Towards Dataset for Extracting Relations in the Climate-Change Domain

Andrija Poleksić[1,2,*], Sanda Martinčić-Ipšić[1,2]

[1]Faculty of Informatics and Digital Technologies (University of Rijeka), Radmile Matejčić 2, Rijeka, 51000, Croatia
[2]Center for Artificial Intelligence and Cybersecurity

Abstract
The impacts of global warming and climate change on ecosystems, weather patterns and human societies pose a significant threat to biodiversity and the sustainability of our planet. Despite the widespread scientific consensus, climate change denial persists among a segment of the population, either due to misconceptions or vested interests. Recent research shows that progress is being made in addressing climate denial as a majority acknowledges man-made climate change. However, the spread of misinformation remains a challenge, often perpetuated by corporate interests. To overcome these challenges, we propose constructing a dataset tailored for automated extraction and structuring of climate change-related scientific findings, focusing on relation extraction (RE) from scientific papers. Our research outlines the steps involved, including the preparation of the dataset for further training of the BERT-based model and downstream relation extraction task formulation. We discuss the process of data collection, preprocessing techniques and preliminary dataset analysis. Additionally, we highlight the need for a specialized Named Entity Recognition model for the climate-change domain and underline the need for annotation of domain-specific relations.

Keywords
dataset, climate change, relation extraction, scientific papers

1. Introduction

Global warming and climate change have profound and far-reaching effects on global ecosystems, weather patterns, sea levels, and human societies, constituting a critical threat to the planet's biodiversity and the prospect of a sustainable future [1]. Despite the widespread acceptance and scientific backing of climate change concepts, there remains a segment of the population that denies human impact on climate change, referred to as climate denial. Climate denial is driven either by misguided beliefs [2] or vested corporate interests [3]. A study by Areni et al. [2] investigates the dynamics between supporter and denier groups of Reddit users. They observe that supporters frequently reference scientific work, whereas deniers tend to rely more on alternative media and sources. Recent comprehensive research conducted by Andre et al. [4] demonstrates significant strides in addressing the issue of climate change denial. Their findings reveal that up to 86% of individuals acknowledge the reality of human-induced climate change and endorse measures aimed at mitigating human impact on the climate. Substantial climate

TEXT2KG 2024: Third International Workshop on Knowledge Graph Generation from Text, May 26 – May 30, 2024, co-located with Extended Semantic Web Conference (ESWC), Hersonissos, Greece
*Corresponding author.
andrija.poleksic@uniri.hr (A. Poleksić); smarti@uniri.hr (S. Martinčić-Ipšić)

denial stems from the dissemination of misinformation by large companies, often driven by vested interests, such as oil companies [5] and false scientific doubt creations, as elaborated by Oreskes and Conway [6]. Furthermore, the ever-increasing amount of data and information, including scientific papers, propels the need for automated information processing to speed up informed research decisions and facilitate fact-checking.

Motivated by both these challenges - information deluge and climate change, in this paper, we propose steps to construct the dataset that is fit to automatically extract and structure climate change-related scientific findings using information extraction (IE) methods. Specifically, we focus on the preliminary steps for relation extraction (RE) from scientific papers. Relation extraction (RE) is tasked with the identification of relations between entities in sentences, paragraphs or larger units of text. Sentence-level relation extraction involves identifying and classifying relations between entities in a single sentence. The goal is to determine the relation or association between two entities, typically represented by nouns or noun phrases such as people, organizations, or locations - named entities [7]. Our overall research plan consists of several steps:

- Preparation of the dataset of scientific papers for a climate-change domain suitable for the training of a BERT-like model;
- Additional pretraining (training with available pretrained weights) of the BERT-like model to adapt to the climate-change domain;
- Definition of relation types for relation extraction and construction of the dataset for the fine-tuning of the newly trained model(s) on the task of sentence-level relation extraction;
- Construction and curation of the climate-change knowledge graph from a high-quality journal.

In the next Section 2 is a short overview of the related work on pertained language models, relation extraction datasets and relation annotation. Section 3 elaborates on data collection, preprocessing and a preliminary analysis of the data. The final Sections 4 and 5, cover the results, discussion and conclusions respectively.

2. Related Work

Recent research efforts [8, 9, 10, 11, 12] report using pretrained models for text classification and sequence labelling tasks. One of the prominent ones is BERT (Bidirectional Encoder Representations from Transformers), an encoder-only transformer model trained on masked language modelling (MLM) task [13]. Although it is shown that encoder-decoder architecture models such as BART [14] and T5 [15] provide comparable and sometimes better results [16], they require the training of a larger number of parameters, which ultimately requires a larger amount of data and computational resources.

Lee et al. [8] perform additional training of the original BERT$_{BASE}$ deep neural model [13] for the biomedical domain - BioBERT. They report that no new WordPeace vocabulary is needed, ensuring the compatibility of the two pretrained models (BioBERT and BERT). BioBERT achieves new SOTA results on benchmarks for relation extraction and named entity recognition.

ClinicalBERT model [11] follows the same principle and further trains the BERT and BioBERT models on a large multicenter dataset.

The other line of research by Beltagy et al. [9] is training a new model SciBERT from scratch, which is also based on the BERT architecture [13], using scientific papers as the training data. For SciBERT they construct a new vocabulary SciVocab. An overall improvement of 0.61 F1-score on the downstream tasks using SciVocab compared to using the original BERT vocabulary is achieved. Additionally, several SOTA results are reported, surpassing also the BioBERT results on the ChemProt [17] benchmark by a fairly large margin. A similar strategy is applied in Chalkidis et al. [12], where a family of LegalBERT models is trained to support legal NLP research, computer-assisted law and legal technology applications.

Webersinke et al. in [10] train the RoBERTa model [18], which was adapted using distillation process [19], on the climate-change domain - ClimateBERT. The model is trained on climate-related news articles and posts on social media.

In our research we will extend our previous research [20], as we plan to perform additional training on two models: for SciBERT additional training for the climate-change domain employing scientific papers; and for ClimateBERT extension of parametrized domain knowledge by carefully curated high-quality dataset, surpassing their drawbacks of either out-of-climate-change-domain vocabulary or improving the quality of media collected information with scientifically obtained facts. To this end, in this paper, we propose the construction of a new dataset for the climate-change domain obtained from scientific papers published in high-quality journals.

For joint entity and relation extraction downstream tasks [21] the model is trained to perform both tasks simultaneously while benefiting from the use of interrelated signals. Relation extraction can be set as a supervised task and requires a huge amount of labelled (i.e. annotated) training data. To speed up the process, many researchers are turning to the idea of distant supervision[1] [22]. This includes datasets such as FewRel [23] and T-REx [24] for RE at sentence level and datasets such as DocRED [25] and Wiki20m [26] for RE on larger text sections.

Recently, the use of Large Language Models (LLMs) for the annotation of relations and entities has been reported [27], either to augment and speed up the annotation process for human annotators [28, 29] or to completely replace human efforts [30]. Besides annotation, LLMs are considered as synthetic data generators [31, 32] or for assessing the LLM-annotation quality [33]. In our research, we plan to engage LLMs for the relation annotation subtask, leveraging of-the-shelf pretrained LLMs to speed up the process, as opposed to training specialised in-house LLMs and using them directly for RE.

3. Dataset Preparation

Adapting one of the BERT models for the RE task for the climate-change domain requires the construction of an appropriate dataset (e.g. scientific and high-quality source). To this end we selected the highest-ranked scientific journals on climate change based on the Scimago

[1]Distant supervision assumes that the presence of a given entity pair in a given text implies a relation between them such that it is found in a Knowledge Graph/Base.

Journal & Country Rank (SJR)[2] and ScienceWatch Rank[3] and open access MDPI journals that are associated with the topic of climate change and in a substantial quantity of available papers and consistent format for parsing. The Table 2 (Appendix A) lists information on 194,673 retrieved research papers from selected journals, where 77.35% (150,583) are available in HTML format, while the remaining 22.65% (44,090) are only available in PDF format.

The PDF documents were first processed with pdfminer.six[4] library [34] for extracting information from PDF documents. They were converted to HTML format retaining the available information for each parsed element, including position, font and font size. This information was obtained with the Layout analysis algorithm[5] that groups characters into words and lines, lines into boxes and finally textboxes hierarchically based on the position of each character. Hence, we developed a parser fine-tuned to each journal formatting style and position information, enabling correct and complete text extraction. For navigation through HTML files, we used BeautifulSoup[6] library [35].

As already mentioned, for each journal a specific parser was needed. Next, we draw a random sample of 100 papers for each journal to evaluate the parsing procedure. Based on the random sample, we create a parser that successfully extracts the content of the papers in 100% of the cases, ranging from pure content to metadata such as authors, affiliations, references and DOI information. The parsing procedure allows extracting data to the full extent. This is manually validated on a random sample of 10 papers per journal by comparing the texts from PDF/HTML with the data stored in Pandas dataframes[7]. Table 3 (Appendix C) lights up some of the most common problems encountered during PDF and HTML parsing. Still, despite many problems, we obtained a well-documented, comprehensive dataset, which is appropriate for further model training. In Table 1 the comparison of the total training data used for each of the neural models (BERT, SciBERT and ClimateBERT) is reported. Our dataset contains ~35% of tokens used for training of SciBERT, and surpasses the number of tokens for ClimateBERT by six times. The average number of sentences per paper in our dataset is ~160% of the average reported for SciBERT. These numbers are encouraging, suggesting that we have collected sufficient high-quality texts for training of BERT-based model.

To further explore the dataset content we report statistics using a readily available part-of-speech (POS) tagger and a named entity recognition (NER) model from flair[8] framework [36]. First, we take a random sample of 10,000 research papers to perform the analysis. Then we tokenize into sentences and perform POS tagging[9] and NER. In each POS-tagged sentence, we determine noun- and verb- phrases. Non traditionally, we define heuristic noun- and verb-phrases as a sequence of words with specific POS tags as listed:

- **Noun phrase**: Cardinal number (**CD**), Adjective (**JJ**), Determiner (**DT**), Noun (**NN**), Foreign word (**FW**), Possessive ending (**POS**), Hyphen (**HYPH**), Symbol (**SYM**) ,

[2]https://www.scimagojr.com/journalrank.php?category=2306
[3]http://archive.sciencewatch.com/ana/st/climate/journals/
[4]https://github.com/pdfminer/pdfminer.six/tree/master
[5]https://pdfminersix.readthedocs.io/en/latest/topic/converting_pdf_to_text.html#id1
[6]https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[7]https://pandas.pydata.org/
[8]https://github.com/flairNLP/flair
[9]The full list of POS tags for the model used can be found here: https://huggingface.co/flair/pos-english.

Table 1
Training data comparison: Data used for model training, number of tokens (CS), and average number of sentences (A#S) per paper if applicable.

Model	Data used	CS	A#S
BERT	BooksCorpus (800M words) and English Wikipedia (2,500M words)	3.30B	/
SciBERT	Random sample of 1.14M papers from Semantic Scholar	3.17B	154
ClimateBERT	Climate related news articles, climate-related papers abstracts and corporate climate and sustainability reports	0.22B[a]	/
OUR	~200,000 climate-related research papers	1.25B[b]	242[c]

[a,b,c] Calculation is reported in Appendix B

- **Verb phrase**: Verb (**VB**), "to" (**TO**), Adverb (**RB**), Modal (**MD**).

This modification, despite being imperfect, allows for analysis of the most frequent verb- and noun- phrases, providing insights into possible types of relations between entities, possible named entities and entity types (e.g. person, organization, location, etc.). With this approximation, we further estimated the number of total and unique triples. Figure 1 shows the total number of verb phrases, noun phrases, entities (tagged by the NER model) and possible triples occurring in the sample of 10,000 papers. The sample consists of 2,406,799 sentences, from which we extracted a total of 15,238,265 noun phrases and 1,790,745 entities. The ratio of noun phrases to extracted entities (~8:1) indicates the need for a NER model, that is better fitted to the climate-change domain vocabulary. Table 4 (Appendix D) lists the top noun phrases consisting of 1, 2 and 3 words respectively. Table 5 (Appendix E) lists the top entities for three entity types: Location Name (LOC), Organization Name (ORG) and Other Name (MISC). Number of entity types will be addressed in the future work, employing more recent methods such as GLiNER [37]. Since the list contains many acronyms and abbreviations the expansion and disambiguation problem needs to be addressed as well.

Similarly, we analyze the occurrence of verb phrases: a total of 5,934,949 verb phrases forming 486,632 unique expressions. Although this is promising, the number of unique expressions needs to be reduced to a feasible set enabling the training of the classifier to extract relations in downstream tasks. Moreover, this is an indication that many climate-change-specific relations are present, which needs to be addressed in the downstream training as well. Table 6 (Appendix F) reports the 30 most frequently occurring verb phrases by number of words (1, 2 and 3 respectively). We observe a high similarity between many unique verb phrases, such as: "is shown", "shows", "are shown" and "has been shown"; indicating the obvious next step of data quality improvement by deduplication.

4. Relation Annotation

To effectively train and evaluate supervised relation extraction models, the annotated data is needed [24]. To this end, we plan to engage the advanced LLM possibilities in the context of

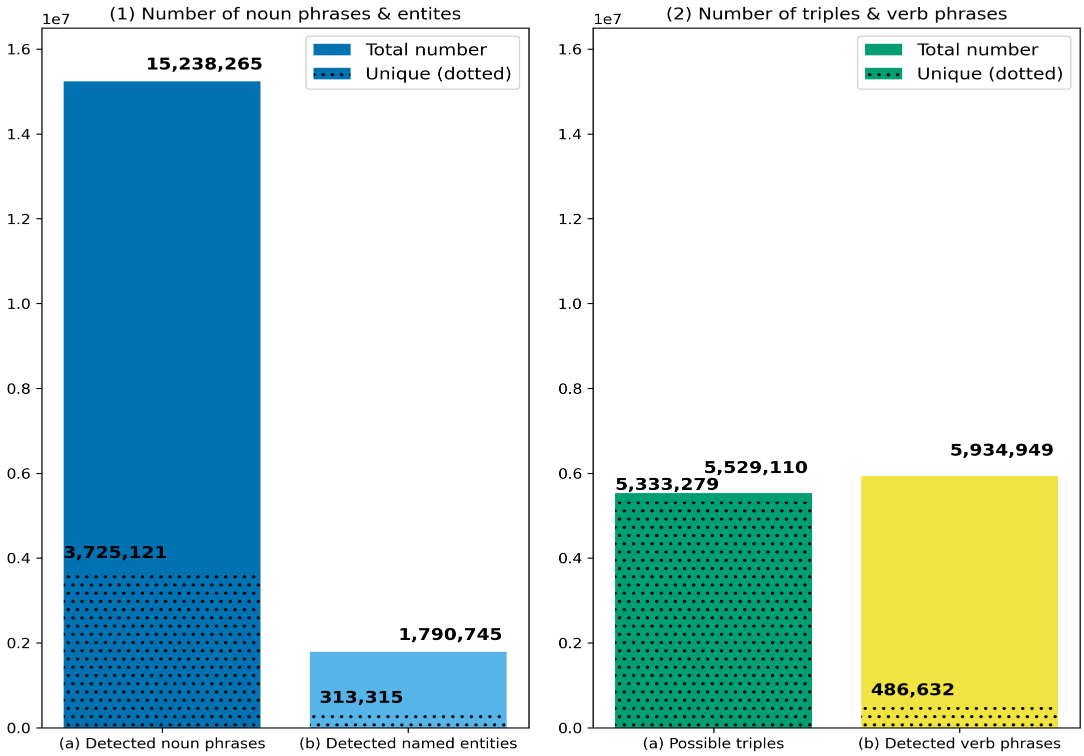

Figure 1: Counts of noun phrases, entities, triples, and verb phrases: Occurrence of noun phrases (1a), named entities (1b), possible triples (2a) and verb phrases (2b) with count of unique expressions (dotted) in the 10,000 papers sample.

automatic or enhanced annotation of relation triples. With POS tagging and NER on the sample of 10,000 papers, we have established the foundation for possible triple detection. We anticipate that a relation is expected to exist if there is a verb between two entities, where entities are either approximated by noun phrases that we have heuristically recognised or named entities recognised by the flair model. Moreover, we hypothesize that this will allow guided annotation by providing better context to LLM-enabled annotation. In the remainder of this section, we preview some examples of possible entities and relations in climate change domain[10], which remains an open question to be addressed in the future:

- 'For example, **Atlantic** cyclones have been well documented as *causing* high surge levels and heavy precipitation.' - (Atlantic cyclones, *cause*, high surge levels)
- 'El Niño–Southern Oscillation (**ENSO**) is another important factor for winter temperature in **China**.' - (ENSO, *affects*, winter temperature in China)
- 'The concentration map captured a significantly *high hazard of* groundwater arsenic in the north and northeast India, particularly in Assam and **West Bengal**, ...' - (West

[10]Underlined words are suggested entities in the sentence, where the bold parts are recognized by the flair NER model. Each sentence has a suggested triple in the form: (entity1, *relation*, entity2)

Bengal, *high hazard of*, groundwater arsenic)

5. Discussion and Conclusion

In this paper, we report on the first steps towards creating a dataset suitable for training the BERT-like model that will subsequently be used for downstream climate-change relation extraction tasks. We have collected and analyzed a set of 200,000 carefully selected scientific papers as the high-quality content of the climate-change domain. We discuss technical details and common pitfalls in parsing PDF and HTML documents as the first steps needed to obtain a sufficient quantity of domain-specific data to train a BERT-based model. Next, we report preliminary statistics of the dataset to ensure its appropriateness for downstream relation extraction. During preliminary analysis, we identified a high number of possible different relations, indicating that further distilling of relations and relation types should be implemented. Moreover, our preliminary findings suggest that the new NER model tailored for the vocabulary of the climate-change domain is required.

With these preliminary results, we open several research directions. First, the collected dataset will be used for additional training of the SciBERT and ClimateBERT models involving different configurations of masked language modelling (MLM) principles. Second, to reduce the abundance of different but similar domain-specific relations we will need to develop a method for fine-tuning annotated relations for training sentence-level relation extraction (RE) model. This will involve the disambiguation of related relations and relation types and LLM-enabled annotation. Finally, as the main goal of this research is the construction and curation of a knowledge graph for the climate-change content captured in a high-quality journal. In future work, we plan to address KG construction-related challenges, relying on existing literature, such as work of Dessi et al [38] and Chessa et al [39].

Acknowledgments

This work has been partially supported by the University of Rijeka under project number uniri-drustv-18-20. Croatian Science Foundation supports AP under the project DOK-2021-02.

References

[1] H.-G. et al, Impacts of 1.5°c global warming on natural and human systems, in: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2018, pp. 175–312. doi:10.1017/9781009157940.005.

[2] C. S. Areni, Motivated reasoning and climate change: Comparing news sources, politicization, intensification, and qualification in denier versus believer subreddit comments, Applied Cognitive Psychology 38 (2024). doi:10.1002/acp.4167, all Open Access, Hybrid Gold Open Access.

[3] J. Farrell, K. McConnell, R. Brulle, Evidence-based strategies to combat scientific misinformation, Nature Climate Change 9 (2019) 191–195. doi:10.1038/s41558-018-0368-6.

[4] P. Andre, T. Boneva, F. Chopra, A. Falk, Globally representative evidence on the actual and perceived support for climate action, Nature Climate Change (2024). doi:10.1038/s41558-024-01925-3.

[5] R. Debnath, D. Ebanks, K. Mohaddes, T. Roulet, R. M. Alvarez, Do fossil fuel firms reframe online climate and sustainability communication? a data-driven analysis, npj Climate Action 2 (2023) 47. doi:10.1038/s44168-023-00086-x.

[6] N. Oreskes, E. M. Conway, Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues From Tobacco Smoke to Global Warming, Bloomsbury Press, 2010.

[7] S. Pawar, G. K. Palshikar, P. Bhattacharyya, Relation extraction : A survey, 2017. arXiv:1712.05191.

[8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.

[9] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.

[10] N. Webersinke, M. Kraus, J. Bingler, M. Leippold, Climatebert: A pretrained language model for climate-related text, SSRN (2022). URL: https://ssrn.com/abstract=4229146. doi:10.2139/ssrn.4229146.

[11] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. B. A. McDermott, Publicly available clinical bert embeddings, 2019. arXiv:1904.03323.

[12] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, Legal-bert: The muppets straight out of law school, 2020. arXiv:2010.02559.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. URL: https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. arXiv:1910.10683.

[16] L. N. Phan, J. T. Anibal, H. Tran, S. Chanana, E. Bahadroglu, A. Peltekian, G. Altan-Bonnet, Scifive: a text-to-text transformer model for biomedical literature, 2021.

arXiv:2106.03598.

[17] J. V. Kringelum, S. K. Kjaerulff, S. Brunak, O. Lund, T. I. Oprea, O. Taboureau, Chemprot-3.0: a global chemical biology diseases mapping, Database (Oxford) 2016 (2016) bav123. doi:10.1093/database/bav123.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, ArXiv abs/1910.01108 (2019).

[20] A. Poleksić, S. Martinčić-Ipšić, Effects of pretraining corpora on scientific relation extraction using bert and scibert, in: Joint Workshop Proceedings of 5th (Sem4Tra) and 2nd NLP4KGC: Natural Language Processing for Knowledge Graph Construction co-located with the 19th International Conference on Semantic Systems (SEMANTiCS 2023), volume Vol-3510 of *CEUR Workshop Proceedings*, Leipzig, Germany, 2023. URL: https://ceur-ws.org/Vol-3510/paper_nlp_3.pdf.

[21] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers, 2023. arXiv:2306.02051.

[22] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: K.-Y. Su, J. Su, J. Wiebe, H. Li (Eds.), Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Association for Computational Linguistics, Suntec, Singapore, 2009, pp. 1003–1011. URL: https://aclanthology.org/P09-1113.

[23] X. Han, H. Zhu, P. Yu, Z. Wang, Y. Yao, Z. Liu, M. Sun, FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 4803–4809. URL: https://aclanthology.org/D18-1514. doi:10.18653/v1/D18-1514.

[24] H. Elsahar, P. Vougiouklis, A. Remaci, C. Gravier, J. Hare, F. Laforest, E. Simperl, T-REx: A large scale alignment of natural language with knowledge base triples, in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (Eds.), Proceedings of the LREC 2018, European Language Resources Association (ELRA), Miyazaki, Japan, 2018. URL: https://aclanthology.org/L18-1544.

[25] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, DocRED: A large-scale document-level relation extraction dataset, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 764–777. URL: https://aclanthology.org/P19-1074. doi:10.18653/v1/P19-1074.

[26] X. Han, T. Gao, Y. Lin, H. Peng, Y. Yang, C. Xiao, Z. Liu, P. Li, J. Zhou, M. Sun, More data, more relations, more context and more openness: A review and outlook for relation extraction, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2020, pp. 745–758. URL: https://aclanthology.org/2020.aacl-main.75.

[27] Z. Tan, A. Beigi, S. Wang, R. Guo, A. Bhattacharjee, B. Jiang, M. Karami, J. Li, L. Cheng, H. Liu, Large language models for data annotation: A survey, 2024. arXiv:2402.13446.
[28] A. Goel, A. Gueta, O. Gilon, C. Liu, S. Erell, L. H. Nguyen, X. Hao, B. Jaber, S. Reddy, R. Kartha, J. Steiner, I. Laish, A. Feder, Llms accelerate annotation for medical information extraction, 2023. arXiv:2312.02296.
[29] J. Li, Z. Jia, Z. Zheng, Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 5495–5505. doi:10.18653/v1/2023.emnlp-main.334.
[30] R. Zhang, Y. Li, Y. Ma, M. Zhou, L. Zou, LLMaAA: Making large language models as active annotators, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 13088–13103. doi:10.18653/v1/2023.findings-emnlp.872.
[31] R. Tang, X. Han, X. Jiang, X. Hu, Does synthetic data generation of llms help clinical text mining?, 2023. arXiv:2303.04360.
[32] Q. Wang, K. Zhou, Q. Qiao, Y. Li, Q. Li, Improving unsupervised relation extraction by augmenting diverse sentence pairs, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 12136–12147. URL: https://aclanthology.org/2023.emnlp-main.745. doi:10.18653/v1/2023.emnlp-main.745.
[33] H. Khorashadizadeh, N. Mihindukulasooriya, S. Tiwari, J. Groppe, S. Groppe, Exploring in-context learning capabilities of foundation models for generating knowledge graphs from text, 2023. arXiv:2305.08804.
[34] Y. Shinyama, P. Guglielmetti, P. Marsman, pdfminer.six, 2018. URL: https://pdfminersix.readthedocs.io/.
[35] L. Richardson, Beautiful soup documentation, 2007. URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.
[36] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, R. Vollgraf, FLAIR: An easy-to-use framework for state-of-the-art NLP, in: NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), 2019, pp. 54–59.
[37] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, 2023. arXiv:2311.08526.
[38] D. Dessí, F. Osborne, D. Reforgiato Recupero, D. Buscaldi, E. Motta, Scicero: A deep learning and nlp approach for generating scientific knowledge graphs in the computer science domain, Knowledge-Based Systems 258 (2022) 109945. URL: https://www.sciencedirect.com/science/article/pii/S0950705122010383. doi:https://doi.org/10.1016/j.knosys.2022.109945.
[39] A. Chessa, G. Fenu, E. Motta, F. Osborne, D. Reforgiato Recupero, A. Salatino, L. Secchi, Data-driven methodology for knowledge graph generation within the tourism domain, IEEE Access 11 (2023) 67567–67599. doi:10.1109/ACCESS.2023.3292153.

A. Data statistics

Table 2
Number of papers: The number of collected research papers in the climate-change domain according to the journal/source.

Journal name	#	Journal name	#	Journal name	#
International Journal of Climatology	3,825	Ecological Applications	4,469	Ecosystem Health and Sustainability	831
Energy Policy	1,023	Journal of Climate	15,325	Climate Dynamics	3,943
Global Change Biology	7,103	Journal of Geophysical Research: Atmospheres	14,512	NPJ Climate and Atmospheric Science	355
NPJ Ocean Sustainability	12	NPJ Climate Action	39	Nature Climate Change	387
Nature Geoscience	560	PNAS	88,534	MDPI water	21,768
MDPI Air	18	MDPI Atmosphere	8.705	MDPI Climate	1,232
MDPI Earth	184	MDPI Ecologies	115	MDPI Energies	8,236
MDPI Hidrology	988	MDPI Forests	10,674	MDPI Fuels	104
MDPI Environments	1,012	MDPI Meteorology	57	MDPI Sustainable Chemistry	116
MDPI Recycling	420	MDPI Oceans	126	**Total**	**194,673**

B. Training data comparison calculations

- **a**: Calculated from reported average number of words [10].
- **b**: Approximation from tokenizer trained on 10,000 papers sample according to The Tokenization pipeline (https://huggingface.co/docs/tokenizers/python/latest/pipeline.html).
- **c**: Approximation from *SegtokSentenceSplitter* (https://github.com/flairNLP/flair/blob/master/flair/splitter.py)

C. Common extraction problems

Table 3
Most frequent problems with text extraction: The left-hand column contains a brief description of the problem, while the explanation or example can be found in the right-hand column. The text in bold indicates what is actually extracted.

Problem Description	Example/Explanation
Text data missing due to unexpected font size/style	Where 2 and *two* always makes up five. **Where and always makes up five.** ... original BERT$_{BASE}$ model with ... **... original BERT model with ...**
Wrong ordering of paragraphs	Layout algorithm heuristics give wrong conclusions based on distance, e.g. bottom right paragraph is ”closer” to top right paragraph then to the top left paragraph due to a figure/table/graph.
Page numbering or similar information abrupt paragraph content	For navigation through HTML files, we used BeautifulSoup library. **For navigation through HTML files, we PAGE 5 AUTHOR ET AL. used BeautifulSoup library.**
Wrong word ordering due to justification	Nature climate change **Nature change climate**
Problems with wrong symbol extraction (Ligatures)	... far-reaching effects on global ecosystems ... **... far-reaching e⊠ects on global ecosystems ...**
First line of paragraph missing	

D. Most common noun phrases

Table 4
Most common noun phrases: Top 30 noun phrases by the number of words (1, 2 and 3) with the corresponding counts (#).

Noun phrase (1)	#	Noun phrase (2)	#	Noun phrase (3)	#
the	205,685	this study	28,914	the other hand	5,417
a	78,714	si appendix	23,542	the study area	4,049
this	51,495	the results	20,506	the present study	3,789
1	37,595	the number	16,034	37 ° c	2,564
2	29,195	the model	15,671	4 ° c	2,540
data	24,962	the presence	14,187	the same time	2,441
that	23,472	table 1	13,144	an important role	2,336
those	22,015	this work	11,452	the united states	2,065
such	21,800	the data	10,638	the time series	2,056
addition	20,902	the authors	10,144	$p < 0.001$	1,982
i.e.	20,685	the effect	9,741	the total number	1,847
3	19,837	these results	9,237	a wide range	1,745
one	19,375	this paper	9,146	the national academy	1,646
consistent	18,599	table 2	8,146	the north atlantic	1,629
c	18,403	the case	7,776	the spatial distribution	1,603
p	18,101	the effects	7,701	$p < 0.05$	1,553
t	17,330	climate change	7,632	a large number	1,376
cells	16,998	an increase	7,187	the standard deviation	1,277
results	16,290	the absence	6,923	the northern hemisphere	1,249
similar	16,193	the use	6,823	the indian ocean	1,095
4	15,813	the difference	6,800	the study period	1,091
changes	15,666	fig. 1	6,747	$p < 0.01$	1,070
e.g.	15,456	the study	6,624	25 ° c	1,011
precipitation	15,163	a result	6,581	the north pacific	985
5	15,128	figure 1	6,571	the current study	979
example	14,729	the surface	6,556	wang et al	977
contrast	14,082	the impact	6,316	30 ° c	939
water	13,820	the analysis	6,084	the plasma membrane	905
b	13,182	figure 2	5,886	the boundary layer	902
time	12,893	the region	5,876	20 ° c	897

E. Most common entities

Table 5
Most common entities: Top 30 entities for three entity types: Location (LOC) name, Miscellaneous (MISC) name and Organization (ORG) name with counts (#).

LOC	#	MISC	#	ORG	#
China	11686	DNA	6599	ENSO	6616
Pacific	9480	Arctic	4182	PNAS	3289
Europe	4213	SI Appendix	3563	EC	2822
United States	4145	F	2786	SST	2775
Atlantic	3653	European	2684	TC	2358
USA	3459	Equation	2603	NAO	2011
North Atlantic	3341	C	2289	ATP	1939
Indian Ocean	2767	Western	1932	N. Institutes of Health	1934
North America	2747	Asian	1930	El Niño	1617
Africa	2665	Chinese	1704	IPCC	1487
CA	2581	SST	1654	NCEP	1457
Australia	2472	Indian	1338	MDPI	1440
Japan	2345	Mediterranean	1308	EU	1396
US	2255	CMIP5	1301	MJO	1326
Germany	2223	Arabidopsis	1246	N. Sci. Fdn.	1236
North Pacific	2098	MJO	1238	SLP	1227
Asia	1987	UTC	1193	WRF	1184
India	1895	Gaussian	1156	NCAR	1181
Canada	1821	African	1148	NOAA	1128
Northern Hemisphere	1810	Bayesian	1089	NIH	1074
South America	1692	North American	973	TP	1040
U.S.	1663	RNA	925	PBL	1010
California	1409	CT	912	PCR	1008
Beijing	1382	GCM	907	Univ. of California	979
Greenland	1290	III	893	The N. Acad. of Sci.	963
MA	1229	ROS	870	ITCZ	941
UK	1228	BC	820	PCA	907
Southern Hemisphere	1212	Eurasian	814	∇	905
Eurasia	1208	DEM	768	RMSE	896
Southern Ocean	1196	PDO	766	WNP	872

Abbreviations: N. - National, Sci. - Science, Fdn. - Foundation, Univ. - University, Acad. - Academy, ∇- N. Nat. Sci. Fdn. of China

F. Most common verb phrases

Table 6
Most common verb phrases: Top 30 occurring verb phrases by number of words (1, 2 and 3) with counts (#).

Verb phrase (1)	#	Verb phrase (2)	#	Verb phrase (3)	#
is	267,521	as well	20,097	can be seen	3,294
are	108,928	is not	9,363	should be addressed	2,764
using	61,697	may be	8,727	can be found	2,396
was	60,454	was supported	8,307	can be used	1,672
however	58,755	are shown	7,242	should be noted	1,609
were	40,883	to determine	7,061	may be addressed	1,585
respectively	33,872	not shown	6,555	have been deposited	1,463
compared	29,998	was used	6,541	can be obtained	1,139
based	29,822	is also	5,826	appears to be	1,097
used	29,393	were used	5,801	was performed using	1,024
shows	27,179	was observed	5,446	can be explained	970
has	24,828	to be	5,444	may not be	968
thus	24,575	is shown	5,364	can be expressed	887
observed	24,346	were obtained	4,836	can be observed	877
including	23,660	was performed	4,819	has been reported	874
therefore	23,036	to identify	4,339	has been shown	802
showed	22,703	to assess	4,210	did not affect	790
’s	22,429	is more	4,167	can be calculated	790
show	21,582	were performed	4,122	have been identified	779
only	21,147	are not	4,053	can be attributed	740
have	20,796	to test	3,991	were performed using	732
increased	20,556	would be	3,923	can be considered	721
found	20,306	have shown	3,780	have been reported	674
more	19,132	were collected	3,737	to better understand	671
following	18,966	can be	3,698	seems to be	665
to	18,603	could be	3,630	did not show	634
see	17,771	to obtain	3,561	has been observed	606
most	17,398	is based	3,551	should be considered	594
associated	16,963	will be	3,540	has been used	568
had	16,785	that is	3,495	has been suggested	566