Towards Enabling Emerging Named Entity Recognition as a Clinical Information and Argumentation Support

Christian Nawroth, Felix Engel, Tobias Eljasik-Swoboda and Matthias L. Hemmje
Lehrgebiet Multimedia und Internetanwendungen, FernUniversität in Hagen, Universitätsstraße 47, Hagen, Germany

Keywords: Emerging Named Entity Recognition, Data Science, Natural Language Processing, Information Retrieval, Clinical Argumentation Support.

Abstract: In this paper we discuss the challenges of growing amounts of clinical literature for medical staff. We introduce our concepts emerging Named Entity (eNE) and emerging Named Entity Recognition (eNER) and show the results of an empirical study on the incidence of eNEs in the PubMed document set, which is the main contribution of this article. We discuss how emerging Named Entities can be used for Argumentation Support, Information Retrieval (IR) Support and Trend Analysis in Clinical Virtual Research Environments (VREs) dealing with large amounts of medical literature. Based on the empirical study and the discussion we derive use cases and a data science and user-feedback based architecture for the detection and the use of eNEs for IR and Argumentation Support in clinical VREs, like the related project RecomRatio.

1 INTRODUCTION

The amount of medical literature is growing, following the global trend of information explosion (Huth, 1989) and Information Overload (Bawden and Robinson, 2009): While in 1980 279.692 citations were added to PubMed / MEDLINE, in 2016 1.178.360 were added, which means that the yearly growth rate increased by the factor 3.6 within 35 years (U.S. National Library of Medicine, 2017). In parallel not only the amount of literature is growing but also the extend of medical vocabularies like Medical Subject Headings (MeSH) (U.S. National Library of Medicine, 1999), which grew by 12226 entries within 10 years from 2007 till 2016. Each of these new entries typically identifies a new medical concept or represents at least a new name for an existing concept. This growth of textual medical data in literature as well as the increase of the medical vocabulary is one major challenge for medical staff today. The sheer amount of new textual data and medical literature cannot be overseen manually for example when recent literature should be used to argument for or against a therapy (individual use case) or when actual trends in public health should be used to support generic planning or argumentation processes in health systems (comprehensive use case). So acquiring and assimilating evidence for decision making is difficult for clinicians and researchers (Hunter and Williams, 2015). This coincides with the observation that (clinical) text data is most pervasive in electronic health records (EHR) (Jensen et al., 2012) while there is a lack of training [of scientists] on processing large unstructured text data (Garmire et al., 2016). The objectives of combining data scientists work with medical experts' knowledge and experience are to a) identify individual recent medical concepts represented by new vocabulary to make them available for daily individual patient-related work of physicians and in nursing (individual use case) and b) identify trends represented by recent vocabulary to be used in health management (comprehensive use case). In this article we present our approach which covers both use cases. To outline the approach, we first define our concept of emerging Named Entities (eNEs) and present experimental results on the incidence of eNEs in medical literature to show the (statistical) relation between eNEs and emerging clinical knowledge. The definition and the experiments are the main contribution of this article. Based on both we then introduce a first approach for a framework for the semi-automated eNE recognition (eNER) for further use in clinical Virtual Research Environments (VREs). Our framework implements the BDMCube Framework to be able to process large amounts of textual data. A related project that utilizes both the

Nawroth, C., Engel, F., Eljasik-Swoboda, T. and Hemmje, M.
Towards Enabling Emerging Named Entity Recognition as a Clinical Information and Argumentation Support.
DOI: 10.5220/0006853200470055
In *Proceedings of the 7th International Conference on Data Science, Technology and Applications (DATA 2018)*, pages 47-55
ISBN: 978-989-758-318-6

individual as well as the comprehensive use cases is RecomRatio, a VRE to support argumentation processes of medical staff in clinical decisions. RecomRatio which is part of the DFG Schwerpunktprogramm "Robust Argumentation Machines" (SPP 1999) (DFG, 2016) primarily is intended to support individual argumentations for individual clinical cases, e.g. for or against a therapy of an individual patient. For this individual use case our approach extracts textual data and related context representing recent medical knowledge based on an individual query from a physician for argumentative support. The aim is to provide the physician the latest medical knowledge to raise awareness for more recent and alternative argumentations beyond his actual individual query. For the comprehensive use case our approach provides a graphical visualization of ongoing trends and emerging concepts represented by emerging within RecomRatio without an individual query but regarding more general information needs e.g. in argumentation processes within clinical planning.

2 STATE OF THE ART AND RELATED WORK

Our work is based on the technique of Named Entity Recognition (NER) (Nadeau and Sekine, 2007), which is a subtask of Natural Language Processing (NLP). Named Entity Recognition is used for Information Extraction (IE) (Sang et al., 2003) and thus for discovering knowledge from free-text (Piskorski and Yangarber, 2013). Amongst others - like Part of Speech (POS) one feature of textual data is the information, whether a text token identifies a name e.g. of a person, a location or in the medical domain of a disease or a drug. Jurafsky and Martin (2009) define the task of Named Entity Recognition (NER) as "the combined task of finding spans of text that constitute proper names and then classifying the entities being referred to according to their type". This is a common definition of NER which is referenced in multiple works analogously such as Grishman (1995). Earlier works also refer to Named Entity Recognition and Classification (NERC) (Nadeau and Sekine, 2007). A more specialized use case for NER is decision making and argumentation mining, in which Named Entities (NEs) besides others are used for argumentation boundary detection (Lippi and Torroni, 2016), which is also addressed by our work. We try to recognize and classify Named Entities that are characteristic for arguments for clinical decisions and provide them for an argumentation support system in our project RecomRatio and identify trends represented by Named Entities. When trying to address this use cases in VREs – like RecomRatio – which contain emerging knowledge (Patel and Ghoneim, 2011) there is a major challenge: Emerging knowledge "arises suddenly and unexpectedly and it cannot be planned and predicted" (Patel and Ghoneim, 2011).

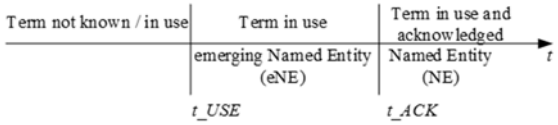

Figure 1: Graphical Definition of eNE.

Although there exist NER methods based on ML that can detect yet unknown NEs in highly specialized fields (like genes) when it comes to broader and interdisciplinary domains, learning-based approaches fail due to the lack of appropriate training data, which creation is resource intensive and often requires know how of both the respective domain and linguistic or Natural Language Processing knowledge. The idea is to utilize user feedback for improving Information Retrieval (IR) processes is not new: User Relevance Feedback (Rocchio, 1971) for a long time is a well-known Information Retrieval (IR) technique for improving IR search result. While traditional User Relevance Feedback refers to IR tasks, Finin et al., (2010) present an approach in which they successfully use feedback through crowdsourcing (Amazon MTurk) for Named Entity Recognition in Twitter messages. They show that crowdsourcing can be used to identify NEs of the traditional categories Person, Organization and Location. We extend their approach to address the specific needs of our project: As the (e)NEs to be identified in VREs are too specific we do not use anonymous annotators via a generic crowdsourcing but domain experts from the respective VREs. Although this leads to a much smaller number of annotators we benefit from their higher domain specific confidence. We also do not limit our approach to the three categories but use categories depending on domain specific needs, for example based on existing taxonomies, like the MeSH tree structure. To the best of our knowledge there do not exist approaches to use user feedback or crowdsourcing approaches to improve the quality of NER models on emerging knowledge in a clinical setup. Upstream to the user feedback we use statistical pattern recognition and classification based on ML. Besides recent Deep Learning based

Table 1: Number of Terms from MSHNEW per year.

2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	SUM
397	740	2147	1426	1913	2279	2489	232	241	335	12199

methods as shown by (LeCun et al., 2015), Support Vector Machines (SVM), are still recognized as robust and efficient ML based methods for classification tasks (Hearst et al., 1998; Joachims, 1998). A related classifier-based approach which focuses on document classification (instead of term classification) but deals with similar challenges (lack of gold-standards) is introduced by (Eljasik-Swoboda et al., 2018). The authors describe a text categorization classifier that does not require a target function to bootstrap text categorization (TC). This way, the need for training examples and gold standards, which are time- and expert work-intensive to create and maintain, is eliminated. Their approach overcomes the supervised learning pattern and provides quick, tangible classification results based on vector space semantics. Besides user feedback and statistical pattern recognition as described our approach also utilizes state of the art NLP techniques. Regarding that, Stanford Core NLP is considered to be one state of the art framework (Manning et al., 2014). It provides a set of generic NLP functions, such as tokenizing and part of speech tagging, which we use for baseline NLP. For NER (Chang and Manning, 2014) propose to complement statistical / supervised learning based NER methods with rule-based approaches, especially when there is no appropriate training data available. That's why our approach for recognizing eNE-candidates combines Conditional Random Field (CRF) based Stanford NER (Finkel et al., 2005) and Stanford TokensRegex for rule based NER. A recent application-example for domain specific adoption of Stanford CRF-NER for smartphone related use case is given by Deubzer et al., (2016). There exist several systems for applying Natural Language Processing and Named Entity Recognition to the clinical sector, such as MetaMap (Aronson, 2001; Aronson and Lang, 2010) and cTakes (Savova et al., 2010). These systems use dictionary-based approaches to identify existing Named Entities, while our approach relies on a combination of user feedback, statistics and rule based / machine learning based NER. The idea of utilizing Named Entities for Information Retrieval support in the clinical domain is already implemented through the MeSH on demand platform. MeSH on demand recognizes NEs from the MeSH vocabulary within user queries, provides additional context for them, uses them to provide related articles and shows related MeSH terms. Our approach goes beyond MeSH on demand by not only identifying and using existing NEs but emerging NEs, collecting user feedback on them and visualize them. For evaluation of (e)NER performance. we use standard metrics Recall, Precision and F-Measure as proposed in CONLL 2003 (Sang et al., 2003) using gold standard corpora lie the GENIA (Ohta, Tateisi and Kim, 2002) and CRAFT (Bada et al., 2012). Our approach implements the BDMCube meta model for big data management (Kaufmann et al., 2017). As related work in the field of clinical argumentation support Hunter and Williams present a framework for the aggregation of clinical evidence using argumentation (Hunter and Williams, 2015). While their approach focuses on combining already extracted evidence from clinical trials, our approach is going to detect evidence in textual document represented through eNEs, which may later be processed through a similar framework like Hunter and Williams presented.

3 EMERGING NAMED ENTITIES

3.1 Definition

We extend the concept of Named Entities and Named Entity Recognition as defined above and define emerging Named Entity (eNE) as follows:

A term, that is in use in domain specific literature since the time t_USE and which is afterwards acknowledged as a Named Entity by respective expert community (e.g. through adding the term to a domain specific vocabulary) at the time t_ACK is defined as an emerging Named Entity (eNE) for the time interval [t_USE, t_ACK[. The aim of emerging Named Entity Recognition(eNER) is to recognize eNEs during the time interval [t_USE, t_ACK[.

3.2 Experimental Setup

The generic aim of the experiments is to find out, if there exists a relation between emerging knowledge in clinical VREs and eNEs as defined before.
To investigate this possible relation we relied on two well-known and trusted sources: The Medical

Subject Headings (MeSH) Vocabulary and the PubMed Central document set (U.S. National Library of Medicine, 1996). Each year MeSH provides a list of terms that are added to the MeSH vocabulary in the respective year (mshnewYYYY.txt), which is supposed to be our source of eNEs for the respective year YYYY. For us the adoption of a term by MeSH proves that a term has been acknowledged by the medical community as a medical Named Entity and it also indicates the time t_ACK. In our study we extracted terms from the MSHNEW lists from the years 2007 to 2016, which respective numbers are shown in Table 1. For each term in a years' MSHNEW-list we performed a search query against the PubMed Central document set through the public PubMed search engine and counted the number of result documents per year given back by the PubMed Search engine dating back to 1980. For example, a search engine query of the term "Sofosbuvir" gives a total result count of 1136 documents, which were published starting in 2010. Before 2010 no document with this term was published and therefore no results were returned by the search for 1980 – 2010. The considered time frame for the result document counts is intentionally longer than the time frame from which the MSHNEW lists were taken. The different time frames are necessary as by definition eNEs are used before t_ACK and so we must consider documents older than t_ACK to find out when and to which extend eNEs are used before t_ACK. Our analysis of the document counts comprises three levels: In a first level we compared the counts and the distribution of selected single terms from MSHNEW2016. In the second level we performed a statistical analysis on all query results of the complete MSHNEW2016 vocabulary. To normalize the results and make them comparable we calculated the median of the derived document counts c for each year y and then calculated the percentage $p_{median_{MSHNEW2016}}(y)$ of a years' median regarding the sum of all medians from 1980 – 2016. We chose the median instead of the average to reduce the impact of single terms that produced a very high count of result documents, because these terms were quite generic and thus are used in a lot of articles although not being discriminative for each of the articles (e.g. "eeking Behavior" from MSHNEW 2016, which creates very "noisy" search results).

$$p_{med_{MSHNEW2007}}(y) = \frac{\widetilde{c(y)}}{\sum_{k=1980}^{2016} \widetilde{c(k)}} \quad (1)$$

For the chosen vocabulary MSHNEW2016 for example there is a median count $c(\widetilde{2010}) = 40$ as a query with a term from MSHNEW2016 in median returns 40 result documents from PubMed published in the year 2010. In relation to the sum of all median counts between 1980 and 2016 – which is 702 – this leads to a percentage $p_{med_{MSHNEW2016}}(2010) = 0.057$. To compare with the overall growth of the PubMed document set we also calculated the percentage $p_{pubmed_{YYYY}}$ of the counts d of all documents in the PubMed document set per year y in relation to the total number of documents in PubMed from 1980 – 2016. Thus, the second formula provides a generic picture of the relative growth of the PubMed corpus from 1980 – 2016.

$$p_{pubmed_{2016}}(y) = \frac{d(y)}{\sum_{k=1980}^{2016} d(k)} \quad (2)$$

The third level of our analysis summarizes the median results of the yearly analyses described before. For each of the MSHNEW-years considered [2007 – 2016] we extracted the respective medians of a 20-year time interval [t_ACK-20, t_ACK]. For example, for t_ACK = 2007 this interval covers $[p_{med_{MSHNEW2007}}(1987), p_{med_{MSHNEW2007}}(2007)]$. We finally calculated the median for each of the generic years [t_ACK-20, t_ACK]. To compare, we also calculated a graph showing the median growth rate of the overall PubMed Collection in the respective years.

Table 2: Example Terms of the MSHNEW 2016 vocabulary.

Term	# of Docs 1980 - 2016	Years T_USE - T_ACK	Relative Distribution 1980 - 2016
Adalimumab	5539	14	
Neuroprotection	18505	29	
Imatinib Mesylate	9773	20	
Cobicistat	225	6	
Rheumatism (comparative, non 2016)		n/a	

3.3 Experimental Results

Equivalent to our experimental setup, our presentation of results follows the three levels "term", "year" and "overall". Table 2 shows a selection of example terms from the MSHNEW 2016 vocabulary and one comparative term. The table shows the number of documents returned by a query of the respective term from the PubMed document collection in the years 1980 – 2016 followed by the time interval in which the term has been emergent and a graphical representation of the relative distribution of documents represented by the term from 1980 – 2016. Although the terms are all considered to be emerging NEs according to our definition it becomes clear that they differ in their level of emergence. The examples "Adalimumab" and "Neuroprotection" show a long lasting and continuous development with thousands of documents returned while the emergence-interval of "Cobicistat" is relatively short as well as the number of returned documents is significantly lower. Compared to that the term "Imatinib Mesylate" shows a different relative distribution which is not typical for an eNE: After an initial increase of documents the distribution drops again before being acknowledged as a NE by the expert community. After having a look on selected terms with characteristic distributions the next step of the analysis is the year-level of all MSHEWN terms from 2016. Figure 2 shows the relative median growth of PubMed Documents returned by queries from the MSHNEW 2016 vocabulary. As a comparison the overall relative growth of PubMed is plotted. It becomes clear that the gradient (1st derivation) of the eNE-graph becomes higher than the gradient of the PubMed overall graph already approximately 1999. In approximately 2005 both the gradient and the growth rate of the eNE-graph become higher than the overall growth with an again significantly growing gradient approximately starting in 2011. Extending the view to ten years, the graph showing the median relative increase of the Years 2007 – 2016. Figure 3 shows a similar gradient as the one of 2016, although it is a bit less distinct than the one from 2016. This is probably the result of applying the medians of ten years which filters out single extreme values. Just as in the 2016 graph it again becomes clear that the gradient (1st derivation) of the eNE graph becomes bigger than the gradient of the overall PubMed relative growth at the time t_ACK – 15. Again at approx. t_ACK- 5 we see both an increase of the gradient, as well as a relative growth rate that becomes bigger than the one of the PubMed document corpus.

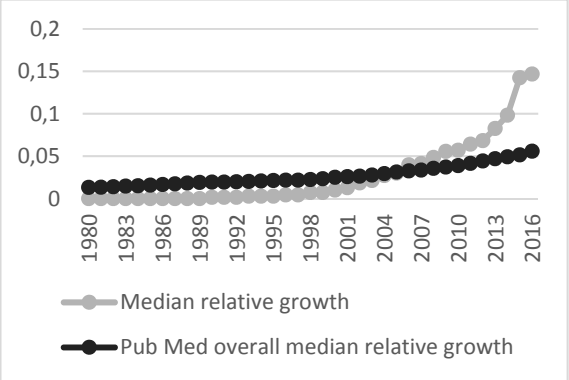

Figure 2: Median Relative growth of PubMed Documents represented by eNE-Queries from MSHNEW 2016.

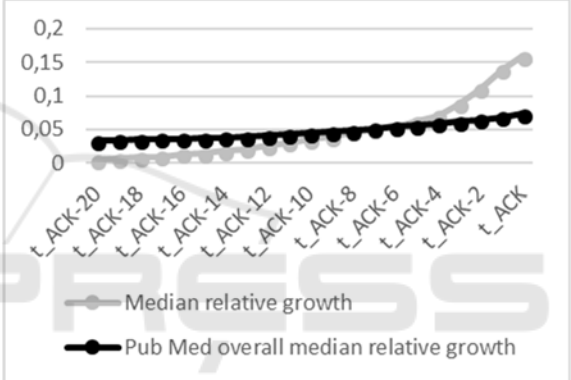

Figure 3: Summarized median relative growth of PubMed Documents represented by eNE-Queries from MSHNEW 2007 – 2016.

4 DERIVED ARCHITECTURE

In this section we discuss our approach for an eNER framework based on statistical methods and expert user feedback. The framework is still in an early status and has not been implemented yet. Our experiments showed main results which are addressed in our derived system architecture. The first result is the fact that the absolute number of documents represented by an eNE differs enormously (compare the examples Neuroprotection and Cobicistat). As our approach must cover the individual use case – in which highly specified knowledge may be needed for an individual therapy – eNEs with a low absolute number must be identified confidently as well as those with high numbers for discovering trends in the comprehensive use case. The second result is that

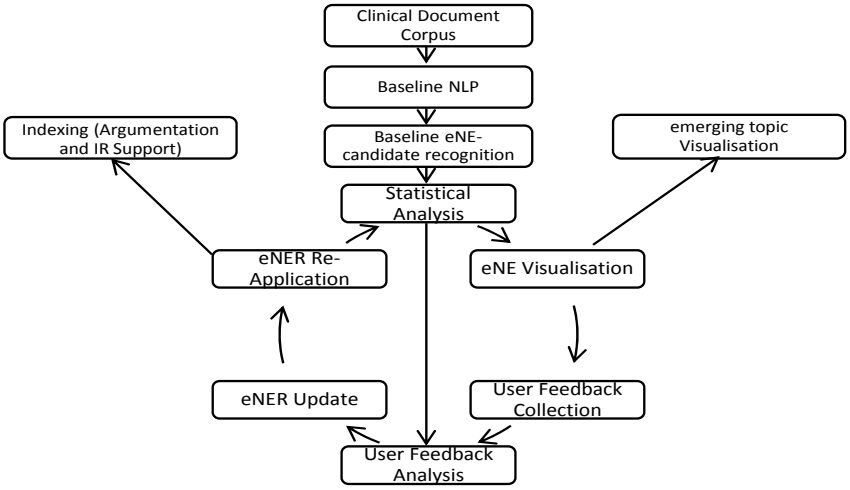

Figure 4: eNE Recognition Cycle.

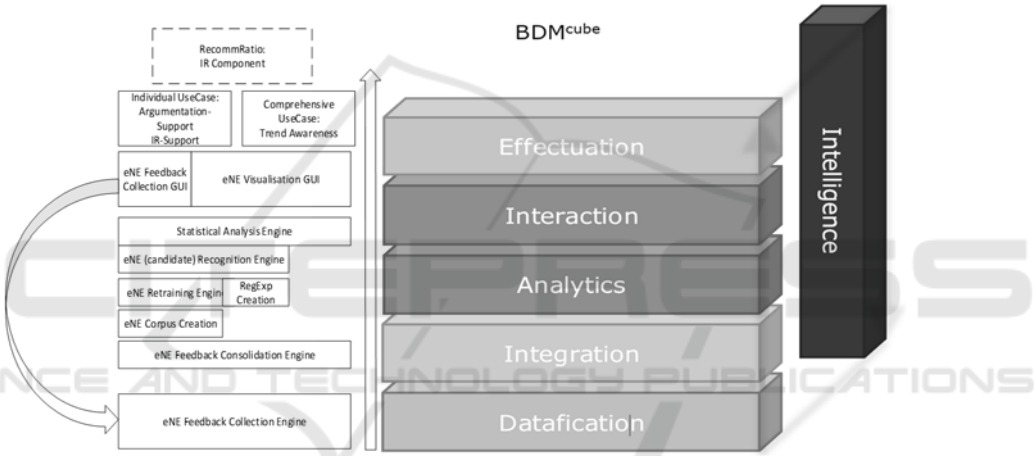

Figure 5: BDMCube Integration.

the timeframe between t_ACK and t_USE is very different as well as the relative distribution of the documents within this timeframe. This leads to the conclusion that the statistical patterns to identify eNEs must be flexible and differentiated. Coming back to our individual use case even patterns as the one shown for Imatinib Mesylate must be considered. This may be an indicator in the argumentation process that the topic was emergent but is not state of the art anymore, which may be a strong argument against a certain therapy. The third relevant result is that statistical features that identify eNEs already appear quite early. The slight increase in the median derivation at t_ACK – 15 shows that our architecture must be able to identify small statistical deviations to recognize eNEs as early as possible. The primary goal of our architecture is to identify eNEs as early as possible and make them usable for individual and comprehensive use cases.

4.1 eNE-Recognition-Cycle

With the empirical study we showed that domain specific eNEs in the medical domain have been used in literature years before they are acknowledged by the expert community. We identified a time frame of about five years before acknowledgement as important, as particularly within this time frame there is a significant increase of documents containing eNEs compared to the general increase of documents in the respective corpus. In this chapter we propose an architecture which addresses the challenge of early detection of eNEs in medical literature through both statistical analysis as well as feedback from medical experts.

Figure 4 shows the principle architecture and feedback workflow which is described in the following. The workflow starts with a collection of clinical textual literature, which is relevant for the

respective VRE, the clinical document corpus. As a first processing step Baseline NLP is done through a State of the Art NLP framework and consists of the generic NLP tasks tokenization, sentence splitting, Part of Speech tagging and generic Named Entity Recognition (Person, Location, Organization). It also comprises the tagging with a domain specific vocabulary (if applicable), like MeSH to distinguish between already existing domain specific NEs and eNE candidates. NLP features created within the Baseline NLP step are needed in the later eNER detection cycle as input for both rule- and ML-based eNER. The second step Baseline eNER is done through a hand-crafted set of Regular Expression (RegExp) rules to detect candidates which could be eNEs based on textual features derived in the baseline NLP step. The rules in this step are quite lenient to achieve a high recall in connection with a low precision to cover a high percentage of eNE candidates in a first row. The following step "Statistical Analysis" is the entry point for the actual detection cycle. This is the step where data scientist method of classification is applied first. As a classification technique for this step we intend to use Support Vector Machine based pattern learning and recognition on the distributional patterns. The aim is to identify those eNE-candidates that have a distribution and increase patterns similar as those demonstrated in the experimental results. The objective is to preserve a high recall from the prior eNE-candidate recognition while increasing the precision compared with the step before. The next step is the visualization of eNE-candidates in a Graphical User Interface (GUI). It will comprise textual visualization of identified eNEs and an integrated visualization of the emergence of them through sparklines, which have been already discussed for displaying data in clinical environments (Radecki and Medow, 2007). An example for the sparklines is shown in Table 2. Based on the visualization the expert users are asked to give feedback on eNE-candidates presented in connection with a search query of a user. It is intended to create a GUI that allows the users an "on-the-fly"-feedback which is integrated strongly into the GUI of the actual VRE. The feedback of the users covers two major questions: a) Is the eNE-candidate a term that is a relevant named entity for your domain and therefore probably acknowledged in the future? b) If yes, please give an estimation about the classification of the eNE-candidate in the classification scheme of your domain (e.g. MeSH tree). b) is intended to collect users' knowledge about the structural context of a term for use in argumentation support tasks. After a defined number of user feedbacks is collected they are consolidated towards one common user feedback which has a high degree of confidence due to the input of several different expert users (User Feedback Analysis component). Based on the consolidated user feedback in the following step an eNER Update is performed which comprises a re-building of Regular-Expressions for the rule-based approach (Whitelisting of Terms that have been identified as eNEs and blacklisting of eNE-candidates which have been discarded through the users.). In parallel an internal training corpus for Machine Based learning NER-approaches is created. For the corpus the sentences containing identified eNEs are extracted from the literature and the eNEs are tagged in these sentences. Beyond the eNE-tagging the corpus comprises the textual features derived in the baseline NLP step. The corpus than is used to re-train a ML based NER-algorithm like CRF to detect yet unknown eNE-candidates. The feedback is also used to evaluate the statistical pattern that relates to the eNE-candidate and update the statistical analysis step. After rebuilding and re-training rule-based and ML-based NE algorithms they are re-applied on the initial clinical document corpus to extract a) eNE-candidates with a higher quality compared to the initial baseline eNE-candidate-recognition and b) to identify eNEs which are no candidates anymore but have been identified by the expert community with a high degree of confidence. From this step in the cycle the candidates are fed to the statistical analysis where the feedback cycle starts again while the identified eNEs with the context information from the users' feedback are indexed for further use in argumentation – and IR support in the VRE.

4.2 BDMCube Integration

The component model implements the BDMCube and follows the design of the eNE recognition cycle. The BDMCube is intended to create data intelligence on Big Data through the layers Datafication, Integration, Analysis, Interaction and Effectuation sources. In our system design the Datafication layer is implemented through a feedback collection engine, which gathers and stores users' feedback on eNE-candidates. The layer Integration is implemented through the consolidation engine and the corpus creation which are the both integration tasks in our eNE recognition cycle. Most tasks of our approach you find the analytics layer, which reflects the analytical and data science-oriented focus of our work. The Analytics layer

covers the engines for the RegExp rebuilding, the eNER retraining (for the ML based eNER), the re-application of the eNER (both ML and RegExp based) and finally the statistical analysis of the newly detected eNE-candidates. On the interaction layer you find the interactive visualization of the statistical results, of identified eNEs and the integrated GUI for collecting feedback and context on the eNE-candidates. The Effectuation layer of the BDMCube is intended to create added value for the user by providing the intelligence for supporting the underlying use cases. In our project this layer contains the interfaces to RecomRatio. It provides the functionalities for the individual and the comprehensive use cases to be integrated into the both projects' IR GUIs.

5 CONCLUSION AND OUTLOOK

In this paper we introduced our concept of emerging Named Entities, eNEs. With the experiments we were able to show how eNEs can represent emerging knowledge in clinical VREs and hence may be used to support IR and Argumentation Support in clinical VREs for both individual and comprehensive use cases. Following these two main contributions – definition of eNEs and the results of the experiments – we discussed our proposal for a framework which can recognize eNEs by combining NLP, statistical methods, ML and expert user feedback and make eNEs usable for individual and comprehensive use cases in clinical VREs. The next steps in our work are the prototypical implementation and evaluation of the proposed framework, including the visualization component, the design of the core component, the development of statistical patterns to identify eNE-candidates and foremost a user survey about search practice of medical staff in clinical VREs. The objective of the user survey is to find typical search patterns used by clinicians when searching for arguments as well as to figure out their expected outcome of the search (ranking and visualization). In addition, with the survey we want to investigate whether clinicians use recent (emergent) vocabulary for search and argumentation or whether they rely on traditional wording. The results of the survey are intended to optimize baseline eNER ("seed") and statistical patterns as well as aligning visualization and ranking principles in the IR GUI based on the expert users' actual needs.

REFERENCES

Aronson, A.R. (2001) 'Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program', *Proceedings of the AMIA Symposium*, pp. 17–21.

Aronson, A.R. and Lang, F.-M. (2010) 'An overview of MetaMap: historical perspective and recent advances', *Journal of the American Medical Informatics Association*, 17(3), pp. 229–236. doi: 10.1136/jamia.2009.002733

Bada, M. *et al.* (2012) 'Concept annotation in the CRAFT corpus', *BMC Bioinformatics*, 13(1), p. 161.

Bawden, D. and Robinson, L. (2009) 'The dark side of information: Overload, anxiety and other paradoxes and pathologies', *Journal of Information Science*, 35(2), pp. 180–191. doi: 10.1177/0165551508095781

Chang, A.X. and Manning, C.D. (2014) *TokensRegex: Defining cascaded regular expressions over tokens* (CSTR 2014-02).

Deubzer, S., Dietrich, K. and Goller, D. (2016) 'Named Entity Recognition mit eBay Auktionsartikeln: Erstellen eines three-class-models mit Smartphonedaten', *Informatik Spektrum*, 39(05), pp. 373–380.

DFG (2016) *Schwerpunktprogramm „Robust Argumentation Machines" (SPP 1999)*, 27 June. Available at: http://www.dfg.de/foerderung/info_wissenschaft/2016/info_wissenschaft_16_38/index.html (Accessed: 7 March 2018).

Eljasik-Swoboda, T., Kaufmann, M. and Hemmje, M. (2018) 'No Target Function Classifier - Fast Unsupervised Text Categorization Using Semantic Spaces', in *Submitted to Proceedings of DATA 2018.*

Finin, T. *et al.* (2010) 'Annotating Named Entities in Twitter Data with Crowdsourcing', *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk.* Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 80–88. Available at: http://dl.acm.org/citation.cfm?id=1866696.1866709.

Finkel, J.R., Grenager, T. and Manning, C. (2005) *Incorporating non-local information into information extraction systems by gibbs sampling*: Association for Computational Linguistics.

Garmire, L.X. et al. (2016) 'The Training of Next Generation Data Scientists in Biomedicine', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22, pp. 640–645.

Grishman, R. (1995) *Namend Entity Task Defintion*, 31 May. Available at: http://www.cs.nyu.edu/cs/faculty/grishman/NEtask20.book_1.html (Accessed: 26 May 2016).

Hearst, M.A. *et al.* (1998) 'Support vector machines', *IEEE Intelligent Systems and their applications*, 13(4), pp. 18–28.

Hunter, A. and Williams, M. (2015) 'Aggregation of Clinical Evidence Using Argumentation: A Tutorial Introduction', in Hommersom, A. and Lucas, P.J.F.

(eds.) *Foundations of Biomedical Knowledge Representation: Methods and Applications*. Cham: Springer International Publishing, pp. 317–337.

Huth, E.J. (1989) 'The information explosion', *Bulletin of the New York Academy of Medicine*, 65(6), p. 647.

Jensen, P.B., Jensen, L.J. and Brunak, S. (2012) 'Mining electronic health records: Towards better research applications and clinical care', *Nature Reviews Genetics*, 13(6), p. 395.

Joachims, T. (1998) *Making large-scale SVM learning practical*.

Jurafsky, D. and Martin, J.H. (2009) *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd edn. (Prentice Hall series in artificial intelligence). Upper Saddle River, N.J.: Pearson Prentice Hall.

Kaufmann, M. *et al.* (2017) 'The Big Data Management Canvas Method', *Submitted to the 6th international Conference on Data Science, Technology and Management DATA*, Madrid.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature*, 521(7553), p. 436.

Lippi, M. and Torroni, P. (2016) 'Argumentation Mining', *ACM Transactions on Internet Technology*, 16(2), pp. 1–25. doi: 10.1145/2850417

Manning, C.D. *et al.* (2014) 'The Stanford CoreNLP Natural Language Processing Toolkit', *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. Available at: http://www.aclweb.org/anthology/P/P14/P14-5010.

Nadeau, D. and Sekine, S. (2007) 'A survey of named entity recognition and classification', *Lingvisticae Investigationes*, 30(1), pp. 3–26.

Nandish V. Patel and Ahmad Ghoneim (2011) 'Managing emergent knowledge through deferred action design principles: The case of ecommerce virtual teams', *Journal of Ent Info Management*, 24(5), pp. 424–439. doi: 10.1108/17410391111166503

Ohta, T., Tateisi, Y. and Kim, J.-D. (2002) *The GENIA corpus: An annotated research abstract corpus in molecular biology domain*: Morgan Kaufmann Publishers Inc.

Piskorski, J. and Yangarber, R. (2013) 'Information Extraction: Past, Present and Future', in Poibeau, T. *et al.* (eds.) *Multi-source, Multilingual Information Extraction and Summarization*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 23–49. Available at: http://dx.doi.org/10.1007/978-3-642-28569-1_2.

Radecki, R.P. and Medow, M.A. (2007) *Cognitive debiasing through sparklines in clinical data displays*.

Rocchio (1971) 'Relevance feedback in information retrieval', *In The SMART Retrieval System: Experiments in Automatic Document Processing*, pp. 313–323.

Sang, Erik F. Tjong Kim and Meulder, F. de (2003) 'Introduction to the CoNLL-2003 shared task: language-independent named entity recognition', in *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 - Volume 4*. Edmonton, Canada: Association for Computational Linguistics, pp. 142–147.

Savova, G.K. *et al.* (2010) 'Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications', *Journal of the American Medical Informatics Association: JAMIA*, 17(5), pp. 507–513. doi: 10.1136/jamia.2009.001560

U.S. National Library of Medicine (1996) *PubMed Central*. Available at: https://www.ncbi.nlm.nih.gov/pmc/ (Accessed: 24 May 2018).

U.S. National Library of Medicine (1999) *Medical Subject Headings*. Public Domain Information. Available at: https://www.nlm.nih.gov/mesh/ (Accessed: 23 May 2018).

U.S. National Library of Medicine (2017) *Yearly Citation Totals from 2017 MEDLINE/PubMed Baseline*. Available at: https://www.nlm.nih.gov/bsd/licensee/2017_stats/2017_Totals.html.