2nd Conference on Production Systems and Logistics

# Feasibility Analysis of Entity Recognition as a Means to Create an Autonomous Technology Radar

Günther Schuh[1], Jan Hicking[1], Max-Ferdinand Stroh[1], Justus Benning[1], Clinton Gnanaraj Charles[1]

[1]*FIR, Institute for Industrial Management at RWTH Aachen University, Aachen, Germany*

**Abstract**

Keeping up to date with the latest technology trends is crucial task for manufacturing companies to remain successful on a globally competitive market. Designing a technology radar is an established, yet mostly manual, process for visualizing recent technology trends.

The challenge of identifying and visualizing technologies is addressed by the project TechRad which uses machine learning to realize an autonomous technology scouting radar. One of its core functionalities is the identification of technologies in text documents. This is implemented via Natural Language Processing (NLP).

This paper aims to summarize the challenges and possible solutions for using entity recognition to identify relevant technologies in text documents. The authors present an early stage of implementation of the entity recognition model. This contains the selection of Transfer Learning as a suitable method, the creation of a dataset consisting of different data sources, as well as the applied model training process. Finally, the performance of the chosen method is benchmarked and evaluated in a series of tests.

**Keywords**

Machine Learning; Technology Management; Natural Language Processing; Entity Recognition

## 1. Introduction

This chapter aims to give a brief introduction to the problem at hand, describing the overarching research goal of the automated technology radar as well as providing motivation and context for the solution chosen in this work.

### 1.1 Challenge

Due to the constantly growing number of technologies available on the market, the number of devices using different digital technologies is growing exponentially [1]. This observation, alongside the fact that the time until a certain technology is known to a large number of users is diminishing [2], indicates that the frequency at which both companies and private users are exposed to new technologies is rising [3].

Organizations have to innovate in order to succeed and stay relevant in the market [4]. Mastering the process of finding technologies and managing innovations is a key success factor to ensure a company's market position [5]. Being unable to oversee the growing technology market endangers companies' long term strategic positions and ultimately their market position, and may even result in bankruptcy [3].

### 1.2 Solution

We are addressing the aforementioned issue by designing an automated technology scouting radar that gives an overlook over recent technological trends while keeping the research effort to a minimum. The software uses recent advancements in the field of Natural Language Processing (NLP), a sub-field of artificial intelligence (AI), to scout for information about technology in various sources. In a previous publication, the authors proposed an architecture for the automated radar [3].

The current paper focuses on a core functionality of the tool, namely identifying technologies in a written document. This functionality is realized using Named Entity Recognition (NER). The authors present the steps involved in the building and evaluation of the proposed Named Entity Recognition model.

### 1.3 Structure of the paper

Following the introduction in section 1, section 2 gives information about the state of the art and basic definitions. Section 3 summarizes the architecture of the automated radar to put the functionality of recognizing technologies into context. Section 4 describes the data sourcing and model building process in detail. After discussing the results in section 5, section 6 summarizes the insights. Section 7 concludes the paper by giving an outlook to future research in the area.

## 2. State of the art

The following paragraphs provide information about the state of the art and basic definitions to ensure a common understanding of the topics and terms used in the solution.

### 2.1 Technology management and visualization

A technology radar is a tool to summarize and visualize the results of a technology scouting process that organizes technologies in a circular diagram. The diagram is divided into sectors for structuring the content and delimiting the search areas, e.g., trends, technology fields, production technologies or product functions [5]. Figure 1 shows an example of a technology radar.

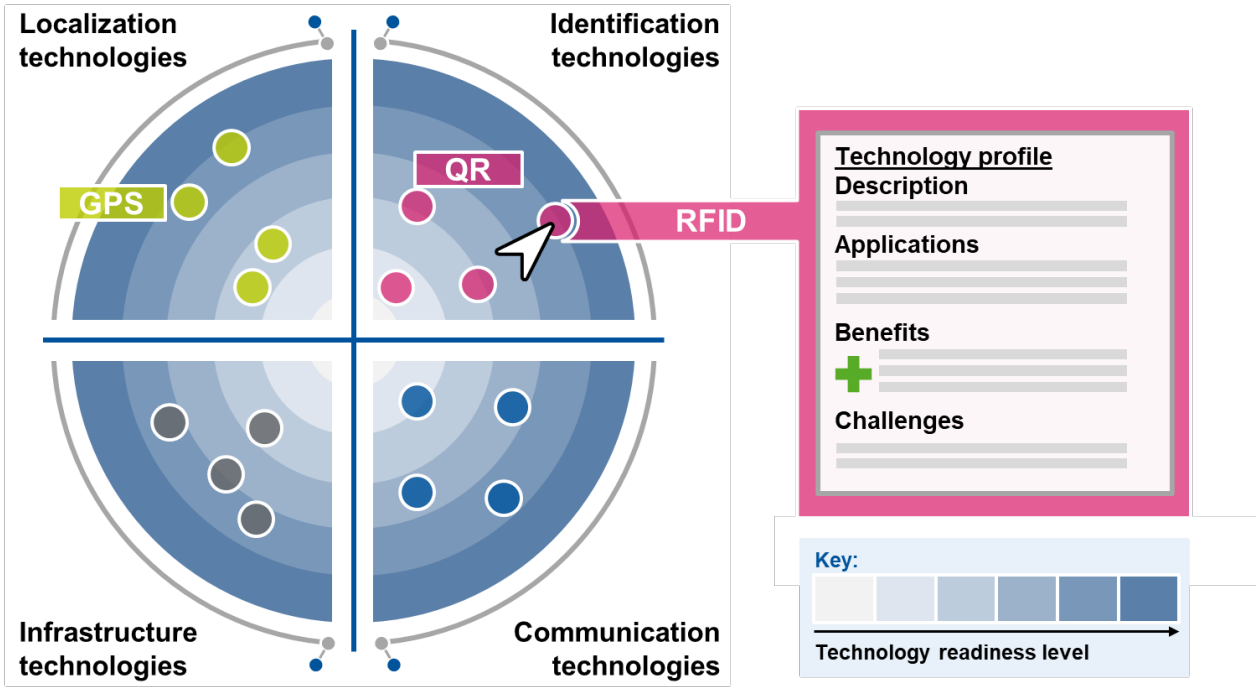


Figure 1: An example of a technology radar [6]

A temporal perspective is mapped on the axes, mostly the technology maturity, which indicates whether a search field or a specific technology is close to market readiness or in a research phase [5]. The underlying architecture proposed in our solution is explained in detail in section 3.

### 2.2 Data collection

Data collection is one of the main tasks when building AI models. Collecting data is not a trivial undertaking and demands a coherent approach. Especially in broadly scoped endeavours like the presented solution, the scientist encounters challenges like the lack of publicly available resources as well as copyrights [7]. The problem discussed in this paper demands a specialized corpus of data about emerging technologies. This is realized by using standardized API queries from scientific publication portals. The approaches and sources of data used are explained in detail in section 4.

### 2.3 Named Entity Recognition

A named entity is defined as a word or a phrase that clearly identifies one item from a set of other items that have similar attributes [8]. Named Entity Recognition (NER) is the problem of identifying sections of a text or certain words that mention named entities, and to subsequently classify them into predefined categories [9]. NER serves is a core functionality of many natural language applications such as translation, context sensitive answering and summarization [7]. In this paper, we present the use of NER in automatically identifying technology terms from a corpus using the advancements in deep learning techniques.

## 3. Architecture including identified functions

The previous work focused on the design process of a possible architecture for the radar [3], as it is a crucial preliminary step of the software and systems engineering process [10]. In this section, a brief summary of the process and results is given to ensure that the context in which Named Entity Recognition is used becomes clear (see also Figure 2).

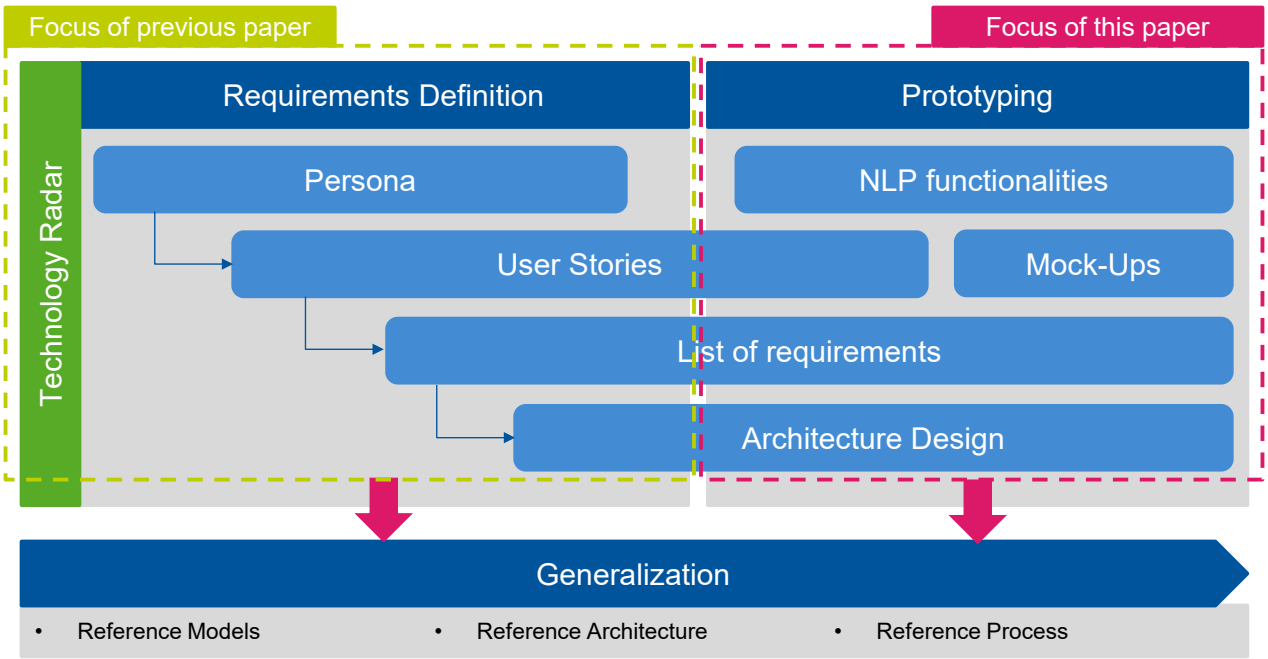


Figure 2: Overall project approach and focus of the paper [3]

### 3.1 Gross Functional Structure and Modules of TechRad

The automated technology radar is a solution consisting of four main steps. In data gathering, API queries and a web crawler are used to gather data from public sources, such as scientific publication portals, blogs and social media [3]. In the training phase presented in this paper, we used data exclusively from scientific publication portals and standardized social media APIs, planning on broadening the scope in further phases of the project. In the second step, data storage, the documents are allocated and organized for further processing. During the analysis the Full text documents will only be stored temporarily (cache). A document's metadata is stored for future reference. The third step is the analysis, in which the text files are checked for technologies using NER, the main focus of the paper. In a successive analysis step the technologies will be classified into technology readiness levels using different NLP methods. The fourth step is largely focused on the front-end and deals with user-friendly visualization of the extracted information. The design of the radar is presented in detail in [3]. A graphical presentation of the steps can be found in Figure 3.

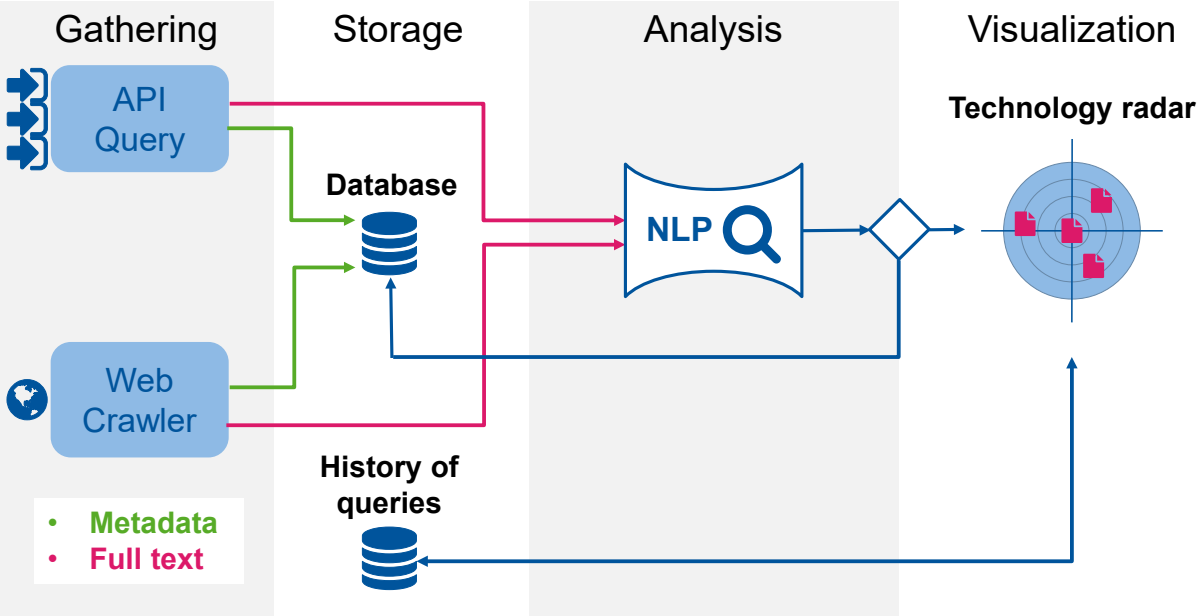


Figure 3: Architecture of the Radar [3]

## 4. Entity recognition for the identification of technologies

The following section describes the stages involved in building the Named Entity Recognition model. The success of the implementation depends on the quality of data sources, the pre-processing of data which is fed to the model and the hyperparameters used in the actual training procedure.

### 4.1 Data sources

With recent advances in machine learning, fostered by techniques such as deep learning, many tasks can be solved once a sufficiently large dataset is available for training. Nevertheless, human-annotated datasets are often expensive to produce, especially when the labels are used in high volume and frequency, as it is the case with word-level-annotation in NER [11].

Various sources were explored to build the necessary training data. As the application demands data rich in emerging technology terms, newly published scientific documents were considered as ideal sources. ArXiv [12] is a free distribution service and an open-access archive for scholarly articles in many emerging fields.

In addition to ArXiv, other major sources of rich scientific information like Google Scholar [13] and the Database and Logic Programming portal (DBLP) [14] were used.

To supplement the above-mentioned data sources, social media data were also collected through standardized API queries. A list of keywords such as Text Mining and Natural Language Processing were used as search terms through API queries in social media sites like Twitter. After the collection of sufficient quantity of data, pre-processing and training were performed.

### 4.2 Pre-processing

As an initial step, pre-processing of the data involves selecting chunks of texts from the collected data which are relevant for the application. The abstracts of the collected scientific papers were extracted and used as the training data. The next step involves preparing our data by annotating our text with "technology" tags. This is a labour-intensive task and is accomplished with the help of an open source text annotation tool for humans called "Doccano".

SpaCy is a free open-source library for Natural Language Processing in Python which features in-built NER [15]. Even though pre-built SpaCy models are good at NER extraction, they are not good enough for customized applications as the training data used is not specific to latest technologies. Therefore, the training data is manually labelled using Doccano.

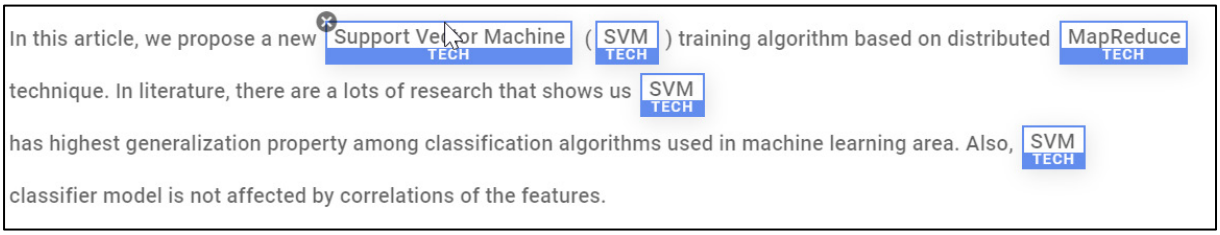


Figure 4: Training data annotation

Figure 4 shows a sample of the training data annotation. A technology entity is defined in the context of our application in the following way: if the entity is an algorithm, a commercial product, library or a framework, it is considered as a technology entity. Standard text corpus, performance evaluation metrics and broad umbrella terms were not considered as technology entities. Doccano provides a GUI for data annotation and the output is stored in a JSONL delimited file with each sentence along with the entity offsets. Figure 5 shows a sentence of the output JSONL file. The task of pre-processing also included procedures like un-latexing.


```
{"id": 133706, "text": "However, the multi-step process still deviates from the unified optimization
generally performed with conventional methods such as k-Means. We propose a pure neural framework for
"meta": {}, "annotation_approver": null, "comments": [], "labels": [[180, 187, "TECH"]]}
```


Figure 5: Sample JSONL file

### 4.3 Training procedure

The following paragraph describes the training procedure of the Named Entity Recognition algorithm. The SpaCy projects repository includes various project templates for different Natural Language Processing tasks, models, workflows and integrations. It allows to manage end-to-end SpaCy workflows for different use cases and domains. For this application, a project template for NER was selected and the pipeline was customized for better performance.

The output files from Doccano were converted into a SpaCy compatible format and stored, which was used as the training data in the pipeline. SpaCy uses a config file that contains all the model training components

to train the model like component type, which is NER in this case, hardware accelerator selection and optimization goals. The algorithm was trained using 'SciBERT' as Transformer with dropout and 'Adam.v1' as the optimizer. Figure 6 shows an exemplary annotation performed by our algorithm on the test data.

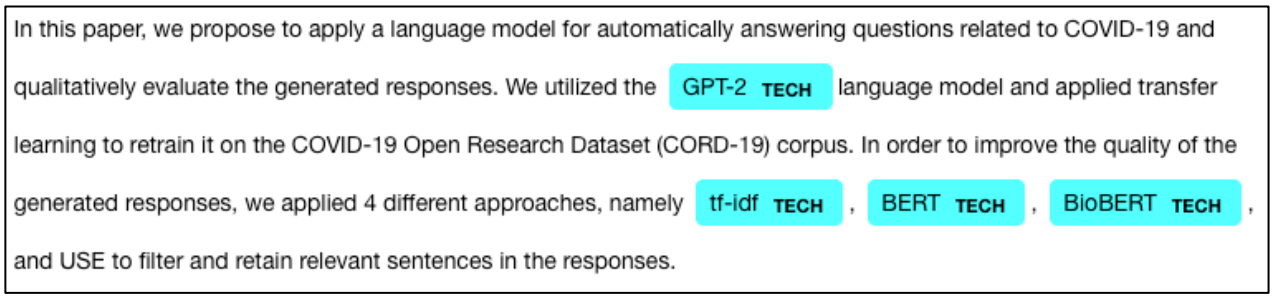


Figure 6: Annotation by trained NER algorithm

## 5. Results and Evaluation

The results, evaluation and benchmarks are presented in this section. Table 1 represents the evaluation of our algorithm on the test data based on a 70/30 test train split. The column 'Epochs' represents the number of epochs in training. The column 'Iterations' represents the number of iterations or steps in that particular epoch, 'F-Score' represents the F-Score, which is the harmonic mean of precision and recall. For a Named Entity Recognition task, the ideal 'F-Score' should be close to 1, which would represent a perfect model. The columns 'Precision' and 'Recall' represent the metrics precision and recall respectively. Precision denotes the percentage of predicted annotations that were correct, while recall denotes the percentage of reference annotations rightly recovered. Both these metrics should increase close to 100 for an ideal model. Our model was trained on a total of 1050 positive samples containing 'TECH' entities.

Table 1: Evaluation of the NER model

| Epochs | Iterations | F-Score | Precision | Recall |
|---|---|---|---|---|
| 4 | 4000 | 0.74 | 70.72 | 77.47 |
| 4 | 4200 | 0.73 | 77.68 | 68.49 |
| 4 | 4400 | 0.71 | 73.47 | 69.62 |
| 4 | 4600 | 0.71 | 65.37 | 78.61 |

### 5.1 Benchmarking in the general context

Transformer-based neural architectures are changing the field of NLP with an attention-based mechanism that outperforms convolutional or recurrent models [16]. Current NLP models are mostly based on deep neural networks which are characterized by great performance but are notoriously opaque in their prediction process [17]. Although different standard metrics have been proposed to standardize the evaluation, the authors stick to the common machine learning metrics like F-Score for evaluation, considering the ease of comparison between algorithms.

The F-Score, Precision and Recall shown in Table 1 are matching expectations of the research team and are close to similar efforts, where transformers have been used to solve NER problems on scientific corpora (see Table 2). These are comparable in so far as similar procedures have been used (fine-tuning a pre-trained model) and are tested on scientific corpora. However, the table is used to support the point that an F-Score of around 0.7 is acceptable with this training method applied to a scientific corpus; a direct performance comparison is not made, as the validation dataset would have to be the same, which is not desired at the current state of the prototype.

Table 2: Named Entity Recognition System evaluation[1]

| Named Entity Recognition System | F-Score | Source |
|---|---|---|
| SpERT | 0.73 | [18] |
| RDANER | 0.69 | [19] |
| Cross-sentence | 0.68 | [20] |

### 5.2 Critical Reflection

NER models are often trained based on formal documents and publications. Informal web documents that would be incorporated into the data basis by using web crawling or more social media sources usually contain noise, as well as incorrect and incomplete expressions. The performance of current NER systems generally decreases as informality increases in web documents [21]. This can be rectified using some post-processing, but could still pose a challenge in the further phases of development.

In general, deep learning techniques are data-hungry and the performance of these models increases with more data. Adding more training data could further improve the performance of our model, however the labelling process proved to be resource intensive. Another lever for better scores is hyperparameter optimization and tuning the config-parameters during the training process. For a prototype of the technology radar, an F-Score of about 0.7 is thought to be sufficient as it demonstrates technical viability and feeds the successive steps with relevant data to process and visualize. Still, the authors plan on increasing the precision of the model in further stages of the project.

## 6. Summary

In this paper, the authors present a prototype of an autonomous technology radar using NLP. The focus of the presented research lies on a core functionality problem that was solved using Named Entity Recognition. In the beginning, the need for automation in technology scouting is explained. The architecture of the autonomous radar is described, followed by the implementation and the evaluation of a prototype using NER to successfully identify technologies in text documents. The authors then evaluate the usefulness of the approach, which is deemed sufficient for the current state of the implementation.

## 7. Outlook

With the prototype showing promising results, the feasibility check for using Named Entity Recognition to develop an autonomous technology radar is performed successfully. The prototype can be used as a first step to build the final visualization of the results. Further research will focus on the steps to improve the accuracy of the algorithm. The architecture and the algorithm will be further evaluated with industry experts and potential users. Based on the feedback, the process will be fine-tuned.

## Acknowledgements

This project has been funded by the Leitmarkt.NRW program (EFRE-0801386) and the European Union in the European Regional Development Fund (EFRE) under the name “TechRad”. The authors wish to

[1] benchmarking conducted with https://paperswithcode.com in March 2021

acknowledge the EFRE for their support. We also wish to acknowledge our gratitude and appreciation to all the “TechRad” project partners for their contribution during the development of various ideas and concepts presented in this paper.

## References

[1] Parasuraman, A., 2000. Technology Readiness Index (Tri). Journal of Serivce Research 2 (4), 307–320.

[2] Statista Research Development, 2018. Informationstechnologien: Entwicklung bis 50 Millionen Nutzer. https://de.statista.com/statistik/daten/studie/298515/umfrage/entwicklung-ausgewaehlter-informationstechnologien-bis-50-millionen-nutzer/. Accessed 21 January 2020.

[3] Schuh, G., Hicking, J., Stroh, M.-F., Benning, J., 2020. Using AI to Facilitate Technology Management – Designing an Automated Technology Radar. Procedia CIRP 93, 419–424.

[4] Hoerlsberger, M., 2019. Innovation management in a digital world. JMTM 30 (8), 1117–1126.

[5] Schuh, G., Klappert, S., 2011. Technologiemanagement. Springer Berlin Heidelberg, Berlin, Heidelberg.

[6] Stroh, M.-F., Benning, J., 2020. Trends finden und überblicken für KMU – Teil 1. MM MaschinenMarkt 126 (8), 60–61.

[7] Zeroual, I., Lakhouaja, A., 2018. Data science in light of natural language processing: An overview. Procedia Computer Science 127, 82–91.

[8] R. Sharnagat, 2014. Named entity recognition: A literature survey: Center For Indian Language Technology.

[9] Nadeau, D., Sekine, S., 2007. Named Entities: Recognition, classification and use. LI 30 (1), 3–26.

[10] ISO, 2011. Systems and software engineering - Architecture description, 2011st ed., 48 pp.

[11] Menezes, D.S., Savarese, P., Milidiú, R.L., 2019. Building a Massive Corpus for Named Entity Recognition using Free Open Data Sources. http://arxiv.org/pdf/1908.05758v1.

[12] Cornell University. https://arxiv.org/. Accessed 17 May 2021.

[13] Alphabet. https://scholar.google.de/. Accessed 17 May 2021.

[14] Schloss Dagstuhl - Leibniz Center for Informatics. Schloss Dagstuhl - Leibniz Center for Informatics. https://dblp.org/. Accessed 17 May 2021.

[15] Industrial-Strength Natural Language Processing, 2021. https://spacy.io/.

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need, 15 pp. http://arxiv.org/pdf/1706.03762v5.

[17] DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C., 2019. ERASER: A Benchmark to Evaluate Rationalized NLP Models. http://arxiv.org/pdf/1911.03429v2.

[18] Eberts, M., Ulges, A., 2019. Span-based Joint Entity and Relation Extraction with Transformer Pre-training. http://arxiv.org/pdf/1909.07755v3.

[19] Yu, H., Mao, X.-L., Chi, Z., Wei, W., Huang, H., 2020. A Robust and Domain-Adaptive Approach for Low-Resource Named Entity Recognition, 297–304.

[20] Zhong, Z., Chen, D., 2020. A Frustratingly Easy Approach for Entity and Relation Extraction. http://arxiv.org/pdf/2010.12812v2.

[21] Kim, M.H., Compton, P., 2012. Improving the Performance of a Named Entity Recognition System with Knowledge Acquisition, in: ten Teije, A., Völker, J., Handschuh, S., Stuckenschmidt, H., d'Aquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N. et al. (Ed.), Knowledge Engineering and Knowledge Management, vol. 7603. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 97–113.

**Biography**

**Günther Schuh** (*1958) is head of the chair of Production Systems (WZL-PS) at RWTH Aachen University and member of the directorates of the Machine Tool Laboratory (WZL) at the RWTH, Fraunhofer Institute for Production Technology (IPT) and Director of the FIR at RWTH Aachen University. He has had entrepreneurial success with multiple start-ups and spin-offs and is a member of the Presidium of the German Academy of Technical Sciences (acatech).

**Jan Hicking** (*1991) has been head of the division Information Management at FIR at RWTH Aachen University since 2020. Starting in 2016, he received his Ph.D. in 2020 in the field of intelligent products. As head of the division, he is responsible for multifaceted consulting and research projects.

**Max-Ferdinand Stroh** (*1991) is a researcher at FIR at RWTH Aachen University since 2017 in the department Information Management. He is leading the group Information Technology Management at FIR. He is also deputy head of the department Information Management. His scientific work is focused on the practical application of AI, smart products and IT-OT-Integration.

**Justus Aaron Benning** (*1995) is a researcher at FIR at RWTH Aachen University in the department of Information Management since 2019. He is leading the group Information Logistics and is head of software development in his department. While his degree is in mechanical engineering and business administration, he spent a semester abroad at Korea University in Seoul to focus on the business applications of artificial intelligence – the main focus of his current research and publications.

**Clinton Gnanaraj Charles** (*1993) is a master student at RWTH Aachen University studying Data Analytics and Decision Science since 2019. He has a Bachelor's degree in Automobile Engineering, followed by 4 years of work experience in the automotive industry in India and United States. Fascinated by the potential of data and artificial intelligence, he decided to pursue further education in this field. He has been associated with the FIR as a research assistant since 2020.