001434	124: 0.001500
145: 0.000903	88: 0.001575	42: 0.001652

were used.

Results

The topic distributions are significantly different between disciplines, $F(162, 59099) = 113.2853$, but the effect is not large with $\eta^2 = 0.08620111678651099$.

For establishing which topics are most differentiating between pairs of disciplines, I have calculated Cohen's d for the probability of each topic. The results are in Tables S1 and S2. The effect sizes are quite small, with highest Cohen's $d \approx 0.20$, which corresponds to the weak effect size for the differences between disciplines. As visualized in Figure S2, this is likely due to the fact that large number of topics results in a low average probability for each individual topic. However, no further topic reduction was performed as the topics generated by BERTopic offer good interpretability at this level of granularity, compared to tested topic reductions.

References

Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." https://arxiv.org/abs/2203.05794.

Honnibal, Matthew, and Ines Montani. 2017. "spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing."

Lawley, Grace O., Peter A. Heeman, Jill K. Dolata, Eric Fombonne, and Steven Bedrick. 2023. "A Statistical Approach for Quantifying Group Difference in Topic Distributions Using Clinical Discourse Samples." In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and*

Dialogue, edited by Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, 55–65. Prague, Czechia: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.sigdial-1.5.

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. "S2ORC: The Semantic Scholar Open Research Corpus." July 6, 2020. http://arxiv.org/abs/1911.02782.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, edited by Regina Barzilay and Mark Johnson, 262–72. Edinburgh, Scotland, UK.: Association for Computational Linguistics. https://aclanthology.org/D11-1024.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084.

Seabold, Skipper, and Josef Perktold. 2010. "Statsmodels: Econometric and Statistical Modeling with Python." In *9th Python in Science Conference*.

Wang, Jianguo, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, et al. 2021. "Milvus: A Purpose-Built Vector Data Management System." In *Proceedings of the 2021 International Conference on Management of Data*, 2614–27. Virtual Event China: ACM. https://doi.org/10.1145/3448016.3457550.

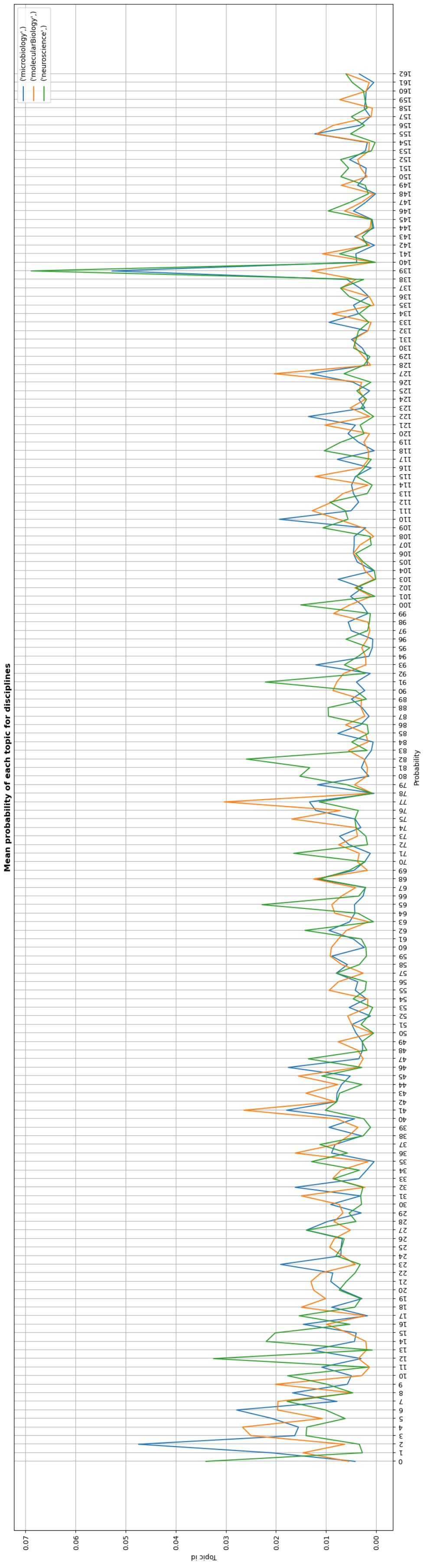

Figure 2: Figure S2. Distribution of topics for each discipline.