Text2Tech - Deep Learning-based Text Mining for Technology Monitoring in Automotive Production

Jan-Tilman Seipp, Felix Köhler, David Harbecke, Leonhard Hennig, Phuc Tran Truong

Background – Technology Monitoring in Automotive Production

The automotive industry is undergoing a transformative phase with the integration of advanced technologies and the rise of intelligent manufacturing systems. To remain competitive in this dynamic landscape, automotive production requires effective utilization of technology monitoring as a part of technology intelligence, which encompasses the acquisition, analysis, and application of relevant technological information. By harnessing NLP techniques, automotive manufacturers can extract valuable insights from vast amounts of unstructured textual data available in the form of patents, research papers, publicly funded projects, and industry news. The goal of the Text2Tech research project is to develop methods for automated extraction of technologies and its relations to other entities from unstructured text sources. We formalize this task as a combination of Named Entity Recognition (NER, Yadav et al., 2018) and Relation Extraction (RE, Bach et al., 2020). Both NER and RE are fundamental, well-researched tasks in Natural Language Processing, however, their application to novel domains such as automotive manufacturing is often hindered by the lack of training and evaluation data. Prior research has shown the promising performance of Large Language Models (LLM) in such low-resource scenarios, e.g. for approaches based on few-shot learning (Fritzler et al., 2019) and instruction-tuning (Wang et al., 2023). In this study, we present preliminary results on the performance of LLM-based few-shot and instruction-based learning for the task of low-resource Named Entity Recognition in the domain of technology monitoring.

Methods

We first define the entity types we are interested in detecting, and describe our data labeling approach. We then give a formal definition of the NER task and present the few-shot and prompt-based NER approaches we implemented.

Entities Definitions, Data Collection and Annotation Process

Table 1 lists the entity types of the technology domain that we aim to detect, and provides a brief explanation as well as examples for each type. We developed annotation guidelines iteratively by collectively labeling and discussing (domain and computational linguistics experts) a small set of test documents. Documents were drawn from a corpus of scientific publications, patent applications, and news reports specifically collected for the task of technology monitoring. Based on the annotation guidelines, and after labeling several training documents, one expert annotator with a background in linguistics labeled a training set of 200 documents. The test set of 10 documents was annotated by a total of 4 annotators. Inter-annotator agreement was moderate (Cohen's κ: Material: 0.29, Method: 0.39, Technological System: 0.55, Technical Field: 0.26).

Table 1. Entity type definitions and examples for Technology Monitoring.

Type	**Definition**	**Examples**
Technological System	any artificial, i.e. man-made artifact (object) or immaterial system	pump, car, software system
Material	raw, preprocessed or immaterial elements used or processed in a technological system	aluminium, glas, carbon nanotube, GPS data
Method	processes that describe the use, generation or creation of a system, a material or a service	injection molding, coating, linear regression

Technological Field	general application areas of technological systems	autonomous driving, solar energy production

NER Task Definition and Implementation

NER is typically formulated as a sequence labeling problem, where the input is a sequence of tokens $\boldsymbol{X} = \{x_1, x_2, \ldots, x_T\}$ and the output is the corresponding T-length sequence of entity type labels $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_T\}$. We implemented several different NER approaches: a) supervised fine-tuning on the training set of 200 documents (Devlin et al., 2019); b) few-shot learning using a pre-trained language model (Chen et al., 2022); and c) few-shot learning using a pre-trained language model fine-tuned on a dataset similar, but not identical to our domain[1]. In addition, we experiment with instruction-based NER, where we instruct a LLM to produce a list of entities in a given input text (Wang et al., 2023).

Experiments & Results

For supervised training, we fine-tune *xlm-roberta-base*[2] on the 200 training documents. For few-shot learning, we use the implementation provided by Chen et al. (2022)[3], using *bert-base-cased*[4] with and without fine-tuning on the FabNER dataset (Kumar & Starly, 2021). We sample instances from the 10 documents in the test split in a 4-way 1-shot setting, and report results averaged over 600 episodes. For all experiments, we report token-level micro-F1 scores, excluding the 'O' (other) class. Further, we conducted zero-shot NER experiments with instruction-tuned models such as ChatGPT (Brown et al., 2020), Llama (Touvron et al., 2023) derivatives and Falcon (Almazrouei et al., 2023) - the corresponding results are still work in progress but will be available at presentation time.

Table 2 lists the results of all four approaches. We observe that fine-tuning works better but few-shot is a suitable approach with limited data, as it requires less samples. *Materials* are easier to identify for few-shot models, whereas the fine-tuned model did much better classifying *Technological Systems*. These early results are promising, but not yet fully meaningful due to the small number of documents used in the evaluation.

Table 2. Token-level micro-F1 scores for different NER approaches.

Model	**F1**				
	Tech. Sys.	*Method*	*Material*	*Tech. Field*	*All*
Fine-tuning	0.60	0.34	0.51	0.29	0.48
Few-shot bert-base-cased	0.36	0.26	0.85	0.30	0.44
Few-shot FabNER-finetuned	0.30	0.29	0.85	0.40	0.47

Discussion & Conclusion

A significant challenge in the realm of this research pertains to the creation of suitably labeled datasets that adhere to the prescribed criteria in order to identify the mentioned technology entity types. This induces problems in training high quality models. Moreover, our preliminary findings offer indications of variations across the used text types (patents, news, scientific publications, research projects).

[1] We use the FabNER dataset from the manufacturing science domain, introduced by Kumar & Starly, 2021

[2] https://huggingface.co/xlm-roberta-base

[3] https://github.com/DFKI-NLP/fewie

[4] https://huggingface.co/bert-base-cased

The next steps in this project will focus on the research and development of models for relation extraction methods. Based on our definition for the different technology types, the models should be able to extract relations between the technological entities and other entity types like organizations.

References

Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., Penedo, G. (2023). "Falcon-40B: an open large language model with state-of-the-art performance". Online.

Bach, N. & Badaskar, S. (2007). "A Review of Relation Extraction". Online.

Brown, T. & Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. & Amodei, D. (2020). "Language models are few-shot learners." *In Proc. NIPS 2020*, pages 1877–1901.

Chen, Yuxuan, Jonas Mikkelsen, Arne Binder, Christoph Alt and Leonhard Hennig. (2022). "A Comparative Study of Pre-trained Encoders for Low-Resource Named Entity Recognition." *CoRR* abs/2204.04980.

Devlin, J. & Chang, M. & Lee, K. & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proc. NAACL-HLT, pages 4171–4186. ACL.

Fritzler, A. & Logacheva, V. & Kretov, M. (2019). "Few-shot classification in named entity recognition task." *In Proc. ACM/SIGAPP Symposium on Applied Computing (SAC '19)*. ACM, pages 993–1000.

Gao, T. & Fisch, A. & Chen, D. (2021). "Making Pre-trained Language Models Better Few-shot Learners." *In Proc. ACL-IJCNLP (Volume 1: Long Papers)*, pages 3816–3830, ACL.

Kumar, A. & Starly, B. (2021). "FabNER": information extraction from manufacturing process science domain literature using named entity recognition." *Journal of Intelligent Manufacturing* 33: 2393 - 2407.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). "LLaMA: Open and Efficient Foundation Language Models." *CoRR* abs/2302.13971.

Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K. R., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I. & Hajishirzi, H. (2023). "How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources." *CoRR* abs/2306.04751.

Yadav, V. & Bethard, S. (2018). "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models." *In Proc. COLING*, pages 2145–2158, ACL.

Noun Extraction Still Resists LLM-based Extractors for Cybersecurity

Maxime Würsch, Andrei Kucharavy, Dimitri Percia-David, Alain

Introduction

In an era dominated by rapid technological advancements, the prevalence and sophistication of cyber threats have similarly surged, necessitating the establishment of secure, reliable systems that can maintain operational continuity over time. This task presents its own set of unique challenges. The brisk pace of innovation in cybersecurity technologies implies a fleeting lifecycle, consequently exposing programs to an evolving landscape of vulnerabilities. This dynamic widens the security gap over time, demanding cybersecurity researchers to remain continually updated to fortify system security. A conventional approach to garner insightful entities pertaining to technological progress involves employing bibliometrics search, combined with entity extractors and embedding similarity. With the advent of Large Language Models (LLMs), the utilization of LLM-based extractors for entity extraction from documents has emerged as a commonplace practice. However, this has also ignited discussions among experts regarding the actual efficacy of these models. To contribute to this discourse, our work complements LLM-based extractors with a noun extraction mechanism, aimed at offsetting the potential shortcomings of LLMs. We will scrutinize the performance of noun extraction as a tool for classifying arXiv preprints in the cybersecurity domain, focusing on its ability to identify specific nouns within a listing and detect novel technologies unbeknownst to LLM-based extractors. This novel approach leverages the inherent structure of sentences, instead of relying solely on pre-trained information. We posit that integrating noun extraction can bolster the LLM-based extractors, offering a more robust tool for scientific bibliometrics. As such, our study provides a crucial step towards expanding the utility of LLMs and presenting a more comprehensive approach for tracking and analyzing advancements in the rapidly evolving field of cybersecurity.

Methods

Our code is accessible at the following link: https://anonymous.4open.science/r/gtm-BCF6. Our study utilized a dataset comprising arXiv preprints categorized under Computer Science (cs). We focus only on English language documents, cleansing the preprints by eliminating the preamble and references. We use spaCy library for noun extraction, including customized matcher rules to identify not only individual nouns but also compound nouns, such as "high school". Subsequent processing was required to correct errors that surfaced during the PDF-to-text conversion process. Each extracted noun was preserved only if it appeared at least four times in a specific preprint and across three separate documents within the same listing. Entities are returned as a set per preprint. To distinguish particular words, we executed a t-test and calculated a fold change that compared the frequency of these words in the specific listings with their frequency across the entire cs category, as visualized in a volcano plot (cf. repository). We assessed the similarity among different listings within the cs category using the spaCy embeddings and computing the average cosine distance. This led us to conduct hierarchical clustering (cf. repository) and employed low-dimensional projection algorithms to visualize the extracted nouns in various common embedding spaces. To facilitate an interpretable representation of our findings, we used subsampling to reduce the number of words within each listing. The results of these processes are illustrated in Figures 1 & 2. This methodology underpins the robustness of our analysis, equipping us to draw nuanced conclusions from the data.

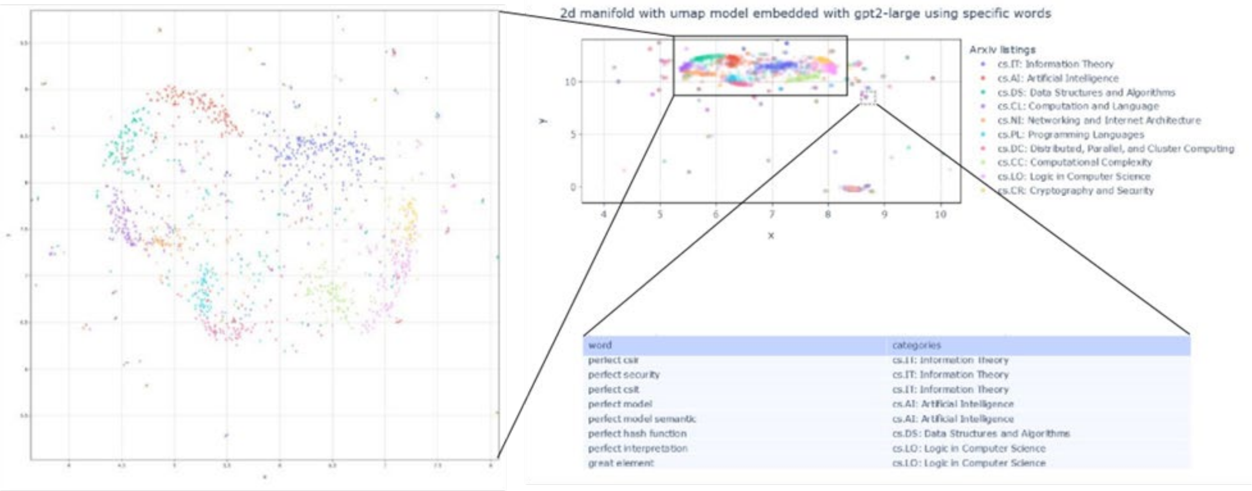
Results and Discussion

Our results validate the potency of noun extractors in obtaining specific nouns pertinent to the cybersecurity domain. As delineated in Table 1, the model demonstrates remarkable proficiency in isolating specific entities within a given listing. Notably, it proves adept at identifying the nomenclature of various "technologies," such as "x509," a widely recognized certificate standard, aligning well with the Cryptography and Security listing (cs.CR). These specific terms serve as accurate representations of each listing. Furthermore, the successful projection to a lower dimension offers valuable insights by identifying clusters representing different listings in the cs category of arXiv. Figure 1 underscores the efficacy of the UMAP algorithm within the GPT-2 embedding space, with clearly identifiable clusters. The smaller satellite clusters surrounding the primary one consist of highly similar entities featured across multiple listings, as visualized in Figure 2. The noun extraction process emerges as a competent tool in retrieving words that are either exceedingly specific or recent, hence not present in the fine-tuned dataset of LLM based extractors. It is our aspiration that this work will stimulate further investigations into this area. To support ongoing exploration, we are providing open-source access to all tools utilized for this project. Additionally, we offer interactive figures showcasing various embeddings and 2D projection algorithms, along with supplementary figures to substantiate our findings. This approach solidifies our commitment to advancing the field by ensuring our resources and discoveries are accessible and easy to build upon.

Table 1. Sample of nouns extracted in the Cryptography and Security (cs.CR) listing.

Raw nouns extraction in cs.CR	specific nouns in cs.CR
Boolean function	Trapdoor
Diagonal matrix	Homomorphic cryptosystem
Recall	Brute force attack
Error probability	Possible collusion
Subroutine	x509

Figures 1 & 2. 2D projection with UMAP embedded with GPT-2 of extracted nouns and a list of selected nouns in the figure.



word	categories
perfect csir	cs.IT: Information Theory
perfect security	cs.IT: Information Theory
perfect csit	cs.IT: Information Theory
perfect model	cs.AI: Artificial Intelligence
perfect model semantic	cs.AI: Artificial Intelligence
perfect hash function	cs.DS: Data Structures and Algorithms
perfect interpretation	cs.LO: Logic in Computer Science
great element	cs.LO: Logic in Computer Science

References

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 'SpaCy: Industrial-Strength Natural Language Processing in Python'. *Zenodo*, 2020. https://doi.org/10.5281/zenodo.1212303.

Ishizaki, Suguru, and David Kaufer. 'Computer-Aided Rhetorical Analysis'. In *Applied Natural Language Processing: Identification, Investigation and Resolution*, 276 96. IGI Global, 2012. https://www.igi-global.com/chapter/content/61054.

McInnes, Leland, and John Healy. 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. *CoRR* abs/1802.03426 (2018). http://arxiv.org/abs/1802.03426.

Percia David, Dimitri, Loïc Maréchal, William Lacube, Sébastien Gillard, Michael Tsesmelis, Thomas Maillart, and Alain Mermoud. 'Measuring Security Development in Information Technologies: A Scientometric Framework Using ArXiv e-Prints'. *Technological Forecasting and Social Change* 188 (2023): 122316.

Dynamic Decision Model Based on Data-driven Agent-Based Modeling - the Case of Licensing Agreement in Life Science

Shihhsin Chen, Duenkai Chen, An-Chen Kao

This study provides a dynamic model-building approach that uses machine learning results to build the agent-based model. Using trained machine learning models as a reference for agent behavior avoids the subjectivity often associated with traditional agent-based model development, which can lead to criticisms of being biased from real-world scenarios. Biases in agent behavior may result in misjudgments and discrepancies between simulation results and expectations. In this experiment, real-world data is used to drive the model construction, and the use of well-trained machine learning models based on data-driven training significantly reduces previously mentioned concerns, focusing solely on the accuracy of model predictions. We are applying this method to analyze licensing agreements within strategic alliances in two biotechnology industry databases. After extracting features from the data, we use the agent-based model to simulate and observe the impact of these features. Through validation, the factors align with those studied in previous literature, supporting the approach used in this experiment.