FoRC4CL: A Fine-grained Field of Research Classification and Annotated Dataset of NLP Articles

Raia Abu Ahmad, Ekaterina Borisova, Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Berlin, Germany
raia.abu_ahmad@dfki.de, ekaterina.borisova@dfki.de, georg.rehm@dfki.de

Abstract

The steep increase in the number of scholarly publications has given rise to various digital repositories, libraries and knowledge graphs aimed to capture, manage, and preserve scientific data. Efficiently navigating such databases requires a system able to classify scholarly documents according to the respective research (sub-)field. However, not every digital repository possesses a relevant classification schema for categorising publications. For instance, one of the largest digital archives in Computational Linguistics (CL) and Natural Language Processing (NLP), the ACL Anthology, lacks a system for classifying papers into topics and sub-topics. This paper addresses this gap by constructing a corpus of 1,500 ACL Anthology publications annotated with their main contributions using a novel hierarchical taxonomy of core CL/NLP topics and sub-topics. The corpus is used in a shared task with the goal of classifying CL/NLP papers into their respective sub-topics.

Keywords: Corpus, taxonomy construction, annotation, field of research classification

1. Introduction

In recent years, there has been an exponential increase in scientific publications (Fortunato et al., 2018; Bornmann et al., 2021). As a result, a wide range of established or emerging repositories, databases, knowledge graphs, and digital libraries aim to capture and effectively manage scientific knowledge. Notable examples include the Open Research Knowledge Graph (ORKG, Jaradeh et al., 2019), the Semantic Scholar Academic Graph (S2AG, Kinney et al., 2023), and the ACL Anthology (Bird et al., 2008), which is tailored to Computational Linguistics (CL). One fundamental task of such repositories involves categorising scientific knowledge into specific research fields and their topics and sub-topics, which is a crucial prerequisite for tracking scientific progress adequately and developing applications such as scientific search engines and recommender systems.

However, when it comes to Computational Linguistics and Natural Language Processing (NLP), we note that although the ACL Anthology is a well-known and comprehensive resource in this domain, it does not follow a classification schema that labels its publications into specific topics or sub-topics. The lack of a classification complicates navigating papers in the ACL Anthology and hinders developers in the field of scholarly information processing from developing effective tools that assist CL/NLP researchers and students.

We address this gap by constructing a corpus of CL/NLP publications extracted from the ACL Anthology (CC BY 4.0) annotated with the topics and sub-topics to which they contribute. The corpus is mainly constructed to deal with the task of Field of Research Classification (FoRC) for the field of CL and is thus named FoRC4CL. For the annotation process, we propose and make use of a novel fine-grained taxonomy that includes specific CL/NLP topics and sub-topics. Our goal is to provide a first stepping stone for a community-driven hierarchical classification model of CL/NLP publications. Our main contributions are:

- **FoRC4CL**, a human-annotated corpus of 1,500 CL/NLP publications according to their topics.[1]
- **Taxonomy4CL**, a novel fine-grained taxonomy of 170 CL/NLP topics and sub-topics in three hierarchical levels that can be used for categorising publications in the CL/NLP field.[2]
- A foundation for the structured categorisation of CL/NLP knowledge and research, which could serve bibliometric studies, as well as the development of CL/NLP research-assisting tools.

We provide these resources to the CL/NLP community and would like to work with their feedback on the expansion and extension of both Taxonomy4CL and the FoRC4CL corpus. If this taxonomy (or a similar one linked to it) is adopted by a CL/NLP research repository, we aim for each author to categorise their own work so that the FoRC4CL corpus can grow naturally. Additionally, the current FoRC4CL corpus has recently been used for training and evaluating a shared task on hierarchical fine-grained FoRC of CL publications (Abu Ahmad et al., 2024).[3]

[1] https://zenodo.org/records/10777674
[2] https://github.com/DFKI-NLP/Taxonomy4CL
[3] https://nfdi4ds.github.io/nslp2024/

LREC-COLING 2024, pages 7389–7394
20-25 May, 2024.

The remainder of this paper discusses related previous work (Section 2), our methods in constructing Taxonomy4CL (Section 3) and annotating the FoRC4CL corpus (Section 4). We address the limitations of our work (Section 5) and conclude with remarks on future research (Section 6).

2. Related Work

Various research fields have developed standard classification schemas that are utilised to label publications in a fine-grained manner. Examples include the Medical Subject Headings (MeSH)[4] used in PubMed (Canese and Weis, 2013) for the field of medicine, the GESIS controlled vocabulary[5] used in the Social Science Open Access Repository (SSOAR)[6] for the field of social sciences, the Physics Subject Headings (PhySH)[7] used in American Physical Society journals[8] for the field of physics, and the Mathematics Subject Classification (MSC)[9] used in the American Mathematical Society[10] publications for the field of mathematics.

While some ontologies and taxonomies that include CL/NLP topics exist, they are not definitive or fine-grained enough to classify CL/NLP publications. For example, the Computer Science Ontology (CSO, Salatino et al., 2018) has a subsection dedicated to NLP, however, it was automatically developed with no human curation and thus contains a lot of noise (e. g., *statistical machine translation* is a child node of *speech transmission* with no clear connection to other types of machine translation). It also misses some core topics (e. g., specific classification tasks such as *hate speech detection*). The ACM Computing Classification System (CSS, Rous, 2012) also has some labels related to NLP, but it is not granular enough and does not cover many core tasks such as *sentiment analysis*. PapersWithCode,[11] an openly available resource that tracks progress in machine learning, has a list of NLP tasks. However, many of its related topics are not interlinked and thus do not show the connections across CL/NLP tasks (e. g., *visual question answering* is separate from *question answering*).

Finally, some efforts have been made to design taxonomies specifically for CL/NLP, however, they are more suitable for educational resources and none of them have been used to label a large corpus of publications in the field. CLICKER (Hingmire et al., 2021) is a recent effort that developed a three-level hierarchical taxonomy by extracting keywords from CL/NLP lecture slides. This was then used to build an educational platform with manually labelled lectures and tutorials. Another resource is nn4nlp-concepts,[12] which tracks concepts related to neural networks and thus does not cover the full research topics of CL/NLP. The NLP Index[13] automatically tracks a limited non-hierarchical list of NLP tasks but is not comprehensive enough, missing topics such as *low-resource languages*.

3. Taxonomy Construction

Since no definitive, fine-grained classification schema specific to CL/NLP exists, we developed our own taxonomy by fetching the ca. 41,500 abstracts classified as *Computation and Language* in the arXiv dataset.[14] We then ran BERTopic (Grootendorst, 2022), a model that clusters similar topic representations from a document corpus, on the abstracts. For BERTopic, the default parameters were used and the random state was set to 42. As a result, more than 300 clusters were extracted,[15] which we manually curated to 170 topics. The criteria for excluding a topic were the following: 1. impossible to define the topic due to the ambiguity of keywords; 2. keywords are misleading and do not reflect the main topic of a paper; 3. the topic is too narrow, appearing in less than ten papers in our corpus. After manually filtering clusters of keywords, we converted them to human-readable labels by prompting ChatGPT[16] as follows: *"I extracted a topic from Computational Linguistics scholarly papers. The topic is described by the following keywords: keyword1, keyword2, keyword3, keyword4. Based on the information above, extract a short topic label in the following format: topic: <topic label>"*. We then constructed the hierarchical taxonomy manually, aligning it with: 1. ACM Subject Classification[17], 2. the Computer Science Ontology (CSO), specifically, its NLP sub-branch[18], 3. the

[4] https://meshb.nlm.nih.gov
[5] https://lod.gesis.org/thesoz/en/
[6] https://www.gesis.org/en/ssoar/home
[7] https://physh.org
[8] https://journals.aps.org
[9] https://mathscinet.ams.org/mathscinet/msc/msc2020.html
[10] https://www.ams.org
[11] https://paperswithcode.com/area/natural-language-processing
[12] https://github.com/neulab/nn4nlp-concepts
[13] https://index.quantumstat.com
[14] https://www.kaggle.com/datasets/Cornell-University/arxiv
[15] The raw output is available at https://github.com/DFKI-NLP/Taxonomy4CL/blob/main/data/results/BERTopic_output.csv, https://github.com/DFKI-NLP/Taxonomy4CL/blob/main/data/results/doc_info_abstracts.csv
[16] https://openai.com/blog/chatgpt
[17] https://dl.acm.org/ccs
[18] https://cso.kmi.open.ac.uk/topics/natural_language_processing

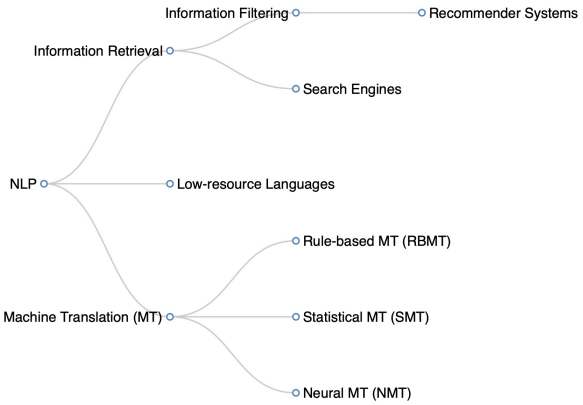

Figure 1: A sample of the developed CL taxonomy

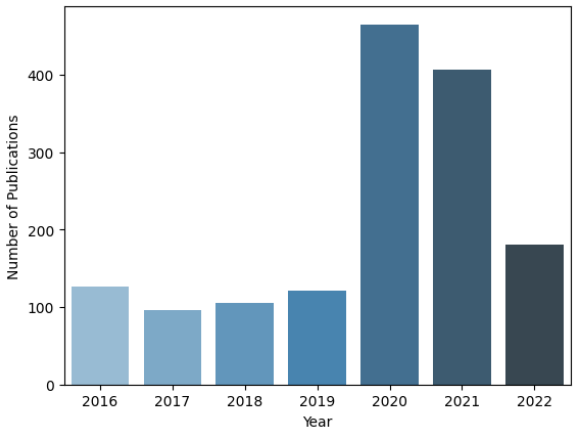

Figure 2: Distribution of publication years in FoRC4CL

PapersWithCode classification of NLP methods,[19] 4. the Oxford Handbook of Computational Linguistics (Mitkov, 2022), and 5. Wikipedia articles on CL/NLP tasks and methods. The final taxonomy[20] has a three-level hierarchical structure, see Figure 1 for a sample.

Importantly, this approach is also applicable to the development of taxonomies in and for other scientific fields. The only prerequisite is at least one larger article repository, which is perceived as being representative of the overall research output produced by the respective field.

4. Corpus Construction

To develop the dataset, we retrieved 1,500 randomly selected English publications from the ACL Anthology,[21] spanning the years 2016 to 2022 (see Figure 2). This timeframe was selected based on our taxonomy construction process, in which we observed a prominent surge in publications during these years, signifying the relevance of the extracted topics to this period. The paper selection process ensured randomness and proportionality to venue size, with smaller venues receiving appropriately scaled representation. The corpus encompasses 255 distinct venues, with the number of papers per venue varying from 30 (main ACL Conference) to one (e. g., GAMESandNLP, AI4HI, NL4XAI, etc.), averaging approximately six papers per venue. Each publication in the corpus is accompanied by the following metadata fields: ACL Anthology ID, title, abstract, author(s), URL to the full text, publisher, publication year and month, proceedings title, Digital Object Identifier (DOI), and venue (see Table 1 for a sample).[22]

To annotate the corpus with labels from Taxonomy4CL, we conducted an annotation project with six Master students of CL. The selection criteria for hiring the students as research assistants were: 1. Excellent theoretical knowledge in CL/NLP based on graded university courses, 2. Proficiency in English at a high level, and 3. Preferably, prior experience in annotation tasks. We used INCEpTION (Klie et al., 2018) as our annotation tool because of its web-based access, which allows for multiple users and roles, as well as its functionalities of overlapping labels and automatic inter-annotator agreement (IAA) calculation (Borisova et al., 2024). Annotators were able to view the title and abstract of each paper before annotating it, with a link to the full paper in PDF. While abstracts typically provide insights into a publication's main contributions, the annotators were required to consult the full PDF file, i. e., the whole paper, as detailed information about constructed datasets (e. g., language specifications) and model architectures is often exclusively available therein. Before starting the project, annotators underwent a training period that involved reading and familiarising themselves with the guidelines and the taxonomy, as well as two rounds of trial annotations in which they worked on 50 random papers from the corpus.

The six annotators were divided into three pairs each assigned 500 random documents to annotate. After annotating, the documents were reviewed by a curator who handled solving annotation conflicts and approving the final annotations. Two curators, the first and second authors of this paper, worked on the project such that each curator was assigned 750 documents. The quality of annotations is evaluated based on IAA scores using Krippendorff's

[19] https://paperswithcode.com/methods/area/natural-language-processing

[20] https://github.com/DFKI-NLP/Taxonomy4CL/blob/main/data/Taxonomy4CL/Taxonomy4CL_v1.0.0.json

[21] https://github.com/shauryr/ACL-anthology-corpus

[22] The complete corpus is publicly accessible at https://zenodo.org/records/10777674

ACL ID	Level 1	Level 2	Level 3
2020.crac-1.12	['Learning Paradigms', 'Data Management and Generation', 'Information Extraction']	['Active Learning', 'Data Preparation', 'Coreference Resolution']	['Annotation Processes']
2021.privatenlp-1.3	['Information Extraction', 'Ethics', 'Model Architectures', 'Domain-specific NLP']	['Transformer Models', 'Medical and Clinical NLP']	–
2021.case-1.4	['Domain-specific NLP', 'Learning Paradigms', 'Information Extraction']	['NLP for News and Media', 'Unsupervised Learning', 'Event Extraction']	–

Table 1: A sample of the FoRC4CL corpus (note that not all metadata fields are shown)

	Level 1	Level 2	Level 3	Average
Pair #1	0.62	0.58	0.52	0.57
Pair #2	0.63	0.59	0.42	0.55
Pair #3	0.62	0.58	0.51	0.57
Average	0.62	0.58	0.48	**0.56**

Table 2: Inter-annotator agreement scores of publications' main contributions using Krippendorff's alpha per annotation pair and taxonomy level.

Alpha for multi-label annotations on each of the three taxonomy levels. According to Krippendorff (2004), a score of at least 0.8 is considered a reliable quality threshold, and tentative conclusions are acceptable with a score of at least 0.667.

However, it is important to note that this annotation task is complex in many aspects. First, although all the data is extracted from the same source (the ACL Anthology), it covers a wide range of different topics and venues, and thus the data is diverse and heterogeneous. Additionally, CL/NLP topics are frequently intertwined and interlinked, resulting in similar labels in the taxonomy with subtle differences (e. g., *Adversarial Attacks and Robustness* vs. *Adversarial Learning* under *Learning Paradigms*), which may confuse annotators and make it more difficult to choose an appropriate label. The varying difficulty of papers is also important, as some annotated papers fit the taxonomy perfectly, while others are more difficult and ambiguous, oftentimes lying in the intersection of two labels. Further, the large number of 170 labels itself is another complex factor, since this demands a larger memory load and can affect annotators' ability to distinguish between different labels or to miss and forget about specific labels. Finally, although we tried to hire annotators who are all on the same level of domain expertise and have similar English language skills, it is inevitable to have individual differences in language understanding skills and expertise in certain CL/NLP areas. These differences can result in biases and diverse outlooks on the same publication, resulting in different annotations. These aspects have all been proven to significantly affect the IAA score, making it challenging to achieve a relatively high number (Bayerl and Paul, 2011; Ide, 2017).

Table 2 shows the IAA scores per pair of annotators on each of the three taxonomy levels, the overall average of which is 0.56. Notably, level 1 of the taxonomy consistently has the highest IAA scores, averaging 0.62, most probably due to the lower number of categories and their relative dissimilarity. The annotators used 163 labels out of the available 170, showing that the taxonomy has good coverage of CL/NLP topics. The five most frequently used labels are, in decreasing order, *Data Management and Generation*, *Low-resource Languages*, *Model Architectures*, *Domain-specific NLP*, and *Data Preparation*.

Upon completing their tasks, the annotators were asked to participate in a questionnaire to clarify their specific annotation methods. The findings reveal that the annotation guidelines[23] were deemed clear and easily comprehensible. However, 50% of the annotators expressed that additional specific examples for each label in the taxonomy could have been beneficial. Notably, all annotators unanimously affirmed that relying solely on the title and abstract was insufficient for assigning labels to a paper. Instead, they consistently consulted the *Introduction*, *Methodology*, and *Conclusion* sections, identifying these as the most frequently referenced when making annotation decisions.

5. Limitations

In terms of the proposed taxonomy, it is important to note that although we believe it to be representative of the current core topics of CL/NLP, we are aware of its limitations and summarise them as follows: 1. the taxonomy was built based on a limited

[23] https://github.com/DFKI-NLP/Taxonomy4CL/blob/main/data/forc4cl-annotation-guidelines.pdf

number of papers from a specific repository, and 2. it was constructed semi-automatically and is thus subjective and cannot keep up with emerging topics in CL/NLP. For this reason, the taxonomy is published as a resource for the community, and we plan to expand it in the future based on feedback. This process has already started during the project, as annotators provided us with label suggestions according to frequently appearing topics not covered by our taxonomy. Examples include *Natural Language Inference (NLI)*, *Software/Toolkit Development*, and *Explainability*. In addition, two CL/NLP experts not familiar with the project reviewed the taxonomy and expressed opinions regarding the naming of specific labels, adding more labels, and re-ordering some hierarchy structures, which were all taken into account. Another idea for addressing these limitations is running the same topic modelling approach on new CL/NLP papers as they are published in different venues every few years and adjusting the taxonomy accordingly.

Regarding the constructed corpus, the main limitation lies in the inherent bias of the annotators' and curators' personal background and expertise in the CL/NLP domain. In the future, if our taxonomy is enriched and deployed on a larger scale, the idea is for authors to label their own papers since they are most familiar with their own topics and contributions, making the FoRC4CL a naturally growing resource as research in CL/NLP develops.

6. Conclusions

In this paper, we introduce our work of constructing FoRC4CL, a corpus of CL/NLP scholarly publications annotated with the topics and sub-topics of their main contributions. In order to annotate publications, we propose Taxonomy4CL, a taxonomy with the most prominent CL/NLP topics utilising topic modelling approaches. We publish this taxonomy publicly to gather feedback from the wider CL/NLP research community and expand it in the future. The FoRC4CL corpus is also publicly available and is used in a shared task (Abu Ahmad et al., 2023) that tackles the multi-label hierarchical classification problem of CL/NLP articles (Abu Ahmad et al., 2024). This paper showcases a first stepping stone to assist researchers working on scholarly information processing to develop systems and applications that assist CL/NLP researchers. Future work can utilise our corpus for tasks such as structured information extraction of scientific artefacts from articles, recommender systems, and scientific search engines.

7. Ethical Statement

FoRC4CL does not contain any sensitive or personal information and is collected from open-source resources. Annotators were compensated through a typical payment scheme and have been informed about the further use of the annotations.

8. Acknowledgements

This publication was supported by the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[24] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The NFDI is funded by the Federal Republic of Germany and its states. The paper received funding through the German Research Foundation (DFG) project NFDI4DS (no. 460234259).

We would like to thank the annotators, Saswat Dash, Uliana Eliseeva, Houda Kaoukab, Melina Plakidis, Greg Shook, and Eunhye Yun, for their contributions to this project. We also thank our colleagues at the DFKI SLT lab in Berlin, Malte Ostendorff and Leonhard Hennig, for reviewing the taxonomy and providing us with feedback, as well as Julian Moreno Schneider and Aleksandra Gabryszak for helping with INCEpTION and the annotation process.

9. Bibliographical References

Raia Abu Ahmad, Ekaterina Borisova, and Georg Rehm. 2024. FoRC@NSLP2024: Overview and Insights from the Field of Research Classification Shared Task. In *Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024)*, Hersonissos, Crete, Greece. 26/27 May. Submitted.

Raia Abu Ahmad, Ekaterina Borisova, Georg Rehm, Stefan Dietze, Saurav Karmakar, Wolfgang Otto, Jennifer D'Souza, Fidan Limani, and Ricardo Usbeck. 2023. NFDI4DS Shared Tasks. In *Proceedings of the Workshop on Research Data Infrastructures for Data Science and AI (RDI4DS)*. 28 September 2023. Co-located with INFORMATIK 2023.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? A meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan R Gibson, Mark Thomas Joseph, Min-Yen Kan,

[24] https://www.nfdi4datascience.de

Dongwon Lee, Brett Powley, Dragomir R Radev, Yee Fan Tan, et al. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*.

Ekaterina Borisova, Raia Abu Ahmad, Leyla Garcia-Castro, Ricardo Usbeck, and Georg Rehm. 2024. Surveying the FAIRness of annotation tools: Difficult to find, difficult to reuse. In *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 29–45, St. Julians, Malta. Association for Computational Linguistics.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Kathi Canese and Sarah Weis. 2013. PubMed: The bibliographic database. *The NCBI handbook*, 2(1).

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379):eaao0185.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Swapnil Hingmire, Irene Li, Rena Kawamura, Benjamin Chen, Alexander Fabbri, Xiangru Tang, Yixin Liu, Thomas George, Tammy Liao, Wai Pan Wong, et al. 2021. CLICKER: A computational linguistics classification scheme for educational resources. *arXiv preprint arXiv:2112.08578*.

Nancy Ide. 2017. Introduction: The Handbook of Linguistic Annotation. In *Handbook of Linguistic Annotation*, pages 1–18. Springer.

Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer. 2019. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 243–246.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Ruslan Mitkov. 2022. *The Oxford handbook of computational linguistics*. Oxford University Press.

Bernard Rous. 2012. Major update to ACM's computing classification system. *Communications of the ACM*, 55(11):12–12.

Angelo A Salatino, Thiviyan Thanapalasingam, Andrea Mannocci, Francesco Osborne, and Enrico Motta. 2018. The computer science ontology: A large-scale taxonomy of research areas. In *The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17*, pages 187–205. Springer.