LLM-Based Entity Extraction Is Not for Cybersecurity

Maxime Würsch[1,2], Andrei Kucharavy[1,3,*], Dimitri Percia-David[1,3] and Alain Mermoud[1]

[1]*Cyber-Defence Campus, armasuisse S+T*
[3]*Institute of Entrepreneurship & Management, HES-SO Valais-Wallis*
[2]*Section of Computer Science, EPFL*

Abstract
The cybersecurity landscape evolves rapidly and poses threats to organizations. To enhance resilience, one needs to track the latest developments and trends in the domain. For this purpose, we use large language models (LLMs) to extract relevant knowledge entities from cybersecurity-related texts. We use a subset of arXiv preprints on cybersecurity as our data and compare different LLMs in terms of entity recognition (ER) and relevance. The results suggest that LLMs do not produce good knowledge entities that reflect the cybersecurity context.

Keywords
NLP, Bibliometrics, NER, LLM, Keyword Extraction, Nouns Extraction, Cyber-security

Secure and reliable information systems have become a central requirement for the operational continuity of the vast majority of goods and services providers [1]. However, securing information systems in a fast-paced ecosystem of technological changes and innovations is hard [2]. New technologies in cybersecurity have short life cycles and constantly evolve [3]. This exposes information systems to attacks that exploit vulnerabilities and security gaps [2]. Hence, cybersecurity practitioners and researchers need to stay updated on the latest developments and trends to prevent incidents and increase resilience [4].

A common approach to gather, cure and synthesize information about such developments is to apply bibliometrics-based knowledge entity extraction and comparison through embedding similarity [5, 6, 7] – recently boosted by the availability of entity extractors based on large language models (LLMs) [8, 9]. However, it is unclear how appropriate this approach is for the cybersecurity literature. We address this by emulating such an entity extraction and comparison pipeline and using a variety of common LLM-based entity extractors to evaluate the relevance of extracted entities to document understanding tasks, using as a proxy the relevance of *arXiv* to cybersecurity (https://arxiv.org)

While LLMs burst into public attention in late 2022, in large part thanks to public trials of conversationally fine-tuned LLMs [10, 11, 12], modern large language models pre-trained on large amounts of data trace their roots back to ELMo LLM, first released in 2018 [13]. The *LLM* term emerged to refer to $>$ 100M parameters and pretrained on $>$ 1B tokens [14, 15], such as BERT or RoBERTa [16, 17]. Smaller LLMs have proven to provide a valuable insight into the behavior and capabilities of larger ones [18, 19, 20], presenting both a weaker version of larger model capabilities, but also milder version of larger model failure modes [21]. In this paper, we focus on such smaller LLMs, ranging from 110M to 350M parameters and fine-tuned for entity extraction tasks, both to evaluate them and gain insight into larger LLMs behavior.

We show that despite the apparent abundance of available models, LLM-based entity extractors perform extremely similarly due to base models and fine-tuning datasets re-use. We then show that such models are ill-suited for bibliometrics tasks not only in cybersecurity-related topics but in computer science research in general, which we argue is related to the nature of fine-tuning datasets. We then show that even if we assume the relevance of extracted terms, their downstream automated processing remains a challenging task, given that it is highly sensitive to the embedding choice.

1. Methods

The complete code to replicate the results presented here with instructions is available at https://github.com/technometrics-lab/0_LLM-based_entity_extraction_CySec.

The dataset used is a copy of arXiv preprints up until late 2022, initially collected by [22]. We focused on the *cs* category, specifically on the cs.CR and cs.NI listings - Cryptography and Security and Network and Internet Architecture, as most relevant to cybersecurity. In addition to them, we added 6 additional unrelated listings (cs.CC, cs.LO, cs.DS, cs.IT, cs.CL, and cs.AI) as compari-

Joint Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informetrics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online
*Corresponding author.
andrei.kucharavy@hevs.ch (A. Kucharavy)

Model Name	Refs	Entities/Doc	Type
spaCy Large*	[23]	99.3 ± 6.93	Noun Extractor
spaCy Transformer	[23]	99.3 ± 6.97	
Yake*	[24]	19.9 ± 1.97	Keyphrase Extractor
KeyBERT	[25]	99.3 ± 7.25	
KBIR kpcrowd	[26, 27]	96.9 ± 14.6	
KBIR inspec	[26, 28]	76.4 ± 27.7	
BERT-base-uncased	[16]	44.7 ± 24.0	
BERT-base-uncased	[16]	43.3 ± 23.3	NER+CON R
XLM-RoBERTa-base Onconotes 5	[29, 30]	36.4 ± 23.4	NER+NUM
ELECTRA-base conll03	[31, 32]	39.9 ± 25.0	NER
BERT-large-cased conll03	[16, 32]	41.7 ± 24.9	
BERT-large-uncased conll03	[16, 32]	33.5 ± 23.3	
DistilBERT-base-uncased conll03	[15, 32]	37.7 ± 24.8	
RoBERTa-large conll03	[17, 32]	28.7 ± 21.1	
XLM-RoBERTa-large conll03	[33, 32]	26.0 ± 19.5	
BERT COCA-docusco	[16, 34]	99.6 ± 6.11	TokC

Table 1
Entity extractors analyzed. Models marked * are non-LLM based. Ent./Doc. is mean ± std.

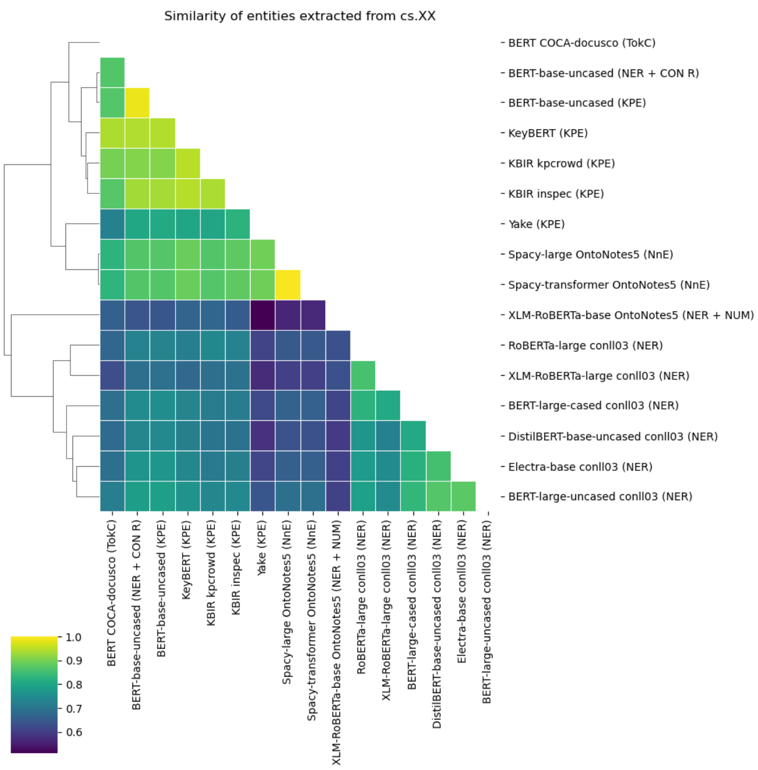

Figure 1: Similarity matrix of extracted terms, using cosine similarity in spaCy embedding.

son domains. The selected listings represented 5000 to 20000 preprints each.

For each of the preprints in the listings, all documents with $<$ 1000 words and not in English were removed. To achieve the latter, we used an XLM-Roberta model fine-tuned on a language identification dataset, available at https://huggingface.co/papluca/xlm-roberta-base-language-detection. Following that, the preamble of the preprint prior to the ”Abstract” keyword and the bibliography following the ”References” keyword were removed.

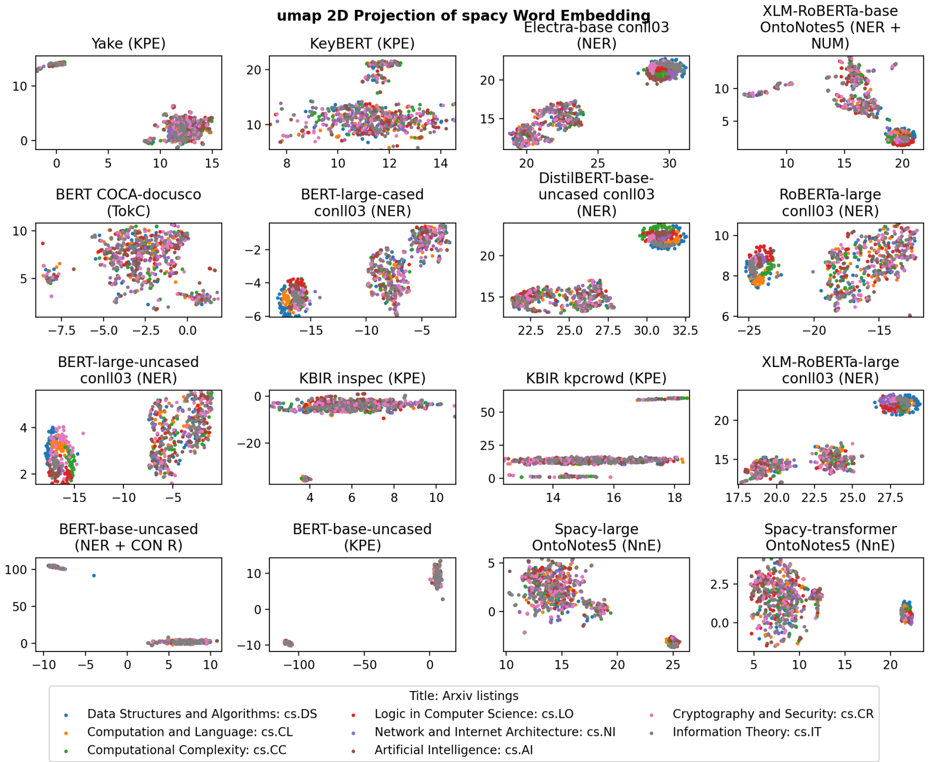

Figure 2: 2D projection with UMAP of spaCy embeddings of extracted entities.

Following that, we applied models described in Table 1 to the documents. Specifically, four major classes of models were used: Noun Extractrs (NnE), Keyphrase Extractors (KPE), Named Entity Recognition (NER), and Token Classification (TokC). Two NER models were augmented: number recognition (NER + NUM) and concept recognition (NER + CON R). Exact model names and sources are available in the code repository.

For LLM models, documents were segmented to fit the attention window. If the number of extracted entities exceeded 100, only 100 entities with the highest activations were retained. Samples of extracted entities are available in the code repository.

We compare the similarity of extractors' outputs on all documents by embedding entities extracted from each document with spaCy and calculating the average cosine similarity between extractors. A hierarchical clustering on cosine similarity was then used to create Fig. 1.

To visualize connections between the extracted entities from different listings, we used common embeddings (spaCy [23], GloVe [35], BERT-Large [16], GPT-2 [36], Fasttext [37], and word2vec [38]) and four low-dimensional projection algorithms (linear, spectral, t-SNE [39], UMAP [40]) to investigate if the entities extracted from preprints would allow arXiv listing identification. To allow the interpretation of the results, we subsampled 100 papers from each listing and, due to high processing time, excluded spaCy Transformer (cf Figs. 2, 3; additional figures in the code repository).

2. Results and Discussion

Our first result is that in computer science bibliometrics, a variety of entity extraction models perform similarly, with performance being mostly defined by their base architecture, task, and dataset used to fine-tune them (Fig. 1. Given that base architectures are predominantly BERT and RoBERTa [16, 17] and fine-tuning datasets are general texts, notably Conll03 newswire [32], we should not expect general LLM-based entity extraction models to perform well on scientific articles. LLM fine-tunes are sensitive to the training data, and only *KBIR-inspec* was fine-tuned using a scientific dataset, consisting of annotated 1998-2002 article abstracts from *Computers and Control and Information Technology* journal [41, 28]. Given the pace of the evolution of computer science, such fine-tunes are unlikely to still be relevant today, which

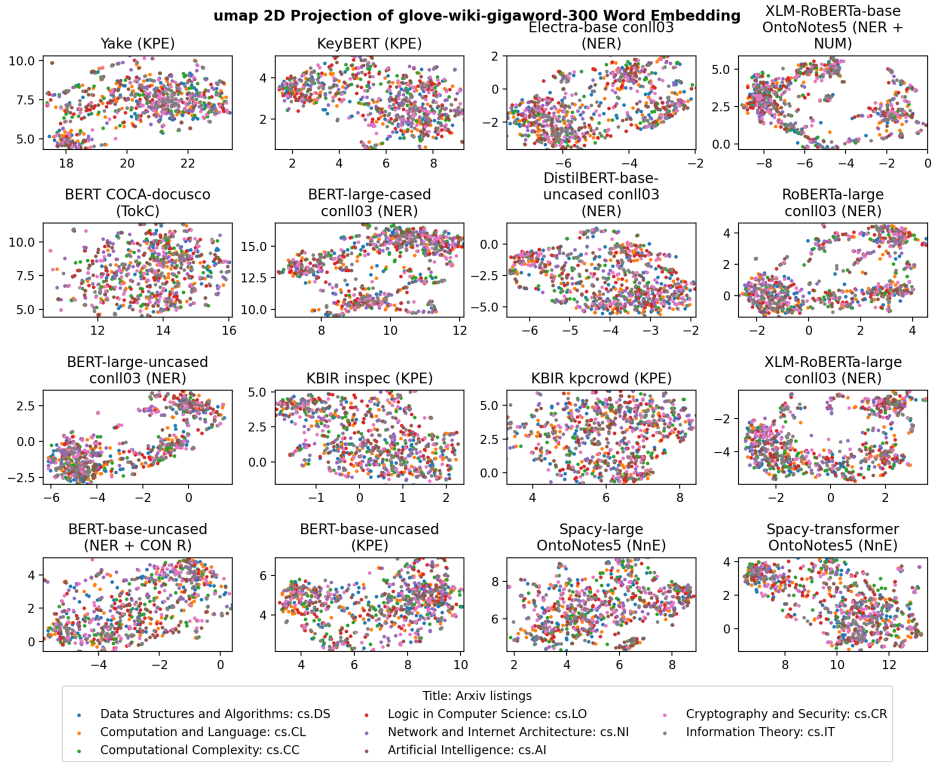

Figure 3: 2D projection with UMAP of GloVe embeddings of extracted entities.

is supported by the lack of thematic clusters in entities extracted by it (Fig. 2 ; NER organized structures are non-informative) suggest that they are indeed not relevant anymore. We hence hypothesize that non-LLM-based Yake [24] and spaCy [23] keywords and nouns extractors could be essential for addressing these issues, especially given that they already give radically different results compared to LLM-based extractors.

Our second result is that similarity of embedding of LLM-extracted entities does not perform well for concept-oriented bibliometrics in computer science. Even 2D projection algorithms known for their tendency to overfit local clusters, t-SNE, and UMAP [42], fail to separate different listings (Fig. 2, 3), with the exception of NERs in spaCy and GPT2 embeddings due to shared theorem names and chapter annotations. We provide interactive 2d projection plots in the code repository for readers to validate this claim.

Our third result is that cosine similarity is highly dependent on the embedding used to infer entity relatedness (Figs. 2, 3). While present here only UMAP 2d projection with spaCy and GloVe, the embedding-dependence of similarity is present across embeddings, which we present in the code repository. We also validate that this is not due to 2d projection algorithms by calculating clustering coefficients (intergroup dispersion vs. intra-group dispersion) across listings in each embedding. Hence, attention is warranted when using word embeddings in entity extraction and analysis pipelines.

Acknowledgments

AK is supported by the CYD Campus, armasuisse W+T, VBS grant (ARAMIS CYD-C-2020015).

References

[1] D. Percia David, Three Articles on the Economics of Information-Systems Defense Capability. Material-, Human-, and Knowledge-Resources Acquisition for Critical Infrastructures, Ph.D. thesis, Université de Lausanne, HEC Lausanne, 2020.

[2] R. Anderson, Security engineering: a guide to building dependable distributed systems, 3 ed., Wiley, 2020.

[3] T. Daim, H. Yalçin, Digital transformations: new

tools and methods for mining technological intelligence, Edward Elgar Publishing, 2022.
[4] T. U. Daim, D. Chiavetta, A. L. Porter, O. Saritas, Anticipating future innovation pathways through large data analysis, Springer, 2016.
[5] X. Chen, H. Xie, Z. Li, G. Cheng, Topic analysis and development in knowledge graph research: A bibliometric review on three decades, Neurocomputing 461 (2021) 497–515.
[6] I. Safder, S.-U. Hassan, Bibliometric-enhanced information retrieval: a novel deep feature engineering approach for algorithm searching from full-text publications, Scientometrics 119 (2019) 257–277.
[7] Y. Zhang, J. Lu, F. Liu, Q. Liu, A. Porter, H. Chen, G. Zhang, Does deep learning help topic extraction? a kernel k-means clustering method with word embedding, Journal of Informetrics 12 (2018) 1099–1117.
[8] A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, A. Jain, Structured information extraction from complex scientific text with fine-tuned large language models, arXiv preprint arXiv:2212.05238 (2022).
[9] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, S. Riedel, Language models as knowledge bases?, arXiv preprint arXiv:1909.01066 (2019).
[10] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, CoRR abs/2203.02155 (2022). URL: https://doi.org/10.48550/arXiv.2203.02155. doi:10.48550/arXiv.2203.02155. arXiv:2203.02155.
[11] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukosiute, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. DasSarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. R. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. Brown, J. Kaplan, Constitutional AI: harmlessness from AI feedback, CoRR abs/2212.08073 (2022). URL: https://doi.org/10.48550/arXiv.2212.08073. doi:10.48550/arXiv.2212.08073. arXiv:2212.08073.
[12] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. Tam, K. Stevens, A. Barhoum, N. M. Duc, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, A. Mattick, Openassistant conversations - democratizing large language model alignment, CoRR abs/2304.07327 (2023). URL: https://doi.org/10.48550/arXiv.2304.07327. doi:10.48550/arXiv.2304.07327. arXiv:2304.07327.
[13] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 2227–2237. URL: https://doi.org/10.18653/v1/n18-1202. doi:10.18653/v1/n18-1202.
[14] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: M. C. Elish, W. Isaac, R. S. Zemel (Eds.), FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021, ACM, 2021, pp. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.
[15] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.
[16] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.
[17] 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, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.
[18] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models, CoRR abs/2001.08361 (2020). URL: https://arxiv.org/abs/2001.08361. arXiv:2001.08361.
[19] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, L. Sifre, Training compute-optimal large language models, CoRR abs/2203.15556 (2022). URL: https://doi.org/10.48550/arXiv.2203.15556. doi:10.48550/arXiv.2203.15556. arXiv:2203.15556.

[20] OpenAI, Gpt-4 technical report, CoRR abs/2303.08774 (2023). URL: https://arxiv.org/abs/2303.08774. arXiv:2303.08774.

[21] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, N. Elhage, S. E. Showk, S. Fort, Z. Hatfield-Dodds, T. Henighan, S. Johnston, A. Jones, N. Joseph, J. Kernian, S. Kravec, B. Mann, N. Nanda, K. Ndousse, C. Olsson, D. Amodei, T. Brown, J. Kaplan, S. McCandlish, C. Olah, D. Amodei, J. Clark, Predictability and surprise in large generative models, in: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022, ACM, 2022, pp. 1747–1764. URL: https://doi.org/10.1145/3531146.3533229. doi:10.1145/3531146.3533229.

[22] D. Percia David, L. Maréchal, W. Lacube, S. Gillard, M. Tsesmelis, T. Maillart, A. Mermoud, Measuring security development in information technologies: A scientometric framework using arxiv e-prints, Technological Forecasting and Social Change 188 (2023) 122316.

[23] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python, Zenodo (2020). doi:10.5281/zenodo.1212303.

[24] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, A. Jatowt, Yake! keyword extraction from single documents using multiple local features, Information Sciences 509 (2020) 257–289.

[25] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: https://doi.org/10.5281/zenodo.4461265. doi:10.5281/zenodo.4461265.

[26] M. Kulkarni, D. Mahata, R. Arora, R. Bhowmik, Learning rich representation of keyphrases from text, in: M. Carpuat, M. de Marneffe, I. V. M. Ruíz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022, Association for Computational Linguistics, 2022, pp. 891–906. URL: https://doi.org/10.18653/v1/2022.findings-naacl.67. doi:10.18653/v1/2022.findings-naacl.67.

[27] L. Marujo, A. Gershman, J. G. Carbonell, R. E. Frederking, J. P. Neto, Supervised topical key phrase extraction of news stories using crowdsourcing, light filtering and co-reference normalization, in: N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (Eds.), Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012, European Language Resources Association (ELRA), 2012, pp. 399–403. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/672.html.

[28] D. Sahrawat, D. Mahata, M. Kulkarni, H. Zhang, R. Gosangi, A. Stent, A. Sharma, Y. Kumar, R. R. Shah, R. Zimmermann, Keyphrase extraction from scholarly articles as sequence labeling using contextualized embeddings, CoRR abs/1910.08840 (2019). URL: http://arxiv.org/abs/1910.08840. arXiv:1910.08840.

[29] A. Ushio, J. Camacho-Collados, T-NER: an all-round python library for transformer-based named entity recognition, in: D. Gkatzia, D. Seddah (Eds.), Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, EACL 2021, Online, April 19-23, 2021, Association for Computational Linguistics, 2021, pp. 53–62. URL: https://doi.org/10.18653/v1/2021.eacl-demos.7. doi:10.18653/v1/2021.eacl-demos.7.

[30] E. H. Hovy, M. P. Marcus, M. Palmer, L. A. Ramshaw, R. M. Weischedel, Ontonotes: The 90% solution, in: R. C. Moore, J. A. Bilmes, J. Chu-Carroll, M. Sanderson (Eds.), Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 4-9, 2006, New York, New York, USA, The Association for Computational Linguistics, 2006. URL: https://aclanthology.org/N06-2015/.

[31] K. Clark, M. Luong, Q. V. Le, C. D. Manning, ELECTRA: pre-training text encoders as discriminators rather than generators, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=r1xMH1BtvB.

[32] E. F. T. K. Sang, F. D. Meulder, Introduction to the conll-2003 shared task: Language-independent named entity recognition, in: W. Daelemans, M. Osborne (Eds.), Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003, ACL, 2003, pp. 142–147. URL: https://aclanthology.org/W03-0419/.

[33] N. Goyal, J. Du, M. Ott, G. Anantharaman, A. Conneau, Larger-scale transformers for multilingual masked language modeling, in: A. Rogers, I. Calixto, I. Vulic, N. Saphra, N. Kassner, O. Camburu, T. Bansal, V. Shwartz (Eds.), Proceedings of the 6th Workshop on Representation Learning for NLP, RepL4NLP@ACL-IJCNLP 2021, Online, August 6, 2021, Association for Computational Linguistics, 2021, pp. 29–33. URL: https://doi.org/10.18653/v1/2021.repl4nlp-1.4. doi:10.18653/v1/2021.repl4nlp-1.4.

[34] S. Ishizaki, D. Kaufer, Computer-aided rhetorical analysis, in: Applied natural language pro-

cessing: Identification, investigation and resolution, IGI Global, 2012, pp. 276–296. URL: https://www.igi-global.com/chapter/content/61054.

[35] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, ACL, 2014, pp. 1532–1543. URL: https://doi.org/10.3115/v1/d14-1162. doi:10.3115/v1/d14-1162.

[36] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[37] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Trans. Assoc. Comput. Linguistics 5 (2017) 135–146. URL: https://doi.org/10.1162/tacl_a_00051. doi:10.1162/tacl_a_00051.

[38] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: Y. Bengio, Y. LeCun (Eds.), 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings, 2013. URL: http://arxiv.org/abs/1301.3781.

[39] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

[40] L. McInnes, J. Healy, UMAP: uniform manifold approximation and projection for dimension reduction, CoRR abs/1802.03426 (2018). URL: http://arxiv.org/abs/1802.03426. arXiv:1802.03426.

[41] A. Hulth, Improved automatic keyword extraction given more linguistic knowledge, in: Proceedings of the 2003 conference on Empirical methods in natural language processing, 2003, pp. 216–223.

[42] T. Chari, J. Banerjee, L. Pachter, The specious art of single-cell genomics, BioRxiv (2021) 2021–08.